EF

By Sunčica Rogić

WORD COUNT	110322	TIME SUBMITTED	07-MAR-2023 02:53PM
		PAPER ID	97361383

UNIVERZITET CRNE GORE
EKONOMSKI FAKULTET PODGORICA

MR SUNČICA ROGIĆ

PREDIKTIVNI MODELI ODLUČIVANJA U DIREKTNOM MARKETINGU BAZIRANI NA *SUPPORT VECTOR MACHINE* METODI

Doktorska disertacija

Podgorica, 2023.

UNIVERSITY OF MONTENEGRO

FACULTY OF ECONOMICS PODGORICA

MSc SUNČICA ROGIĆ

PREDICTIVE DECISION SUPPORT MODELS IN DIRECT MARKETING BASED ON SUPPORT VECTOR MACHINE METHOD

PhD thesis

Podgorica, 2023.

3

PODACI O DOKTORANDU

Ime i prezime: Sunčica Rogić

Datum rođenja: 23.10.1992.

Naziv završenog studijskog programa i godina završetka: Ekonomski fakultet Podgorica, magistarske studije - smjer Marketing i biznis, 2018.

UDK, OCJENA I ODBRANA DOKTORSKE DISERTACIJE

Datum sjednice Senata Univerziteta na kojoj je prihvaćena teza:

Mentor:

Prof. dr Ljiljana Kašćelan, redovni profesor, Univerzitet Crne Gore, Ekonomski fakultet

Komisija za ocjenu podobnosti teze i kandidata:

Prof. dr Ivan Luković, redovni profesor, Univerzitet u Beogradu, Fakultet organizacionih nauka
Prof. dr Ljiljana Kašćelan, redovni profesor, Univerzitet Crne Gore, Ekonomski fakultet
Prof. dr Boban Melović, redovni profesor, Univerzitet Crne Gore, Ekonomski fakultet

Komisija za ocjenu doktorske disertacije:

Komisija za odbranu doktorske disertacije:

Datum odbrane:

Datum promocije:

UDK broj:

Rezime

Digitalna revolucija, koja je uticala na rast količine raspoloživih podataka, odrazila se i na marketing sferu i komunikaciju kompanija s kupcima. Kako savremeni potrošači zahtijevaju odgovor na svoj upit, dvosmjernu komunikaciju i personalizovanu ponudu, upravo je *online* direktni marketing jedan od najznačajnijih alata za direktnu komunikaciju s kupcima, njihovu retenciju i plasiranje diferenciranih poruka, kreiranih na osnovu specifičnosti pojedinačnih kupaca ili segmenata kupaca. Uzimajući u obzir činjenicu da sve veći broj savremenih kompanija svoje poslovne odluke zasniva na podacima, organizacije koje efikasno koriste ove podatke stiču značajnu konkurentsku prednost i veće šanse za uspjeh na tržištu. Samim tim, *big data* analitika i prediktivni modeli za podršku odlučivanju, koji pružaju uvid u ponašanje potrošača, predstavljaju aktuelan izazov ne samo za poslovne subjekte, već i za naučnu zajednicu.

U ovom radu predloženi su konceptualni modeli za unapređenje sistema selekcije i targetiranja kupaca u direktnom marketingu, bazirani na predikciji pripadnosti segmentu zavisno od kupovnog ponašanja, predikciji odgovora na kampanju i predikciji njihove profitabilnosti. Predloženim prediktivnim modelima prevaziđeni su određeni nedostaci modela iz prethodne literature, prije svega kroz rješavanje problema nebalansiranosti klasa i asimetrične distribucije profitabilnosti respondenata, koji nastaju zbog manjeg broja respondenata i visokoprofitabilnih kupaca u odnosu na ostale. Zbog ovih problema, prediktivni modeli pokazuju loše prediktivne performanse za klasu najvrednijih i najvažnijih kupaca za kompaniju. Zbog toga je predloženo balansiranje podataka pomoću *Support Vector Machine* (SVM) metode, koja je i ranije u literaturi primjenjivana u tu svrhu, ali u direktnom marketingu po prvi put u ovom radu. Balansirani podaci se zatim predviđaju pomoću različitih *data mining* (DM) klasifikatora. Da bi se povećala prediktivna tačnost, primjenjuju se *ensemble* metode, koje kombinuju više slabih klasifikatora u jaki, koji značajno poboljšava prediktivne performanse. Za predikciju profitabilnosti predložena je *Support Vector Regression* (SVR) metoda, čije su izvanredene

mogućnosti generalizacije već potvrđene u literaturi i zbog čega se uspješno prevazilazi problem asimetrične distribucije.

Empirijsko testiranje predloženih modela na podacima iz direktnih kampanja crnogorske kompanije, kao i na javno dostupnim skupovima podataka, pokazalo je značajno povećanje prediktivnih performansi za kategoriju najvrednijih kupaca, a samim tim i efikasnost selekcije i targetiranja kupaca, kao i potencijalnog profita ostvarenog u kampanjama direktnog marketinga. Pored toga, predloženi modeli obezbjeđuju profilisanje segmenta najvrednijih kupaca – visokoprofitabilnih respondenata, stvarajući tako pretpostavke za kreiranje relevantnijih ponuda i prilagođenih poruka, te uspostavljanje baze lojalnih kupaca.

U ovom radu nisu samo predloženi koncepti za navedene prediktivne modele, već i način njihove realizacije u alatu *Rapid Miner*. Modeli su realizovani u vidu gotovih procesa koje kompanije mogu, uz prilagođavanje sopstvenoj bazi kupaca, primjenjivati u direktnom marketingu. Stoga, rezultati ove disertacije mogu biti od koristi marketing menadžerima da povećaju povjerenje u donošenje odluka na osnovu podataka, kao i efikasnost i profitabilnost direktnih kampanja. Takođe, mogu biti podsticaj i putokaz za primjenu analitičkih alata kao što je *Rapid Miner* u marketingu.

Ključne riječi: direktni marketing, selekcija i targetiranje kupaca, profilisanje kupaca, *data mining*, *Support Vector Machine*

Naučna oblast: Poslovna ekonomija

Uža naučna oblast: Poslovna inteligencija u marketingu

Abstract

The digital revolution, which influenced the growth of the amount of available data, also affected the marketing sphere and the communication of companies with customers. As modern consumers demand an answer to their inquiry, two-way communication and a personalized offer, online direct marketing is one of the most important tools for direct communication with customers, their retention and differentiated messages, created based on the specifics of individual customers or customer segments. Taking into account the fact that an increasing number of modern companies base their business decisions on data, organizations that effectively use this data gain a significant competitive advantage and greater chances of success on the market. Therefore, big data analytics and predictive models for decision support, which provide insight into consumer behavior, represent a current challenge not only for businesses, but also for the scientific community.

In this paper, conceptual models are proposed for improving the system of selection and targeting of customers in direct marketing, based on the prediction of affiliation to a segment depending on purchasing behavior, the prediction of the campaign response and the prediction of their profitability. The proposed predictive models overcome certain limitations of models from previous literature, primarily by solving the problem of class imbalance and asymmetric distribution of respondents' profitability, which arise due to a smaller number of respondents and highly profitable customers compared to others. Because of these problems, predictive models show weaker predictive performance for the class of most valuable and important customers for the company. Therefore, in this dissertation, data balancing using the Support Vector Machine (SVM) method, which has been used for this purpose in the literature before, but in direct marketing for the first time in this paper, is proposed. The balanced data are then predicted using different data mining (DM) classifiers. In order to increase the predictive accuracy, ensemble methods are applied, which combine several weak classifiers into a strong one, which significantly improves the predictive performance. For profitability prediction, the Support Vector Regression (SVR) method is proposed, whose extraordinary

generalization capabilities have already been confirmed in the literature, which is why it successfully overcomes the problem of asymmetric distribution.

Empirical testing of the proposed models on data from direct campaigns of the Montenegrin company, as well as on publicly available data sets, showed a significant increase in predictive performance for the category of the most valuable customers, and thus the effectiveness of selection and tagging of customers, as well as the potential profit achieved in direct marketing campaigns . In addition, the proposed models provide profiling of the segment of the most valuable customers - highly profitable respondents, thus creating prerequisites for creating more relevant offers and customized messages, and establishing a base of loyal customers.

In this dissertation, not only are the concepts proposed for these predictive models, but also the method of implementation in the Rapid Miner tool. The models are realized in the form of ready-made processes that companies can, with adaptation to their own customer base, apply in direct marketing. Therefore, the results of this dissertation can be useful to marketing managers to increase confidence in decision-making based on data, as well as the efficiency and profitability of direct campaigns. Also, they can be an incentive and guide for the application of analytical tools like Rapid Miner in marketing.

Keywords: direct marketing, customer selection and targeting, customer profiling, data mining, Support Vector Machine

Scientific field: Business Economics

Narrow scientific field: Business intelligence in marketing

Spisak tabela

Spisak slika

SADRŽAJ

Spisak skraćenica

Skraćenica	Puni naziv
AHP	Analitički hijerarhijski proces *(eng. Analytic Hierarchy Process)*
AI	Vještačka inteligencija *(eng. Artificial Intelligence)*
ANN	*Vještačka neuronska mreža* (eng. *Artificial Neural Network*)
ARIMA	Autoregresivni integrisani pokretni prosjek *(eng. Autoregressive Integrated Moving Average)*
AUC	*Area Under the Curve*
BNN	*Bagging* neuronska mreža *(eng. Bagging Neural Network)*
BPNN	*Back Propagation Neural Network*
CART	Klasifikaciona i regresiona stabla odlučivanja *(eng. Classification and Regression Trees)*
CHAID	*Chi-Square Automatic Iteration Detection*
CRISP	Iterativni postupak segmentacije zasnovan na odgovoru korisnika *(eng. Customer Response-based Iterative Segmentation Procedure)*
CRISP-DM	Međuindustrijski standardni proces za rudarenje podataka *(eng. Cross-Industry Standard Process for Data Mining)*
CRM	Menadžment odnosa sa kupcima (eng. *Customer Relationship Management*)
CUE	*Clustering Balanced Undersampling and Ensemble*
CV	Vrijednost kupca *(eng. Customer Value)*
DB	*Davies Bouldin*
DL	Duboko učenje *(eng. Deep Learning)*
DM	Rudarenje podataka (eng. *Data Mining*)
DMEF	Edukativna fondacija za direktni marketing (eng. *Direct Marketing Education Foundation)*
DT	Drvo odlučivanja (eng. *Decision tree*)
EU	Evropska unija (eng. *European Union*)
FN	Lažno negativni *(eng. False Negative)*
FP	Lažno pozitivni *(eng. False Positive)*
FPR	Stopa lažno pozitivnih primjera *(eng. False Positive Rate)*
GA	Genetski algoritam *(eng. Genetic Alogrithm)*
GBT	*Gradient Boosted Trees*
GDPR	Opšta uredba o zaštiti podataka EU *(eng. General Data Protection Regulation)*
ID3	*Iterative Dichotomiser 3*
k-NN	Metoda k-najbliži susjed *(eng. k-Nearest Neighbour)*
LDA	Linearna diskriminantna analiza *(eng. Linear Discriminant Analysis)*
LR	Logistička regresija (eng. *Logistic Regression*)
LTV	Cjeloživotna vrijednost *(eng. Lifetime Value)*

MAE	Srednja apsolutna greška *(eng. Mean Absolute Error)*
MAPE	Srednji apsolutni procenat greške *(eng. Mean Absolute Percentage Error)*
MARS	*Multiple Adaptive Regression Splines*
ME	Srednja greška *(eng. Mean Error)*
ML	Mašinsko učenje *(eng. Machine Learning)*
MLE	Procjena maksimalne vjerovatnoće *(Maximum Likelihood Estimation)*
MLPNN	*Multilayer Perceptron Neural Network*
MPE	Procenat srednje greške *(eng. Mean Percentage Error)*
MSE	Srednja kvadratna greška *(eng. Mean Square Error)*
NMSE	Normalizovana srednja kvadratna greška *(eng. Normalized Mean Square Error)*
NN	Neuronska mreža (eng. *Neural Network*)
OLS	Obični najmanji kvadrati *(eng. Ordinary Least Squares)*
PPV	Pozitivna predviđena vrijednost *(eng. Positive Predicted Value)*
R^2	Koeficijent determinacije *(eng. Coefficient of Determination)*
RBF	*Radial Basis Function*
RF	*Random Forest*
RFM	*Recency Frequency Monetary*
RMSE	Korijen srednje kvadratne greške *(eng. Root Mean Square Error)*
RNN	Rekurentne neuronske mreže *(eng. Recurrent Neural Networks)*
ROC	*Receiver Operating Characteristic*
SEM	Marketing u pretraživačima *(eng. Search Engine Marketing)*
SMM	Markeitng na društvenim mrežama *(eng. Social Media Marketing)*
SMOTE	*Synthetic Minority Oversampling Technique*
SOM	Samo-organizujuće mape *(eng. Self Organizing Maps)*
SSE	Suma kvadratne greške *(eng. Sum of the Squared Error)*
STP	Segmentacija-targetiranje-pozicioniranje (*eng. Segmentation-Targeting-Positioning*)
SVM	Metoda potpornih vektora *(eng. Support Vector Machine)*
SVM-RE	*Support Vector Machine Rule Extraction*
SVR	Regresija potpornih vektora (eng. *Support Vector Regression*)
TN	Stvarno negativni *(eng. True Negative)*
TNR	Stopa stvarno negativnih primjera *(eng. True Negative Rate)*
TP	Stvarno pozitivni *(eng. True Positive)*
TPR	Stopa stvarno pozitivnih primjera *(eng. True Positive Rate)*
WSS	Suma kvadrata unutar klastera *(eng. Within-cluster Sum of Square)*
XGB	*Extreme Gradient Boosting*

1. UVOD

Dinamičan razvoj e-trgovine, komunikacija putem interneta, društvenih mreža i direktnog marketinga, stvorili su mogućnosti za stvaranje nove vrijednosti u *online* tržišnom okruženju. Pored toga, ovi faktori su uticali na promjene u oblasti direktnog marketinga, koji je u posljednjoj deceniji dobijao sve više karakteristika digitalnog marketinga. Savremeni potrošač je prisutan na mreži, kupuje u *online* prodavnicama, posjećuje stranice svojih omiljenih brendova na društvenim mrežama i tako za sobom ostavlja digitalni trag i podatke o svojim interesovanjima i potencijalnim željama (Rogić & Kašćelan, 2022). Za razliku od tradicionalne trgovine, digitalni marketing i e-trgovina generišu ogromnu količinu vrijednih podataka o proizvodima, namjerama i željama kupaca, kao i njihovom ponašanju, kao što su podaci o tome odakle kupci dolaze, koje uređaje koriste, koje proizvode pregledaju i kupuju, koliko dugo ostaju na pojedinim stranicama, te da li odgovaraju na promotivne sadržaje i poruke (Esmeli et al., 2020). Skorija istraživanja su pokazala da generičke ponude plasirane potrošačima predstavljaju veoma neefikasnu strategiju oglašavanja (Behera et al., 2020; Stewart-Knox et al., 2016). S tim u vezi, prilagođene ponude i potreba za razvijanjem marketing aktivnosti koje su dominantno digitalne, nameće se kao imperativ u savremenom tržišnom okruženju.

Konkurentna kompanija koja može da predvodi digitalnu transformaciju u određenoj zemlji, trebalo bi da ima zaposlene koji su u mogućnosti da iskoriste svoje digitalne vještine na način da dodaju vrijednost organizaciji (Majovski et al., 2018). Mnogi istraživači ističu činjenicu da se u savremenom poslovnom okruženju dešava ekspanzija korišćenja analitike, uzimajući u obzir prisustvo ogromne količine podataka, razvoj digitalnog marketinga na društvenim mrežama, kao i marketing analitike uopšte (Iacobucci et al., 2019; Verhoef et al., 2016). U tom smislu, kombinovanjem i korišćenjem svih aspekata ovakvih promjena i otkrivanjem skrivenih informacija o svojim klijentima i njihovim korišćenjem u poslovnim odlukama, kompanije mogu značajno da unaprijede svoje performanse.

Ove tehnike kompanije mogu koristiti s ciljem da identifikuju tržišne segmente i prepoznaju kupovne i navike i preferencije svojih kupaca, prilagode ponude za jačanje odnosa s kupcima, poboljšaju efikasnost svojih marketing inicijativa korišćenjem preciznijeg targetiranja, što sve zauzvrat može dovesti do profitabilnije marketing aktivnosti i većeg povrata ulaganja (Kašćelan & Rogić, 2022).

U ovom radu biće predstavljen koncept sistema za efikasnu segmentaciju, selekciju i targetiranje kupaca, kao i za procjenu njihove profitabilnosti. Uzimajući u obzir trendove koji nameću personalizovan i tzv. "jedan na jedan" odnos s kupcima, u radu će biti predstavljen model za unapređenje jednog dijela marketing sistema, s fokusom na direktni marketing.

U ovom poglavlju, biće ukratko opisan koncept direktnog marketinga, zatim funkcije i metode za *data mining* - DM (prevod za ovaj pojam bi bio rudarenje podatka i ova dva termina će ravnopravno biti korišćena u radu), čije će tehnike biti korišćene za kreiranje predloženog sistema. Takođe, biće ukazano na probleme koji se javljaju kod realizacije ovakvih sistema i način na koji predloženi koncept rješava ove probleme. U skladu s tim, na kraju poglavlja biće definisani ciljevi istraživanja, hipoteze, kao i doprinosi istraživanja.

1.1 Koncept i primjena direktnog marketinga

U današnjem tržišnom okruženju, koje karakteriše nedostatak formalne privrženosti kompanijama, kao i jednostavnost pretrage i informisanja o proizvodima u uslugama konkurenata, proces izgradnje lojalne baze kupaca sve je značajniji, ali i zahtjevniji i izazovniji za menadžere. S tim u vezi, kvalitet proizvoda i konkurentnost cijena više se ne mogu smatrati dovoljnim faktorima za uspjeh na tržištu. Savremeni potrošač je danas u prilici da pretražuje ponudu, uporedi cijene, sazna za potencijalne promocije, obavi kupovinu (transakciju), podijeli je sa zajednicom i ostavi svoju recenziju - u bilo kom trenutku, s mobilnog uređaja i bilo koje lokacije. Jasno je da je, više nego ikad, potrošač u mogućnosti da kontroliše proces, te da najveći izazov, posebno u oblasti marketinga, više nije pronaći

potrošača (akvizicija), već zadržati ga (retencija), kroz izgradnju dugotrajnih, smislenih i dubljih veza i odnosa.

Marketing u 21. vijeku odlikuje ubrzana ekspanzija tehnologije i digitalne sfere, te promjena gotovo svih dimenzija životnog stila potrošača. Naime, individualne navike potrošača su redefinisane i izmijenjene, kao i načini na koje oni traže informacije i provode slobodno vrijeme (Jackson & Ahuja, 2016). Stoga je jasno da kompanije ne mogu ostati indiferentne, te da je potrebno da preispitaju načine obavljanja svakodnevnih aktivnosti i pristup marketingu. Prosječno vrijeme koje internet korisnici širom svijeta provedu *online*, zaključno sa trećim kvartalom 2020. godine je šest sati i 45 minuta, s trendom rasta (Statista, 2021a). U skladu s tim, jasna je težnja i potreba kompanija da dio svojih aktivnosti (ako ne i njihov najveći dio) prenesu u digitalnu sferu, te da koriste *online* medije za potrebe komunikacije s potrošačima, promocije proizvoda i jačanja brenda.

Digitalna transformacija je nezaustavljiv i sveobuhvatan proces. Čak 89% kompanija planira da usvoji ili su usvojile strategiju digitalne transformacije, a prosječan iznos direktnih investicija u digitalnu transformaciju za period 2022-2024. se procjenjuje na 6,3 biliona dolara (Carosella et al., 2021). S jedne strane, digitalna transformacija otvara različite poslovne prilike, dok, s druge, predstavlja značajan izazov u kontekstu implementacije neophodnih promjena u postojećim poslovnim modelima, koji moraju da se prilagode, kako bi bili održiviji, efikasniji i konkurentniji (Stojanovic & Kostic, 2018).

Dosadašnja istraživanja su pokazala da korišćenje *big data* mogućnosti i tehnologija ubrzava promjene u marketingu, kroz implementaciju preciznih strategija i efikasno korišćenje ograničenih marketing resursa i njihovo usmjeravanje na vrijedne kupce, što značajno utiče na razvoj tržišta u eri nove maloprodaje (Zhu & Gao, 2019).

U tako dinamičnom i promjenljivom okruženju, koje sve više postaje digitalizovano, kompanije treba da pokušaju da razumiju svoje potrošače, sticanjem uvida u njihove potrebe, stavove, želje i ponašanje u procesu kupovine. Idealno, svaki potrošač trebalo bi da bude tretiran kao specifična individua, te bi na taj način kompanije komunicirale „jedan na jedan" sa svim svojim kupcima. Međutim, kako ovaj pristup

očigledno nije moguć u praksi, kao efikasna alternativa nameće se segmentacija kupaca u grupe prema različitim karakteristikama, kao i razvoj diferenciranih strategija, koje će na najbolji način adresirati specifičnosti pojedinačnih segmenata. Dakle, različite segmentacione strategije mogu se sprovoditi u skladu sa specifičnim poslovnim ciljevima kompanije, poput, recimo, segmentacije na osnovu vrijednosti ili profitabilnosti kupaca, kako bi se identifikovali visokovrijedni kupci koji bi bili prioritetna grupa za buduće marketing napore. S druge strane, kupci se mogu podijeliti i prema vjerovatnoći odgovora na kampanju, u skladu s njihovim prethodnim kupovnim ponašanjem, kako bi se efikasnije sprovelo targetiranje u budućim kampanjama direktnog marketinga.

Potrebe savremenih kupaca ne mogu se zadovoljiti masovnim marketing aktivnostima (Dibb & Simkin, 1996), već raznovrsnim i personalizovanim pristupom. Prvi korak u kreiranju prilagođene strategije je segmentacija kupaca. Teorija segmentacije pretpostavlja da će *„grupe kupaca sa sličnim potrebama i kupovnim ponašanjem vjerovatno pokazati homogeniji odgovor na marketing programe koji ciljaju na specifične grupe potrošača"* (Dibb & Simkin, 1996). Međutim, s obzirom na to da svaki segment tržišta nema istu vrijednost za kompaniju, za najvrednije segmente se nameću specifične strategije, kao logičan put ka isplativom marketing planu.

S obzirom na to da živimo u eri masovnih podataka (eng. *big data era*), u kojoj se procjenjuje da se na dnevnom nivou kreira oko 2,5 milijarde gigabajta podataka, sve veći izazov za kompanije je kako na efikasan način te podatke valorizovati i staviti ih u funkciju svojih poslovnih ciljeva (Carter, 2021). *Sigma Computing* u svom istraživanju navodi da skoro 2/3 kompanija ne može da dobije relevantne, adekvatne i pravovremene informacije iz svojih IT rješenja (Sigma, 2021). S tim u vezi, uz razvoj tehnologije i digitalnog društva, izazov s kojima se kompanije susreću je kreiranje robustnog i sveobuhvatnog sistema za analitički pristup potrošačkoj bazi, kao i za automatizaciju velikog broja procesa, prije svega kroz *data mining* tehnike. Naime, procesiranjem podataka o kupcima, njihovom ponašanju i transakcijama i sinhronizacijom strategija na osnovu rezultata takvih analiza, veza s kupcima postaje čvršća, profitabilnija i dugoročnija, što sve zajedno doprinosi

dodatnoj vrijednosti za kompaniju. Prediktivni procesi u marketingu, korišćenjem *data mining* alata predstavljaju benefit i za kompanije i za potrošače. S jedne strane, kompanije na osnovu rezultata kreiraju profitabilnije, preciznije i efikasnije strategije, a s druge strane, potrošačima se nudi personalizovana ponuda, prilagođena njihovim potrebama, očekivanjima i afinitetima, te na taj način potrošač dobija utisak da kompanija zaista „osluškuje" potrebe tržišta i kreira relevantan sadržaj i usluge. Upravo je direktni marketing, kao dio cjelokupnog marketing sistema, najpotentniji alat za direktnu komunikaciju s kupcima, njihovu retenciju i plasiranje diferenciranih poruka, kreiranih na osnovu specifičnosti tržišnih segmenata.

Sa razvojem internet tehnologija, direktni marketing je dobijao na značaju uporedo sa ubrzanim padom troškova komunikacije. Kao posljedica niskih troškova pristupa internetu, kompanijama se otvorila mogućnost da kreiraju direktne veze s hiljadama potrošača na način koji je do tada bio nezamisliv (Palmer & Koenig-Lewis, 2009). Dalji pad troškova komunikacije je, pored mogućnosti za kompanije, donio i određene izazove za direktni marketing. Jedan od izazova odnosi se na kontrolu komunikacije u situaciji kada veliki broj potrošača može komunicirati međusobno, uzimajući u obzir sve konkurentnije okruženje, te nadmetanje za pažnju potrošača od strane sve većeg broja kompanija. Ispitivanja eksperata iz oblasti digitalnog marketinga navode da je u SAD u 2021. godini, korisnik interneta podvrgnut plasiranju do 10 hiljada *online* reklama dnevno (Forbes, 2017), u poređenju sa sedamdesetim godinama prošlog vijeka, kada je prosječna osoba vidjela između 500 i 1 600 oglasa (AdAge, 2003; Carr, 2021). Kada su potrošači konektovani i prisutni na mreži, može se reći i da su prisutni na tržištu i dostupni za komunikaciju (Deighton & Kornfeld, 2009). Samim tim, potreba za svrsishodnim, smislenim, efikasnim komuniciranjem u cilju stvaranja dubljih i trajnijih veza i odnosa sa potrošačima, mora da zauzme ključno mjesto u kreiranju strategije komunikacije savremenih kompanija.

Najznačajnija karakteristika savremenog direktnog marketinga je orijentisanost na konkretnu akciju. Stoga, u cilju prodaje, ili koraka prema prodaji, praktičari

direktnog marketinga obezbjeđuju poziv na akciju, odnosno mogućnost za jednostavan odgovor – klikom na link, korišćenjem kupona za popust i slično. Pored orijentisanosti na akciju, važne karakteristike direktnog marketinga su: targetiranje, personalizacija, mjerljivost, mogućnost testiranja i fleksibilnost (Csikósová et al., 2014).

Strategije direktnog marketinga u savremenim tržišnim uslovima su efikasnije uz korišćenje metoda poslovne inteligencije (eng. *business intelligence*). Na primjer, metode poslovne inteligencije mogu olakšati i unaprijediti proces odabira potrošača koji imaju visoku vrijednost za kompaniju u strategijama akvizicije; stvoriti interakciju s profitabilnim potrošačima, uz kreiranje personalizovanih iskustava u cilju njihovog zadržavanja; povećati vrijednosti i profitabilnosti potrošačke korpe kroz različite prodajne strategije preporučivanja proizvoda (Stone & Woodcock, 2014).

1.2 Data mining funkcije i metode

Potreba za razumijevanjem velikih, složenih i informacijama bogatih skupova podataka je zajednička za gotovo sva polja poslovanja, nauke i inženjerstva (Kantardzic, 2020). U korporativnom svijetu, podaci o klijentima, njihovom ponašanju i obavljenim transakcijama prepoznati su kao strateško dobro. U današnjem svijetu hiperkonkurencije, postaje sve važnija sposobnost izvlačenja korisnog i vrijednog znanja skrivenog u ovim podacima i tržišnog djelovanja baziranog na tom znanju.

U korak s porastom količine podataka i izazovima za njihovo skladištenje i analizu, razvija se i *data mining* - jedna od najbrže rastućih oblasti u kompjuterskoj industriji. Stoga će se na početku ovog poglavlja definisati dva pojma koja će često biti navođena u radu – *data mining* i mašinsko učenje.

Naime, *data mining* predstavlja proces identifikovanja obrazaca koji postoje u velikim skupovima podataka i njihove primjene na druge skupove podataka. S druge

strane, mašinsko učenje se definiše kao skup algoritama za otkrivanje tih obrazaca, koji omogućavaju rješavanje problema od strane programa, poput klasterizacije, klasifikacije, prediktivne analize i asocijativnih pravila, bez potrebe da se kreiraju eksplicitne programske instrukcije, koje bi algoritmu sugerisale kako da obavi zadatak (Jamsa, 2021). Na ovaj način, mogućnosti koje nudi mašinsko učenje korišćenjem podataka, mogu pomoći u rješavanju kompleksnih problema.

Dakle, *data mining* se može označiti kao cjelokupan proces primjene kompjuterski zasnovane metodologije za otkrivanje znanja iz podataka. *Kantardžić* (2020) ističe da je *data mining* najkorisniji u scenariju istraživačke analize, u kojem ne postoje unaprijed određene ideje o tome šta će zapravo biti ishod. U tom smislu, *data mining* je potraga za novim, vrijednim i netrivijalnim informacijama u velikim količinama podataka.

S obzirom na to da *data mining* često koristi algoritme mašinskog učenja, ova dva pojma se često u literaturi javljaju zajedno.

Najčešće korišćeni alati za *data mining* su (Jamsa, 2021):

- Sistemi za upravljanje bazama podataka (eng. Database Management System – DBMS), kao što su *MySQL* i *MongoDB*,
- *Excel*,
- Alati za vizuelizaciju, poput *Tableau* softvera,
- Alati za poslovnu inteligenciju, kao što je *Microsoft Power BI*,
- Rješenja korišćenjem programskih jezika,
- Specifični *data mining* alati, kao što su *Rapid Miner*, *Orange* i *Weka*.

S druge strane, za sprovođenje operacija mašinskog učenja, najčešće se koriste (Jamsa, 2021):

- Programski jezici *R* i *Phyton*,
- Alati za vizuelno programiranje, kao što su *Rapid Miner* i *Orange*,
- *Excel* dodaci *(eng. add-ins)*, kao što je *Solver*.

Data mining je aktuelan istraživački domen koji je zbog značajne mogućnosti primjene zaokupio pažnju kako naučne javnosti, tako i industrije. S obzirom na

veliku količinu podataka koja je organizacijama na raspolaganju, stvara se potreba da se oni pretvore u korisne informacije i znanje (Binu & Rajakumar, 2021). Znanje dobijeno iz *data mining* procesa ima široku primjenu, od kontrole proizvodnog procesa, naučnih istraživanja, menadžmenta, analize tržišta i potrošača, tehnologija za prepoznavanje lica, predikciju bankrotstva u kompanijama, predikciju napuštanja potrošača i slično.

Data mining se smatra rezultatom ubrzanog razvoja baza podataka, kao i njihove ekspanzije u domenu informacionih tehnologija. Razvoj ovog sistema uključuje i napredak u sferama formiranja skupova podataka, prikupljanja podataka i nadzora tehnologija baza podataka za skladištenje i pronalaženje podataka, kako bi se postigla efikasna analiza podataka za bolje razumijevanje istraživane pojave (Binu & Rajakumar, 2021). U literaturi se ovaj proces navodi kao zajednički napor ljudi i kompjutera (Kantardzic, 2020), pri čemu se najbolji rezultati postižu upravo balansiranjem znanja analitičara i stručnjaka u opisivanju problema i ciljeva s mogućnostima efikasne analize koju omogućavaju kompjuteri.

U praksi, dva primarna cilja *data mininga* su predviđanje i opisivanje. Predviđanje uključuje korišćenje određenog broja varijabli u skupu podataka za predviđanje nepoznatih ili budućih vrijednosti drugih promjenljivih od interesa. Opisivanje je fokusirano na pronalaženje obrazaca koji opisuju podatke, kako bi se omogućilo njihovo tumačenje. Stoga se *data mining* aktivnosti mogu klasifikovati u jednu od navedene dvije osnovne kategorije (Kantardzic, 2020):

1. Prediktivni *data mining*, koji proizvodi model sistema opisan datim skupom podataka ili
2. Deskriptivno istraživanje podataka, koje proizvodi nove, netrivijalne informacije na osnovu raspoloživog skupa podataka.

32

U zavisnosti od toga da li se radi o prediktivnom ili deskriptivnom zadatku, *data mining* tehnike se mogu označiti kao nenadgledane (eng. *unsupervised*) ili nadgledane (eng. *supervised*). Prva kategorija tehnika koristi se za rad sa skupovima podataka koji nemaju ciljni atribut (eng. *label*), pri čemu je cilj izvući najviše mogućih informacija iz dostupnih podataka, dok se druga koristi kada ciljni atribut

postoji, te je cilj da se dati podaci koriste za predviđanje vrijednosti tog atributa za slučajeve koji još nisu viđeni. Detaljan opis ovih funkcija i *data mining* metoda koje im pripadaju biće predstavljen u poglavlju 4.

U skladu s prethodno navedenim, kao osnovne *data mining* funkcije mogu se navesti (Bramer, 2020):

1. Nadgledano učenje – klasifikacija i regresija (prediktivne tehnike),
2. Nenadgledano učenje – klasterizacija i asocijativna pravila (deskriptivne tehnike).

Relativni značaj predviđanja i deskripcije za određene *data mining* primjene može značajno da varira. S tim u vezi, različiti autori na različite načine grupišu i navode primarne zadatke *data mining* procesa. Na primjer, *Kantardžić* (2020) ističe sljedeće zadatke kao najznačajnije:

- Klasifikacija - otkrivanje funkcije prediktivnog učenja koja klasifikuje stavku podataka u jednu od nekoliko unaprijed definisanih klasa;
- Regresija - otkrivanje prediktivne funkcije učenja koja mapira stavku podataka u promjenljivu predviđanja realne vrijednosti;
- Klasterizacija – deskriptivni zadatak, u kome se nastoji identifikovati konačan skup kategorija ili klastera da bi se opisali i grupisali podaci;
- Rezimiranje - dodatni deskriptivni zadatak, koji uključuje metode za pronalaženje kompaktnog opisa za skup (ili podskup) podataka;
- Modeliranje zavisnosti - pronalaženje lokalnog modela koji opisuje značajne zavisnosti između promjenljivih ili između vrijednosti obilježja u skupu podataka ili u dijelu skupa podataka;
- Otkrivanje promjena i odstupanja - otkrivanje najznačajnijih promjena u skupu podataka.

Na sličan način osnovne *data mining* zadatke opisuju *Berry* i *Linoff* (2004) i navode sljedeće zaključke: klasifikacija, regresija, predikcija, grupisanje afiniteta, klasterizacija i opis i profilisanje.

U poslovnoj sferi, *data mining* se može koristiti za otkrivanje novih trendova kupovine, planiranje strategija ulaganja, procjenu profitabilnosti klijenata, poboljšanje efikasnosti marketing kampanja, pri čemu se rezultati mogu koristiti da se klijentima pruži fokusiranija podrška i pažnja koju oni zahtijevaju i očekuju na savremenom tržištu. *Data mining* može pomoći donosiocima odluka u identifikaciji ključnih klijenata na koje se mogu fokusirati, pružanju postojećim potrošačima preporuka za dodatne artikle tokom procesa kupovine, kao i identifikaciji vrijednih kupaca koji su spremni da napuste kompaniju (Sheshasaayee & Logeshwari, 2018). *Data mining* tehnike mogu pružiti uvid u to koliko će potencijalni klijenti postati profitabilni kada postanu aktivni kupci, kao i koliko dugo će ostati aktivni i koliko je vjerovatno da će napustiti kompaniju zbog konkurentske ponude.

Upravo zbog značajne primjene u poslovnom okruženju, u literaturi se navodi da *data mining* nije samo tehnološki proces, već proces kombinovanja biznisa i tehnologije, odnosno poslovni proces koji tehnologija omogućava (Xiahou et al., 2016). Naglasak na upravljanju odnosima s klijentima tokom posljednje decenije, marketing funkciju čini idealnom oblasti za primjenu *data mining* alata, koja potencijalno može imati velike koristi u procesu podrške odlučivanju. Dakle, *data mining* metode omogućavaju dobijanje odgovora na pitanje koji faktori utiču na posmatranu poslovnu činjenicu i zašto, što predstavlja ključni momenat u rješavanju poslovnih dilema (Kastratović & Kašćelan, 2009).

Pored navedenih prednosti i očiglednih motiva za široku primjenu *data mining* alata, potrebno je navesti i određene nedostatke ovih procesa. S tim u vezi, ističu se sljedeći: neophodna ljudska interakcija, prekomjerno prilagođavanje modela podacima (eng. *overfitting*), postojanje ekstremnih vrijednosti (eng. *outliers*), ogromna količina podataka, visoka dimenzionalnost i problem bezbjednosti (Kantardzic, 2020).

Uprkos sve jednostavnijem interfejsu i automatizaciji procesa, rudarenje podataka je pogodnije za one sa iskustvom u oblasti statistike, matematike, menadžmenta i operacionih istraživanja, pa se neophodna interakcija analitičara i sistema navodi kao jedan od nedostataka u poslovnim primjenama *data mining* tehnika. S druge

strane, problemi poput prekomjernog prilagođavanja modela podacima, ekstremnih vrijednosti i dimenzionalnosti mogu biti prevaziđeni korišćenjem adekvatnih *data mining* alata, što će biti predmet analize u okviru poglavlja 4. Konačno, važno etičko pitanje pri rudarenju podataka je nepostojanje saglasnosti za prikupljanje i korišćenje podataka ukoliko osobe nisu svjesne da se informacije prikupljaju i kako će se one koristiti (van Wel & Royakkers, 2004). S tim u vezi, u literaturi i praksi aktuelna su pitanja bezbjednosti podataka na mreži, kao i pitanja etike, privatnosti i kontrole ličnih podataka, što se preslikava i na njihovu upotrebu korišćenjem *data mining* tehnika za efikasnije donošenje poslovnih odluka.

1.3 Metode za selekciju - targetiranje kupaca i prediktivni modeli odlučivanja u direktnom marketingu

Identifikovanje i selekcija kupaca za targetiranje u budućim kampanjama, kao i analiza njihovih potreba i obrazaca ponašanja predstavljaju neke od osnovnih ciljeva direktnog marketinga. S tim u vezi, *Jonker et al.* (2006) navode dvije osnovne metode za izbor i targetiranje kupaca: segmentacione i bodovne metode.

Logika segmentacionih metoda je kreiranje homogenih segmenata kupaca na osnovu pretpostavljene reakcije na kampanju direktnog marketinga. Nakon utvrđivanja segmenata, ponuda se plasira onim segmentima za koje se procijeni da imaju najveću vjerovatnoću odgovora. Segmentiranje se najčešće vrši po kodiranim atributima prethodnog ponašanja u kupovini, poput: datuma posljednje transakcije, broja obavljenih transakcija i ukupnog iznosa, odnosno vrijednosti transakcija, tj. *Recency, Frequency* i *Monetary* (RFM) atributa. Segmentaciju kupaca na osnovu RFM atributa uveo je *Hughes* (1994), a ova metoda će detaljno biti opisana u sekciji 3.1.1, gdje će biti utvrđeni sistemi za poboljšanje efikasnosti ove metode, korišćenjem *data mining* tehnika.

Prednost pristupa predloženog u ovom radu ogleda se kroz mogućnost prediktivne segmentacije, odnosno predviđanja pripadnosti novog kupca određenom segmentu na osnovu njegovih karakteristika (Khalili-Damghani et al., 2018; Sarvari et al.,

2016). Prediktivna segmentacija je, dakle, od posebnog značaja za akviziciju novih kupaca, čije podatke kompanija nema u svojoj bazi.

Prema bodovnoj metodi, svaki od kupaca se ocjenjuje određenim brojem bodova na osnovu definisanih kriterijuma, poput predviđene vjerovatnoće odgovora na ponudu ili predviđenog profita koji će biti ostvaren. Nakon toga, vrši se rangiranje kupaca na osnovu ukupno ostvarenog broja bodova, te se za targetiranje u narednoj kampanji biraju oni koji imaju najveći skor. S tim u vezi, najveći broj dosadašnjih istraživanja zasniva se na predikciji vjerovatnoće odgovora kupca na kampanju, odnosno razvoju modela odgovora na kampanju (eng. *customer response model*).

Kao i kod prve grupe metoda, korišćenje *data mining* tehnika može unaprijediti efikasnost modela odgovora na kampanju, a u tu svrhu se najčešće koriste metode logističke regresije, vještačke neuronske mreže (eng. *Artificial Neural Networks* - ANN) i drvo odlučivanja (eng. *Decision Tree* - DT) (Bose & Chen, 2009; Coussement et al., 2015; Han et al., 2012). Naime, kao osnovni cilj razvoja predloženih modela ističe se distinkcija između respondenata, tj. kupaca koji su prethodno odgovorili na kampanju kupovinom i nerespondenata, tj. kupaca koji nemaju zabilježenu transakciju ovim putem.

Međutim, neophodno je naglasiti da visoka stopa odgovora ne podrazumijeva i visoku profitabilnost kupaca (Kim et al., 2008). S tim u vezi, uz procjenu odgovora na kampanju, predlaže se i sprovođenje analize, odnosno predikcije profitabilnosti kupaca – razvijanje modela za maksimizaciju profita (eng. *profit-maximization model*). Na taj način, mogu se identifikovati segmenti visokoprofitabilnih i niskoprofitabilnih respondenata (Cui et al., 2015; Otter et al., 2006).

U narednom dijelu rada biće ukratko opisani problemi koji se najčešće javljaju kod prediktivnih modela odlučivanja – minorna klasa respondenata i asimetrična distribucija njihove profitabilnosti.

1.4 Problemi kod prediktivnih modela odlučivanja: minorna klasa respondenata i asimetrična distribucija njihove profitabilnosti

U cilju razumijevanja potciljeva i obrazloženja hipoteza, potrebno je uvesti i ukratko objasniti probleme koji su karakteristični za prediktivne modele. S tim u vezi, u ovom dijelu rada biće istaknute prednosti *Support Vector Machine* (SVM) metoda za rješavanje ovih problema, tj. biće obrazložen istraživački jaz (eng. *research gap*).

Kada su u pitanju teme prediktivne segmentacije, kao jedan od osnovnih problema ističe se nebalansiranost klasa (eng. *class imbalance*), uzimajući u obzir da je skup pozitivnih primjera po pravilu značajno manji od negativnih. Ukoliko ovu analogiju preslikamo na tržište i kupce, možemo zaključiti da je segment najvrednijih kupaca po pravilu najmanji, a upravo je predikcija segmenta ovih kupaca najznačajnija za kompaniju, po pitanju profitabilnosti kupaca. Osim toga, kod razvijanja modela odgovora na kampanju, broj kupaca koji odgovore na kampanju kupovinom, odnosno stopa konverzije (eng. *response rate/conversion rate*) je relativno mala u odnosu na cjelokupan skup potencijalnih kupaca kojima je ponuda plasirana. Stoga, i u ovom slučaju, uz stopu odgovora obično manju od 5%, možemo zaključiti da se javlja problem minorne klase. U literaturi koja tretira ovaj problem ističe se da kod najvećeg broja prediktivnih klasifikacionih metoda dolazi do pristrasnosti i do pogrešne klasifikacije minorne klase (eng. *misclassification*) (Kim et al., 2013; Miguéis et al., 2017).

Rješenja za problem nebalansiranosti klasa u literaturi najčešće se odnose na uzorkovanje (eng. *sampling*), tj. metode poduzorkovanja (eng. *undersampling*), koje obuhvataju uzimanje podskupova veće klase slučajnim izborom, kao i metode preuzorkovanja (eng. *oversampling*), kojima se generišu sintetički primjeri kao dopuna manjoj klasi. Međutim, pomenute metode imaju određene nedostatke. Naime, primjenom poduzorkovanja, iz analize se u potpunosti isključuje određeni broj podataka iz negativne klase, što može dovesti do toga da konačni skup za analizu adekvatno ne odslikava kompletnu klasu negativnih primjera. S druge strane, preuzorkovanjem se vještački generišu pozitivni primjeri, koji ne moraju

nužno biti realni, te ne moraju odgovarati stvarnim karakteristikama primjera pozitivne klase.

Kako bi se izbjegli problemi izazvani uzorkovanjem, u literaturi se ističe korišćenje SVM metoda, posebno u situacijama kada bazu podataka karakteriše nebalansiranost klasa i linearna neseparabilnost. U ovim slučajevima, SVM metoda pokazuje značajno bolje klasifikacione performanse i može se upotrijebiti i kao pretprocesor za prečišćavanje i odvajanje klasa za druge klasifikatore (Barakat & Bradley, 2010; Diederich, 2008; Farquad & Bose, 2012; Kašćelan et al., 2015; Martens et al., 2007, 2008).

Prethodna istraživanja su ukazala na prednosti SVM metode za rješenje problema preklapanja klasa, kao i za dopunjavanje minorne klase novim, relevatnim primjerima iz veće klase (Farquad & Bose, 2012). Dodatno je utvrđeno da je u poređenju s poduzorkovanjem i preuzorkovanjem, SVM metod značajno efikasniji u balansiranju klasa, te da se SVM može koristiti da prečisti podatke za druge klasifikatore, što čini da standardni klasifikatori pokazuju bolje performanse. Međutim, kao jedan od nedostataka SVM metode, ističe se njegova karakteristika "crne kutije" (eng. *black box*), odnosno njegova nemogućnost da generiše model koji se može interpretirati. S tim u vezi, predlaže se korišćenje hibridnog pristupa, gdje se za ekstrakciju pravila iz SVM klasifikacije koristi DT metod (Barakat & Bradley, 2010; Kašćelan et al., 2015; Rogic & Kascelan, 2019).

Pored problema malog broja pozitivnih primjera za proces klasifikacije, kod regresionih modela za predikciju profitabilnosti kupaca, često se javlja problem malog broja visokoprofitabilnih kupaca u poređenju sa ostalim. S tim u vezi, dolazi do asimetričnosti, odnosno zakrivljenosti (eng. *skew*) distribucije zavisne varijable. Ovaj problem utiče na nekonzistentnost standardnih regresionih metoda, kako linearnih, tako i polinomnih. Prethodna istraživanja su ovaj problem asimetrične zavisne varijable tretirala korišćenjem neke vrste parametarskog regresionog metoda, poput *Ridge* ili *Quantile* regresije ili generalizovanih linearnih modela (Malthouse, 1999; Zhang, 2009). Međutim, ovi modeli zahtijevaju ispunjavanje određenih preduslova, a problemi poput multikolinearnosti regresora, prvenstveno

RFM atributa, često ne zadovoljavaju te zahtjeve. Uz to se javlja i problem specifikacije funkcionalne forme, posebno za interaktivne članove kojih može biti veliki broj za veći broj regresora. U cilju rješavanja pomenutih problema, u literaturi se ističe da *Support Vector* regresija (eng. *Support Vector Regression* - SVR) predstavlja robustan metod za modeliranje nelinearnih zavisnosti sa asimetričnom distribucijom zavisne varijable (Basak et al., 2007; Christmann, 2004).

Uzimajući u obzir istaknute prednosti SVM metode, u ovom radu je predložena za rješavanje problema minorne klase kod modela prediktivne segmentacije, kao i kod modela odgovora na kampanju. Ovaj metod, kao pretprocesor podataka, uvećava broj tačno selektovanih i targetiranih najvrednijih kupaca i respondenata u postojećoj bazi podataka, što utiče na smanjenje nepotrebnih troškova u kampanji, a samim tim i na profitabilnost kampanje. Dodatno, kreiranjem hibridnog *SVM-Rule Extraction* (SVM-RE) metoda, kojim se na osnovu SVM izlaza generiše DT model, dobijaju se eksplicitna pravila o kupcima, ponudi koju preferiraju i njihovom kupovnom ponašanju. Na osnovu generisanih pravila olakšava se uspostavljanje interakcije s kupcima, prilagođavanje ponude i stvaranje baze lojalnih i zadovoljnih kupaca. Dakle, na ovaj način, predloženi modeli odlučivanja ne podržavaju samo selekciju i targetiranje kupaca za kampanju direktnog marketinga, već i jačanje odnosa s njima.

Dodatno, za predikciju profitabilnosti kupaca u ovoj disertaciji predložena je SVM regresiona metoda, koja je u literaturi prepoznata kao efikasna u rješavanju problema nelinearnih zavisnosti i asimetričnosti distribucije profita. Korišćenje SVR metode, s jedne strane, omogućava efikasan izbor najprofitablnijih respondenata u postojećoj bazi podataka o kupcima, a s druge, predikciju očekivane profitabilnosti za potencijalne respondente, odnosno nepoznate kupce, na osnovu njihovih karakteristika. Na ovaj način, dolazi do značajnih ušteda izbjegavanjem masovnog i iracionalnog slanja ponuda, i u krajnjem – povećanja prihoda u kampanji, ako postojeći i potencijalni visokoprofitabilni kupci odgovore na kampanju kompletiranjem transakcije.

Korišćenjem senzitivne analize moguće je utvrditi nelinearne zavisnosti prediktora i profitabilnosti, tj. može se opisati profil najprofitabilnijih kupaca, koji mogu, iako ih je malo, činiti najveći dio prihoda od kampanje. Dakle, *profit-maximization* model baziran na SVM metodi, omogućava izbor specifičnih aktivnosti prilagođenih profilu najprofitabilnijih respondenata i time jača strateške odnose s najvažnijim segmentom kupaca.

Predloženi modeli povećavaju efikasnost procesa odlučivanja u direktnom marketingu, usljed čega se očekuje profitabilnija kampanja, bolja interakcija s najznačajnijim kupcima, lakše kreiranje specifične ponude i veći stepen zadržavanja kupaca.

1.5 Ciljevi istraživanja i hipoteze

Procesi odlučivanja u direktnom marketingu zahtijevaju efikasne modele za selekciju i profilisanje kupaca, kako bi se maksimizirao ostvareni profit od kampanje i poboljšali odnosi s kupcima.

Kvalitet podataka s kojima donosioci odluka u marketingu raspolažu u bazi i izbor efikasnog metoda su ključni za uspješne modele odlučivanja. S obzirom na veliki broj podataka u bazama podataka kompanija, sve više se koriste *data mining* metode za realizaciju efikasnih modela i pomoć u donošenju poslovnih odluka. Ovo posebno važi za *online* kampanje za koje je broj podataka još veći, jer se u bazi evidentiraju svi *online* pristupi korisnika s pripadajućim pokazateljima, kao što su broj klikova, vrijeme provedeno na sajtu i slično. Svi ti podaci su izuzetno važni kod targetiranja *online* kupaca jer govore o jednom novom modalitetu kupovnog ponašanja. Na primjer, dok je kod klasičnog direktnog marketinga veoma važan datum posljednje trgovine, kod *online* targetiranja je podjednako važno kada je kupac posljednji put pristupio sajtu, koliko se zadržao, koje proizvode je pregledao i slično. Takođe, kod *online* kampanje od velikog broja inicijalnih pristupa, veoma mali broj se završi kupovinom (često manje od 1%), što dovodi do ekstremne nebalansiranosti između klase respondenata i nerespondenata.

Osnovni cilj ovog istraživanja je da se definišu efikasni modeli odlučivanja u direktnom marketingu zasnovani na *data mining* metodama, s dobrim prediktivnim performansama i mogućnostima balansiranja klasa, što podrazumijeva da imaju dobre prediktivne performanse, bez obzira na visoku nebalansiranost klasa, tj. da dobro predviđaju *online* kupce koji će najvjerovatnije odgovoriti na direktnu kampanju.

Ovaj cilj biće ostvaren kroz sljedeće pomoćne ciljeve:

1. Definisanje prediktivnog RFM segmentacionog modela koji prevazilazi problem nebalansiranosti klasa (tj. problem najmanjeg segmenta najvrednijih kupaca);
2. Definisanje modela odgovora na kampanju koji prevazilazi problem minorne klase (tj. problem izuzetno malog broja respondenata u odnosu na nerespondente);
3. Definisanje modela za predikciju profitabilnosti respondenata koji prevazilazi problem asimetrične distribucije profitabilnosti (tj. problem malog broja visokoprofitabilnih respondenata);
4. Definisanje modela iz tačke 1 i 2, u kombinaciji sa *ensemble* metodama, kako bi se unaprijedile njihove prediktivne performanse.

Kada je riječ o prvom cilju, problem klasične RFM segmentacije kupaca je što korisnik manuelno definiše na koliko segmenata će kupci biti podijeljeni i koji segmenti će biti selektovani za targetiranje u kampanji direktnog marketinga. U literaturi je potvrđeno da *data mining* klasterizacija automatski identifikuje homogene grupe kupaca, odnosno klastere (Cheng & Chen, 2009), a broj klastera se kod ove metode određuje na osnovu indikatora o performansama klasterizacije, kao što je *Davies-Bouldin* (DB) indeks (Abdi & Abolmakarem, 2019; Khalili-Damghani et al., 2018; Rogic & Kascelan, 2019), čime se obezbjeđuje objektivna segmentacija kupaca. U vezi sa klasterizacijom, autori *Cheng* i *Chen* (2009), u svom istraživanju navode da su RFM atributi su uniformno kodirani, tako da 20% najrecentnijih kupaca dobija ocjenu 5 (najčešće 20%, mada ovaj procenat korisnik može i ručno mijenjati), sljedećih 20% ocjenu 4 i tako dalje. Na isti način su kodirani i atributi za

frekvenciju trgovanja i monetarnu vrijednost kupca. Međutim, s obzirom na to da klasterizacija operiše s numeričkim atributima, ovi atributi, koji su numerički po definiciji, mogu se koristiti bez kodiranja, čime se eliminiše subjektivnost i smanjuje gubitak značajnih informacija, odnosno finih razlika između kupaca po vrijednosti ovih atributa (Rogic & Kascelan, 2019).

Kada su kupci podijeljeni u segmente, sljedeći korak prediktivnog segmentacionog modela je predviđanje pripadnosti kupca odgovarajućem segmentu, na osnovu njegovih karakteristika i podataka o ponudi u kampanji (podaci o proizvodima, popustu i slično). *Cheng* i *Chen* (2009) su kao prediktore uključili i RFM atribute, što dovodi do povećanja tačnosti predikcije, uzimajući u obzir da su segmenti već definisani na bazi ovih atributa. S tim u vezi, njima se može apsorbovati uticaj ostalih prediktora i mogu se izgubiti važne informacije potrebne za targetiranje novih kupaca (Rogic & Kascelan, 2019).

S obzirom na to da je klaster najvrednijih kupaca po pravilu i najmanji, većina prediktivnih metoda dovodi do pogrešne klasifikacije takvih kupaca, zbog čega se javlja problem minorne klase. Pogrešna klasifikacija minorne klase obično dovodi do niskih vrijednosti (koje nekada iznose i nula) za prediktivne pokazatelje ove klase, kao što su *class recall* (procenat tačno klasifikovanih aktuelnih primjera klase) i *class precision* (procenat tačno predviđenih primjera klase).

S druge strane, za interpretaciju SVM klasifikacije potreban je metod koji generiše pravila iz SVM izlaza. Najčešće se u tu svrhu primjenjuje DT metod, koji generiše sveobuhvatni skup jednostavnih „ako-onda" (eng. *if-then*) pravila. Kombinovanjem SVM i DT metoda na ovaj način dobija se hibridni SVM-RE metod, koji povećava tačnost klasifikacije male klase i generiše eksplicitna pravila klasifikacije (Kašćelan et al., 2015; Rogic & Kascelan, 2019). Imajući u vidu ove, u prethodnoj literaturi potvrđene mogućnosti SVM-RE metoda, u cilju povećanja efikasnosti RFM segmentacionog modela, definisane su hipoteze H1 i H2 (sa odgovarajućim pothipotezama):

H1: Primjenom *data mining* klasterizacije povećava se efikasnost RFM segmentacije kupaca i izbjegava se subjektivnost pri izboru broja segmenata;

H1.1: Primjenom *k-means* klasterizacije na nekodiranim RFM atributima i *Davies-Bouldin* indeksa za izbor broja klastera, automatski se realizuje segmentacija kupaca s maksimalnom homogenošću unutar klastera, maksimalnom heterogenošću između različitih klastera i optimalnim brojem klastera;

H2: Primjenom hibridnog SVM-RE metoda povećava se efikasnost targetiranja i predikcije najvrednijih kupaca kod RFM segmentacionog metoda, čime se smanjuju nepotrebni troškovi kampanje, povećavaju ukupni prihodi i na osnovu generisanih pravila, formira se profil segmenta najvrednijih kupaca, koji omogućava efikasniju interakciju s njima;

H2.1: Hibridni SVM-RE metod, koristeći karakteristike kupaca i podatke o proizvodima, predviđa pripadnost kupca RFM klasteru najvrednijih kupaca (minorna klasa) sa *class precision* i *class recall* većim od 50%, tj. prevazilazi problem pogrešne klasifikacije minorne klase;

H2.2: SVM metod, pretprocesiranjem podataka o kupovnim transakcijama, tj. eliminisanjem preklapanja i nebalansiranosti klasa (klastera kupaca), povećava prediktivne performanse DT metoda;

H2.3: DT interpretira SVM model, tj. generiše pravila iz SVM izlaza s visokim stepenom povjerenja (sa *confidence* većim od 80%).

Što se tiče drugog definisanog cilja, kod modela odgovora na kampanju ekstremno je izražen problem minorne klase, jer je procenat korisnika koji odgovore na kampanju obično niži od 5% (Chaffey, 2012; Sprague, 2022). Kod većine klasifikatora, ovaj problem dovodi do izuzetno niskih vrijednosti (često su ove vrijednosti nula) za senzitivnost (procenat tačno klasifikovanih aktuelnih primjera pozitivne klase – u ovom slučaju pozitivna klasa je klasa respondenata, tj. minorna) i za specifičnost (procenat tačno predviđenih primjera pozitivne klase). Takođe, pokazatelj prediktivnih performansi *Area Under the Curve* – AUC (kriva koja

prikazuje odnos između tačno klasifikovanih aktuelnih primjera i pogrešno predviđenih primjera pozitivne klase) može imati vrijednosti koje su jednake ili svega neznatno veće od 0,5 što znači da se model malo razlikuje od nasumičnog pogađanja.

Uzimajući u obzir prethodno obrazložene mogućnosti SVM-RE metoda, da bi se prevazišao ovaj problem, kao i u slučaju hipoteze H2, biće testirane njegove mogućnosti da balansira klase i poboljša prediktivne performanse, u smislu smanjenja pogrešne klasifikacije minorne klase. Model odgovora na kampanju kao prediktore obično uključuje karakteristike kupaca i RFM atribute. Međutim, uključivanjem *web* metrika (*online* ponašanja kupaca) kao prediktora, može se unaprijediti identifikovanje najvjerovatnijih respondenata. Uzimajući ovo u obzir, za povećanje efikasnosti modela odgovora podataka biće testirana hipoteza H3 (tj. odgovarajuća pothipoteza).

H3: Primjenom hibridnog SVM-RE metoda povećava se efikasnost targetiranja i predikcije kupaca koji će najvjerovatnije odgovoriti na direktnu kampanju, čime se postiže ušteda nepotrebnih troškova, povećanje ukupnog profita ostvarenog u kampanji i formiranje profila respondenata na osnovu generisanih pravila koji omogućava efikasniju interakciju s njima;

H3.1: Hibridni SVM-RE metod, koristeći karakteristike kupaca, podatke o proizvodima, podatke o kupovnom ponašanju (RFM atribute) i *web* metrike (u slučaju *online* kampanje) predviđa vjerovatnoću odgovora na kampanju (eng. *binary choice 1/0*) s metrikama senzitivnost i specifičnost većim od 50% i AUC značajno većim od 0,5, tj. prevazilazi problem pogrešne klasifikacije minorne klase.

S obzirom na multikolinearnost RFM atributa kao i zakrivljenu distribuciju profitabilnosti kao zavisne varijable, a imajući u vidu mogućnosti SVM regresije potvrđene u prethodnim istraživanjima, za realizaciju trećeg cilja biće testirana hipoteza H4.

H4: Primjenom *Support Vector Regression* povećava se efikasnost targetiranja i predikcije najprofitabilnijih respondenata, čime se postiže povećanje ukupnog profita od direktne kampanje i formiranje profila najprofitabilnijih respondenata, kako bi se ostvarila efikasnija interakcija sa ovom najvažnijom kategorijom kupaca za kompaniju;

H4.1: SVM regresija, koristeći karakteristike kupaca, podatke o proizvodima, RFM atribute i *web* metrike (u slučaju *online* kampanje), predviđa profitabilnost kupca sa dobrim prediktivnim performansama (*Root Mean Squared Error* – RMSE, tj. greškom manjom od 10% prosječnog iznosa profita i R^2, tj. koeficijentom determinacije većim od 0,5) i na taj način prevazilazi probleme asimetričnosti distribucije zavisne varijable i multikolinearnosti nezavisnih varijabli.

Na kraju, za realizaciju četvrtog postavljenog cilja biće sprovedena komparacija primijenjenog SVM-RE metoda sa *ensemble* metodama (*Bootstrap Aggregating - Bagging, Adaptive Boosting - AdaBoost i Random Forest*), koje su u literaturi prepoznate kao efikasne u povećanju prediktivnih performansi slabih klasifikatora (Dietterich, 2002; Miguéis et al., 2017; Zhang & Ma, 2012). Osnovna ideja ovih metoda je da se generiše više klasifikatora na slučajnim podskupovima podataka, uz model s vraćanjem, što znači da isti podatak može biti uključen kod sljedećeg uzorkovanja. Rezultati se na kraju agregiraju najčešće tako što modeli glasaju i uzima se onaj rezultat za koji je glasalo najviše modela. Kod nebalansiranih klasa, iz veće klase se slučajnim uzorkom bira broj primjera jednak manjoj klasi i tako se pokušava riješiti problem nebalansiranosti. S obzirom na to da se podaci koji ulaze u veću klasu biraju slučajnim izborom, ove metode smanjuju problem gubitka informacija koje mogu biti važne za diferencijaciju između klasa. U literaturi je već potvrđeno da daju bolje rezultate od metoda poduzorkovanja i preuzorkovanja (Galar et al., 2012; Miguéis et al., 2017). Autori *Farquad* i *Bose* (2012) su potvrdili da *Random Forest* (RF) metoda, primijenjena poslije SVM pretprocesiranja, daje značajno bolje rezultate nego prije SVM procesiranja, odnosno ako se primijeni samostalno. I *Kang et al.* (2012) su u svom radu pokazali da se primjenom *ensemble* metoda unapređuju prediktivne performanse kod modela odgovora na kampanju.

Dakle, postavlja se pitanje da li SVM-RE metoda u domenu razmatrane problematike direktnog marketinga daje bolje rezultate kod balansiranja klasa od *ensemble* metoda i može li se i u kojoj mjeri rezultat dobijen SVM-RE metodom popraviti kombinovanjem sa *ensemble* pristupom. S tim u vezi, za četvrti cilj nisu definisane hipoteze, već je formulacija data u vidu sljedećih istraživačkih pitanja:

IP1: Da li se rezultat dobijen SVM-RE metodom može popraviti i u kojoj mjeri, kombinovanjem sa *ensemble* pristupom?

IP2: Da li SVM-RE metoda u domenu razmatrane problematike direktnog marketinga daje bolje rezultate kod balansiranja klasa od *ensemble* metoda?

Uzimajući u obzir da će testiranje biti sprovedeno u domenu digitalnog direktnog marketinga, u cilju definisanja atributa koji imaju najveći uticaj na predikciju, formulisano je i treće istraživačko pitanje:

IP3: Da li je za predikcije kod *online* direktnog marketinga važnije *web* ili kupovno ponašanje korisnika i koji od atributa koji opisuju kupovno ponašanje je najvažniji?

U narednoj sekciji biće predstavljena metodologija istraživanja.

1.6 Kraći prikaz metodološkog pristupa istraživanja

U empirijskom dijelu ovog istraživanja korišćeni su primarni i sekundarni podaci. Primarni podaci su preuzeti iz kompanije *Sport Vision*, koja je regionalni lider u distribuciji sportske opreme. Kao sekundarni podaci za validaciju modela korišćeni su javno dostupni skupovi podataka preuzeti sa interneta. Riječ je o longitudinalnim podacima koji se odnose na duži vremenski period (više kampanja ili više godina). Detaljan opis skupova podataka biće predstavljen u sekciji 5.1.

U istraživanju je korišćen *multi*-metod i *mixed*-metod pristup (Brewer & Hunter, 1989, 2006; Cresswell, 1999; Malina et al., 2011), jer se primjenjuje i kombinuje više

različitih prediktivnih metoda. Korišćen je dizajn istraživačkog eksperimenta, tj. testiranje modela na podacima. Detaljan opis korišćenih metoda biće predstavljen u poglavlju 4, dok će razvijeni koncepti modela biti predstavljeni u sekcijama 4.5.1-4.5.5.

Hipoteze su dedukovane iz postojeće teorije i testirane na različitim skupovima kvantitativnih podataka, što znači da je pristup u istraživanju deduktivni i kvantitativni.

1.7 Doprinos istraživanja

Pregledom literature iz oblasti prediktivnih analiza u marketingu, primjetan je jaz između praktične primjene savremenog direktnog marketinga i zastupljenosti ove problematike u akademskim istraživanjima. Naime, uprkos novoj paradigmi direktnog marketinga, većina istraživanja i dalje poistovjećuje direktni marekting s tradicionalnim kanalima komunikacije – prije svega direktnom poštom i telemarketingom. Tek u posljednjoj deceniji, pojavljuju se istraživanja koja se bave digitalnim oblicima direktne komunikacije, i to, gotovo isključivo, *e-mailom*. Prema saznanjima autora, u ovom trenutku, postoji manji broj istraživanja koja uključuju podatke iz kampanja direktnog marketinga plasiranih putem društvenih mreža, uprkos činjenici da je ovo danas dominantan vid komunikacije za direktni marketing u praksi. Ovaj rad u tom smislu predstavlja dopunu postojeće teorije, jer se upravo bavi modelima za selekciju kupaca kod *online* direktnih kampanja plasiranih putem društvenih mreža.

Očekivani naučni doprinos ovog rada je potvrda novog koncepta tri poznata modela za selekciju kupaca u direktnom marketingu (modela prediktivne RFM segmentacije, modela predikcije odgovora na kampanju i modela za predikciju profitabilnosti kupaca), koji je baziran na SVM metodi. Ovaj koncept treba da poveća efikasnost targetiranja kupaca i profit ostvaren u kampanjama direktnog marketinga, kao i da unaprijedi interakciju i odnose s kupcima. Dakle, u ovom istraživanju biće definisani efikasni prediktivni modeli odlučivanja u direktnom

marketingu, zasnovani na *data mining* metodama, koji prevazilaze nedostatke postojećih modela, nastalih zbog problema minorne klase i asimetrične distribucije profitabilnosti respondenata, što predstavlja i osnovni naučni doprinos ovog rada.

Koncept prediktivne klasifikacije za realizaciju RFM segmentacije i modela odgovora na kampanju, primjenom standardnih klasifikacionih metoda, kao što su logistička regresija, stabla odlučivanja ili neuronske mreže, već je prepoznat u literaturi. Međutim, prisutan je problem minorne klase (mali broja kupaca koji odgovore na direktnu kampanju ili mali broj visokovrijednih kupaca), koji kod većine klasifikatora dovodi do pogrešne klasifikacije ove klase, a upravo je ona najvažnija za uspjeh kampanje. Ovaj problem je u literaturi rješavan na različite načine, korišćenjem tehnika poduzorkovanja, preuzorkovanja *ili ensemble* metodama, što će detaljno biti predstavljeno u okviru sekcije 4.1.4.

S obzirom na to da je SVM u velikom broju prethodnih studija potvrđen kao klasifikator koji može najbolje riješiti problem nebalansiranosti klasa (što je detaljno obrazloženo u sekciji 4.1.3), po prvi put je u ovom radu predložen za rješenje problema minorne klase kod modela za selekciju kupaca u direktnom marketingu. Predloženi koncept definiše efikasnije modele za selekciju kupaca, tako što povećava tačnost klasifikacije za minornu klasu (eng. *class recall*) i na taj način omogućava tačno i precizno targetiranje većeg broja kupaca, čija je vjerovatnoća odgovora na direktnu kampanju najveća. Samim tim, raste i vjerovatnoća za ostvarenje većeg prihoda od kampanje. Takođe, povećanjem prediktivne tačnosti modela za minornu klasu (eng. *class precision*) smanjuje se broj pogrešno predviđenih respondenata, čime se smanjuju nepotrebni troškovi kampanje, koji bi nastali zbog realizacije direktne kampanje prema kupcima koji vjerovatno neće odgovoriti kupovinom (Rogic & Kascelan, 2019). Na ovaj način, kompanije koje koriste *online* direktni marketing mogu povećati ukupni profit od kampanje. Od velikog značaja je i da marketari imaju eksplicitni opis profila kupca koji sa visokom vjerovatnoćom odgovara na kampanju (preko njegovih karakteristika, kupovnog ponašanja, *online* ponašanja i ostalih dostupnih podataka u bazi), da bi mogli efikasno da targetiraju nove potencijalne respondente. Predloženi koncept, baziran na ekstrakciji eksplicitnih *if-then* pravila iz SVM izlaza, omogućava efikasno

targetiranje novih kupaca, koji će vrlo vjerovatno odgovoriti na plasiranu kampanju (Rogic & Kascelan, 2019).

Generisanjem eksplicitnih pravila povećava se efikasnost interakcije s najvažnijim segmentima kupaca: sa onim koji su skoro odgovorili na kampanju, koji najčešće odgovaraju i koji najviše potroše, kao i sa onim koji će najvjerovatnije odgovoriti na kampanju, što rezultira većim stepenom povjerenja, lojalnosti kupaca i jačanjem strateških odnosa s njima. Naime, veoma je važno da marketari imaju eksplicitni opis profila kupca koji sa visokom vjerovatnoćom odgovara na kampanju (preko njegovih karakteristika, kupovnog ponašanja, *online* ponašanja i ostalih dostupnih podataka u bazi), kako bi mogli efikasno da prepoznaju potencijalne respondente, prilagode im ponudu i tako povećaju efikasnost interakcije s njima.

Dakle, uzimajući u obzir problem minorne klase u bazama podataka u direktnom marketingu, u ovom istraživanju SVM se koristi kao pretprocesor koji prečišćava podatke, tj. separira i balansira klase. Za ekstrakciju pravila prediktivne klasifikacije primjenjuje se DT metod na izlazu koji generiše SVM.

Predikcija profitabilnosti kupaca je poznati model za selekciju kupaca, koji se u prethodnim istraživanjima realizovao uglavnom putem standardnih ili specijalnih parametarskih regresionih modela (linearna, generalizovana linearna, *quantile* ili *ridge* regresija). Specijalne regresije mogu u određenoj mjeri riješiti problem zakrivljene distribucije prihoda kao zavisne varijable (veoma mali broj kupaca od kojih se ostvaruje visok prihod). Međutim, i dalje ostaje problem što parametarske metode zahtijevaju ispunjenost određenih preduslova (nepostojanje multikolinearnosti, koreliranosti regresora i slično), što kod ove problematike često nije ispunjeno. Takođe, kategoričke varijable se kod ove predikcije moraju pretvoriti u vještačke (eng. *dummy*) varijable kojih ima onoliko koliko varijable ukupno imaju vrijednosti. Specifikacija funkcionalne forme za parametarske regresione metode u tom slučaju postaje veoma kompleksna, posebno ako je potrebno testirati interaktivni uticaj regresora. Ukoliko zavisnosti između prediktora i prihoda nisu linearne, to dodatno komplikuje izbor nelinearne funkcionalne forme.

S druge strane, SVR je neparametarski regresioni metod (ne zahtijeva specifikaciju funkcionalne forme), koji uspješno modelira nelinearne veze i, zahvaljujući mogućnostima generalizacije (adekvatnim podešavanjem parametara), pokazuje najbolje prediktivne performanse, što je u literaturi potvrđeno i obrazloženo u okviru sekcije 4.1.7. S obzirom na to da je SVR neparametarska metoda, pojedinačni interaktivni uticaj prediktora se može interpretirati senzitivnom analizom. U osnovi SVR metoda je rješavanje optimizacionog problema, zbog čega ova metoda nema probleme koji su karakteristični za metodu najmanjih kvadrata. S obzirom na prisutnu multikolinearnost regresora kod predikcije profitabilnosti kupaca i sve pomenute prednosti SVR metode, ona je u ovom radu predložena za povećanje efikasnosti modela za selekciju kupaca na osnovu njihove očekivane profitabilnosti. Iako je SVR već korišćena u literaturi za ovu problematiku, po prvi put se u ovom radu predlaže za analizu i predikciju profitabilnosti kupaca za *online* marketing preko društvenih mreža, što podrazumijeva njeno testiranje na drugačijem i specifičnom skupu prediktora. Predloženi metod povećava efikasnost modela za predikciju profitabilnosti tako što, zahvaljujući izvanrednim mogućnostima generalizacije, generalno povećava tačnost predikcije visokoprofitabilnih kupaca. Upućivanjem direktne ponude takvim kupcima, odnosno targetiranjem određene grupe najvrednijih kupaca, iako ih je po pravilu veoma malo, može se ostvariti veći prihod od *online* kampanje, nego kampanjom prema većem broju kupaca koji su relativno skoro i često obavljali transakcije. Često je profitabilnost kupca obrnuto proporcionalna broju odgovora na kampanju, tj. visokoprofitabilni kupci kupuju rjeđe, ali troše više novca u pojedinačnim transakcijama. Jasno je da je tačna predikcija takvih kupaca od velikog značaja za kompaniju (jedan tačno predviđeni kupac može donijeti ogroman profit za kampanju, dok jedan pogrešno predviđeni visokoprofitabilni kupac, prema kome neće biti upućena ponuda, može značajno smanjiti prihod od kampanje). Senzitivnom analizom se mogu eksplicitno utvrditi karakteristike takvih kupaca i njihovo kupovno ponašanje, što je donosiocima odluka u marketingu izuzetno važno za targetiranje novih kupaca. Za targetiranje kupaca preko društvenih mreža od ključnog je značaja *online* ponašanje kupaca, što je uključeno u koncept modela selekcije kupaca na osnovu profitabilnosti koji je predložen u ovom radu.

Koncept je predložen za modele selekcije kupaca u *online* direktnom marketingu preko društvenih mreža, koji, pored standardnih prediktora, podrazumijeva uključivanje *web* metrika tj. *online* ponašanja kupaca preuzetih sa *Google Analytics* alata i *Facebook Ads Manager* platforme (podaci, kao što su: broj pristupa sajtu, prosječno vrijeme provedeno na sajtu i slično), kao i dodatnih karakteristika kupaca koje se odnose na tehnologiju (tip uređaja s koga se pristupa, operativni sistem i slično) i lokaciju sa koje su pristupili *online* prodavnici. Kod *online* marketinga ovi parametri su podjednako važni, ako ne i važniji za selekciju kupaca od standardnih, kao što su demografske karakteristike kupaca i njihovo kupovno ponašanje. Prema saznanju autora, po prvi put se u ovom radu definiše koncept modela selekcije kupaca za direktni marketing preko društvenih mreža, uz povećanje efikasnosti putem SVM metoda.

Kao rješenje za pomenute nedostatke, kako u klasifikacionim, tako i u regresionim modelima korišćenim u prethodnim istraživanjima, u ovom radu je predložen SVM metod. Ovaj metod je u prethodnim istraživanjima potvrđen kao efikasan u slučaju nebalansiranosti i linearne neseparabilnosti klasa, kao i kod neregularne distribucije zavisne varijable i nelinearnih zavisnosti u slučaju regresije. Međutim, po prvi put se u ovom istraživanju i na ovaj način primjenjuje na probleme iz oblasti direktnog marketinga.

Pored naučnog doprinosa rad ima aplikativni, praktični i društveni značaj.

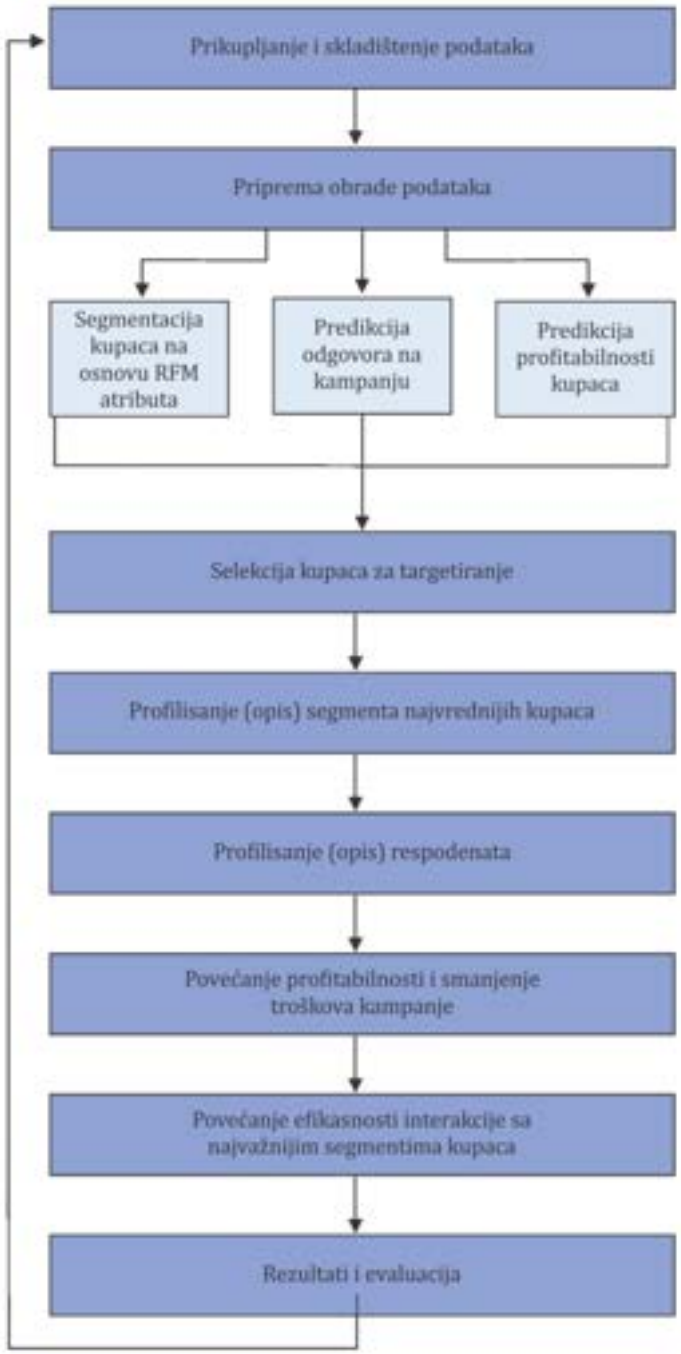
Naime, u radu će biti razvijeni prediktivni modeli u vidu gotovih operatora u *Rapid Miner* alatu, koje kompanije mogu direktno implementirati i koristiti kao podršku pri donošenju odluka u direktnim kampanjama. Međutim, treba napomenuti da upotreba prediktivnih modela u direktnom marketingu uvijek zahtijeva sofisticiranog marketing analitičara koji poznaje *data mining* metode, kako bi se iskoristili svi potencijali ovog pristupa.

S obzirom na to da se pristup bazira na metodama poslovne inteligencije i *data mining* metodama, može se reći da će u ovom istraživanju biti predložen novi konceptualni model, koji daje osnovu da se definiše dio sistema poslovne inteligencije u *online* direktnom marketingu, koji bi povećao efikasnost selekcije

kupaca i omogućio ostvarivanje većeg profita od direktnih kampanja. Benefite od ovakvog sistema nemaju samo kompanije, čiji će se poruke i ponude plasirati najzainteresovanijim kupcima, već i kupci, kojima će se, dok koriste društvene mreže, prikazivati promotivni materijali za proizvode ili usluge koje su direktno povezane s njihovim interesima i potrebama. Dakle, ovako osmišljen sistem za podršku odlučivanju u direktnom marketingu doprinosi i kompaniji i kupcima na različite načine, počev od materijalnih koristi, pa sve do kvalitetnijih odnosa između njih.

Realizacijom ovog koncepta u *Rapid Miner* softveru, tj. kreiranjem odgovarajućih procesa za obučavanje i primjenu modela obezbjeđuje se jednostavnija implementcija ovog koncepta za zainteresovane kompanije.

Kao rezultat predloženih modela i prediktivnih procedura, može se uspostaviti efikasniji sistem za direktni marketing u kompanijama, čije su aktivnosti prikazane na Slici 1.



Slika 1. Ilustracija sistema direktnog marketinga

Kako je na početku ilustracije navedeno, kompanija tokom svojih aktivnosti prikuplja i skladišti podatke o kupcima, njihovom kupovnom ponašanju, transakcijama i *online* metrikama. Prilikom planiranja kampanje, kompanija generiše podatke iz svojih baza i priprema ih za analizu korišćenjem predloženih *data mining* modela. Naredni korak odnosi se na primjenu *data mining* modela u cilju segmentacije kupaca, procjene njihove profitabilnosti i procjene vjerovatnoće odgovora na planiranu kampanju direktnog marketinga. Rezultati predloženih modela, detaljno opisanih u sekcijama 5.2–5.6, omogućavaju uvid u znanje skriveno u velikoj količini sirovih podataka, na osnovu kojeg se mogu izvući važni zaključci o odabiru potencijalnih respondenata za targetiranje u narednoj kampanji. Važno je istaći da širina obuhvata kupaca za targetiranje prevashodno zavisi od ciljeva kampanje i budžeta.

Nakon odabira kupaca za targetiranje, plasira se kampanja direktnog marketinga, što je praćeno rezultatima kampanje i njihovom evaluacijom. Nakon svake sprovedene kampanje, baza podataka o kupcima se dopunjava, tako da se ovaj proces može ponoviti za svaku značajniju kampanju, ili, u unaprijed predviđenim vremenskim razmacima, kada se može očekivati da je baza pretrpjela značajne izmjene, te da rezultati modela mogu biti drugačiji i precizniji. Kvalitet podataka u bazi i njihovo pretprocesiranje u velikoj mjeri će uticati na efikasnost modela, a samim tim i na rezultate kampanje, te se kao zadaci od posebnog značaja ističu upravljanje bazama podataka kompanije i priprema podataka za analizu.

Na kraju, s obzirom na to da se u Crnoj Gori tek pojavljuju kompanije koje koriste *online* kampanje direktnog marketinga preko društvenih mreža na osnovu dostupnih podataka o kupcima, rezultati ovog istraživanja ih mogu podstaći da uzmu u razmatranje sve mogućnosti i prednosti ove vrste direktne marketing aktivnosti, u kombinaciji s podrškom u donošenju odluka pomoću metoda mašinskog učenja.

1.8 Struktura teze

Nakon uvodnog poglavlja, u kome su predstavljeni osnovni postulati savremenog direktnog marketinga, *data mining* metode i tehnike, kao i ciljevi, hipoteze i doprinos istraživanja, u narednom poglavlju biće detaljnije opisan direktni marketing s teorijskog aspekta.

Naime, drugo poglavlje obuhvata sveobuhvatan opis direktnog marketinga i definisanje njegove pozicije u cjelokupnom marketing sistemu, kao i razvoj tehnika direktnog marketinga, koji je u ovom istraživanju podijeljen na dva najznačajnija perioda – direktni marketing prije i nakon nastanka digitalnih medija. U okviru drugog poglavlja predstavljeni su i trendovi koji u najvećoj mjeri determinišu razvoj direktnog marketinga u savremenom digitalnom dobu, kao i faktori koji ga oblikuju, poput razvoja baza podataka, društvenih mreža i menadžmenta odnosa s kupcima (eng. *Customer Relationship Management* – CRM). Posljednji dio drugog poglavlja odnosi se na konkretne aktivnosti koje se sprovode u direktnom marketingu, gdje su u okviru posebnih sekcija opisani: prikupljanje i priprema podataka, segmentacija i kreiranje profila kupaca, odabir kupaca za targetiranje, *cross-selling* i *up-selling* tehnike, planiranje strategija direktnog marketinga i procjena odgovora na kampanju, zaključno sa evaluacijama performansi kampanje, kao posljednjeg koraka u ovom procesu.

U trećem poglavlju detaljnije su predstavljene metode za segmentaciju, selekciju i targetiranje kupaca, koje su od posebne važnosti za sistem direktnog marketinga, kako tradicionalnog, *offline*, tako i savremenog, *online* direktnog marketinga. U okviru ovog poglavlja detaljno su opisana tri podsistema direktnog marketinga, koja će biti predmet empirijskog istraživanja ove doktorske disertacije: metode za segmentaciju kupaca, metode za predikciju odgovora na kampanju, kao i analiza profitabilnosti kupaca. S tim u vezi, analizirane su kako tradicionalne, tako i savremene *data mining* metode za segmentaciju i targetiranje kupaca.

Četvrto poglavlje predstavlja detaljan prikaz *data mining* metoda, koje će biti korišćene u empirijskom dijelu ovog rada. U posebnim sekcijama, opisane su

metode: *k-means* klasterizacije, drvo odlučivanja, *Support Vector Machine*, *Support Vector Regression* i *ensemble* metode. U ovom dijelu rada detaljnije su objašnjena i dva problema koja se javljaju u bazama podataka o kupcima koje se koriste za direktni marketing – problem nebalansiranosti klasa i problem asimetrične distribucije. S tim u vezi, naznačeni su i načini za prevazilaženje ovih problema korišćenjem SVM odnosno SVR metoda. Nakon navedenih sekcija, navedeni su i opisani i pokazatelji koji se koriste za evaluaciju klasifikacionih i regresionih modela. Pored toga, u ovom poglavlju je predstavljen i proces kros-validacije, korišćen za testiranje prediktivnih performansi modela, kao i *Grid-Search* tehnika, korišćena za izbor optimalne kombinacije parametara u modelu.

Završni dio četvrtog poglavlja odnosi se na predstavljanje koncepta prediktivnih modela koji se predlažu u ovom istraživanju u cilju unapređenja sistema direktnog marketinga:

1. Koncept modela prediktivne RFM segmentacije, tj. segmentacije na osnovu stepena vrijednosti kupaca, baziran na klasterizaciji i *SVM-Rule Extraction* metodi;
2. Koncept modela prediktivne RFM segmentacije, baziran na klasterizaciji, *SVM-Rule Extraction* i *ensemble* metodama;
3. Koncept modela odgovora na kampanju, baziran na SVM-RE metodi i *web* metrikama;
4. Koncept modela odgovora na kampanju, baziran na balansiranim *ensemble* metodama (kombinovanjem metode poduzorkovanja i *ensemble* pristupa) i *web* metrikama;
5. Koncept metoda za targetiranje najprofitabilnijih kupaca, baziran na SVR metodi i *web* metrikama.

Peto poglavlje se odnosi na realizaciju i empirijsko testiranje modela čiji su koncepti predloženi u četvrtom poglavlju. Na početku ovog poglavlja detaljno su opisani svi skupovi podataka koji su korišćeni u empirijskom dijelu disertacije. Nakon opisa podataka, predstavljeni su rezultati obučavanja, testiranja i validacije predloženih prediktivnih modela. U okviru pojedinačnih sekcija, pored dobijenih rezultata i

njihove diskusije, navedeni su zaključci, prednosti i doprinosi svakog od razvijenih modela. Konačno, u posljednjem dijelu petog poglavlja, diskutovani su cjelokupni rezultati istraživanja doktorske disertacije i potvrđene hipoteze.

2. KONCEPTUALNI OKVIR DIREKTNOG MARKETINGA

Marketing, kao oblast, doživio je svoje „zlatno doba" tokom šezdesetih i sedamdesetih godina prošlog vijeka (American Association of Advertising Agencies, 2022; Nemhauser, 2014), zbog toga što kompanije nisu mogle da prodaju velike količine proizvoda masovne proizvodnje na tržištu. U ovom periodu, okarakterisanom dubokim sociološkim, političkim i ekonomskim preokretima, kreativnost je bila veoma tražena. Savremene kompanije se susriječu sa sličnim izazovima, iako je tržišno i konkurentsko okruženje u potpunosti izmijenjeno. S tim u vezi, istraživači iz *McKinsey* konsultantske agencije istakli su da se svijet nalazi na rubu novog „zlatnog doba" marketinga, pri čemu nauka, podaci i modeliranje imaju jednu od ključnih uloga (Gordon & Perrey, 2015). Takođe, sam koncept marketinga je promijenjen, pa u fokusu više nije samo privlačenje, već zadržavanje i stvaranje baze lojalnih kupaca (Ravald & Grönroos, 1996). *Kotler* i *Armstrong* (2008) su marketing definisali kao *"proces stvaranja vrijednosti i izgradnje odnosa s potrošačima, u cilju ostvarivanja prodaje, profita i dugoročnog kapitala u vidu lojalnih kupaca"*.

Sa razvojem cjelokupnog sistema marketinga, a u cilju stvaranja održive konkurentske prednosti, došlo je do razvijanja marketinga odnosa, a s njim i savremenog direktnog marketinga, u odnosu na prethodni sistem, poznatiji kao transakcioni marketing. Poređenja radi, *Kotler* je 1972. godine marketing definisao kao *"nauku koja pokazuje kako se transakcije stvaraju, podstiču, olakšavaju i vrednuju"* (Kotler, 1972), što se značajno razlikuje od savremenog definisanja marketinga i definicije navedene u prethodnom pasusu. Orijentisanost na transakcije je i razlog zašto se ova epoha marketinga u literaturi još naziva i "transakcionim marketingom". Cilj transakcionog marketinga bio je, prije svega, kratkoročna akvizicija kupaca i maksimizacija vrijednosti transakcija, uz jednako tretiranje svih kupaca, bez obzira na njihovu vrijednost za kompaniju ili prethodno ponašanje u kupovini (Jüttner & Wehrli, 1994; Stone & Mason, 1997).

Promjena fokusa u marketingu dovela je do potrebe za stvaranjem veza i dubljih odnosa s kupcima. Ovaj proces je u savremenom marketingu podržan informacionim tehnologijama, odnosno bazama podataka o kupcima, na osnovu kojih se mogu otkriti profili profitabilnih kupaca, pratiti njihovo ponašanje i stavovi. Korišćenjem znanja iz baza, u velikoj mjeri je olakšan i objektivizovan proces donošenja odluka u direktnom marketingu.

Direktni marketing predstavlja metod marketinga zasnovan na bazi podataka, koju čine zapisi o individualnim kupcima. Ova baza predstavlja osnovu za marketing analizu, planiranje, implementaciju programa i kontrolu svih povezanih aktivnosti. S druge strane, marketing u tradicionalnom smislu bio je orijentisan na stvaranje brendova za sve pojedinačne proizvode iz asortimana, te naknadno obezbjeđivanje tržišnog učešća za te proizvode. Ukoliko uporedimo ova dva pristupa, važna prednost direktnog marketinga ogleda se u korišćenju baze podataka, koja prenosi fokus s proizvoda na potrošače (Tapp et al., 2014). Pored ove razlike, može se navesti i dodatna – potencijalni potrošač se kroz direktni marketing kontaktira s ciljem dobijanja direktnog odgovora. Naime, ova forma marketinga obuhvata korišćenje komunikacije "jedan na jedan", te se odgovori svih pojedinačnih potencijalnih potrošača mogu mjeriti. Cilj savremenog direktnog marketinga je obezbjeđivanje podataka o kupcu nakon prve kupovine, što, kroz prilagođavanje ponude u skladu s karakteristikama kupca, ponašanja u kupovini i samih proizvoda, omogućava stvaranje veze s njim. Kompanije počinju da se udaljavaju od svojih tradicionalnih strategija masovnog marketinga u korist ciljanih marketinških akcija (Burez & Van den Poel, 2009).

Dizajniranjem komunikacija zasnovanih na analiziranim preferencijama potrošača, uz postojanje jasno definisanih ciljnih grupa, optimizuje se trošak medija, izbjegava rasipanje poruka, kao i masovni marketing.

2.1 Pozicija direktnog marketinga u sveobuhvatnom marketing sistemu

Na početku poglavlja je važno napraviti razliku između pojmova koji se u literaturi nerijetko poistovjećuju s direktnim marketingom. Jedna od najvećih zabluda odnosi se na izjednačavanje direktnog marketinga i direktne pošte (Tapp et al., 2014). Naime, direktni marketing je metod u okviru marketinga i predstavlja cjelovit sistem marketinga. Direktna pošta predstavlja veoma značajan medij, koji se zbog toga koristi u sistemu direktnog marketinga. Pored direktne pošte, direktni marketing se sprovodi i *online*, preko društvenih mreža, telefonskim putem, kao i putem drugih elektronskih i digitalnih medija.

Pored ove zablude, direktni marketing se često poistovjećuje i sa neželjenom poštom (eng. *junk/spam mail*). Neželjena pošta je vrsta direktne pošte koja je loše targetirana, neadekvatnog kvaliteta i snishodljiva, a često i kombinacija sve tri karakteristike. Nažalost, ova pojava je široko rasprostranjena, što urušava ugled savremenog sistema direktnog marketinga.

Dodatno, direktni marketing se često u literaturi navodi kao dio integrisanih marketing komunikacija. Direktni marketing svoje korijene vuče iz naručivanja putem pošte, što zapravo predstavlja sistem distribucije, a ne komunikacije. Savremeni sistem direktnog marketinga funkcioniše kroz baze podataka, s ciljem prikupljanja, održavanja i analize podataka, na osnovu kojih se omogućava stvaranje efikasnih strategija komunikacije s kupcima.

U literaturi i praksi, pojam direktnog marketinga se često koristi u kombinaciji sa sljedećim pojmovima i aktivnostima: marketing zasnovan na bazama podataka (eng. *database marketing*) (Blattberg et al., 2008; Ładyżyński et al., 2019; Seller & Gray, 1999; Zahay et al., 2009), menadžment odnosa s kupcima (eng. *customer relationship management*) (Fletcher & Peters, 1997; Hasouneh & Ayed Alqeed, 2010; Reutterer et al., 2006; Winer, 2001), marketing s direktnim odgovorom (eng. *direct-response marketing*) (Mallin & Finkle, 2009), marketing vođen podacima (eng. *data driven marketing*) (Jeffery, 2010; Mulvenna et al., 1998), "jedan na jedan" marketing

(eng. *one-to-one marketing*) (Peters, 1998), personalizovani marketing (eng. *personalised marketing*) (Cheung et al., 2003; Kaniewska-Seba & Pilarczyk, 2014) i interaktivni marketing (eng. *interactive marketing*) (Frank Mulhern, 2010; Shankar & Malthouse, 2006, 2007; Wang, 2021). U nastavku biće navedene definicije direktnog marketinga vodećih svjetskih autora iz ove oblasti.

Jednu od najcitiranijih definicija dao je *Bird* (1989), koji je direktni marketing definisao kao *"svaku aktivnost koja stvara i koristi direktan odnos između kompanije i kupca kao pojedinca"*. Ovom definicijom, *Bird* je proširio koncept direktnog marketinga, od jednostavnog dijela komunikacionog miksa, do sistema koji može da stvori odnos i održava vezu s kupcima (Tapp et al., 2014).

Tapp, Whitten i Housden (2014) dali su sljedeću definiciju: *"Direktni marketing je način za privlačenje, zadržavanje i razvoj kupaca i, pri tome, zadovoljavanje potreba kupaca, kao i organizacije koja ih opslužuje. To se ostvaruje kroz pružanje okvira za tri aktivnosti: analiza podataka o pojedinačnim kupcima, formiranje strategije i njena implementacija, tako da kupci direktno mogu reagovati kroz različite online i offline kanale i medije."*

Fill i Turnbull (2016) su direktni marketing definisali kao *"strategiju koja se koristi za stvaranje ličnog i neposrednog dijaloga s kupcima. To bi trebalo da bude mjerljiva aktivnost, koja se često zasniva na medijima, a koristi se s ciljem stvaranja i održavanja odnosa sa obostranim koristima."*

Iz navedenih definicija zaključuje se da direktni marketing zahtijeva određenu klasifikaciju i odabir potrošača, prema kojima će se usmjeriti odgovarajući marketing napori, kroz personalizovano oglašavanje i prilagođavanje ponude.

Direktni marketing i generalni marketing razvijaju se sa istim ciljem – zadovoljenje potreba potrošača kroz kreiranje superiorne ponude na tržištu. Stoga, direktni marketing, zajedno sa ostalim marketing metodama, radi na stvaranju konkurentske prednosti. Međutim, kao osnovna prednost direktnog marketinga, navodi se mogućnost obezbjeđivanja specifičnih proizvoda specifičnim grupama

kupaca, kao i snižavanja transakcionih troškova, uz širok spektar komunikacionih kanala (Liao et al., 2011), posebno uz razvoj elektronske trgovine i društvenih mreža, koji u sve značajnijoj mjeri determinišu marketing strategije. Analitičari su procijenili da će finalni potrošači u digitalnim transakcijama tokom 2020. godine potrošiti oko 933 milijarde američkih dolara, dok će kompanije u istom periodu potrošiti preko 9,1 biliona (Laudon & Traver Guercio, 2017).

2.2 Koncept, značaj i primjena savremenog direktnog marketinga

Tržišno orijentisane kompanije u sve većoj mjeri prihvataju direktni pristup u marketingu, fokusirajući se na potencijalne kupce, koji će s većom vjerovatnoćom obaviti kupovinu podstaknutu ponudom iz plasirane kampanje. S druge strane, u okviru tradicionalnog pristupa, promotivne aktivnosti se adresiraju starim i potencijalnim kupcima, bez sprovedene segmentacije i targetiranja. Značajan rast investicija u direktni marketing tokom posljednje dvije decenije uslovljen je većom mogućnosti za maksimizaciju profita, kroz targetiranje potencijalih kupaca s najvećom vjerovatnoćom odgovora na kampanju (Barwise & Farley, 2005).

Sistem direktnog marketinga u različitim organizacijama ima različite oblike, koji odslikavaju strategiju i njihove konkretne ciljeve. *Fill* i *Turnbull* (2016) navode četiri osnovna oblika direktnog marketinga:

- **Komplementarni alat** - mediji koji omogućavaju direktni odgovor koriste se kao komplement ostalim elementima i aktivnostima komunikacionog miksa, uz primarni cilj generisanja lidova (eng. *lead*). Generisanje lidova predstavlja proces pretvaranja potencijalnih u stvarne kupce. S tim u vezi, osobe koje su vidjele oglas direktnog marketinga koji ih je zainteresovao, te su kontaktirali kompaniju sa upitima u vezi s cijenom, uslovima plaćanja ili slično, mogu se označiti kao lidovi;

- **Primarni diferencijator** - mediji koji omogućavaju direktni odgovor predstavljaju primarnu formu komunikacije s potencijalnim potrošačima, a koriste se u cilju stvaranja jasne diferencijacije od konkurentskih marketing napora. Ova forma omogućava snižavanje troškova, izbjegavanje korišćenja posrednika i doseg do precizno targetirane publike;
- **Prodajni kanal** - treći tip direktnog marketinga sprovodi se s ciljem razvoja veće efikasnosti, kao i načina za unapređenje trenutnih usluga – različiti prodajni kanali se mogu koristiti za zadovoljavanje potreba različitih grupa kupaca;
- **Jačanje brenda** - na ovom nivou se, uz prepoznatu tržišnu šansu, razvijaju brendovi, stvaranjem cjelokupne organizacione kulture zasnovane na stvaranju odnosa s potrošačima.

Kompanije traže nove načine za diferencijaciju svog brenda, proizvoda i usluga, pa je kreiranje superiornog i efektivnog korisničkog iskustva, koje je, pri tome, troškovno isplativo, jedan od osnovnih izazova na savremenom tržištu (Deloitte Consumer Review, 2016; Lemon & Verhoef, 2016).

Direktni marketing podrazumijeva procjenu veličine tržišta, konkurentskog okruženja i preferencija potrošača, kao i dizajniranje profila potrošača (pojedinaca ili segmenata), u cilju razvijanja strategije promocije proizvoda i usluga namijenjene tim potrošačima. Direktno komuniciranje s kupcima ili prodajnim prilikama, korišćenjem telefona, *online* kanala ili direktne pošte, omogućava odabir potrošača za komunikaciju, u odnosu na poruke plasirane putem masovnih medija, gdje se komunicira s velikim brojem ljudi. Dodatno, poruka plasirana na ovaj način omogućava direktan odgovor – „poziv na akciju" (eng. *call to action*). Naime, poziv na akciju predstavlja cilj sprovođenja kampanje (najčešće obavljanje kupovine). Još jedna važna karakteristika direktnog marketinga je mjerljivost. Ovaj marketing sistem najčešće zahtijeva od potencijalnog kupca akciju koja se može kvantifikovati, poput: klika na link *web* sajta, naručivanja proizvoda *online* putem, pozivanja besplatnog telefonskog broja za više informacija, korišćenja kupona ili promotivnog koda pri obavljanju transakcije i slično (Chun, 2012). S tim u vezi, svaka od

navedenih akcija kupca, koja se može smatrati odgovorom na kampanju, može se pratiti, mjeriti i kvantifikovati, a rezultati se najčešće čuvaju u određenoj bazi podataka. Dakle, uspjeh svake direktne kampanje može se mjeriti pomoću različitih metrika u zavisnosti od cilja kampanje. Za potrebe empirijskog istraživanja u okviru ovog rada, kao odgovor na kampanju smatraće se obavljena kupovina, kao konačni cilj sprovođenja kampanje. Međutim, u literaturi i praksi, kao odgovor na kampanju može se uzeti u obzir i set drugih metrika, kao što su: klik na oglas, kontaktiranje kompanije putem društvenih medija, posjeta *web* stranice kompanije i druge prethodno navedene akcije.

Uzimajući u obzir karakteristiku mjerljivosti, direktni marketing je u jednom dijelu kvantitativna disciplina. S tim u vezi, *Basye* (2008) navodi osnovne matematičke koncepte koji se moraju poznavati u cilju dizajniranja efikasne direktne kampanje:

- **Trošak medija ili trošak na hiljadu** (eng. *cost per mille* - CPM) – ovaj pokazatelj računa se tako što se podijeli iznos novca utrošen u kampanji na svakih hiljadu prikaza;
- **Stopa odgovora** – procentni pokazatelj broja transakcija u odnosu na ukupan broj kontaktiranih potrošača. U praksi se, odgovorom, opet u zavisnosti od cilja komunikacije, mogu smatrati različiti ishodi (klik, posjeta, zadržavanje na stranici, poziv i sl);
- **Trošak po odgovoru** – ukupan budžet kampanje podijeljen s brojem ostvarenih transakcija[1];
- **Životna vrijednost potrošača** – definiše vrijednost potrošača za kompaniju, što determiniše iznos novca koji se može investirati u akviziciju tog potrošača (Blattberg et al., 2008). Postoji više načina za procjenu životne vrijednosti potrošača, kao i sličnih pokazatelja, kao što su profitabilnost potrošača, vrijednost potrošača i druge. U empirijskom dijelu ovog rada biće predstavljen model za procjenu i predviđanje profitabilnosti potrošača u direktnom marketingu;

[1]Kao i za prethodnu tačku, osim transakcija, pod odgovorom se, u zavisnosti od cilja komunikacije, mogu smatrati različiti ishodi.

- **RFM model** - jedna je od najkorišćenijih tradicionalnih metrika za evaluaciju kupaca. RFM modele više od trideset godina primjenjuju praktičari direktnog marketinga, kako bi targetirali specifične grupe kupaca, smanjili troškove kampanje i povećali profit (Aghdaie, 2016). Detaljni opis RFM metodologije biće predstavljen u sekciji 3.1.1.

Mjerljivost kao karakteristika kampanja direktnog marketinga je prirodno dovela do razvijanja različitih kvantitativnih modela. *Nash* (1984) je definisao osnovne faktore uspjeha kampanje direktnog marketinga: ponuda, elementi komunikacije, vremenski okvir i selekcija kupaca. Navedeni elementi predstavljali su motivaciju za istraživanje kvantitativnih modela u ovoj oblasti (Bose & Chen, 2009). Kvantitativni modeli koji se primjenjuju u direktnom marketingu pretežno se odnose na aktivnosti vezane za: inpute (prikupljanje podataka o kupcima), procesiranje (selekcija ciljne grupe, kreiranje profila kupaca i prodajnih strategija zasnovanih na njima) i autpute (evaluacija rezultata) (Wasson, 2005). Tačnost predikcije i adekvatna interpretacija rezultata omogućavaju donosiocima odluka da s većom preciznošću izvrše segmentaciju i targetiraju potencijalne potrošače, kao i da identifikuju varijable koje mogu imati značajan uticaj na stopu odgovora na kampanju. Dakle, analiza prethodnog ponašanja potrošača omogućava predviđanje buduće profitabilnosti (Leick, 2007).

2.3 Razvoj direktnog marketinga do pojave digitalnih medija

Direktni marketing nastao je kao forma distribucije za izdavačke kuće, književne klubove i druge kompanije koje su omogućavale poručivanje putem pošte (McCorkell, 1992). Ekspanzija direktnog marektinga, od poručivanja putem pošte do savremenog sistema marketinga, počela je u Sjedinjenim Američkim Državama tokom sedamdesetih godina prošlog vijeka. Od nastanka ovog koncepta, direktni marketing je evoluirao od specifičnog procesa čiji je cilj kreiranje transakcije od strane potrošača u aktivnost u koju su uključene praktično sve velike organizacije.

Najbrži stepen razvoja ovog koncepta ostvaren je nakon osnivanja IDM instituta (eng. *Institute for Data and Marketing*) (Webber, 2013).

Jedna od malobrojnih kompanija koja je uočila potencijal ovog sistema u to vrijeme bila je *American Express*. Iako je danas prepoznata kao kompanija iz oblasti finansijskih usluga, kompanija *American Express* je osnovana 1850. godine kao kurirska služba koja je dostavljala dragocjenosti u državi *New York*, dok je od 1891. godine počela da nudi putne čekove (American Express, n.d.). Najznačajnija promjena u kompaniji desila se 1915. godine, kada je počela da nudi usluge agencije za putovanje (Roque, 2014). Zatim, od 1968. godine ova kompanija razvija putnički časopis na bazi pretplate, koji je, između ostalog, promovisao njihove finansijske usluge, dok su postojeći korisnici finansijskih usluga bili upućivani na časopis. Od tog momenta, za kompaniju *American Express* poštansko sanduče nije više bilo samo mjesto za ostavljanje pisama i kataloga, već i važan kanal za razmjenu informacija s klijentima (Case, 2015). U saradnji sa agencijom *Wunderman, Ricotta & Kline*, dizajniran je prvi program lojalnosti s ciljem zadržavanja lojalnih kupaca (Centro de Documentación Publicitaria, 2020; Markethink, n.d.).

Sve do kraja sedamdesetih godina prošlog vijeka, telefonska tehnologija nije bila dovoljno sofisticirana i rasprostranjena da bi se formirali centralizovani *call*-centri koji bi bili ekonomski opravdani (AdAge, 2013). Uz razvoj naprednih tehnologija, eliminisana je potreba za manuelnim upravljanjem i transferom poziva, te je automatski telefonski sistem sve više dobijao na značaju. Uporedo sa ovim promjenama, telemarketing je zauzimao sve veće učešće u marketing budžetima kompanija, uprkos visokim troškovima i slaboj kompjuterskoj pismenosti. S tim u vezi, telemarketing je tokom osamdesetih godina prošlog vijeka po prvi put prevazišao troškove direktne pošte (Air Marketing Group, 2018). Dakle, ovaj period karakteriše telemarketing, kao dominantan medij za sprovođenje aktivnosti direktnog marketinga.

Od ranih osamdesetih godina prošlog vijeka počele su da se odvijaju i druge značajne promjene u marketing teoriji i praksi, što se posebno ogledalo kroz usvajanje pristupa stvaranju dugotrajnijih, dubljih i smislenijih odnosa s kupcima (Robert &

Shelby, 1994). S obzirom na to da je satisfakcija potrošača predstavljala prepoznati pokretač uspjeha kompanija, u fokusu novog marketing sistema nalazio se lojalni kupac, tj. stvaranje i održavanje dugoročnih veza s kupcima, koje je moguće ostvariti kroz stvaranje superiorne vrijednosti za potrošača (Berger & Nasr, 1998). Prepoznavanje potreba potrošača, njihova selekcija i kreiranje jedinstvenih ponuda za specifične segmente, omogućeno je kroz sprovođenje direktnog marketinga, što je i ubrzalo njegov razvoj.

U evropskim zemljama nije došlo do značajnije ekspanzije direktnog marketinga do 1980-ih godina, kada su u Velikoj Britaniji kompanije iz sektora finansija počele da koriste evidencije o kupcima (Henley Centre, 1995). Finansijski sektor slijedile su i telekomunikacione kompanije, a direktni marketing su krajem osamdesetih godina u svoj program uvrstile i humanitarne organizacije.

Najznačajniji razvoj direktnog marketinga, koji je zasnovan na podacima, desio se tokom 1990-ih i 2000-ih godina, zahvaljujući fleksibilnijim i pristupačnijim tehnologijama, kao i sve obrazovanijim marketing menadžerima (Tapp et al., 2014). Od 90-ih godina, revolucionarni napredak kompjuterske tehnologije imao je najznačajni uticaj na stvaranje digitalne ekonomije, a upravo u tom periodu personalni računari su bili sve prisutniji u domaćinstvima i kompanijama, što je uslovljeno padom njihovih cijena i sveukupnom dostupnošću tehnologija (Mutula, 2009). Dodatno, tokom devedesetih godina direktni marketing je po prvi put donosiocima odluka mogao obezbijediti novu vrijednost, koja se ogledala kroz preciznost i mjerljivost.

U to vrijeme, kompanije iz sektora avio prevoza, proizvodnje automobila, trgovine, kao i hotelske industrije, uočile su i prihvatile trend korišćenja direktnog marketinga u cilju stvaranja i očuvanja veze s kupcima. S tim u vezi, *Nagel* (2007) ističe da direktna komunikacija s potrošačima podstiče stvaranje dugoročnih odnosa s kupcima zasnovanih na povjerenju.

Sve do pojave digitalnih medija, direktna pošta je bila najzastupljeniji metod sprovođenja direktnog marketinga. Direktna pošta je pružala određene prednosti

kompanijama, poput komunikacije personalizovanih ponuda, nepostojanja direktne konkurencije za pažnju kupca i sposobnosti da se kupac uključi u jednostavan proces (Verhoef, 2003). Pored ovih prednosti, navodi se i fleksibilnost u pogledu formata, vremena i procesa testiranja (Vriens et al., 1998), kao i stimulacija interesovanja, mogućnost uključivanja dodatnih materijala i karakterisanje iskustva kao opipljivog za potencijalnog kupca. S druge strane, osnovni nedostaci se ogledaju u visokim troškovima direktne pošte za svakog pojedinačnog potrošača (posebno, ukoliko se uporede sa alternativnim digitalnim medijima danas), što zahtijeva značajnu stopu odgovora, kako bi se osigurala profitabilna implementacija kampanje (Gázquez-Abad et al., 2011).

Tokom posljednje dvije decenije, stvaranje inovativnih komunikacionih kanala u marketingu omogućilo je jednostavnije i efikasnije dostavljanje personalizovanih poruka odabranoj grupi kupaca. Novi, digitalni kanali postali su značajne komponente programa direktnog marketinga u mnogim organizacijama (Roach, 2009; Trappey & Woodside, 2005; Xu, 2005). Praksa i primjena direktnog marketinga je proširena, tako da danas obuhvata veći broj različitih aktivnosti u fazama akvizicije i zadržavanja kupaca (Webber, 2013).

2.4 Trendovi razvoja direktnog marketinga u digitalnoj eri

Uz razvoj novog, virtuelnog svijeta, mnogi aspekti ljudskih aktivnosti su postali digitalizovani, poput čuvanja zapisa i podataka, procesa kupovine i komunikacije. Tokom 2019. godine, 14,1% ukupne globalne maloprodaje obavljeno je elektronskim putem, a procjenjuje se da će do 2023. godine e-trgovina obuhvatati oko 22% svih maloprodajnih transakcija (Statista, 2020). Zajedno s promjenom okruženja, mijenja se i potrošač. Dvadeset prvi vijek stvorio je novi tip potrošača – osnaženog i osposobljenog da koristi novu tehnologiju. Ovakav potrošač ima pristup informacijama u digitalnom svijetu i biva targetiran od strane kompanija u veoma konkurentskom okruženju. Po posljednjim istraživanjima, preko dvije milijarde kupaca svoje kupovine obavlja *online* preko digitalnih kanala (Statista, 2020).

Dodatno, više od polovine (51%) *online* potrošača obavlja transakcije korišćenjem mobilnog uređaja (Finances Online, 2019). Posljedice ovih promjena su dalekosežne – ne samo da mijenjaju sferu tehnologije i njenog razvoja, već utiču i na domene poslovne strategije i marketinga (Constantinides & Fountain, 2008). Zadatak koji se postavlja pred marketing praktičare je pronalaženje načina za pridobijanje pažnje ovakvog potrošača. Studija koju je sprovela kompanija *Microsoft* ukazala je na promjenu u rasponu pažnje savremenog potrošača, tj. vrijeme koje jedna osoba provede u potpunosti fokusirana na određenu aktivnost. Sa 12 sekundi tokom 2000. godine, raspon pažnje u 2013. godini je smanjen na osam sekundi (Microsoft, 2015). Istraživanja pokazuju da kolektivni globalni raspon pažnje opada zbog opsega informacija koje se plasiraju (Lorenz-Spreen et al., 2019). Velika količina poruka usmjerava se ka potrošaču u realnom vremenu, što dovodi do situacije da se sve manje vremena može posvetiti pojedinačnim porukama. Tokom sedamdesetih godina prošlog vijeka, prosječna osoba bila je izložena broju od 500 do 1.600 promotivnih oglasa dnevno, a uzimajući u obzir da digitalni marketing nije postojao, najveći dio ovih oglasa bio je na bilbordima, u novinama, časopisima i na televiziji. Uz razvoj digitalnog marketinga, ova brojka posljednjih nekoliko godina iznosi između 6.000 i 10.000 promotivnih oglasa dnevno (Forbes, 2017; Protect, 2020). U cilju iznalaženja efikasnih načina komunikacije sa savremenim potrošačima, procesi segmentacije i profilisanja kupaca su od velikog značaja za pripremu strategije.

Tokom posljednje dvije decenije i mediji su bili pogođeni digitalnom revolucijom. Razvoj tehnologije uklonio je mnoge nedostatke štampanih i drugih tradicionalnih medija, što je uticalo i na transformaciju marketing komunikacija. Iako kreativne poruke ostaju srž komunikacije (Mulhern, 2009), sistemi poput postavljanja oglasa koji se zasnivaju na podacima, kao i oglašavanja u okviru *Google* pretrage, odnosno društvenih medija, u potpunosti mijenjaju proces planiranja marketing kampanja. Stoga, donosioci odluka ne mogu ostati indiferentni u vremenu čestih promjena poslovnog okruženja. Kao što je u prethodnom poglavlju navedeno, najznačajniji razvoj ova forma marketinga dostigla je nakon što je tehnologija potrebna za skladištenje i obradu podataka o kupcima postala dostupna širokom krugu korisnika.

Tapp sa grupom autora (2014) faktore koji su doveli do ekspanzije rasta primjene direktnog marketinga u digitalnoj eri dijeli u dvije osnovne grupe – socijalne i tehničke faktore, s jedne, te poslovne faktore, s druge strane.

U socijalne i tehničke razloge za rast primjene direktnog marketinga spadaju (Tapp et al., 2014):

- **Razvoj interneta.** Krajem devedesetih godina dvadesetog vijeka, autori su predviđali da će internet u potpunosti izmijeniti postojeću i ustaljenu marketing praksu (Peterson et al., 1997). Potencijal interneta kao alata za direktan marketing proizilazi iz njegove sposobnosti da poboljša, a ne da zamijeni postojeću vezu između kompanije i njenih kupaca. Imajući u vidu da se tokom prvih 20 godina 21. vijeka najveći dio komunikacije obavlja preko interneta, jasno je da je njegov razvoj ubrzao i razvoj direktnog načina komunikacije s potrošačima. Dodatno, internet je smanjio operativne troškove sprovođenja direktnog marketinga, pa se čak stopa odgovora od 0,5% može smatrati profitabilnom kampanjom sprovedenom putem *e-maila* (Direct Marketing Association, 2009);
- **Fragmentacija društva.** Uz ovu pojavu, jača i težnja za individualizmom, te potrošači zahtijevaju informacije koje su lične, relevantne i pravovremene u određenim situacijama. Ovaj fenomen direktnom marketingu daje prednost, s obzirom na to da je ovaj sistem fleksibilan i dozvoljava diferenciranje ponude za pojedinačne potrošače. Prema istraživanju *DuckDuckGo*, rivala kompanije *Google* koji se fokusira na privatnost, *Google* prati posjetioce *web* sajtova na 86% od 50.000 najbolje rangiranih *web* sajtova na svijetu, dok *Facebook* prati 35% sajtova (Koetsier, 2020). Logika je jednostavna - navedene mreže uče o željama i potrebama korisnika prateći na šta on klikne i kuda se kreće po internetu. Nakon toga, plasiraju oglase za pretraživane stavke ili relevantne usluge na korisnikovoj internet ruti;
- **Širenje medija**. Kao što je pojava kablovske i satelitske televizije doprinijela razvoju masovnog marketinga, tako je i razvoj društvenih mreža, kao što su *Facebook, Instagram, Twitter, LinkedIn* i *Pinterest*, kreirao velika tržišta sa

značajnom mogućnošću segmentacije i targetiranja pojedinačnih korisnika. Novi mediji komunikacije, poput elektronske pošte i društvenih mreža, zamijenili su direktnu poštu i telefon u komunikaciji s potrošačima. S tim u vezi, 45% potrošača dijeli svoja loša iskustva s kompanijama na društvenim mrežama, dok oko 30% dijeli svoja dobra iskustva iz korisničkih usluga (Dimensional Research, 2013);

- **Veća sofisticiranost potrošača**. Sve veća želja potrošača da se tretiraju u skladu sa sopstvenim zahtjevima, dovodi do nužnosti kompanija da segmentaciju tržišta obave na objektivan i efikasan način, dok je u određenim situacijama poželjna i direktna i personalizovana komunikacija sa svim pojedinačnim kupcima. Kupci širom svijeta istakli su da imaju veća očekivanja od korisničkih usluga nego u prethodnom periodu, što smatra 54% svih potrošača (Microsoft, 2017). Pored toga, u istraživanju kompanije *Microsoft*, 96% potrošača navodi da je usluga korisnicima važan faktor u njihovom izboru brenda i lojalnosti prema istom (Microsoft, 2017). S tim u vezi, 89% potrošača ističe da je prešlo na poslovanje s konkurentom nakon lošeg korisničkog iskustva (Nextiva, 2021);
- **Želja potrošača da kontrolišu proces**. Napredak tehnologije stvorio je doba u kojem su kupci osposobljeni za komunikaciju, istraživanje i kupovinu gdje god i kad god žele. Današnji kupci očekuju da kompanije brzo uvode inovacije u skladu s njihovim promjenljivim preferencijama - u suprotnom će proizvod potražiti kod konkurenata. Dodatno, sveprisutnost pametnih uređaja stvorio je kulturu neposrednosti, pri čemu kupci kao pravovremeni odgovor vrednuju isključivo trenutni odgovor – čak 64% potrošača očekuje da kompanije na upite odgovaraju u realnom vremenu (Salesforce Research, 2016).

Parise, Guinan i *Kafka* (2016) su u svom radu opisali pojam "kriza neposrednosti" kao potrebu da potrošači dobiju sadržaj, ekspertizu i personalizovana rješenja u realnom vremenu u procesu kupovine. Današnji potrošač je informisan o proizvodima i uslugama koje želi da kupi i sposoban je da koristi tehnologiju. Pored toga, kako bi obavili jednu transakciju, sve veći broj potrošača kupuje i istražuje

preko više kanala, kao što su tradicionalne prodavnice, web sajtovi, socijalne platforme i mobilne aplikacije (Parise et al., 2016). *McPartlin* i *Dugal* (2012) ističu da čak 86% globalnih potrošača kupuje bar preko dva kanala, dok potrošači koji kupuju *online* troše više novca (Maxwell, 2013) i smatraju se profitabilnijim (Graeber, 2013). Uzimajući ovo u obzir, jasna je težnja kompanija da s potrošačima stvore interakciju, kreiraju ponudu na individualnom nivou i ulažu u direktnu komunikaciju. Kompanije na savremenom tržištu ne mogu biti pasivni posmatrači u nadi da će njihov proizvod pronaći put do potrošača, već moraju ulagati napore u neposrednu komunikaciju i personalizovani sadržaj (Parise et al., 2016). Kompanije, dakle, prepoznaju potrebu potrošača za "instant" informacijama, a digitalni mediji kroz direktno komuniciranje podižu komunikaciju s potrošačima na novi nivo.

Prednosti korišćenja interneta za B2C transakcije su očigledne, uzimajući u obzir otvorenost i mogućnost izbora kako širih, tako i užih tržišta, sve do mikro niša. Kompanije koriste internet za komunikaciju, transakcije i kao distributivni kanal, a razvoj elektronske trgovine rezultirao je stvaranjem novog tipa prisustva kroz novu formu - virtuelna prodavnica (Chen et al., 2002), koja podržava instant komunikaciju i informacije.

U poslovne razloge za razvoj direktnog marketinga ubrajaju se: jačanje konkurencije, kritike tradicionalnih marketing metoda, povećano interesovanje za zadržavanje kupaca i stvaranje grupe lojalnih kupaca, kao i kontinuirani pad troškova kompjuterske obrade podataka (Tapp et al., 2014).

Marketing putem interneta uglavnom postaje jedna od vodećih strategija velikog broja kompanija, s ciljem promocije proizvoda i interakcije s kupcima. Uz dobro razumijevanje potrošača koji kupuju *online*, njihovog ponašanja i stavova, kompanije mogu stvoriti efikasne i efektivne strategije za privlačenje novih kupaca (Suki & Suki, 2013) i zadržavanje postojećih. Brojna istraživanja su pokazala da akvizicija novih potrošača košta pet do šest puta više od zadržavanja postojećih (Athanassopoulos, 2000; Bhattacharya, 1994; Colgate & Danaher, 2000; Verbeke et al., 2011). U prilog tome, *Sabbeh* (2018) navodi da je zadržavanje postojećih kupaca od pet do čak 20 puta troškovno efikasnije od privlačenja novih. Dodatno,

Rosenbloom (2003) ističe da su kompanije koje su u svoje strategije uključile direktni marketing upravo one koje su izabrale da koriste direktnu komunikaciju, kako bi upotpunile svoje tradicionalne marketing metode. Jačanje konkurencije navodi kompanije da ulažu sve veći napor u kreiranje kreativnog advertajzing sadržaja, kao i da pronađu inovativne pristupe u optimizaciji i minimiziranju troškova *online* oglašavanja (Jackson & Ahuja, 2016; Semerádová & Weinlich, 2019). Razvoj elektronske trgovine još više je podstakao borbu za pažnju potrošača, s obzirom na to da je konkurent udaljen samo jedan "klik". S druge strane, razvoj interneta i e-trgovine proširio je mogućnosti za sprovođenje kreativnih marketing aktivnosti i stvorio veliku količinu podataka o kupcima, koja je raspoloživa kompanijama za analizu. Kroz analizu podataka o potrošačima, kompanije mogu definisati njihove preferencije i, posredno, unaprijediti sistem za donošenje odluka (Liu & Shih, 2005). U tom smilsu, personalizacija poruka, zasnovana na znanju generisanom iz podataka o kupcima, može biti ključ uspješne kampanje. *Shanahan* s grupom autora (2019) navodi da personalizovani sadržaj pozitivno utiče na stvaranje interakcija s brendom, jačanje veze, kao i lojalnost. Pored personalizacije, važni faktori uticaja na atraktivnost kampanje za kupce su i stepen poznavanja brenda, vizuelna atraktivnost oglasa, kvalitet pruženih informacija i veća uključenost proizvoda, što doprinosi povećanju namjere korisnika da nastavi dalju interakciju s brendom (Yu et al., 2020). Pravila tradicionalnog direktnog marketinga i novog, koji se sprovodi preko digitalnih *online* kanala, veoma su slična, međutim, digitalni mediji su efikasniji u pogledu troškova i targetiranja (Vest, 2013).

Kao osnovne grupe novih, digitalnih medija, *Shankar* i *Hollinger* (2007) su naveli: nametljive – gdje je korisnik "prekinut" u nekoj aktivnosti od strane oglašivača; nenametljive – gdje potrošač odlučuje da li želi da odobri dalju komunikaciju sa oglašivačem; generisane od strane korisnika – gdje sami korisnici kreiraju komunikaciju. U okruženju bogatom podacima, kao što je internet, od posebne je važnosti zaštita privatnosti potrošača. S tim u vezi, nenametljivi sistemi komunikacije (eng. *opt-in campaign*), koji se sprovode nakon odoborenja potrošača da prihvati dalju komunikaciju, sve više dobijaju na značaju. Jedan od osnovnih primjera *opt-in* kampanja su *e-mail* kampanje, poput *newslettera*, koje uključuju

opciju za odustajanje od naredne komunikacije. Dakle, postoji jasna razlika između kampanje direktnog marketinga koja je plasirana putem direktne pošte – gdje se promotivni materijal dostavlja svim lidovima sa određene liste i *opt-in* poruka, koje se dostavljaju isključivo potencijalim kupcima koji su odobrili dalju komunikaciju s kompanijom. Ovaj sistem *opt-in* kampanja je *Seth Godin* (1999) nazvao "marketing odobrenja" (eng. *permission marketing*). Uzimajući sve navedeno u obzir, jedan od glavnih izazova prilikom kreiranja kampanje direktnog marketinga je pronalaženje kombinacije kanala koja će biti personalizovana, pravovremena, adekvatna i lišena nametljivosti, koja se povezuje s komercijalnim aspektom brendova.

Korišćenjem mogućnosti koje nudi informaciona tehnologija, donosioci odluka mogu dizajnirati marketing strategije zasnovane na sposobnosti predviđanja ponašanja u kupovini potrošača i potencijalnih potrošača koji pripadaju relativno homogenom segmentu (Guido et al., 2013).

2.5 Faktori koji oblikuju direktni marketing danas

Vernon (2019) je podijelio marketing u digitalnom dobu na direktni marketing i brend marketing. U direktni marketing ubraja *Google, Facebook, YouTube* i *e-mail* oglašavanje, odnosno svaki tip odglašavanja koji se može mjeriti sve do transakcije. S druge strane, u brend marketing ubraja bilborde, sponzorstva, organizacionu kulturu, dizajn uniforme, organizaciju događaja u cilju umrežavanja i druge slične aktivnosti. Slično njemu, *Fill* i *Turnbull* (2016) navode da se oglašavanje i odnosi s javnošću sprovode s ciljem razvijanja vrijednosti brenda i podizanja svijesti potrošača, dok se prodajna promocija i strategija direktnog marketinga sprovode s ciljem dobijanja odgovora kroz ponašanje potrošača koje se može mjeriti. Dakle, osnovni element i bitan znak prepoznatljivosti direktnog marketinga danas je konkretna ponuda s jasnim pozivom na akciju.

Uzimajući u obzir trend rasta usko targetiranog i čak "jedan na jedan" marketinga, koji je podstaknut brzim napretkom tehnologija baza podataka i novih medija, sve veći broj kompanija prihvata direktni marketing kao način direktnog komuniciranja

s pažljivo odabranim i targetiranim potrošačima (Liao et al., 2011). Razvoj informacionih tehnologija je izmijenio način sprovođenja marketinga, kao i cjelokupan sistem upravljanja podacima o kupcima. S jedne strane, dostupnost podataka o kupcima i novi alati za njihovu analizu kreirali su nove mogućnosti, a s druge, i mnoge izazove u korišćenju tih podataka u cilju stvaranja konkurentske prednosti na tržištu (Shaw et al., 2001). *Big data* era omogućava kompanijama da izgrade dublje, smislenije odnose s kupcima, uz jednostavniji proces donošenja odluka i unapređenje potrošačkog iskustva. Međutim, s obzirom na izazove analiziranja, interpretiranja i smislene prezentacije (Vadrevu et al., 2016), prikupljeni podaci od strane kompanija u najvećem broju slučajeva ne donose vrijednost. Osim toga, tokom vremena nastaju i novi izvori, pa se kao jedan od izazova nameće i efikasna integracija novih kanala, koja može da omogući generisanje ekonomske vrijednosti iz novih podataka (Breur, 2011). Sama dostupnost podataka ne podrazumijeva nužno i njihovu korisnost, te se akcenat stavlja na proces ekstrakcije znanja i korisnih informacija pogodnih za donošenje odluka.

Osim pomenutih, *big data* era donosi i sljedeće izazove:

- **Pravna regulativa u Evropskoj uniji** (eng. *General Data Protection Regulation* – GDPR) umnogome je regulisala mehanizme za prikupljanje podataka koje posjetioci interneta ostavljaju za sobom. Uprkos činjenici da je GDPR osmislila i usvojila Evropska unija (EU), on nameće obaveze svim kompanijama koje targetiraju ili prikupljaju podatke o građanima EU (Wolford, n.d.). Kao takav, GDPR je najstroži zakon o privatnosti i bezbjednosti podataka na svijetu;
- **Pritisak javnosti za većom kontrolom prikupljanja podataka.** Posljednjih godina razvijao se snažan pokret za jaču zaštitu privatnosti potrošača ili, u najmanju ruku, više transparentnosti o podacima koje kompanije dobijaju o svojim klijentima. Kao rezultat toga, realizovana su značajna prilagođavanja velikih IT kompanija (Newman, 2021):

- *Apple* je krajem 2020. godine omogućio korisnicima da biraju koje aplikacije imaju pristup njihovim podacima i koji su to podaci. Ovo je uklonilo značajnu količinu podataka za praćenje, koje su koristile aplikacije za *online* targetiranje, poput kompanije *Facebook*;
- *Google Chrome* će prestati da prikuplja kolačiće (eng. *cookies*) s *web* sajtova 2023. godine, iako je prvobitno najavljeno da je postepeno ukidanje ovog monitoringa trebalo da počne 2022. godine. Međutim, ovo odlaganje motivisano je omogućavanjem kompanijama da razmotre kako da funkcionišu u budućnosti bez kolačića i da osmisle nove načine za targetiranje svojih korisnika;
- *Facebook* je u novembru 2021. godine izjavio da više neće pratiti „osjetljiva pitanja", kao što su: rasa, etnička pripadnost, politička opredjeljenja, zdravstvena zaštita, religija ili seksualna orijentacija. Ove reforme mogu imati značajan uticaj na targetiranje korisnika zdravstvenih organizacija, potencijalnih simpatizera političkih partija i još mnogo drugih organizacija koje su se oslanjale na navedene podatke;

- **Veća upotreba aplikacija za blokiranje reklama** (eng. *ad-blockers*). Preko 38% korisnika interneta blokira oglase da bi zaštitili svoju privatnost (Kirsch, 2021). Ovo predstavlja izazov za kompanije koje na ovaj način targetiraju svoje potencijalne kupce.

Međutim, i pored navedenih izazova, prikupljanje i analiza podataka o korisnicima kompanijama donosi značajne benefite. Naime, analizom podataka o kupcima i njihovim prethodnim transakcijama, mogu se dizajnirati obrasci kupovina za pojedinačne kupce ili grupe kupaca, a na osnovu njih, mogu se razvijati jedinstvene marketing strategije (Liu & Shih, 2005). Dakle, digitalni direktni marketing prevazilazi bezličnu prirodu tradicionalnog marketinga, pa korišćenjem tehnologije omogućava kompanijama da slične grupe kupaca tretiraju na jedinstven način. S obzirom na to da omogućava prilagođavanje ponude postojećim potrošačima, benefiti direktnog marketinga ogledaju se kroz precizno targetiranje, minimiziranje

rasipanja poruka i profitabilnije kampanje, bez stvaranja stalnih troškova i pronalaska novih potrošača.

Kao značajni faktori razvoja direktnog marketinga u savremenom poslovnom okruženju ističu se: napredak sistema i tehnologija baza podataka, potreba za kreiranjem efikasnog sistema za odnos s kupcima, ekspanzija društvenih mreža i njihovog korišćenja u svrhu oglašavanja, kao i automatizacija i vještačka inteligencija. U nastavku rada biće detaljnije predstavljena sva četiri faktora i njihov značaj za razvoj direktnog marketinga.

2.5.1 Baze podataka o kupcima kao stimulans razvoja direktnog marketinga

U eri velike količine podataka (eng. *big data*), eksponencijalni razvoj i popularnost digitalnih medija dovode do stvaranja baza podataka sa ogromnom količinom podataka. Količina podataka koja se stvara na mreži podstakla je stvaranje novih rješenja za njihovo prikupljanje, skladištenje i distribuciju. *Big data* koncept predstavlja jedan od proizvoda digitalne ekonomije, tj. jednu od ekonomskih i tehnoloških karakteristika informatičkog društva (Lazović & Đuričković, 2018). Uz to, dinamika savremenih digitalnih tržišta, poput niskih troškova ulaska, sveopšte dostupnosti alata, niskih cijena oglašavanja i jake konkurencije, stvara izazov selekcije, obrade i strukturiranja podataka u cilju dobijanja uvida u ponašanje potrošača i njihove karakteristike. Napredak tehnologija baza podataka i sistema za njihovo upravljanje utiče da marketing postaje direktniji, fokusiraniji i interaktivniji (Csikósová et al., 2014).

Za razliku od tradicionalne ekonomije, u kojoj su podaci o kupcima bili neintegrisani i na osnovu kojih je bilo izrazito teško praviti dugoročne strategije i predviđati njihovo ponašanje u procesu kupovine, digitalna ekonomija je, sa interaktivnom komunikacijom i sistemom poslovanja na mreži, ponudila u potpunosti nova rješenja i mogućnosti. Rast mreže je uzrokovao eksponencijalni rast obima podataka, kao i tehnologija za njihovo prikupljanje, obradu i pretraživanje, koje kroz

big data sisteme mogu generisati važne i korisne informacije za budući odnos prema kupcima (Lazović & Đuričković, 2018).

Napredak tehnologija baza podataka, kao i njihov značaj za sistem direktnog marketinga u velikoj mjeri su uslovljeni razvojem interneta – povećana je količina podataka dostupna kompanijama, a sam proces prikupljanja podataka je pojednostavljen. Veliki broj organizacija prihvatio je činjenicu da je znanje sadržano u velikim bazama podataka ključni faktor u procesu donošenja odluka, konkretno, znanje o potrošačima iz ovih baza je krucijalno za marketing funkciju (Shaw et al., 2001). Međutim, znanje u bazama često ostaje neotkriveno – uz raspoloživost velike količine podataka, javljaju se i određeni izazovi: filtriranje, sortiranje, procesiranje, analiziranje i upravljanje podacima u cilju generisanja informacija koje mogu biti korisne donosiocima odluka (Shaw et al., 2001).

Pretpostavka razvoja pristupa u marketingu koji je zasnovan na podacima, a samim tim i direktnog marketinga, predstavlja stvaranje detaljne baze podataka o kupcima. Pojava digitalnih medija i društvenih mreža u značajnoj mjeri je povećala raspoloživost podataka o kupcima kroz obezbjeđivanje podataka u realnom vremenu (Mulhern, 2009). *Nunan* i *Di Domenico* (2013) opisali su značaj stvaranja velike količine podataka kroz tri perspektive:

- Tehnološki izazovi - povezani sa skladištenjem, zaštitom i analiziranjem ogromne količine različitih tipova podataka u organizacijama, koji uključuju i razvoj "*cloud*" skladišta podataka i koji zahtijevaju nove tipove analize;
- Dodatna komercijalna vrijednost – kroz stvaranje znanja o potrošačima u cilju kreiranja prodajnih i marketing strategija;
- Društveni uticaj – poput privatnosti, slobode govora, regulacije i odgovarajućih mjera u segmentu etične komercijalne upotrebe ovih podataka, uzimajući u obzir mogućnost praćenja i analize ponašanja potrošača.

Iz fokusa na dobijanje korisnih informacija iz baza podaka o kupcima razvio se pojam "marketing baza podataka". Ova forma marketinga primjenjuje statističku

analizu i informatičke modele na pojedinačnim bazama podataka (Drozdenko & Drake, 2002). *Schafer* s grupom autora (2001) navodi da je marketing baza podataka nastao s ciljem da se pruži personalizovana usluga kupcima u vremenu kada su kompanije počele naglo da se razvijaju i da gube lični kontakt s pojedinačnim kupcima. S tim u vezi, neke kompanije su na ovaj problem odgovorile tako što sve potrošače tretirale na isti način, dok su druge koristile podatke poput zip koda, nivoa dohotka i zanimanja, kako bi prilagodile ponudu određenim segmentima (Schafer et al., 2001).

Raspoloživost ovih i sličnih podataka u kompanijama olakšava i unapređuje primjenu (Mulhern, 2009):

- **Procjene vrijednosti i segmentacije kupaca**. Raspoloživi podaci o kupcima omogućavaju procjenu profitabilnosti kupaca i dizajniranje tržišnih segmenata. Zahvaljujući digitalnim medijima, tradicionalne analize kupaca koje su se sprovodile polugodišnje ili godišnje, zamijenjene su češćim i tačnijim analizama vrijednosti kupaca na osnovu prethodnog i trenutnog kupovnog ponašanja, što omogućava blagovremeno prilagođavanje strategija i taktika komuniciranja (Schafer et al., 2001);
- **Analize odgovora na kampanju**. Mjerenje odgovora na kampanje direktnog marketinga unaprijeđeno je u digitalnom svijetu, s obzirom na mogućnost detaljnog praćenja aktivnosti i svih faza u prodajnom lijevku. Pored toga, kroz konstantno praćenje troškova i stope odgovora mogu se redefinisati taktike komuniciranja u cilju povećanja efikasnosti;
- **Tržišne inteligencije**. Podaci iz digitalnih medija obezbjeđuju brojne mogućnosti za bolje razumijevanje potrošača, konkurenata i samog tržišta. Sa istim ciljem se razvija veliki broj *data mining* metoda i sistema za upravljanje znanjem. *Data mining* alatima mogu se pronaći važne informacije i znanje sakriveno u bazama u cilju boljeg razumijevanja kupaca, dok sistematsko upravljanje tim znanjem omogućava njegovo pretvaranje u efektivne marketing strategije. Unapređenje efikasnosti ovih metoda u dinamičnim i konkurentnim uslovima poslovanja, postaje imperativ

savremenih kompanija i predmet je velikog broja naučnih istraživanja u oblasti direktnog marketinga i *big data* analitike (Cheng & Chen, 2009; Chen et al., 2015). Dakle, istraživanja iz sfere ekstrakcije znanja iz baza podataka i upravljanja tim znanjem je od posebne važnosti za marketing;

- **Finansijskih modela**. Raspoloživi podaci poboljšavaju finansijske marketing metrike, pa se dominantno korišćene marketing metrike, poput svijesti o brendu i stavova potrošača, zamjenjuju inkrementalnim profitom, povratom na investicije (eng. *return on investment* – ROI) i vrijednošću potrošača. Korišćenje finansijskih metrika omogućava preciznije budžetiranje i kreiranje komunikacionih planova.

Podaci o kupcima koji se najčešće prikupljaju, čuvaju i obrađuju su sociodemografski i podaci o kupovnom ponašanju, poput *web* metrika i podataka o transakcijama (Guido et al., 2011; Bose & Chen, 2009). Upravo se podaci o transakcijama najčešće koriste u direktnom marketingu u cilju formiranja klasičnog RFM modela (Malthouse, 1999).

S obzirom na to da su baze podataka o kupcima preplavljene podacima, metode direktnog marketinga postaju sve relevantnije za naučna istraživanja i praksu i u velikoj mjeri se poklapaju sa istraživanjima u oblasti poslovne inteligencije i *data mining* metoda. S tim u vezi, veći broj raspoloživih podataka u bazi poboljšava performanse kampanje (Heilman et al., 2003; Kovčo et al., 2019). Osim toga, korišćenjem istorijskih podataka koji opisuju odgovor na prethodne kampanje, može se procijeniti i predvidjeti broj ili stopa odgovora za planiranu kampanju, što može biti važna informacija donosiocima odluka.

Međutim, velika količina podataka o oglašavanju i komunikacionim aktivnostima i dalje je neistražena, jednim dijelom iz razloga što se ovi podaci nalaze u formatima koji se ne mogu tretirati, procesirati i koristiti (Muñoz-García, 2015).

2.5.2 Menadžment odnosa s kupcima kao faktor razvoja direktnog marketinga

Pri dizajniranju jedinstvenih strategija i korisničkih iskustava u direktnom marketingu, kompanije se mogu osloniti na koncept upravljanja odnosima s potrošačima (CRM). Ključ uspjeha u konkurentskom poslovnom okruženju predstavlja sposobnost identifikovanja profitabilnih kupaca i razvoja dugoročnih odnosa s njima, a CRM omogućava preciznije identifikovanje potreba potrošača zahvaljujući upotrebi tehnologije. Dakle, CRM sistem podržava marketing kroz selekciju i targetiranje potrošača u cilju stvaranja troškovno efikasnih odnosa s njima. S tim u vezi, CRM predstavlja proces otkrivanja obrazaca u ponašanju potrošača i obezbjeđivanja informacija koje će služiti kao podrška u procesu odlučivanja pri akviziciji, zadržavanju i procjeni profitabilnosti kupaca (Sabbeh, 2018; Sin et al., 2005). Dakle, u tradicionalnom CRM okviru, organizacije posjeduju značajne podatke o kupcima, koje koriste u cilju izgradnje odnosa s njima (Wang & Kim, 2017; Verhoef et al., 2010).

Bull (2003) je CRM definisao kao *„sistem koji koristi tehnologiju u cilju prikupljanja, analiziranja i diseminacije informacija o postojećim i potencijalnim kupcima, preciznog identifikovanja potreba kupaca i stvaranja kvalitetnih odnosa s njima"*.

Swift (2001) definiše CRM kao *„poslovni pristup za razumijevanje i uticaj na ponašanje potrošača, kroz smisleno komuniciranje, a u cilju unapređenja akvizicije i zadržavanja potrošača, kao i njihove lojalnosti i profitabilnosti."*

Dodatno, *Parvatiyar* i *Sheth* (2001) su CRM definisali na sljedeći način: *„sveobuhvatna strategija i proces akvizicije, zadržavanja i razvoj partnerstva sa odabranim potrošačima, s ciljem kreiranja superiorne vrijednosti za kompaniju. Uključuje integraciju marketinga, prodaje, korisničkog servisa i funkcija u lancu nabavke organizacije za postizanje veće efikasnosti i efektivnosti u procesu isporuke vrijednosti za potročaša."*

Nešto novija definicija, nastala nakon ekspanzije društvenih mreža, CRM definiše kao *„integraciju aktivnosti orijentisanih na potrošače, uključujući procese, sisteme i*

tehnologije, kao i aplikacije društvenih medija, u cilju stvaranja i jačanja kolaborativnih veza s potrošačima" (Trainor, 2012).

Navedene definicije naglašavaju značaj posmatranja CRM-a kao značajnog i sveobuhvatnog procesa zadržavanja postojećih i akvizicije novih kupaca, najčešće uz pomoć poslovne inteligencije. Dakle, CRM koristi poslovnu inteligenciju za selekciju i targeritanje najprofitabilnijih kupaca, što povećava stopu zadržavanja kupaca i maksimizira vrijednost potrošača. Osim toga, savremeni koncept menadžmenta odnosa s kupcima uključuje i nove aspekte, omogućene razvojem tehnologije i društvenih mreža. Osnov CRM-a predstavlja potreba za ekstenzivnim znanjem o potrošačima, koje će se koristiti u operativnim i analitičkim procesima, dok se strategijski ciljevi CRM-a ostvaruju kroz marketing, prodajne i uslužne procese kompanija. S tim u vezi, ovaj sistem podrazumijeva usklađivanje i integraciju strategijskih ciljeva s poslovnim procesima i informaciono-telekomunikacionom tehnologijom.

Četiri dimenzije CRM-a su (Chagas et al., 2020; Ngai et al., 2009):

- **Identifikacija kupaca** (eng. *customer identificaiton*) – prvi korak CRM okvira, razvijen s ciljem targetiranja potencijalnih kupaca kroz identifikovanje najprofitabilnijih, upotrebom klasifikacionih i segmentacionih metoda. Dakle, korišćenjem metoda klasterizacije i segmentacije, na osnovu istorijskih podataka o kupcima, mogu se kreirati segmenti (klasteri) sličnih kupaca. Segmentacija omogućava podjelu tržišta na grupe, kako bi se omogućile različite analize, vrednovanje i upravljanje različitim grupama. Često se u ovom prvom koraku CRM sistema teži identifikovati najprofitabilnija grupa kupaca (Deepa, 2017);
- **Privlačenje kupaca** (eng. *customer attraction*) – alokacija resursa u cilju targetiranja identifikovane ciljne grupe upotrebom određenih metoda i medija za komunikaciju. S tim u vezi, analiziraju se definisani klasteri iz prve dimenzije. Nakon što se utvrde jedinstvena obilježja i karakteristike segmenata, razvijaju se različite marketing strategije s ciljem privlačenja

kupaca, među kojima su targetirano oglašavanje ili direktni marketing (Kazemi & Babaei, 2011);

- **Zadržavanje kupaca** (eng. *customer retention*) – ulaganje napora u cilju ispunjavanja očekivanja potrošača i zadovoljavanja njihovih potreba. Uzimajući u obzir konkurentnost i kompleksnost savremenog tržišta, kompanije traže načine i razvijaju nove aktivnosti u cilju zadržavanja potrošača (Chuang & Shen, 2008) i obezbjeđenja kvalitetnih usluga, kako tokom pocesa kupovine, tako i nakon kupovine. Zadržavanje kupaca omogućava značajne benefite za organizacije, kao što su troškovna efikasnost u odnosu na akviziciju novih kupaca (Verbeke et al., 2011) i dugoročna profitabilnost lojalnih kupaca (Van den Poel & Larivière, 2004), uz pozitivan „marketing od usta do usta" (eng. *word of mouth - WOM*);
- **Razvoj kupaca** (eng. *customer development*) – težnja ka ostvarivanju povećanog volumena transakcija, vrijednosti transakcija i individualne profitabilnosti potrošača. Najznačajniji djelovi ove dimenzije su analiza cjeloživotne vrijednosti kupca, analiza potrošačke korpe i *up-selling* i *cross-selling* (Oliveira, 2012; Drew et al., 2001). *Cross-selling* predstavlja praksu predlaganja sličnih proizvoda ili usluga kupcima koji će s velikom vjerovatnoćom ostvariti kupovinu, dok *up-selling* označava praksu sugerisanja proizvoda ili usluga većeg kvaliteta, boljih performansi i viših cijena onim kupcima koji već razmatraju određenu kupovinu (Bose & Chen, 2009), poput mobilnog telefona ili televizora najnovije generacije ili, na primjer, dodavanja pomfrita i gaziranog pića uz naručeni burger u *McDonald'su*. Da bi se ove strategije efikasno sprovele, moraju biti zasnovane na predviđenom ishodu ponašanja kupaca, korišćenjem tehnika, kao što su analiza potrošačke korpe, procjena profitabilnosti kupca i slično. Najvećim dijelom se rezultati ovih analiza dobijaju korišćenjem *data mining* tehnika. Dakle, cilj razvoja kupaca, kao dimenzije CRM-a, predstavlja precizno definisanje jedinstvenih strategija targetiranja za različite kategorije kupaca.

Neki od faktora koji utiču na razvoj CRM-a su: razvoj interneta, baza podataka i elektronske trgovine, kao i mogućnost personalizacije promotivnih materijala.

Iz prethodno navedenog, može se zaključiti da direktni marketing predstavlja dio CRM-a. Dakle, CRM uključuje veliki broj aktivnosti, pri čemu direktni marketing predstavlja jednu od njih. Dodatno, iako se CRM i direktni marketing u određenom dijelu tehnologija i alata koje koriste poklapaju, cilj CRM-a je stvaranje korisnih informacija za cijelu kompaniju, a ne samo marketing odjeljenje. Na primjer, neke od razlika ogledaju se kroz aktivnosti praćenja zadovoljstva kupaca kroz CRM i stvaranje strategija u cilju ponovnog pridobijanja kupaca koji su izabrali konkurentske kompanije (Tapp, 2008).

CRM kompanijama olakšava razumijevanje vrijednosti potrošača, selekciju i targetiranje najprofitabilnijih kupaca, kroz razvoj odgovarajućih marketing i prodajnih strategija, čiji je cilj povećanje lojalnosti i profita (Lee & Park, 2005). Najčešće CRM sistemi koriste alate i modele napredne analize podataka u cilju dobijanja informacija o potrošačima i njihovom ponašanju u kupovini, kako bi se predvidio uspjeh kampanje, koji se ogleda kroz stopu odgovora i stopu zadržavanja potrošača. S tim u vezi, koncepti poput vještačke inteligencije (eng. *artificial intelligence* - AI), mašinskog učenja (eng. *machine learning* - ML) i *data mininga* (eng. *data mining* - DM) nude nove tehnike i algoritme, koji mogu procesirati i koristiti velike količine heterogenih podataka (de Sousa et al., 2019). Izlazne vrijednosti ovih sistema, kroz analizu velikog broja raspoloživih podataka, mijenjaju način na koji kompanije stvaraju interakciju sa svojim kupcima (Chagas et al., 2020). Dobijene predikcije iz modela mašinskog učenja koriste se za kreiranje budućih marketing planova i aktivnosti kojim će se targetirati selektovani kupci.

Konačno, jedna od važnih karakteristika CRM-a je mogućnost mikro segmentacije. Ovaj napredni sistem segmentacije omogućava naprednu podjelu tržišta na grupe s dobro definisanim karakeristikama, pri čemu je vizija mikro targetiranja često upravljanje strategijama za pojedinačne kupce (jedan kupac = jedan segment). CRM podržava ovu viziju kroz sisteme personalizacije i prilagođavanja proizvoda (Leick, 2007). S jedne strane, personalizacija se može ogledati kroz jednostavno oslovljavanje kupaca njihovim imenom i prezimenom pri obraćanju. S druge strane, kupcima se kroz direktni marketing može predstaviti ekskluzivna i jedinstvena

ponuda kreirana za svaki segment pojedinačno. Personalizacija predstavlja jedan od važnih faktora diferencijacije na tržištu (Robertson, 2007), što u uslovima tržišne zasićenosti ima poseban značaj.

Istraživanje koje su sproveli *Wang* i *Kim* (2017) naglašava značaj uključivanja društvenih mreža u CRM sistem. Njihovi rezultati ukazuju da je stvaranje novog, socijalnog CRM-a od posebnog značaja kada kompanije uključuju društvene mreže u svoje marketing aktivnosti, posebno za stvaranje interakcije s kupcima, kao i za performanse kompanije.

U empirijskom dijelu ovog rada, u poglavljima 5.2-5.5 biće predstavljeni modeli za segmentaciju kupaca, kao i za predikciju odgovora na kampanje direktnog marketinga, plasiranih putem društvenih mreža *Facebook* i *Instagram*. S tim u vezi, u narednom dijelu rada biće predstavljene osnovne karakteristike društvenih mreža kao kanala komunikacije u direktnom marketingu.

2.5.3 Društvene mreže kao akcelerator razvoja online direktnog marketinga

U literaturi se često navodi da je *web* drugačiji od ostalih medija u smislu fokusa ka budućnosti, interakcije i društvenih karakteristika. *Calder, Malthouse* i *Schaedel* (2009) smatraju *online* iskustvo aktivnijim, participativnijim i interaktivnijim, dok *Rappaport* (2007) navodi da je i sam internet društvene prirode, uzimajući u obzir da se koristi za komunikaciju i razmjenu informacija, zbog čega podstiče društveni angažman.

Društvene mreže omogućavaju (geografski udaljenim) korisnicima da se međusobno povežu i da razmjenjuju tekstualne i multimedijalne poruke i informacije, koristeći tehnologije bazirane na mreži (Oliverio, 2018). *Boyd* i *Ellison* (2008) su društvene mreže definisali kao "*servise zasnovane na mreži, koji omogućavaju pojedincima da kreiraju javni ili privatni profil u okviru sistema, definišu listu ostalih korisnika sa kojima žele da podijele vezu, te da pregledaju listu konekcija kreiranih od strane ostalih korisnika sistema*". Ova definicija obuhvata dva važna aspekta društvenih mreža – kreiranje individualnih profila i aktivno angažovanje u

komunikaciji s drugim korisnicima. Platforme društvenih mreža su fokus svojih razvojnih strategija postavile na dijeljenje sadržaja i povezivanje s drugima (Alaimo, 2014). Uporedo sa ekspanzijom društvenih mreža, razvijala se i nova era mreže – *Web 2.0*, koji se bazira na sadržaju koji generišu sami korisnici. Ovaj sistem *Constantinides* i *Fountain* (2008) u svom radu definišu kao "*kolekciju otvorenih, interaktivnih online aplikacija koje kontrolišu korisnici, u cilju širenja iskustava, znanja i tržišne moći samih korisnika, kao učesnika u poslovnim i društvenim procesima*". Dakle, kao i u novoj eri marketinga, korisnik se nalazi i u fokusu nove ere mreže.

Uz razvoj društvenih mreža značajno se izmijenio način komunikacije kako između samih potrošača, tako i između potrošača i kompanija. S tim u vezi, društvene mreže su otvorile mogućnost kompanijama da pristupe potrošačima na novi način, te da kroz ovaj medij dizajniraju ponudu koja će zadržati potrošače. Društveni mediji su nepovratno izmijenili marketing komunikacije, mijenjajući načine na koji potrošači biraju, dijele i ocjenjuju informacije. S obzirom na to da su društvene mreže stvorile komunikacioni medij i platformu za oglašavanje s niskim troškovima, uz veliku bazu podataka o kupcima, rapidno su se razvile i mogućnosti za kompanije da sprovode personalizovani advertajzing (Ertugan, 2017), uz interakciju i komunikaciju s potrošačima u realnom vremenu (Mize, 2009; Palmer & Koenig-Lewis, 2009). Potrošači su postali fragmentisani i proaktivni, što predstavlja izazov za donosioce odluka i kreatore komunikacionih strategija. Međutim, kompanijama su na raspolaganju *online* aplikacije za prikupljanje, analiziranje i korišćenje podataka o potrošačima u cilju preciznijeg targetiranja i prilagođavanja poruka. Potrošači, uz to, sve češće koriste društvene mreže u cilju dobijanja informacija prilikom odličivanja o kupovini (Vollmer & Precourt, 2008).

Brzina internet komunikacije i brojni izvori informacija čine oglašavanje na tradicionalnim izvorima sve manje relevantnim (Duffett, 2015). Ekspanzija društvenih mreža i njihova efikasnost u komuniciranju, dovela je do toga da štampani časopisi i novine smanjuju tiraže iz godine u godinu. Osim toga, karakteristike novih komunikacionih procesa s potrošačima, poput interaktivnosti i

mogućnosti komuniciranja u realnom vremenu, omogućava kompanijama da blagovremeno analiziraju odgovor korisnika i da na osnovu njega prilagode dalje operativne korake (Gurău, 2008).

Bilo je samo pitanje vremena kada će kompanije i marketing agencije za oglašavanje prepoznati društvene mreže kao vodeći medij za direktni marketing. Kao marketing alat, društveni mediji pružaju kompanijama jedinstvenu mogućnost da pristupe potrebama, stavovima, interesovanjima, sklonostima i željama svojih potencijalnih potrošača, kao i njihovim obrascima kupovine (Maurer & Wiegmann, 2011; Oliverio, 2018). Na ovaj način, omogućeno im je da objektivno segmentiraju tržište, prilagode sadržaj pojedinačnim segmentima, te da izgrade pouzdan i lojalan odnos s kupcima. *Rosemann* je s grupom autora (2012) predložio tri koraka za korišćenje podataka s društvenih mreža u cilju donošenja strategijskih odluka u kompaniji: društveno slušanje (šta se komunicira na kanalima društvenih medija); analiza podataka (otkrivanje obrazaca u podacima); društveno angažovanje (donošenje odluka i sprovođenje akcionog plana).

U izvještaju "*Digital 2020*", u januaru 2020. godine godine zabilježeno je da više od 4,5 milijardi ljudi koristi internet, dok su društvene mreže prevazišle 3,8 milijardi korisnika (We Are Social & Hootsuite, 2020). Prosječni korisnik interneta provodi 6 sati i 43 minuta na mreži svakog dana.

Dio budžeta predviđen za društvene mreže u savremenim kompanijama ima trend rasta s obzirom na vrijeme koje korisnici provode na društvenim mrežama (Lee et al., 2018). Tokom 2019. godine, prosječno vrijeme koje korisnici interneta provode na ovim platformama globalno je 144 minuta, što je za dva minuta više nego prosjek iz 2018. godine (Statista, 2020). Osim toga, donosioci odluka sve češće biraju "bihejvioralno targetiranje" kroz fokus na razvijanje personalizovanih poruka, baziranih na ponašanju korisnika na internetu i njihovoj lokaciji dobijenoj preko GPS sistema mobilnih telefona (Winer, 2009), što predstavlja podatke koji se mogu dobiti iz baza društvenih mreža.

Neki autori su naveli da društveni mediji imaju važnu ulogu i uticaj na proces donošenja odluka o kupovini (Rrustemi et al., 2020). Na primjer, sa 2,91 milijardi aktivnih korisnika mjesečno, *Facebook* predstavlja najveću društvenu mrežu na svijetu (Statista, 2021b). Svi ti korisnici su subjekti marketinga na društvenim mrežama i svakodnevno su targetirani u raznim oglasnim kampanjama.

Međutim, postavlja se pitanje – da li su te kampanje i marketing aktivnosti na društvenim mrežama zasnovane na subjektivnoj menadžerskoj procjeni ili na objektivnom pristupu zasnovanom na podacima? U ovom radu je predstavljen pristup segmentaciji kupaca zasnovan na podacima i *data mining* metodama. Kao jedna od brojnih upotreba *data mining* metoda, ova studija predstavlja primjer procesa odlučivanja zasnovanog na podacima u podjeli tržišta i procjenjivanju različitih segmenata tržišta pojedinačno, uzimajući u obzir specifičnosti svakog od njih, što bi omogućilo kompanijama da kreiraju diferencirane i prilagođene marketing strategije za sve definisane segmente.

S obzirom na to da će se empirijski dio istraživanja bazirati na podacima iz baze podataka o *Facebook* i *Instagram* oglašavanju, u nastavku poglavlja biće opisane karakteristike *Facebooka* kao platforme za oglašavanje, tipovi i formati objava koje se mogu kreirati za ostvarivanje specifičnih ciljeva, kao i prednosti koje platforma nudi.

2.5.3.1. Karakteristike i prednosti Facebook platforme za oglašavanje

Koristeći *Facebook*, kompanije mogu komunicirati sa svojim potrošačima na dva načina: organsko komuniciranje (besplatno) i plaćeno oglašavanje (Curran, Graham & Temple, 2011). Imajući u vidu opseg dostupnih demografskih i drugih podataka, *Facebook* omogućava detaljnu segmentaciju, targetiranje adekvetnih segmenata i pristup željenoj širini tržišta, što ga čini jednom od najpopularnijih platformi za oglašavanje na društvenim mrežama (Bannister, Kiefer & Nellums, 2013). Targetirani korisnik *Facebooka* može kliknuti na plaćeni oglas, koji ga najčešće preusmjerava na *web* stranicu kompanije (Cvijikj & Michahelles 2013; Logan 2014), gdje je omogućena *online* kupovina. Pored toga, kroz direktnu komunikaciju s

ciljnim tržištem, *Facebook* omogućava interakciju i otkrivanje potreba i želja potrošača (Sanne & Wiese, 2018; Bannister, Kiefer & Nellums, 2013). *Ertugan* (2017) je u svom istraživanju naveo da ispitanici smatraju da je *Facebook* efikasan u stvaranju veza između kompanije i potrošača, te da predstavlja kvalitetan alat za promociju proizvoda.

Kako bi se izbjeglo rasipanje poruke, što predstavlja jedan od osnovnih nedostataka masovnog marketinga, sprovodi se precizno targetiranje potrošača, kome prethodi objektivna segmentacija tržišta. *Facebook* ima mogućnost da obezbijedi visok stepen kvaliteta usko targetiranih oglasa na svojoj platformi, što ga, u poređenju sa ostalim društvenim mrežama, čini najboljom platformom za oglašavanje (Pace Technical, n.d.). U prvom kvartalu 2020. godine osam miliona aktivnih oglašivača je koristilo *Facebook* kao platformu za promociju svojih proizvoda i usluga, što je više u odnosu na sedam miliona oglašivača u prvom kvartalu 2019. godine (Statista, 2020). Korišćenjem *Facebook* plaforme u svrhe oglašavanja, marketing menadžeri mogu da dosegnu jednu trećinu svjetske populacije starije od 18 godina, odnosno više od polovine odraslih ljudi širom svijeta između 18 i 34 godine (We Are Social & Hootsuite, 2020).

Facebook kao platforma za oglašavanje nudi različite opcije za postavljanje, targetiranje i formatiranje, koje omogućavaju kompanijama da eksperimentišu sa stotinama kombinacija pri podešavanju oglasa (Facebook, 2019). Međutim, u zavisnosti od kreativnog sadržaja koji se koristi, kao i samih kriterijuma za targetiranje, uspjeh pojedinačne kampanje može varirati uprkos suženom i preciznom targetiranju (Ládrová, 2018). Osim toga, nisu svi tipovi targetiranja univerzalno pogodni za sve ciljeve marketinga – konkretno, advertajzing sistem *Facebooka* trenutno razlikuje tri kategorije marketing ciljeva (Facebook, 2019): svijest (eng. *awareness*), razmatranje (eng. *consideration*) i konverzija (eng. *conversion*). S tim u vezi, ukoliko targetiranje nije precizno odrađeno, veoma kreativna i originalna ponuda može rezultirati niskom stopom odgovora, dok, s druge strane, slabo formulisana i osrednje kreativna ponuda koja je upućena pravoj ciljnoj grupi može umanjiti, ali ne i eliminisati poželjan odgovor potrošača (Stone &

Jacobs, 2008). Stoga, razumijevanje preferenci i potreba potrošača predstavlja značajniji faktor u kreiranju kampanje od samog kreativnog procesa i načina komunikacije ponude.

U zavisnosti od ciljeva kampanje, *Facebook* nudi 11 marketing ciljeva (Newberry, 2019):

- Svijest o brendu (eng. *brand awareness*): upoznati ciljanu publiku s brendom;
- Doseg (eng. *reach*): izlaganje oglasa što većem broju korisnika;
- Promet (eng. *traffic*): usmjeriti korisnike ka određenoj *web* stranici ili aplikaciji;
- Interakcija (eng. *engagement*): podstaći korisnike na interakciju na stranici kompanije („lajk" objave ili stranice), povećati broj prisutnih na događaju koji organizuje kompanija ili podstaći korisnike da iskoriste specijalnu ponudu;
- Instaliranje aplikacije (eng. *app installs*): podstaći korisnike da instaliraju aplikaciju kompanije;
- Video pregledi (eng. *video views*): povećati broj pregleda video sadržaja;
- Generisanje lidova (eng. *lead generation*): dodavanje informacija o novim prodajnim prilikama u prodajni lijevak, tj. proces, čiji je cilj pretvaranje potencijalnog klijenta u lojalnog kupca;
- Poruke (eng. *messages*): ohrabriti korisnike da kontaktiraju kompaniju putem *Facebook Messengera*;
- Konverzije (eng. *conversions*): podstaći korisnike na određenu akciju na sajtu – najčešće da obave kupovinu;
- Kataloška prodaja (eng. *catalog sales*): povezati *Facebook* oglas s katalogom proizvoda, kako bi se korisnicima prikazivali oglasi za proizvode koje bi najvjerovatnije željeli da kupe;
- Promet ka prodavnici (eng. *store traffic*): usmjeriti korisnike ka lokaciji tradicionalne (*offline*) prodavnice.

Dakle, uprkos shvatanju određenog dijela praktičara iz oblasti marketinga da je komuniciranje putem *Facebooka* isključivo pogodno za razvoj svijesti o brendu, *Tate* (2014) navodi da se kombinacijom brend marketinga i direktnih kampanja mogu

ispuniti oba cilja – jačanje brenda i stvaranje konverzije. Kroz stvaranje WOM i viralnog marketinga, jača se imidž brenda, a posredno i podstiče namjera za kupovinu kod potencijalnih potrošača (Dehghani & Tumer, 2015).

Prethodno navedeni ciljevi mogu biti realizovani kroz dizajniranje oglasa, koji se može prikazati na *News Feedu*, *Messengeru*, sekciji *Videos on Facebook*, kao i na *Instagram News Feedu*. (Facebook, 2019). Bilo koju sekciju za prikazivanje oglasa da kompanija izabere, mora se pridržavati pravila koja se tiču sadržaja i formata. Na ovaj način, *Facebook* teži da osigura održavanje svoje osnovne uloge – socijalizacije, uz visok kvalitet oglasa i sprečavanje širenja ilegalnog biznisa (Ngo, 2019).

2.5.3.2. Načini i formati Facebook oglašavanja

Prema podacima sa *Facebooka*, postoji sedam formata oglasa koji se mogu plasirati preko ove platforme (Facebook, 2020):

- Slika (eng. *image*) - omogućava kompanijama da kroz jednu sliku predstave željeni sadržaj. Slika se mora sastojati od manje od 20% teksta, dok se oglas može povezati URL adresom do stranice proizvoda ili početne stranice *web* sajta kompanije;
- Karusel (eng. *carousel*) – ovaj format omogućava da se u okviru jednog oglasa prikažu dvije ili više slika i/ili video zapisa, ili poziva na akciju. Korisnici se mogu kretati kroz karusel kartice prevlačenjem na mobilnim telefonima ili tabletima ili klikom na strelice na ekranu kompjutera;
- Kolekcija (eng. *collection*) - format kolekcije uključuje trenutno iskustvo i olakšava korisnicima da vizuelno otkriju, pregledaju i kupuju proizvode i usluge sa svog telefona. Prikazani oglas sadrži četiri proizvoda pod slikom osnovnog vizuala ili video zapisa, koji se otvara u trenutnom doživljaju preko cijelog ekrana kada se klikne na oglas;
- Trenutni doživljaj (eng. *instant experience*) – u skladu s prethodno navedenim, trenutni doživljaj je format prikazivanja sadržaja preko cijelog ekrana, koji se otvara nakon što korisnik klikne na oglas koristeći mobilni uređaj;

- Video – ovaj format omogućava prikazivanje proizvoda, usluga ili brenda korišćenjem video sadržaja;
- Priče (eng. *stories*) - ovaj kreativni format korisnicima na *Facebooku, Instagramu* i *Messengeru* omogućava gledanje i dijeljenje svakodnevnih trenutaka kroz fotografije i video zapise koji nestanu (osim ako nisu sačuvani) u roku od 24 sata;
- Brendirani sadržaj (eng. *branded content*) – ovaj format podrazumijeva sadržaj autora ili izdavača koji sadrži proizvod, uslugu ili informacije poslovnog partnera radi razmjene vrijednosti. Kreatori ili izdavači su odgovorni za označavanje stranice poslovnog partnera prilikom postavljanja brendiranog sadržaja.

U članku *New Yorker* magazina (Cassidy, 2014), *Facebook* je opisan kao najveća kompanija za direktni marketing. Neki autori očekuju da *Facebook* u narednim godinama postane konkurent kompaniji *Amazon* u segmentu prodaje, s obzirom na broj korisnika i već razvijenu praksu kompanija u sferi komunikacije s potrošačima – direktna prodaja preko *Facebook* platforme nameće se kao naredni korak (Kh, 2017; Kleinberg, 2019).

Uzimajući u obzir sve navedeno – broj korisnika, kvalitet platforme za oglašavanje, broj kompanija koje se oglašavaju ovim putem, kao i razvoj infrastrukture potrebne za procesiranje transakcija, opravdan je izbor *Facebook* platforme za empirijsko istraživanje, koje će biti predstavljeno u posljednjem dijelu rada.

2.5.4 Vještačka inteligencija i automatizacija procesa donošenja odluka u direktnom marketingu

Bucklin, Lehmann i *Little* (1998) su procijenili da će se do 2020. godine funkcija tehnologije u marketingu pomjeriti s „podrške odlučivanju" na "automatizaciju odlučivanja". Zbog usmjerenja kompanija ka masovnom prilagođavanju, poboljšanim procesom donošenja odluka i povećanom efikasnošću, očekivali su da će određeni procenat marketing izbora biti automatizovan. Ova ideja je za određeni broj savremenih kompanija već postala realnost i automatizacija marketinga

ubrzano dobija na značaju kao korporativni alat, s jedne, i istraživački fenomen, s druge strane. Sadašnja literatura iz ove oblasti se uglavnom koncentriše na opisivanje operativne logike marketing automatizacije (Järvinen, 2016; Vecchia & Peter, 2018) i studija slučaja o njenoj primjeni (Hong & Park, 2020; Mero et al., 2020; Zumstein et al., 2021).

Osnovna ideja marketing automatizacije je da se procesi koji se ponavljaju dizajniraju na takav način da se automatski pokreću, i da su, korišćenjem alata vještačke inteligencije i sistematizovanih baza podataka o kupcima, prilagođeni svakom pojedinačnom kupcu. Fokus na donošenju odluka zasnovanih na podacima u marketingu je ključna komponenta današnje automatizacije marketinga (Vecchia & Peter, 2018). Pored podataka, ovaj koncept uključuje i podršku toka posla (eng. *workflow support*), analizu ponašanja kupaca, podršku odlučivanju, upravljanje sadržajem i upravljanje kanalima. Cilj automatizacije marketinga je da ubrza, olakša i pomogne kompanijama da segmentiraju klijente, pokrenu višekanalne marketing kampanje i pruže personalizovane informacije i ponude svojim potencijalnim kupcima.

Da bi se procesi u marketingu automatizovali, kompanije moraju obezbijediti dvije najznačajnije komponente: podatke o kupcima i njihovom kupovnom ponašanju i softver za automatizaciju marketing operacija. S tim u vezi, marketing automatizacija najčešće obuhvata softversko rješenje koje služi za automatizaciju osnovnih marketing operacija koje se ponavljaju i koje bi kompanije u suprotnom izvršavale manuelno, kao što su: bilteni e-pošte, zakazivanje objava na društvenim mrežama, ažuriranje liste kontakata, pronalaženje potencijalnih klijenata i bilježenje podataka o njima, praćenje kampanja i izvještavanje o njihovoj realizaciji.

Marketing automatizacija se generalno sve više koristi za uspostavljanje platforme za uspješniji i efikasniji marketing, što dovodi do povećane produktivnosti i povrata na marketing investicije. Ovo se postiže temeljnom analizom situacije, isticanjem uspješnih i neuspješnih inicijativa sprovedenih u prošlosti i pomaganjem u preraspodjeli napora ka najperspektivnijim mogućnostima. Ovakvi rezulati mogu se obezbijediti korišćenjem prediktivnih *data mining* tehnika.

Na ovaj način se mogu otkriti grupe kupaca koje ne reaguju na marketing napore i izostaviti u narednim kampanjama, što rezultira značajnim uštedama. Upotreba analitike će također olakšati identifikaciju manjih grupa za dodatnu komunikaciju, kao što je, na primjer, promocija unakrsne prodaje, a mogu se otkriti i mogućnosti za jedinstvene *ad hoc* šanse. Konačno, povećana efikasnost automatizacije marketinga će osnažiti trgovce da prepoznaju i iskoriste jedinstvene mogućnosti za nove i trenutne potrošače (LeSueur, 2007).

U istraživanju sprovedenom u Švajcarskoj navedeno je da većina organizacija koristi upravljanje odnosima s klijentima (CRM), digitalnu analitiku i sistem za upravljanje sadržajem, dok skoro sve koriste e-poštu, marketing na društvenim mrežama i marketing na pretraživačima. Marketing automatizaciju koristi 40% ispitanika, pri čemu je četvrtina koristi u velikoj mjeri, a trećina je koristi samo rijetko. U ovoj studiji je istaknuto da ove marketing alate najviše koriste „uspješne" organizacije – one koje su bolje u ispunjavanju svojih ciljeva *online* strategije. Uspješne kompanije sa tri puta većom vjerovatnoćom od neuspješnih koriste platformu za automatizaciju marketinga, a preduzeća sa značajnim budžetima je koriste dvostruko češće od onih sa ograničenim budžetima (Zumstein et al., 2021). Međutim, jedan od faktora koji sprečavaju kompanije da uvedu automatizaciju marketinga je obično nedostatak interne ekspertize.

U prethodnim poglavljima istaknuta je potreba kompanija da svakog potrošača tretiraju jedinstveno, u skladu s njegovim prethodnim postupcima, željama i potrebama. Kako bi se steklo povjerenje i održala baza lojalnih kupaca, odluke kompanija koje se odnose na plasiranje promotivnih ponuda i sadržaja moraju biti prilagođene i individualizovane za svakog kupca (ili grupu sličnih kupaca) i potrebno ih je donositi u realnom vremenu i na osnovu unaprijed definisanih pravila ili kriterijuma. U *big data* eri, kompanije mogu razvijati prediktivne modele na osnovu podataka o prethodnom kupovnom ponašanju i karakteristikama kupaca i izgraditi pravila koja će biti vodič za buduće marketing aktivnosti. Ovaj proces se može u velikoj mjeri automatizovati, nakon inicijalnog kreiranja modela na osnovu dostupnih podataka. S tim u vezi, *Ammerman* (2019) procjenjuje da će poruke koje su usmjerene na pojedinačne potrošače sve više biti diktirane kombinacijom

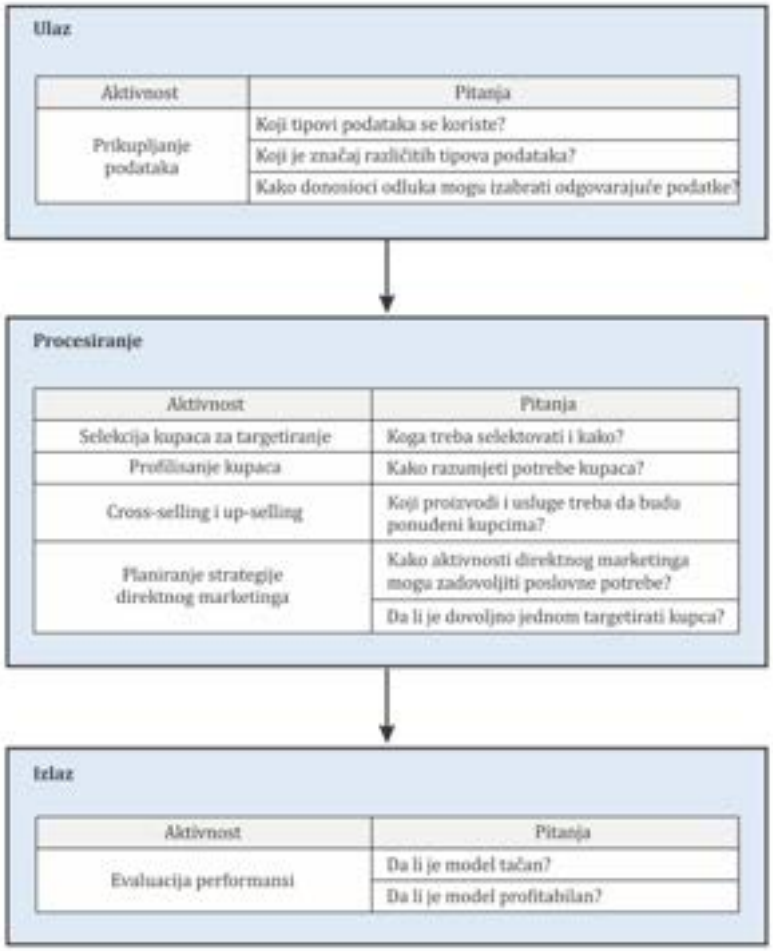
marketing automatizacije i mašinskog učenja, pri čemu autor ističe da je *"sve dizajnirano da nas ubijedi na kupovinu na sve sofisticiranije i neprimjetnije načine"*.

2.6 Aktivnosti u direktnom marketingu

S obzirom na to da empirijski dio rada predstavlja nove modele za povećanje efikasnosti aktivnosti segmentacije i predviđanja odgovora na kampanju direktnog marketinga, u ovom dijelu rada biće prikazane osnovne aktivnosti koje se sprovode u okviru sistema direktnog marketinga.

Veza između kompanije i potrošača počinje nakon prve interakcije inicirane u cilju prodaje proizvoda ili usluga (Bose & Chen, 2009). Na osnovu dobijene ponude, potrošači procjenjuju da li će obaviti kupovinu ili ne, dok, s druge strane, donosioci odluka i marketing menadžeri na osnovu rezultata inicijalne kampanje definišu ciljeve i strategije za buduće kampanje. S tim u vezi, prva aktivnost u okviru direktnog marketinga odnosi se na prikupljanje podataka, na osnovu kojih se procjenjuje koliko kupaca je odgovorilo na kampanju i koje su osnovne karakteristike grupe selektovanih kupaca koji su obavili transakciju. Na osnovu navedenih podataka, definiše se nova ciljna grupa, odnosno targetira se novi skup potencijalnih kupaca za naredne kampanje. Targetiranje kupaca se zasniva na definisanju segmenata tržišta i profilima profitabilnih kupaca i odgovarajućih segmenata, kako bi se ovaj proces efikasno izvršio, uzimajući u obzir da profitabilnost kampanje zavisi od stope odgovora i preciznosti selektovanja kupaca. Konačno, posljednja aktivnost koja zatvara cjelokupni proces odnosi se na evaluaciju rezultata kampanje i svih prethodnih aktivnosti direktnog marketinga.

Na slici koji slijedi (Slika 2) prikazan je redosljed aktivnosti u sistemu direktnog marketinga.



Slika 2. Sistemska perspektiva modela direktnog marketinga (Bose & Chen, 2009)

U narednim djelovima rada biće detaljnije predstavljene osnovne aktivnosti direktnog marketinga (prilagođeno prema: Bose & Chen, 2009).

2.6.1 Prikupljanje i priprema podataka

Razvoj interneta je omogućio akumuliranje velike količine podataka, a dodatno je i stvorena mogućnost za njihovo čuvanje i skladištenje. Digitalno okruženje je podstaklo razvoj direktnog marketinga, uz tehnologiju koja otvara mogućnosti za generisanje i pretragu velike količine podataka različite kompleksnosti (Aliabadi & Berenji, 2013).

Baze podataka o kupcima postaju preplavljene podacima, s tim što svi raspoloživi podaci nisu neophodni za analizu pojedinačnih kupaca i tržišnih segmenata. S tim u vezi, prva aktivnost marketinga odnosi se na prikupljanje i selekciju podataka koji mogu otkriti značajne informacije o preferencijama potrošača. Osim toga, s obzirom na veliku količinu različitih podataka, praktičari direktnog marketinga moraju osigurati da su prikupljeni podaci prečišćeni, da mogu obezbijediti značajnu i smislenu analizu i da su dovoljni da pomognu u donošenju odluke (Bose & Chen, 2009).

Donošenje blagovremenih i efektivnih odluka zavisi od kvaliteta raspoloživih podataka, pa je pretprocesiranje podataka važan korak u procesu otkrivanja znanja iz podataka. Upravo ovaj korak najčešće zahtijeva najviše vremena u cjelokupnom procesu analize, a čisti i konzistentni podaci predstavljaju preduslov za sprovođenje *data mining* aktivnosti, na kojima se zasniva empirijski dio ovog istraživanja. U praksi, podaci o kupcima koji se čuvaju u bazama obično su nekompletni, nekonzistentni i u formatu koji nije prikladan za analizu i otkrivanje znanja. Stoga, važni procesi u okviru ove aktivnosti su čišćenje podataka, konverzija, grupisanje ili razdvajanje podataka, agregiranje, formatiranje i drugi. Čišćenje podataka odnosi se na postupke rješavanja problema nedostajućih vrijednosti, kao i uklanjanja "šuma" iz podataka (eng. *noise*), kako bi se dobili čisti podaci, čije je korišćenje podržano u *data mining* algoritmima. U zavisnosti od karakteristika podataka, problem nedostajućih vrijednosti u bazama može se riješiti na različite načine: zapisi sa nedostajućim vrijednostima mogu se brisati iz baze, mogu se popuniti statističkim metodama ili na osnovu procjene (Xiahou, 2016). Konačno, kao i u bazama podataka generalno, podaci se za *data mining* algoritme moraju posebno pripremiti. Priprema i pretprocesiranje podataka neki su od najznačajnijih aspekata *data mining* projekata. Faza pripreme podataka obuhvata: redukciju podataka, izbor varijabli, transformaciju podataka i normalizaciju (Khajvand et al., 2011).

Dodatni problem koji se javlja pri pripremi podataka za analizu odnosi se na činjenicu da je većina ličnih podataka o kupcima nepoznata, pa stoga prikupljanje određenih podataka zahtijeva dodatni napor i vrijeme. Pored toga, čak i u slučaju da

kompanije odluče da prikupe ove podatke, oni su promjenjivog karaktera – zanimanje, nivo dohotka i bračni status potrošača danas, ne moraju važiti za dvije godine (Tsai & Chiu, 2004). Stoga, kompanije najčešće za analizu koriste osnovne podatke o kupcima, poput pola, regiona i godina starosti, a za precizniju segmentaciju i definisanje strategija targetiranja, ove podatke dopunjavaju podacima o kupljenim proizvodima i ponašanju u kupovini. Ova aktivnost najčešće zahtijeva spajanje različitih baza podataka, kao što su baze podataka o kupcima, transakcijama, proizvodima, kampanjama i slično.

Prikupljanje kompleksnih podataka i njihova analiza omogućavaju donosiocima odluka da razumiju ponašanje i obrasce kupovine pojedinačnih potrošača, što doprinosi kreiranju personalizovane strategije zasnovane na znanju i prethodnom iskustvu. Uz uzlazni trend prihvatanja podataka kao ključnih faktora za kreiranje strategija direktnog marketinga, raste i uloga tehnologije u sferi izgradnje prediktivnih modela. Ovi modeli, zasnovani na *data mining* metodama, pomažu organizacijama da stvore strategije i definišu ciljeve s fokusom na kupcima, s obzirom na to da pravovremena i kvalitetna analiza podataka omogućava identifikovanje potreba i faktora koji utiču na ponašanje potrošača u svim fazama procesa donošenja odluke o kupovini (Grandhi et al., 2017). Na ovaj način, osigurava se bogatije iskustvo za kupca kroz efikasnu personalizaciju, što doprinosi kako privlačenju novih, tako i zadržavanju postojećih kupaca. Posredno, direktni marketing zasnovan na znanju snižava troškove, te povećava produktivnost i efikasnost kompanije (Grandhi et al., 2017).

Veliki broj organizacija prikuplja i čuva podatke o svojim kupcima, potencijalnim kupcima, dobavljačima i poslovnim partnerima, ali, takođe, veliki broj kompanija nema mogućnosti i vještine pretvaranja ovih podataka u vrijedno i korisno znanje koje će pomoći u donošenju odluka (Berson et al., 2000; Ngai et al., 2009). Čak 74% kompanija nije u mogućnosti da angažuje analitičara podataka zbog toga što postoji oskudica analitičara i statističara koji traže poslove, što pokazuje da se savremene kompanije susriječu sa izazovom istovremenog analiziranja i interpretiranja

podataka (Grandhi et al., 2017). S tim u vezi, ovim kompanijama bi *data mining* alati mogli pomoći da otkriju znanje u podacima koje skladište.

2.6.2 Segmentacija i kreiranje profila kupaca

Marketing analitiku čine dva osnovna procesa – segmentacija i targetiranje kupaca. Segmentacija dijeli bazu kupaca na grupe kupaca sličnih karakteristika, dok se targetiranjem teži identifikovati grupa potrošača s visokom vjerovatnoćom odgovora na kampanju. Oba procesa unapređuju performanse marketinga i alociraju resurse najprofitabilnijim kupcima (Yang et al., 2016). Iako povezani, ovi procesi se sprovode odvojeno – za segmentaciju se koriste metode klasterizacije, dok se targetiranje kupaca sprovodi metodama klasifikacije (Yang et al., 2016).

U cilju pripreme efikasnih kampanja, donosioci odluka zahtijevaju informacije o karakteristikama kupaca, demografiji i njihovom ponašanju u kupovini. Razvijen je veliki broj tehnika, modela i algoritama koji se koriste za klasifikaciju kupaca, odnosno njihovu podjelu u određene tržišne grupe. Uzimajući u obzir da razvoj informaciono-komunikacionih tehnologija, sistema upravljanja bazama podataka kao i primjena *data mining* metoda mijenjaju i poboljšavaju efikasnost marketing procesa, izbor tehnika za segmentaciju u savremenom tržišnom okruženju je od posebnog značaja za stepen preciznosti rezultata (Sarvari, 2016).

Rezultat procesa segmentacije pruža donosiocima odluka važne informacije za pripremu budućih marketing aktivnosti i razvoj strategija u skladu s potrebama različitih segmenata. Grupisanje kupaca sa sličnim potrebama, željama i ponašanjem omogućava bolje razumijevanje tržišta, na osnovu čega donosioci odluka mogu prilagođavati marketing aktivnosti, cijene i izbor komunikacionih kanala (Dogan et al., 2018). Dakle, segmentacija kupaca i sama potreba za razumijevanjem različitih segmenata sprovodi se, između ostalog, u cilju omogućavanja personalizovanog iskustva u kupovini i dizajniranja specifičnih strategija, koje će s većom vjerovatnoćom konvertovati potencijalnog u lojalnog kupca (Sheng & Subramanian, 2019).

U digitalnom okruženju sprovođenje kako segmentacije, tako i personalizacije prozvoda i iskustava olakšano je razvojem raličitih alata, koji omogućavaju sticanje sveobuhvatnog znanja o prethodnom ponašanju kupaca.

Dakle, segmentacija kupaca dijeli cjelokupnu bazu kupaca na grupe pojedinačnih kupaca, koji su sličnih profila, kupuju slične proizvode i sličnih su prioriteta i vrijednosti za kompaniju (Sarvari et al., 2016; Yang et al., 2016). Na ovaj način, omogućen je bolji uvid u karakteristike različitih segmenata, pa se može kreirati jedinstveni profil kupca za sve segmente (eng. *buyer persona*). Pretpostavka u procesu segmentacije je da će kupci koji pripadaju istom segmentu pokazati slično ponašanje u kupovini (Tian et al., 2006). Kvalitet sprovedene segmentacije najčešće se mjeri homogenošću unutar segmenta i heterogenošću između segmenata (Wedel & Kamakura, 2000).

Tradicionalne metode segmentacije obuhvatale su proces podjele kupaca u grupe na osnovu demografskih i psihografskih atributa, kao i podataka o njihovim stavovima. Lee i Park (2005) navode da su ovi modeli bili isuviše jednostavni, te nisu obezbjeđivali zadovoljavajuću tačnost, posebno u kompleksnom savremenom poslovnom okruženju. Stoga, ovi autori navode da se nove metode segmentacije baziraju na podacima o transakcijama i prethodnom kupovnom ponašanju, poput vrste proizvoda, vrijednosti kupovine, aktivnosti na mreži, broju raklamacija i slično (Lee & Park, 2005).

S obzirom na to da je objektivna i adekvatna segmentacija preduslov za efikasno targetiranje kupaca u kampanjama direktnog marketinga, u poglavlju 3 biće detaljno predstavljene različite metode za segmentaciju kupaca, kao i modeli primijenjeni u dosadašnjim istraživanjima iz ove oblasti.

2.6.3 Odabir kupaca za targetiranje

Targetiranje potencijalnih kupaca sprovodi se s ciljem definisanja ciljne grupe za konkretnu kampanju. Efikasno targetiranje doprinosi snižavanju troškova i povećanju prodaje i profita kampanje, što predstavlja srž moderne marketing

analitike (Antipov et al., 2010; Chou et al., 2000). U kampanjama se mogu targetirati pojedinačni kupci, domaćinstva ili definisani segmenti. U okviru ove aktivnosti, donosioci odluka definišu koga će targetirati u narednoj kampanji i koje tehnike će koristiti za selekciju ciljnog tržišta. Dakle, pretpostavka efikasnog targetiranja u kampanji je adekvatna i objektivna segmentacija tržišta. Segmentacija omogućava kreiranje profitabilnih segmenata, međutim, veliki broj marketing menadžera u ovoj sferi nailazi na izazov pri identifikovanju pravih segmenata kupaca u cilju organizacije kampanja (Mohammadian & Makrani, 2016), što dovodi do neuspješnih programa lojalnosti i neprofitabilnih promocionih aktivnosti, kao i rasipanja resursa predviđenih za marketing aktivnosti (Dogan et al., 2018).

Sistemi za oglašavanje i algoritmi koje oni koriste konstantno napreduju i proširuju mogućnosti za dostizanje do potencijalnih kupaca. Pored toga, razvija se i pristup mikro targetiranja, koji koriste detaljne podatke o kupcima i sistem automatizacije digitalnog marketinga u cilju dostavljanja targetiranih i personalizovanih poruka putem različitih kanala komunikacije (Semerádová & Weinlich, 2019). Kako bi se omogućila personalizacija poruke, važno je precizno i objektivno definisati segmente.

Međutim, važno je napomenuti da segmentacija kupaca obuhvata mnogo više od spajanja grupe kupaca sa specifičnim proizvodima i uslugama. Ovaj proces obuhvata i način komunikacije s kupcima, koji se definiše na osnovu informacija koje kompanija posjeduje o njima, kao i kreiranje jedinstvenog iskustva u kupovini za sve definisane ciljne grupe, kako bi se izgradila njihova lojalnost brendu (Magento, 2019). U tom smislu, personalizacija i prilagođavanje proizvoda, usluga ili iskustava ima poseban značaj u direktnom marketingu. *Semerádová* i *Weinlich* (2019) istraživali su performanse različitih postavki targetiranja, koristeći skup podataka iz 840 *Facebook* oglasa sa različitim nivoima personalizacije i uporedili su rezultate, kao što su: doseg, broj reakcija, učestalost prikazivanja, broj klikova, prosječno vrijeme provedeno na *web* sajtu, broj pregledanih stranica, broj konverzija i profitabilnost. Otkrili su da personalizovani oglasi imaju znatno veću vrijednost od onih koji nisu personalizovani.

S obzirom na to da selekcija kupaca za targetiranje predstavlja osnovnu aktivnost direktnog marketinga (Bose & Chen, 2009), u trećem poglavlju ovog rada biće detaljnije predstavljene metode za selekciju i targetiranje kupaca u direktnom marketingu.

2.6.4 Cross-selling i up-selling

U savremenom poslovnom okruženju, fokus kompanija usmjeren je ka lojalnosti i profitabilnosti kupaca, u cilju povećanja zadovoljstva potrošača i, posredno, uvećanja tržišnog učešća. U okviru poglavlja 2.5.2, gdje su predstavljene osnovne dimenzije CRM-a, *cross-selling* i *up-selling* su navedeni kao elementi dimenzije "razvoj kupaca". S tim u vezi, razvoj uspješnog CRM sistema zasnovan je na identifikovanju vrijednosti potrošača za kompaniju, što predstavlja informaciju koja se dobija kroz proces segmentacije. Uvid u vrijednost, odnosno profitabilnost potrošača za kompaniju važan je element razvoja personalizovanih i targetiranih strategija u direktnom marketingu (Kim et al., 2006).

Podaci o prethodnim kupovinama, koji uključuju vrstu proizvoda, namjenu, brend, cijenu i slično, važni su inputi za proces dizajniranja *cross-selling* i *up-selling* strategija. Ovi podaci omogućavaju uočavanje sličnosti između kupljenih proizvoda i dodatnih proizvoda koji mogu zainteresovati kupce, a koji se mogu preporučiti u okviru navedenih strategija (Liu & Shih, 2005). Ovi podaci se nalaze u zapisima o transakcijama u bazama podataka kompanije, a stvaraju mogućnost za dizajniranje direktne i precizne strategije izbora proizvoda koji se mogu ponuditi kupcima u narednoj kampanji.

Dakle, podaci o karakteristikama proizvoda su od posebne važnosti za *cross-selling* u *up-selling*, uz pretpostavku da se kupcima u narednim kampanjama ponude proizvodi slični onim koje su ranije kupili (Li et al., 2005; Weng & Liu, 2004). Međutim, ponuda isključivo istih ili sličnih proizvoda nije dovoljna, pa kompanije moraju pronaći načine da istraže promjene u potrebama i željama potrošača, na osnovu čega će preciznije definisati strategije za *cross-selling* i *up-selling*. Jedan od načina za istraživanje ponašanja potrošača je, pored pregleda prethodnih

transakcija, istraživanje aktivnosti i ponašanja na sajtu e-prodavnice, što može ukazati na preferencije potrošača i njihova interesovanja. Podaci prikupljeni na ovaj način uključuju: tip operativnog sistema koji se koristi, ključne riječi korišćene za pretragu, tipove stranica koje je kupac posjećivao, vrijeme provedeno na tim stranicama (Bose & Chen, 2009), podatke o proizvodima sačuvanim u okviru "željenih proizvoda", kao i podatke o proizvodima koje je kupac "stavio u korpu", ali nije obavio transakciju.

2.6.5 Značaj procjene odgovora kupca u planiranju direktne kampanje

Saturacija i hiperkonkurentnost tržišta usmjeravaju donosioce odluka ka novim metodama za selekciju kupaca, koji prevazilaze korišćenje demografskih i psihografskih atributa. Direktni marketing predstavlja jedan od marketing sistema koji aktivno primjenjuje individualni marketing koncept, uzimajući u obzir da su sve aktivnosti u direktnom marketingu zasnovane na podacima iz baza podataka o kupcima.

Istraživači iz ove oblasti navode neke od značajnih odluka koje se donose u direktnom marketingu, među kojima se targetiranje i predikcija ističu kao ključne (Desarbo & Ramaswamy, 1994; Zahavi & Levin, 1995). Kada je u pitanju predikcija u direktnom marketingu, jedan od najznačajnijih elemenata planiranja kampanje odnosi se na predviđanje odgovora na kampanju. S tim u vezi, model predviđanja odgovora na kampanju, koji se koristi i za aktivnosti targetiranja i predikcije, predstavlja jedan od najznačajnijih zadataka u direktnom marketingu (Suh et al., 1999). Dakle, dizajniranje ovog modela obezbjeđuje vrijedne informacije donosiocima odluka, zasnovane na predviđanju da li će potrošač odgovoriti na kampanju direktnog marketinga ili ne (u narednoj sekciji rada opisane su metrike za evaluaciju kampanje direktnog marketinga, pri čemu se kao odgovor na kampanju u praksi najčešće uzima obavljena transakcija). U modelu predviđanja odgovora na kampanju, potrošači se dijele u dvije grupe – respondenti i nerespondenti (Daneshmandi & Ahmadzadeh, 2013). Rezultati ovog modela preciznije i objektivnije pokazuju koje potrošače treba targetirati u kampanji, u odnosu na subjektivnu procjenu donosioca odluka. Predviđanje odgovora na

kampanju od posebnog je značaja za kreiranje strategije u direktnom marketingu za sve kampanje i ponude individualno. Na ovaj način, uz informacije iz modela, kompanija će alocirati marketing resurse potrošačima s najvećom vjerovatnoćom odgovora, odnosno onim potrošačima koji imaju najveću potencijalnu vrijednost za kompaniju (Sun et al., 2014). Prilagođavanje marketing aktivnosti definisanim segmentima koji se razlikuju po interesovanjima, profitabilnosti, vrijednosti za kompaniju ili nekim drugim karakteristikama, čini cjelokupnu strategiju direktnog marketinga efikasnijom (Donio et al., 2006). Kako bi se resursi za marketing efikasno investirali, ovaj proces, uz razvoj društvenih mreža koje su omogućile preciznije targetiranje kupaca više nego ikada do sada, još više dobija na značaju.

Modeliranje odgovora na marketing kampanju identifikuje grupu respondenata i njihove karakteristike, te samim tim snižava marketing troškove i povećava profitabilnost kampanje. Međutim, s obzirom na specifične karakteristike zapisa u bazama podataka o respondentima i nerespondentima, poput nebalansiranosti klasa (detaljnije u sekciji 4.1.5), kompanije se suočavaju sa izazovom da dizajniraju efektivne modele (Daneshmandi & Ahmadzadeh, 2013). S tim u vezi, model predviđanja odgovora na kampanju prepoznat je kao jedan od najatraktivnijih aspekata direktnog marketinga za akademska istraživanja (Blattberg, 1987). U empirijskom dijelu rada, u sekcijama 5.4 i 5.5 korišćenjem *Support Vector Machine* i *ensemble* metoda, biće testiran novi predloženi koncept modela za predviđanje odgovora na kampanju.

2.6.6 Evaluacija performansi kampanje

Uzimajući u obzir činjenicu da kompanije stiču nove informacije u toku i nakon svake pojedinačne kampanje, kako bi kreirali preciznije i efikasnije kampanje u budućnosti, konačna aktivnost direktnog marketinga predstavlja evaluaciju performansi kampanje. U savremenom poslovnom okruženju nije dovoljna subjektivna procjena efikasnosti pojedinačnih marketing kampanja, već je neophodno korišćenje alata i modela za praćenje i evaluaciju rezultata, posebno kada je u pitanju elektronska trgovina i digitalni direktni marketing. Godinama unazad, kompanije nisu mogle objektivno procijeniti efikasnost marketing

kampanja, sa izuzetkom analiza ostvarenih prodaja. Međutim, u digitalnom okruženju, čak i ako kampanja ne ostvari cilj u aspektu prodaje, njenom analizom se mogu dobiti važni podaci i informacije, što može unaprijediti proces donošenja odluka za naredne kampanje (Belyh, 2019). Dakle, za efikasnost budućih kampanja, ključno je korišćenje raspoloživih marketing alata za analizu dosadašnjih marketing napora.

Kampanja se najčešće ocjenjuje na osnovu stečenih prihoda (Bose & Chen, 2009), a mogu se posmatrati i drugi elementi, kao što su: stopa odgovora, broj ostvarenih interakcija, razvoj profitabilnosti kupaca i slično. Dakle, efikasnost kampanje prije svega zavisi od preciznosti selekcije i targetiranja kupaca za konkretnu ponudu.

Kada je riječ o digitalnom direktnom marketingu, evaluacija kampanja može se obaviti još efikasnije kroz uvid u različite metrike koje se mogu pratiti u realnom vremenu, poput: konverzije lidova, broja individualnih posjetilaca, broja novih posjetilaca, broja stranica koje je korisnik posjetio, stope konverzije, broja otvorenih *mailova* i drugih.

Jedan od analitičkih alata koji se najčešće koristi u marketing svrhe je *Google Analytics* (Belyh, 2019). Iz tog razloga, u empirijskom dijelu rada (poglavlje 5), za predikciju profitabilnosti kupaca i predviđanje odgovora na kampanju biće korišćeni podaci iz ovog alata.

3. METODE ZA SELEKCIJU I TARGETIRANJE KUPACA

U ovom poglavlju će biti predstavljene metode za selekciju i targetiranje kupaca na osnovu dostupne literature. Pored toga, biće predstavljen proces segmentacije kupaca, odnosno njihove podjele u homogene grupe sličnog ponašanja i afiniteta, što se nameće kao preduslov za selekciju kupaca, odnosno njihovo targetiranje, kao i proces opisivanja tih segmenata, koji je važan za efikasnu interakciju s njima. Biće ukazano na prethodnu praksu i sisteme za selekciju i segmentaciju, kako u cilju targetiranja kupaca, tako i u cilju opisivanja segmenata kupaca, radi preciznije selekcije kupaca i plasiranja relevantnih ponuda pojedinačnim grupama. S tim u vezi, biće ukazano na njihove određene nedostatke i načine za prevazilaženje navedenih nedostataka kreiranjem novih, hibridnih metoda baziranih na *data mining* tehnikama.

Izbor potencijalnih kupaca za targetiranje u budućim marketing kampanjama odnosi se na proces podjele tržišta u dvije grupe – korisnike koje će kompanija targetirati u kampanji i one koji neće primiti poruku iz kampanje. U tom smislu, segmentacija kupaca za potrebe direktnog marketinga ima značajnu ulogu u procesu planiranja budućih kampanja, uzimajući u obzir da će kampanje biti efikasne ukoliko poruka dođe do korisnika koji će na nju odgovoriti s visokom vjerovatnoćom. Metode za izbor kupaca za targetiranje autori najčešće dijele u dvije osnovne grupe: segmentacione i bodovne (eng. *scoring*) metode (Jonker et al., 2004; Kaymak, 2001).

Uvid u misli, želje i potrebe potrošača ima veliki značaj za generisanje održivih strategija na konkurentskim tržištima. U tom procesu, cilj je grupisati potrošače sličnih karakteristika, interesovanja i potreba, a na osnovu sveobuhvatnog razumijevanja svake pojedinačne grupe. S tim u vezi, segmentacija potrošača je jedan od kritičnih procesa u organizacijama koje teže razvoju adekvatnih promotivnih strategija za različite kategorije kupaca (Hsu et al., 2012). Osim toga,

segmentacija potrošača se nameće kao prvi korak u procesu selekcije kupaca za targetiranje u narednim marketing aktivnostima.

Segmentacija i targetiranje su dva važna elementa jedne od najznačajnijih koncepcija u marketingu: segmentacija – targetiranje – pozicioniranje (STP), koja omogućava dizajn i implementaciju ciljno usmjerenog marketing programa organizacije (Melović et al., 2019). Dakle, nakon segmentacije i identifikacije svih potrošačkih grupa, u procesu targetiranja se biraju relevantne grupe za definisanu ponudu, nakon čega slijedi pozicioniranje, tj. odabir adekvatne kombinacije marketing alata za dizajniranje ponude, tako da ona zauzme značajnu i jasnu poziciju u umu ciljnog kupca (Fahy & Jobber, 2022; Melović et al., 2019).

S tim u vezi, u prvom dijelu STP procesa, jedna od ključnih faza u postupku segmentacije kupaca je izbor atributa, koji se najčešće dijele u dvije osnovne grupe: opšti atributi i transakcioni atributi (Tsai & Chiu, 2004). Prva grupa uključuje demografske i psihografske podatke, kao i podatke o stavovima i životnom stilu (Huang & Tzeng, 2007). Iako su opšti atributi važni za analizu i opisivanje različitih tržišnih segmenata i jednostavni za analizu i razumijevanje, oni imaju i određene nedostatke. Naime, kupci sličnih demografskih karakteristika nemaju nužno isto ili slično ponašanje u kupovini. Uz to, podaci o stavovima potrošača i životnom stilu se teško prikupljaju i integrišu s podacima o kupovnom ponašanju (Hsu et al., 2012). Stoga, najveći dio istraživanja u ovoj oblasti koristi atribute vezane za transakcije, odnosno ponašanje u kupovini (Cheng & Chen, 2009; Hosseini et al., 2010; Sarvari et al., 2016; Tsai & Chiu, 2004), što će biti fokus i ovog istraživanja. Na ovaj način, jasno se definišu slične grupe kupaca, koje će se na slične načine tretirati u narednim marketing aktivnostima.

Otkrivanje preferencija i ponašanja kupaca u okviru segmenata omogućava dizajniranje adekvatnih marketing strategija, što doprinosi jačanju odnosa s potrošačima i povećanju njihove lojalnosti.

3.1 Segmentacione i bodovne metode

Segmentacione metode funkcionišu po principu podjele kupaca u segmente, odnosno grupe, pri čemu se koriste eksplanatorne varijable. Na ovaj način se obezbjeđuje sličnost kupaca u segmentima, tj. homogenost u smislu očekivanog odgovora na plasiranu kampanju direktnog marketinga, te se na osnovu rezultata mogu odabrati i targetirati segmenti s najvećom vjerovatnoćom odgovora. Najčešće korišćena segmentaciona metoda je *Recency Frequency Monetary* (RFM), koja će detaljno biti opisana u narednoj sekciji. Informacije dobijene RFM metodom mogu se uključiti u prediktivne modele, na primjer, kao eksplanatorne varijable u modelima odgovora na kampanju. Takođe, korišćenjem DT metode ili neuronskih mreža mogu se povezati RFM vrijednosti s kupovnim ponašanjem (Cui et al., 2006; Rhee & Russell, 2009).

Za razliku od segmentacionih, kod bodovnog pristupa svakom pojedinačnom kupcu se dodjeljuje odgovarajuća ocjena, koja se dobija na osnovu predviđene vjerovatnoće odgovora na kampanju ili kupovine određenog proizvoda, s jedne ili predviđenog profita koji će biti ostvaren, s druge strane. Nakon definisanja ocjene, kupci se sortiraju na osnovu dodijeljenih rezultata, pa se u narednim kampanjama targetiraju oni s najvećim vrijednostima. Bodovne metode su razvijene s ciljem predviđanja budućeg ponašanja potrošača i često se definišu u formi regresionih modela (Baesens et al., 2002; Berry & Linoff, 2004; Parr Rud, 2001). Autori iz ove oblasti navode da se upravo prethodno ponašanje potrošača može uzeti kao osnovni prediktor njihovog budućeg ponašanja, te se u literaturi iz direktnog marketinga najčešće u ove svrhe koristi RFM model. U narednom poglavlju biće opisana primjena RFM metodologije za dodjelu ocjena potrošačima.

U istraživanjima u kojima je primijenjen bodovni pristup, najčešće se uzima u obzir isključivo procjena, odnosno predikcija odgovora na kampanju. S tim u vezi, podjela kupaca na respondente ili nerespondente u kampanji predstavlja binarni klasifikacioni problem. U modelima odgovora na kampanju se kao eksplanatorne varijable najčešće koriste karakteristike kupaca i podaci o njihovom ponašanju u

kupovini. S druge strane, zavisna varijabla je binarna i sadrži podatak o tome da li je kupac u prethodnom periodu odgovorio na kampanju ili ne. Na ovaj način, modelima odgovora na kampanju se, na osnovu podataka o prethodnom kupovnom ponašanju, može predvidjeti vjerovatnoća odgovora na buduću kampanju. U cilju predviđanja odgovora na kampanju, autori iz ove oblasti primjenjivali su *data mining* metode linearne i logističke regresije, kao i vještačke neuronske mreže i drvo odlučivanja (Bose & Chen, 2009; Coussement et al., 2014; Guido et al., 2013; Kang et al., 2012).

Linearna regresija, kao jedan od tradicionalnih (statističkih) metoda korišćenih u direktnom marketingu, predstavlja pristup modeliranju veze između zavisne varijable y i skupa nezavisnih varijabli X. Rezultati, tj. predviđanja dobijaju se na osnovu linearne funkcije u formi neprekidnih rezultata. Za primjenu u direktnom marketingu definiše se prag, na osnovu kojeg se kupci dijele u dvije grupe. Kupci čiji je rezultat iznad definisane vrijednosti mogu biti targetirani u narednoj kampanji, dok grupa kupaca s nižim rezultatom od definisanog neće biti selektovana (Wang, 2013). Pored linearne regresije, u literaturi se ističe i logistička regresija, kao jedan od često korišćenih metoda, koji je popularnost u ovoj oblasti stekao jednostavnošću i pogodnošću eksplikacije i opisivanja rezultata. Konkretno, u direktnom marketingu se koristi za predviđanje potencijalih potrošača, tj. u formi modela odgovora na kampanju.

Međutim, visoka stopa odgovora ne znači nužno i visoku profitabilnost respondenata (Kim et al., 2008). S tim u vezi, prilikom primjene bodovnog pristupa, veoma je značajno obuhvatiti i predikciju profitabilnosti respondenata, odnosno kupaca koji imaju visoku vjerovatnoću odgovora na kampanju. Na ovaj način, primjenom modela za maksimizaciju profita (eng. *profit-maximization model*), omogućava se identifikovanje kategorije visokoprofitabilnih i niskoprofitabilnih respondenata (Cui et al., 2015; Otter et al., 2006). Definisanje profitabilnih segmenata važan je input za dizajniranje budućih marketing aktivnosti i kampanja, koje će, kao relevantan sadržaj za odabrane i targetirane kupce, ojačati njihovu vezu s kompanijom i uticati na održavanje baze lojalnih kupaca.

U ovom radu biće predstavljeni predlozi za nove prediktivne modele u svim navedenim sekcijama: model za RFM segmentaciju kupaca, model za predviđanje odgovora na kampanju i model za procjenu profitabilnosti kupaca u cilju targetiranja potencijalno najprofitabilnijih kupaca. Međutim, prethodno će u narednim sekcijama biti predstavljeni modeli iz sve tri kategorije na osnovu postojeće literature, kao i njihova primjena, prednosti i nedostaci, kako bi se identifikovali njihovi doprinosi predloženi u ovoj disertaciji.

3.1.1 Segmentacioni Recency-Frequency-Monetary modeli

Zadovoljstvo potrošača za savremene kompanije predstavlja osnov za njihov rast. Stoga, fokus kompanija je na prepoznavanju najvrednijih potrošača u cilju povećanja njihovog zadovoljstva, a posredno i tržišnog udjela. To se postiže sistematizovanjem i procesiranjem podataka o potrošačima i njihovim navikama, preferencijama i potrebama, da bi se, na kraju, na osnovu ovih podataka donijele poslovne odluke koje utiču na stvaranje dubljih i trajnijih veza s potrošačima. Poslovne odluke su bazirane na modelima za segmentaciju potrošača, a strategije komunikacije pripremljene prema definisanim segmentima predstavlju ultimativni rezultat ovog procesa. Segmentacija je ključna zato što omogućava da se svaki kupac ili grupa kupaca tretira na poseban način, što povećava njihovo zadovoljstvo i lojalnost (Safari et al., 2016).

Tradicionalni modeli segmentacije kupaca bili su bazirani na demografskim i psihografskim atributima kupaca, što je, uz subjektivnu procjenu marketing menadžera, rezultiralo niskim stepenom tačnosti, posebno u savremenom i komplikovanom poslovnom okruženju (Lee & Park, 2005). Stoga, segmentacija se danas zasniva na transakcionim podacima i podacima o ponašanju u kupovini, poput: vrste kupljenih proizvoda, obima transakcija u posmatranom periodu, broja poziva upućenih *call*-centru, broja reklamacija i zabilježene aktivnosti na *web* sajtu kompanije.

Jedna od najčešće korišćenih analiza u direktnom marketingu je RFM analiza, koja se u sličnim istraživanjima sprovodi već nekoliko decenija, a koju je u njenom najrasprostranjenijem obliku definisao *Hughes* (1994). RFM model predstavlja model baziran na prethodnom ponašanju potrošača u kupovini, zbog čega spada u grupu bihejvioralnih modela, a koristi se za predviđanje budućeg ponašanja zasnovanog na ponašanju evidentiranom u bazi (Hughes, 1996). Uzimajući u obzir dostupnost podataka potrebnih za RFM analizu, ova metoda predstavlja jednu od najčešće korišćenih od strane marketing profesionalaca (Jonker et al., 2006; Olson & Chae, 2012; Verhoef et al., 2002). Dodatna prednost ovog modela, pored dostupnosti podataka, odnosi se na jednostavnost upotrebe i kreiranja modela, posebno ako se radi o prvobitnom i tradicionalnom tipu RFM modela, zasnovanom na prostom tabelarnom prikazu.

Recency predstavlja dužinu vremenskog perioda od posljednje kupovine ili datum posljednje kupovine, *Frequency* označava broj obavljenih kupovina u navedenom periodu, dok *Monetary* definiše ukupnu monetarnu vrijednost transakcija kupca u tom periodu (Cheng & Chen, 2009; Hosseini & Shabani, 2015). Dakle, ova metoda omogućava jednostavno kvantifikovanje prethodnog ponašanja potrošača u kupovini.

Tradicionalna RFM analiza počinje sortiranjem podataka o transakcijama potrošača prema datumu kupovine (*Recency*) – baza podataka se dijeli na pet jednakih djelova, pri čemu 20% kupaca koji su najskorije kupili proizvod dobijaju broj 5, sljedećih 20% broj 4, itd. (McCarty & Hastak, 2007). Sljedeći korak uključuje sortiranje potrošača u okviru svih kvantila prema frekventnosti, te im se kao i u prvom koraku, dodjeljuju brojevi od pet do jedan. Važno je napomenuti da autori veću frekventnost posmatraju kao lojalnost potrošača (Wei et al., 2010). Nakon drugog koraka, baza je podijeljena u 25 grupa, pa se u posljednjem koraku svaka od njih dijeli na pet djelova - prema pokazatelju vrijednosti utrošene u svim posmatranim transakcijama u navedenom vremenskom periodu, što će, u konačnom, rezultirati bazom podijeljenom u 125 grupa, prema RFM vrijednostima (McCarty & Hastak, 2007). S

tim u vezi, najbolji segment potrošača u tabelarnom prikazu imaće ocjene 555, dok će najgori segment imati ocjene 111 (Wei et al., 2010; Yao & Xiong, 2011).

Na osnovu RFM rezultata, potrošači mogu biti grupisani u segmente, čija se profitabilnost može posebno i dodatno analizirati. Na primjer, ukoliko potrošač ima RFM rezultat 155, to znači da je u posmatranom periodu obavio veći broj kupovina visoke monetarne vrijednosti, ali nije kupovao već duži vremenski period. Ovaj rezultat može ukazati na situaciju prelaska u konkurentsku kompaniju, pa se strategija formulisana za ovu grupu potrošača mora bazirati na programu reaktivacije (Birant, 2011), najčešće korišćenjem značajnih sniženja cijene ili drugih promocionih taktika za privlačenje pažnje potrošača. S druge strane, potrošačima sa RFM rezultatom od 551 pristupa se sa *up-selling* strategijama, dok je grupi s rezultatom 515 potreban podsjetnik, zbog čega *Birant* (2011) predlaže kreiranje specijalnog paketa koji uključuje personalizovano pismo, listu benefita koje nudi kompanija, kao i neku formu podsticaja za obavljanje kupovine u narednih 30 dana. U skladu s prethodno navedenim, osnovni tip RFM modela ne zahtijeva softver za statističku obradu podataka. Tabelarni prikaz rezultata nakon dodjeljivanja bodova od 1 do 5 daje bazični uvid u prethodno ponašanje kupaca, na osnovu čega se mogu planirati naredne poslovne aktivnosti usmjerene ka njima.

Doğan et al. (2018) predložili su dva modela za segmentaciju potrošača distributera sportske opreme, korišćenjem RFM atributa. Prvi korak u njihovoj analizi bio je definisanje vrijednosti za RFM atribute, pri čemu su, za razliku od nekih prethodnih istraživanja koja su bazu potrošača dijelila na 5 jednakih djelova, ovi autori podijelili na tri dijela, te atributima R, F i M dodijelili vrijednosti od 1 do 3 (veća vrijednost – bolji rezultat). Nakon definisanja vrijednosti ovih atributa, predložili su dva modela za klasterizaciju: prvi se zasniva na *log-likelihood* metodi za mjerenje distance i *Shwarz's Bayesian* kriterijumu (*BIC*) za klasterizaciju, dok se drugi zasniva na *k-means* klasterizaciji. Prvi model je kao rezultat dao tri klastera, dok je drugi definisao četiri klastera potrošača.

Na sličan način, podjelom baze podataka na tri dijela, *Fader et al.* (2005) definisali su vrijednosti RFM atributa. U ovom radu, autori su povezali RFM atribute i procjenu

cjeloživotne vrijednosti potrošača (CLV) kroz vizuelizaciju „*iso-value*" krivih, koje prikazuju *trade-off* između ove dvije kategorije.

Klasična RFM analiza sprovedena je i u radu *Biranta* (2011), gdje autor predlaže pristup od tri koraka, koji koristi RFM analizu u *data mining* okvirima: klasterizacija na osnovu RFM atributa, s ciljem definisanja segmenata sličnih kupaca; klasifikacija na osnovu demografskih atibuta i RFM vrijednosti, za predviđanje budućeg kupovnog ponašanja i generisanje asocijativnih pravila za kreiranje preporuka budućim kupcima. U ovom istraživanju primijenjen je klasični pristup u definisanju vrijednosti za RFM atribute, podjelom skupa podataka na pet jednakih djelova po svim atributima pojedinačno, te dodjeljivanjem ocjene od 1 do 5. Jedan od nedostataka ovog rada je uključivanje RFM vrijednosti u proces klasifikacije, odnosno generisanja pravila klasifikacije korišćenjem C4.5 DT metode. Naime, vrijednosti RFM atributa već su korišćene za definisanje segmenata u procesu klasterizacije, što će vještački dovesti do veće tačnosti klasifikacije. Posljednji korak - generisanje preporuka, autor je realizovao korišćenjem *Frequent Pattern Growth* algoritma.

Cheng i *Chen* (2009) koristili su *k-means* klasterizaciju (MacQueen, 1967) za RFM segmentaciju. Podijelili su podatke na segmente od po 20% (ujednačeno kodiranje) i, kako bi testirali pristup, formirali su tri, pet i sedam klastera. Nedostatak ovog pristupa sastoji se u tome što ujednačeno kodiranje dovodi do gubitka razlika između vrijednosti RFM atributa (npr. kupci koji imaju pet ili devet transakcija ili oni čija vrijednost transakcija iznosi 5.000 eura ili 7.000 eura mogu biti postavljeni u isti rang). Dodatno, da bi razvili set pravila za targetiranje kupaca na osnovu njihovih karakteristika (region i kreditni dug), autori su koristili *rough set* i metodu ekstrakcije pravila *LEM2*. Prediktivni atributi takođe uključuju RFM atribute, međutim, na ovaj način se postiže visoka stopa tačnosti, jer su klasteri već formirani na osnovu RFM atributa. S tim u vezi, generisana pravila možda ne pokazuju neke značajne karakteristike kupaca za targetiranje, jer mogu biti apsorbovana efektom RFM atributa. Pored toga, RFM atributi su nepoznati za nove potencijalne kupce koji još uvijek nisu u bazi, pa se ovaj model ne može koristiti za njihovo predviđanje.

McCarty i *Hastak* (2007) su uporedili performanse RFM metoda s performansama CHAID (*Chi-square Automatic Interaction Detection*) pristupa i logističke regresije, koristeći isključivo varijable *Recency, Frequency* i *Monetary* zbog uporedivosti rezultata. U ovom procesu koristili su dvije studije – jednu sa stopom odgovora od 2,46%, a drugu sa znatno većom stopom - 27,4%. Njihovi rezultati su pokazali da RFM daje slične rezultate kao i pristupi s kojima je poređen u slučaju kada je stopa odgovora relativno visoka ili kada marketing menadžeri kampanju ili ponudu plasiraju većem dijelu baze kupaca. S druge strane, CHAID i logistička regresija dali su bolje rezultate u slučaju studije s nižom stopom odgovora. Autori naglašavaju da se prednost metoda poput CHAID i logističke regresije ogleda u činjenici da ove metode mogu u analizi da koristi veći broj nezavisnih varijabli, koje nisu ograničene na RFM varijable, poput podataka o motivima, stavovima, vrijednostima i životnom stilu.

Najveći broj istraživanja koji prati metodologiju klasične ili tradicionalne RFM segmentacije (Birant, 2011; Tsai & Chiu, 2004; Yao & Xiong, 2011) usmjerava pažnju na „555" potrošače, odnosno potrošače s najvećim ocjenama za sve atribute. S tim u vezi, interesantan je pristup koji je primijenio *Miglautsch* (2002), kako bi opisao kupce s najgorim ocjenama – „111" kupce. Naime, autor navodi da kupci koji rijetko kupuju, troše malo novca i nisu kupovali skoro, često čine gotovo 50% baze podataka o kupcima. Autor ističe da je za sveobuhvatniju analizu ovih, naizgled sličnih kupaca, potrebno uključiti i dodatne atribute, poput podataka iz baze transakcija. Dodatno, *Miglautsch* (2002) zapaža da je upravo ova grupa kupaca najveći neiskorišćeni i neistraženi potencijal svake kompanije.

Pored klasične RFM analize, realizovane kodiranjem varijabli vrijednostima od 1 do 5 (ili od 1 do 3), različiti autori su koristili različite verzije RFM analize. U ponderisanoj RFM verziji, sve tri vrijednosti atributa se množe s ponderima (wR, wF i wM) na osnovu relativne važnosti svakog atributa. Ovu verziju RFM analize primijenili su *Sarvari et al.* (2016) na podacima iz globalnog lanca picerija. *Khajvand et al.* (2011) su koristili ovu metodu za analizu segmenata kupaca drogerije na osnovu 7.000 transakcija, pri čemu su relativni značaj svakog atributa definisali

korišćenjem metode analitičkog hijerarhijskog procesa (eng. *Analytic Hierarchy Process* - AHP), kao i drugi autori (Liu & Shih, 2005; Shen & Chuang, 2009; Yao & Xiong, 2011), koji razliku u ponderima za RFM atribute pripisuju specifičnostima industrija i proizvoda.

Naime, *Hughes* (1994) je pristupio analizi dajući RFM varijablama jednak značaj, što je pratio i određeni broj drugih autora (Cheng & Chen, 2009). Međutim, jedan broj autora navodi da bi važnost RFM varijabli trebalo procjenjivati pojedinačno, što bi rezultiralo pridavanjem različite težine svakoj od njih (McCarty & Hastak, 2007; Stone, 1995). Proces određivanja težine svake promjenljive može biti različit - od jednostavnog menadžerskog subjektivnog ocjenjivanja, do njegove kombinacije s metodom AHP (Liu & Shih, 2005; Monalisa et al., 2019; Yao & Xiong, 2011). Ponderi daju različitu važnost svakoj od RFM varijabli, što znači da one, u konačnom, nemaju isti uticaj na vrijednost potrošača, s obzirom na to da se RFM metoda u velikoj mjeri koristi u analizi vrijednosti i profitabilnosti kupaca (Doğan et al., 2018; Heldt et al., 2019; Khajvand et al., 2011; Rogic & Kascelan, 2019; Rogić et al., 2022).

Kako je i ranije navedeno, autori *Liu* i *Shih* (2005) su primijenili AHP metodu u cilju dobijanja pondera RFM varijabli za analizu životne vrijednosti kupaca, koristeći prosuđivanje tri grupe evaluatora: administrativnih i poslovnih menadžera, marketing konsultanata i kupaca. Na osnovu njihove procjene, relativne težine RFM varijabli bile su: 0,731, 0,188 i 0,081, respektivno, pri čemu je najvažnija varijabla bila *Recency*. Grupa donosilaca odluka je takođe utvrdila značaj svake od RFM varijabli za kategorizaciju vrijednosti kupaca u radu autora *Monalisa et al.* (2019): menadžer, supervizor i administrativni radnik, što je bio input za AHP metodu za određivanje pondera: 0,058, 0,546 i 0,395, respektivno. U ovom slučaju, *Frequency* dobija najveću težinu. Dodatno, u radu *Khajvanda et al.* (2011), AHP metoda je, na osnovu mišljenja eksperata iz odjeljenja prodaje, odredila najveću vrijednost pondera za promjenljivu *Frequency* 0,637, zatim *Monetary* sa 0,258 i konačno *Recency* sa 0,105. Ove težine su kasnije korišćene za izračunavanje cjeloživotne vrijednosti kupaca, kako bi se kreirale i prilagodile različite marketing strategije za svaki definisani segment.

Slično su *Safari et al.* (2016) u svom radu procijenili težine RFM varijabli, gdje su težine postavljene na: 0,17, 0,47, 0,35, respektivno, kao i u radu autora *Chen et al.* (2017): 0,3, 0,6 i 0,1, što je u potpunosti zasnovano na subjektivnoj procjeni evaluatora. U oba slučaja, utvrđeno je da je atribut *Frequency* od najveće važnosti. Ove informacije su korišćene za izračunavanje RFM atributa i identifikaciju kupaca visoke vrijednosti, kao i pružanje uvida u njihove preferencije i obrasce kupovine. U Tabeli 1 prikazane su RFM težine dobijene u prethodnim studijama.

Tabela 1. RFM ponderi dobijeni u prethodnim studijama

Autori	*Recency* težina	*Frequency* težina	*Monetary* težina
Liu i *Shih* (2005)	0,731	0,188	0,081
Monalisa et al. (2019)	0,058	0,546	0,395
Khajvand et al. (2011)	0,105	0,637	0,258
Safari et al. (2016)	0,170	0,470	0,350
Chen i Wang (2017)	0,300	0,600	0,100

Pored dodjele pondera za svaki od RFM atributa, određeni autori su težili unapređenju metoda kroz uključivanje dodatnih varijabli. Na primjer, *Heldt et al.* (2019) su umjesto klasične RFM analize, sproveli analizu sa izmijenjenim *Monetary* atributom. Naime, njihov model sadrži standardne atribute *Recency* i *Frequency*, dok se *Monetary* vrijednost računa po proizvodima, te su tako formirali RFM/P model. Svojim modelom, autori su prvo procjenjivali vrijednost potrošača po kupljenim proizvodima, a zatim su rezultate agregirali kako bi dobili sveobuhvatnu sliku vrijednosti kupaca. Rezultati njihovog rada pokazali su da informacije o kupljenim proizvodima mogu unaprijediti proces procjene vrijednosti kupaca. Dodatno, dimenziju kupljenih proizvoda su u svoj prilagođeni RFM model uključili i *Chang* i *Tsai* (2011) u svom GRFM (*Group RFM*) modelu.

Alizadeh Zoeram i *Karimi Mazidi* (2018) su u cilju obuhvatanja mjere za lojalnost potrošača, kao dodatnu varijablu definisali *Length* (LRFM), dok su *Sheikh et al.* (2019) osim ovog, dodali i atribut *Periodicity*, te su tako kreirali model LRFMP. Konačno, *Horita* i *Yamashita* (2019) su proširili osnovni model atributom *Distance*

(RFMD), kako bi proces segmentacije kupaca obuhvatio i mjeru udaljenosti potrošača od mjesta kupovine. Iz navedenih primjera se može zaključiti da je RFM model izuzetno fleksibilan, te da se može prilagoditi i primijeniti u različitim industrijama.

Uopšteno govoreći, kvalitet segmentacije mjeri se homogenošću u okviru segmenta i heterogenošću između segmenata (Wedel & Kamakura, 2000). S tim u vezi, jedan od nedostataka RFM metode odnosi se na ograničen broj varijabli koji se uzima u obzir za analizu segmenata. Naime, za sveobuhvatnu analizu kupovnog ponašanja potrošača, potrebno je razmotriti i dodatne karakteristike koje mogu imati uticaja na mogućnost odgovora na kampanju ili profitabilnost kupca. Stoga, autori predlažu da se, pored RFM atributa, u opis i definiciju segmenata uključe i drugi parametri, kao što su demografski podaci, što dovodi do generisanja tačnijih i detaljnijih pravila za buduće targetiranje korisnika (Sarvari et al., 2016). *Buckinx* i *Van den Poel* (2005) ističu da RFM varijable predstavljaju najznačajnije podatke za procjenu mogućnosti napuštanja kupca, kao i da, zajedno s demografskim podacima, čine idealnu kombinaciju za predviđanje odliva kupaca. Ovaj rezultat autora ukazuje na značaj RFM varijabli u pocesu segmentacije kupaca, pri čemu se posebno ističe prepoznavanje segmenta onih kupaca koji s velikom vjerovatnoćom mogu napustiti kompaniju u korist konkurenata. Stoga se, uz uvid u definisane segmente, određenim marketing aktivnostima može uticati da se stopa odliva kupaca svede na minimum.

Segmenti mogu biti opisani kao grupa kupaca koji imaju slične demografske karakteristike, vrijednosti i ponašanje u kupovini, a napredak informacione i komunikacione tehnologije, sistema za upravljanje bazama podataka i *data mining* tehnika u potpunosti mijenjaju način sprovođenja ovog procesa na savremenom tržištu. Prepoznavanjem obrazaca i učenjem na osnovu istorijskih podataka, vještačka inteligencija i mašinsko učenje mogu, uz veliku tačnost, generisati preporuke za definisane segmente i ponuditi proizvode koji će zadovoljiti potrebe potrošača (Rogić & Kašćelan, 2021).

RFM metoda je često korišćena uz različite *data mining* metode u cilju klasterizacije i klasifikacije kupaca. *Chan* (2008) je u svojoj studiji za segmentaciju preko 4.000 kupaca *Nissan* automobila prvo primijenio RFM model za analizu ponašanja potrošača, a zatim analizu cjeloživotne vrijednosti potrošača (eng. *lifetime value – LTV*) za evaluaciju predložene segmentacije. Osim toga, u ovoj studiji je predloženo korišćenje genetskih algoritama (eng. *Genetic Algorithm* - GA) za selekciju odgovarajućih potrošača za svaku narednu kampanju i strategiju. S druge strane, *Tsai* i *Chiu* (2004) su koristili GA kao dio algoritma za klasterizaciju kupaca, odnosno njihovu segmentaciju, a zatim su na osnovu RFM modela analizirali relativnu profitabilnost svakog od segmenata.

Korišćenjem podataka iz kompanije koja se bavi maloprodajom, *González Martínez et al.* (2019) su uporedili klasični RFM model sa RFM modelom baziranim na *2-tuple fuzzy linguistic* pristupu,. Primijenili su *k-means* algoritam za klasterizaciju na oba modela i zaključili da *2-tuple RFM* može jasnije da identifikuje klastere od klasičnog. *Liu* i *Shih* (2005) su predložili kombinaciju ponderisanog RFM modela, korišćenjem AHP metode, klasterizacije i metode zasnovane na asocijativnim pravilima, u cilju preciznijeg definisanja segmenata i odgovarajućih marketing strategija. *K-means* algoritam se često u literaturi koristi uz RFM metodu za poboljšanje CRM sistema (Hosseini & Shabani, 2015; Hosseini et al., 2010), procjenu vrijednosti i lojalnosti potrošača, kao i predviđanja odlaska korisnika (eng. *churn prediction*) (Abdi & Abolmakarem, 2019), a biće detaljno opisan u sekciji 4.1.1.

Među najznačajnijm prednostima RFM modela ističu se troškovna efikasnost u prikupljanju podataka i jednostavnost kvantifikovanja prethodnog ponašanja potrošača, kao i njegovog opisivanja kroz samo tri varijable. Uz to, transformacija podataka ne zahtijeva napredna znanja, a kroz primjenu ovog modela mogu se dobiti značajne informacije o vrijednosti potrošača, te ovo može biti adekvatna polazna metoda za sveobuhvatniju analizu profitabilnosti potrošača. S druge strane, nedostaci ovog modela ogledaju se u već pomenutom fokusu na „najbolju" grupu potrošača i fokusiranju na samo tri varijable pri opisivanju segmenata. Međutim, kako je prethodno navedeno, ovi nedostaci su akcentirani u naučnoj i stručnoj

literaturi. Pored ovih nedostataka, autori iz ove oblasti navode i nedostatak koji se odnosi na mogućnost analize isključivo postojećih potrošača i činjenicu da se ova metoda ne može primijeniti za akviziciju novih kupaca, jer podaci o njima ne postoje u bazi. Konačno, preciznost ovog modela se dovodi u pitanje zbog njegove jednostavnosti, kao i značaj primjene u određenim industrijama (Yeh et al., 2009).

Iako je originalno razvijena za potrebe kompanija koje se bave maloprodajom, RFM metoda je tokom godina modifikovana kako bi se mogla primijeniti i u drugim industrijama. Počeci primjene ove metode vezuju se za osamdesete godine prošlog vijeka, kada je i dokazana njena efikasnost u targetiranju kupaca u kampanjama direktnog marketinga (Tsiptsis & Chorianopoulos, 2010). Međutim, usložnjavanjem tržišnog okruženja, RFM podaci postali su nedovoljni za dobijanje potpune slike o kupcima, tako da je u literaturi predloženo da se ovaj pristup kombinuje s drugim važnim podacima o kupcima, poput preferencija proizvoda koje su ranije kupili, njihovog opisa, kao i *web* ponašanja, što je realizovano u empirijskom dijelu ove disertacije. Na ovaj način, korišćenjem većeg broja atributa koji opisuju ponašanje u kupovini, kompanije mogu dobiti značajne informacije koje im mogu pomoći u donošenju poslovnih odluka i prilagođavanju njihovih marketing aktivnosti i komunikacionih strategija u skladu s jasnim profilima svakog definisanog potrošačkog segmenta.

Uzimajući sve ovo u obzir, donosioci odluka mogu iskoristiti adekvatnu formu RFM modela i u cilju efikasnije analize problema i izazova s kojim se suočavaju u svakodnevnim aktivnostima, poput precizne analize ponašanja potrošača, njihove segmentacije, procjene profitabilnosti ili cjeloživotne vrijednosti.

U narednom dijelu rada biće predstavljeni modeli odgovora na kampanju.

3.1.2 Modeli odgovora kupaca

Nove tehnologije, promjene obrazaca potrošnje i strukture tržišta uslovile su dinamiziranje marketing inovacija u 21. vijeku. Era masovnog marketinga i proizvodne orijentacije je prošla, kao i vrijeme pojednostavljenih segmentacionih

modela, pa, uz postojanje hiperkonkurencije, marketing menadžeri ne mogu više da se oslanjaju na jednostavne i individualne segmentacione faktore (poput demografskih i psihografskih). Osim toga, u novijim istraživanjima sve relevantniji postaju modeli odgovora na kampanju koji uključuju *web* metrike (Baumann et al., 2019; Chaudhuri et al., 2021; Lee et al., 2021). Međutim, u tim slučajevima je posebno izražen problem nebalansiranosti klasa, s obzirom na to da je stopa odgovora na kampanju mnogo niža zbog velikog broja korisnika koji pristupe sajtu, a čija se sesija ne završava transakcijom.

Promjena marketing fokusa s proizvoda na potrošača posebno je ubrzana tokom posljednje decenije zbog sve većeg interesovanja za poslovnu inteligenciju (eng. *business intelligence*) uopšte i posebno za upravljanje odnosima s kupcima. Uzimajući u obzir važnu ulogu koju odluke marketing sistema imaju u trenutnom okruženju usmjerenom na kupca, nameće se potreba za jednostavnim i integrisanim okvirom za sistematsko upravljanje raspoloživim podacima o kupcima. Savremeni potrošači su obrazovani i sofisticirani, što uslovljava potrebu za razvojem marketing strategija koje će zadovoljiti njihove zahtjeve (Hauser et al., 2011).

Marketing kampanje, kao jedna od najčešće korišćenih aktivnosti koje su usmjerene ka kupcima, predstavljaju osnov za razvoj i poboljšanje odnosa s njima. Međutim, one se često sprovode za ispunjavanje kratkoročnih ciljeva i ostvarivanje profita, dok su u tom procesu odnosi s kupcima zapostavljeni. Ovom problemu dodatno doprinosi i nerazumijevanje potreba potrošača i njihovog ponašanja u kupovini. Važnost informacija o potrebama, željama i kupovnim navikama potrošača za planiranje i realizaciju aktivnosti orijentisanih ka potencijalnim kupcima posebno je naglašena u prethodnim istraživanjima iz ove oblasti (Jayachandran et al., 2004; Kumar et al., 2008; Shaw et al., 2001).

Jedan od ciljeva koji opredjeljuje uspješnost kampanje direktnog marketinga predstavlja procjenu broja i strukture kupaca koji će odgovoriti na kampanju, što je upravo problematika koju tretiraju modeli odgovora na kampanju. Ovi modeli dijele grupu potencijalnih kupaca na respondente i nerespondente, odnosno grupu koja će s većom vjerovatnoćom odgovoriti na kampanju direktnog marketinga kupovinom,

naspram one grupe koja ima nižu vjerovatnoću odgovora. S tim u vezi, modeliranje odgovora na kampanju od strane potrošača predstavlja važan segment direktnog marketinga, s obzirom na to da identifikovanje potrošača s većom vjerovatnoćom odgovora može smanjiti troškove marketinga i povećati promet tokom kampanje. Ovi modeli, kreirani na osnovu istorijskih podataka o kupovini, primjenjuju se s ciljem predviđanja vjerovatnoće da će potencijalni kupac odgovoriti na kampanju direktnog marketinga kupovinom (Blattberg, 1987).

Ovaj prediktivni model dodjeljuje buduće vjerovatnoće odgovora kupcima na osnovu istorijskih podataka o kupovini. Distribucija marketing poruka ne obavlja se korišćenjem masovnog pristupa, a targetiranje kupaca ne zasniva se na bazičnim demografskim karakteristikama, već se optimizacija poruka i izbor ciljne grupe definiše na osnovu prethodnog kupovnog ponašanja (Jonker et al., 2004; Rowe, 1989). Dakle, prije lansiranja kampanje, ključni izazov je efikasan izbor kupaca koji će se targetirati.

Međutim, udio potrošača koji odgovore na kampanju kupovinom je mali, tj. često je stopa odgovora u kampanjama veoma niska. Kada je procenat manje klase manji od 5%, može se konstatovati da se javlja tzv. „rijetki događaj" (eng. *rare event*) (Au et al., 2010). Aktuelnost ovog problema u akademskoj zajednici zasniva se na činjenici da algoritmi mašinskog učenja obično imaju poteškoća da uče iz podataka s neuravnoteženim klasama. Naime, u cilju minimiziranja ukupne stope greške (Vassiljeva et al., 2017), često dolazi do pristrasnosti modela prema većinskoj klasi. Ovo bi, kao rezultat toga moglo dovesti do pogrešne klasifikacije svih instanci minorne klase, što bi rezultiralo lošim performansama klasifikacije (He & Garcia, 2009; Wang & Pineau, 2016). Pitanje koje je manje istraženo, a koje je veoma relevantno u kontekstu rasta *online* trgovine, jeste da li je i u kojoj mjeri ponašanje korisnika na internetu važno za predviđanje odgovora korišćenjem metoda mašinskog učenja (Rogić & Kašćelan, 2022).

Dakle, podaci u bazama su nebalansirani, pa se dizajniranje efikasnog modela odgovora na kampanju nameće kao jedan od izazova direktnog marketinga (Daneshmandi & Ahmadzadeh, 2013). Međutim, ako se donosioci odluka suočavaju

s nedostatkom kvalitetnih informacija o kupcima, to može prouzrokovati određeni gubitak kupaca, pa je predviđanje odgovora na kampanju aktuelan problem i predmet interesa stručnjaka za marketing i analitičara.

Informacije koje model odgovora na kampanju pruža donosiocima odluka od posebne su važnosti za alociranje marketing resursa aktivnim kupcima s visokom potencijalnom vrijednošću za kompaniju. Nakon vremena direktne pošte, danas se, uz razvoj interneta i društvenih mreža, kao i mogućnosti plasiranja kampanja na ovaj način, otvara polje za istraživanje efikasnosti ovog medija. Internet je omogućio akumuliranje velike količine podataka i postao je jedan od najefikasnijih sistema za njihovo skladištenje. U tom virtuelnom okruženju, koncept direktnog marketinga posebno dobija na značaju, s obzirom na to da obuhvata informacione tehnologije koje mogu generisati i skladištiti mnoštvo složenih i kompleksnih podataka (Aliabadi & Berenji, 2013). Dodatno, društvene mreže imaju značajnu ulogu u procesu razvoja brenda, a pored toga, obilje podataka o korisnicima koje je dostupno na ovim platformama, omogućava precizniju selekciju i targetiranje za direktni marketing. U modeliranju odgovora na kampanju putem društvenih medija, inputi uključuju različite tipove i strukture podataka, a svrha ovog procesa ostaje ista – efikasno identifikovati respondente (Sun et al., 2014). Adekvatan izbor ciljne grupe respondenata utiče na smanjenje ukupnih troškova marketing aktivnosti. Ovo implicira da se efikasnije marketing odluke mogu donijeti prihvatanjem naprednih prediktivnih modela i tehnika, u odnosu na tradicionalne, deskriptivne modele (Olson & Chae, 2012).

S obzirom na to da srž CRM-a predstavljaju upravo podaci o kupcima, problematika modeliranja odgovora na kampanju uklapa se u filozofiju „jedan na jedan" marketinga i potrebe razvoja CRM sistema (Mahdiloo et al., 2014). Ovaj sistem omogućava efikasno prikupljanje i skladištenje, te jednostavno analiziranje podataka, što pruža mogućnosti za izvlačenje značajnih zaključaka iz podataka o kupcima. S tim u vezi, autori *Danaher* i *Rossiter* (2011), kao i *Ngai et al.* (2009) navode da su CRM i modeliranje odgovora na kampanju neodvojivi i kritični faktori uspjeha na savremenom tržištu. Uz pravu količinu i vrstu informacija iz ovih izvora,

donosici odluka u marketingu mogu s većom tačnošću da odrede grupe kupaca koje će targetirati u promocijama, kao i učestalost targetiranja.

Kada je u pitanju izbor podataka za analizu i predviđanje odgovora na kampanju, autori su u prethodnim istraživanjima koristili različite skupove i baze podataka. *Hauser et al.* (2011) su analizirali efikasnost sjedinjavanja tzv. „tvrdih" podataka iz kompanijske baze podataka (podaci o prethodnim kupovinama) s „mekim" podacima (demografskim i psihografskim) iz sekundarnih izvora. Rezultati ovog istraživanja su pokazali da su „tvrdi" podaci u značajno većoj mjeri izabrani kao efikasni prediktori u odnosu na drugi tip. Ovi zaključci ukazuju na to da se precizni modeli odgovora na kampanju mogu kreirati korišćenjem isključivo istorijskih podataka o transakcijama na nivou kupca, bez uključivanja demografskih i psihografskih podataka (Hauser et al., 2011), što uklanja dodatne troškove prikupljanja i procesiranja dodatnih podataka iz sekundarnih izvora.

Model odgovora na kampanju se najčešće obučava na skupu podataka u kojem se posmatraju nezavisne promjenljive, koje opisuju i profilišu određenog kupca i zavisna promjenljiva (odgovor na kampanju), koja pokazuje da li je kupac odgovorio na posmatranu kampanju. Nakon toga, obučeni model se primjenjuje na novom skupu - testnom skupu podataka. Kao rezultat ovog procesa, dobija se vjerovatnoća odgovora svakog kupca u testnom skupu, u zavisnosti od njegovog ili njenog prethodnog ponašanja u kampanjama (Coussement et al., 2015). Ako se posmatra upravljačka strana, rezultat modela odgovora na kampanju treba donosiocima odluka da obezbijedi informaciju o targetiranju određenog procenta kupaca s najvećom vjerovatnoćom odgovora, u skladu s raspoloživim budžetom za predviđenu buduću kampanju.

Yao et al. (2014) su razvili konceptualni model za otkrivanje promjena u ponašanju kupaca tokom trajanja marketing (prodajne) kampanje. U svojoj studiji, autori su dokazali da sprovođenje kampanje ima direktan pozitivan efekat na prosječni dnevni prihod, kao i da kampanje najviše uticaja ostvaruju upravo na segment najvrednijih kupaca za kompaniju. S tim u vezi, autori navode da integracija sistema segmentacije kupaca i modeliranja odgovora na kampanju može poboljšati

efikasnost analitičkog CRM modela radi boljeg upravljanja kampanjama i odnosima s kupcima. Dodatno, njihov model segmentacije kupaca je dinamički, tj. obuhvata promjene u okviru segmenata, što u srednjem i dužem roku uključuje promjene pripadnosti kupca određenom segmentu, kao i promjene strukture segmenata tokom vremena, što unapređuje tačnost modela.

U tradicionalne tehnike za modeliranje odgovora kupca, karakteristične za ranija istraživanja, spadaju regresione tehnike, prvenstveno linearna i logistička regresija. *Levin* i *Zahavi* (1998) su u modelu linearne regresije za predviđanje odgovora kupaca primijenili analizu profitabilnosti, kriterijume dobrog uklapanja i tačnost predviđanja. Takođe, model linearne regresije koristi *Malthouse* (1999), radi predviđanja koliko će kupac potrošiti u predstojećoj promotivnoj ponudi. Donosioci odluka u marketingu bi mogli da kalibrišu modele logističke regresije u malim bazama podataka, koristeći pristupe maksimalne vjerovatnoće da bi otkrili relevantne faktore i izmjerili vjerovatnoću reakcije kupaca (Naik & Tsai, 2004). S obzirom na to da je varijabla odgovora diskretna, logistička regresija znatno olakšava izgradnju modela (Wang, 2013). S tim u vezi, u ranijim radovima iz oblasti predikcije odgovora na kampanju, često su se koristili modeli logističke regresije.

Neki od često korišćenih modela odgovora na kampanju, koji ne koriste *data mining* tehnike su *Probit* i *Logit*, odnosno regresioni modeli. Navedeni modeli su jednostavni za upotrebu, međutim, nemaju zadovoljavajuću eskplanatornu moć (Bose & Chen, 2009). *Probit* i *Logit* modeli kreiraju binarnu ocjenu izbora (Bult et al., 1997) ili kategoričke vrijednosti ako postoji više od dva izbora. Jedan od nedostataka ovih modela je pretpostavka simetrije troškova pogrešne klasifikacije, odnosno dodjeljivanje istih pondera lažno negativnim i lažno pozitivnim greškama klasifikacije. Međutim, ovi troškovi nisu simetrični – lažno negativne su „skuplje" od lažno pozitivnih grešaka.

DeSarbo i *Ramaswamy* (1994) su predstavili postupak segmentacije nazvan CRISP (*Customer Response-based Iterative Segmentation Procedures*) za simultano generisanje tržišnih segmenata i kreiranje modela odgovora kupaca u svakom od definisanih segmenata. Autori su procjenjivali višestruke segmente kupaca na

osnovu podataka o odgovoru na kampanju, uz procjene veličine svakog izvedenog segmenta, vrijednosti parametara na nivou segmenta i njihove statističke značajnosti, kao i vjerovatnoće pripadnosti segmentu. U ovom radu model je testiran na skupu podataka izdavača časopisa, koji uključuje binarne podatke o odgovoru na ponudu domaćinstvima da se pretplate na časopis.

Za model odgovora na kampanju *Deichmann et al.* (2002) su predložili primjenu *Multiple Adaptive Regression Splines* (MARS) i logističku regresiju. Rezultati su pokazali da je MARS model pružio bolje rezultate, pri čemu su kao osnovne prednosti modela naveli: sposobnost identifikovanja relativno malog broja prediktorskih varijabli, koje nastaju složenom transformacijom početnih varijabli i kao takve sadrže značajnu eksplanatornu moć; sposobnost otkrivanja nelinearnosti, koje mogu postojati u odnosu između zavisne varijable (odgovora na kampanju) i prediktora; identifikovanje interakcije, kao i mogućnost kreiranja grafikona koji pomažu u vizualizaciji i razumijevanju interakcija među varijablama.

S obzirom na to da se baze podataka o kupcima u realnom vremenu dopunjavaju s novim podacima, tradicionalne tehnike za modeliranje odgovora na kampanju mogu pokazati određene nedostatke. *Genkin et al.* (2007) navode da zbog nedostatka konvergencije, velikih procijenjenih varijansi koeficijenata, niske tačnosti predviđanja i ograničene moći za testiranje hipoteza o procjeni modela, upotreba procjene maksimalne vjerovatnoće (eng. *Maximum Likelihood Estimation* - MLE) u skupu podataka s velikim brojem prediktorskih varijabli može biti neuspješna. Pored toga, isti autori navode da ove tehnike mogu dovesti do pojave problema pretjeranog prilagođavanja, kao i problema specifikacije modela.

U studijama novijeg datuma, sve više su aktuelne *data mining* tehnike za predviđanje odgovora kupaca. Tako, na primjer, autori *Kang et al.* (2012) dizajnirali su model odgovora kupaca koristeći klasterizaciju, balansirano poduzorkovanje i *ensemble* (eng. *Clustering, Balanced Undersampling and Ensemble* - CUE), kako bi riješili problem neravnoteže klasa respondenata, uparujući ga s nekoliko klasifikacionih algoritama za predviđanje – logistička regresija (LR), *multi-layer perceptron* neuronska mreža, *k-Nearest Neighbors* klasifikacija (k-NN) i SVM. Autori su koristili

metodu poduzorkovanja za nerespondente iz svakog klastera, kako bi izbjegli gubitak informacija relevantnih za klasifikaciju, do kojih može doći pri primjeni slučajnog poduzorkovanja. Njihov CUE pristup pokazao je bolje rezultate u balansiranju respondenata i nerespondenata u poređenju sa slučajnim poduzorkovanjem, slučajnim preuzorkovanjem, SMOTE tehnikom uzorkovanja, jednostranom selekcijom i rezultatima na originalnom skupu bez uzorkovanja. Pored toga, autori su zaključili da je SVM pokazao najbolje performanse modela u neuravnoteženim okolnostima. Međutim, autori su se fokusirali na metode balansiranja podataka, pa stoga nisu detaljnije prikazali rezultate tačnosti za segment respondenata, već samo ukupnu tačnost modela.

U cilju rješavanja problema neravnoteže klasa, koristeći skup podataka osiguravajućeg društva sa stopom odgovora od 6%, *Farquad* i *Bose* (2012) testirali su SVM kao pretprocesor podataka, zajedno s tehnikama uzorkovanja (100% preuzorkovanje, 200% preuzorkovanje, 25% poduzorkovanje, 50% poduzorkovanje i SMOTE). Nakon prethodne obrade podataka i zamjene ciljne varijable sa SVM predikcijom, takav modifikovani skup podataka je korišćen za obuku MLP, LR i RF modela. Autori su se fokusirali na metriku senzitivnosti, tj. proporciju *True Positive* (TP) primjera. Rezultati pokazuju da predloženi pristup balansiranja podataka poboljšava performanse klasifikacije u svakom slučaju. S tim u vezi, MLP, LR i RF su dobili sljedeće rezultate senzitivnosti na originalnim neuravnoteženim podacima: 5,88%, 1,26% i 7,14%, respektivno, dok su u kombinaciji sa SVM pretprocesiranjem rezultati bili sljedeći: 65,31%, 63,03% i 63,03%, respektivno. Najbolji učinak u ovoj studiji postignut je korišćenjem 25% poduzorkovanih podataka u RF modelu, koji je postigao senzitivnost od 71,01%.

U svom istraživanju, *Daneshmandi* i *Ahmadzadeh* (2013) predložili su novi pristup za tretiranje problema neravnoteže klasa i pokazali su da najveću tačnost predikcije ostvaruje hibridni model vještačkih neuronskih mreža (ANN). Da bi kreirali hibridni model, autori su primijenili *Bagging* neuronsku mrežu (BNN) na izlazu *k-means* klasterizacije, nakon čega slijedi agregiranje rezultata. Dobijeni rezultat senzitivnosti za hibridni model bio je 89%, dok je AUC rezultat iznosio 0,985, u

poređenju sa samostalnim BNN, sa senzitivnošću od 55% i AUC jednakim 0,88. Stoga su autori naveli da se hibridne tehnike obično predlažu kao efikasnije od osnovnih klasifikatora. Ovaj pristup je testiran na skupu podataka sa stopom odgovora od 19,81%.

Aliabadi i *Berenji* (2013) su predstavili hibridni model za predviđanje odgovora na kampanju, koji se sastoji od četiri faze. Ovaj predloženi model zasniva se na vještačkim neuronskim mrežama i genetskom algoritmu, uz metode pretprocesiranja podataka, kao što su redukcija podataka i izbor instanci u cilju poboljšanja tačnosti modela. Predstavljeni rezultati ukazali su na efikasnost ovog pristupa i visoku tačnost predviđanja, a kao osnovne prednosti autori navode sposobnost relativno brzog obučavanja modela i visok stepen preciznosti. Model su primijenili na podacima Edukativne fondacije za direktni marketing (eng. *Direct Marketing Educational Foundation dataset* - DMEF2). *Asare-Frempong* i *Jayabalan* (2017) su takođe analizirali modele za predviđanje odgovora na kampanju direktnog marketinga na podacima iz bankarskog sektora. Primijenili su četiri klasifikatora: *Multilayer Perceptron Neural Network* (MLPNN), drvo odlučivanja (C4.5), logističku regresiju and *Random Forest*. Njihovi rezultati su pokazali da RF ima najbolje prediktivne sposobnosti, sa tačnošću od 86,8% i AUC vrijednosti 0,927. Ostvarena TP stopa u njihovoj studiji bila je 90,2%, za skup podataka sa 11,63% respondenata. Ovaj skup podataka bio je poduzorkovan u odnosu 1:1 prije primjene modela.

Kim et al. (2013) su koristili tri skupa podataka *Direct Marketing Education Foundation* (DMEF) (1, 2 i 4) da testiraju svoj pristup, sa stopama odgovora od 27,42%, 2,46% i 9,42%, respektivno. Primijenili su dva nasumična pravila poduzorkovanja (2:1 i 1:1) da uravnoteže podatke. Za skup podataka s najvišom klasnom neravnotežom (DMEF2), bez balansiranja podataka, SVM je postigao najbolji rezultat senzitivnosti (7,3%) i ukupne tačnosti (95,3%), dok su DT, LR i NN ostvarili 0% senzitivnosti, pokazujući čest i najznačajniji problem u klasifikaciji neuravnoteženih podataka – svi modeli su bili pristrasni prema dominantnoj klasi. Nakon poduzorkovanja od 2:1, SVM je i dalje postigao najbolje performanse

klasifikacije sa 23,8% senzitivnosti, dok je njegova efikasnost smanjena nakon poduzorkovanja u odnosu 1:1, gdje je ostvario najmanju stopu senzitivnosti od 9,5%, u poređenju sa DT, LR i NN sa 41,1%, 56,5% i 62,9%, respektivno.

U svojoj studiji *Mandapaka et al.* (2014) primijenili su DT, NN, LR i SVM model za predviđanje odgovora na plasiranu ponudu. Njihovi rezultati su pokazali da *Stepwise Logistic Regression Model* ima najmanju stopu pogrešno klasifikovanih klijenata. Izborom tri grupe korisnika na osnovu rezultata najboljeg modela, kumulativna stopa odgovora povećala se na 14,5%, u odnosu na osnovnu stopu odgovora od 5%. Pored ovih autora, SVM su, u cilju predviđanja odgovora na kampanju, koristili i *Shin* i *Cho* (2006). Ovi autori su ukazali na probleme s kojima se susreću analitičari podataka u ovoj sferi, poput velikih skupova podataka za obučavanje, binarnog SVM autputa i problema nebalansiranosti klasa.

Na skupu podataka direktnog marketinga iz portugalske banke sa stopom odgovora od 11,2%, *Migueis et al.* (2017) primijenili su RF metodu na reuzorkovane skupove podataka, i to: sa prekomjernim uzorkovanjem (SMOTE) i poduzorkovanim (*EasyEnsemble*). Autori su postigli najbolje rezultate kombinacijom poduzorkovanja i RF metode – AUC je iznosio 0,989, za razliku od preuzorkovanih i originalnih rezultata skupa podataka od 0,945 u oba slučaja. Ovi rezultati dobijeni RF modelom upoređeni su sa LR, NN i SVM modelima, ali je u svakom slučaju RF i dalje nadmašio druge tehnike. Međutim, poduzorkovanje je jedino značajno poboljšalo rezultate kada je korišćen RF kao klasifikator. U drugim slučajevima se pokazalo da to nije univerzalno pogodna metoda za prevazilaženje problema neravnoteže klasa.

Marinakos i *Daskalaki* (2017) su tretirali problem neravnoteže klasa upoređujući statističke algoritme, algoritme za klasifikaciju zasnovane na udaljenosti, indukciju i mašinsko učenje, koristeći javno dostupan skup podataka sa stopom odgovora od 11,7% na ponudu direktnog marketinga. Najbolji učinak je postignut kombinovanjem tehnike poduzorkovanja, zasnovane na klasterima i k-NN – TP stopa je bila 88%, dok je SVM postigao TP stopu od 71%. Autori su naveli da su, bez obzira na izabrani algoritam, poduzorkovanje zasnovano na klasterima i SMOTE dobili sličan rezultat TP ≈ 70%.

Kada je riječ o *online* direktnim marketing kampanjama i ukupnom predviđanju kupovine preko interneta pomoću podataka *web* evidencije, postoji niz novijih radova koji tretiraju ovaj problem. Na primjer, autori *Lee et al.* (2021) pokazali su da je algoritam *eXstreme Gradient Boosting* (XGB), u kombinaciji sa SMOTE preuzorkovanjem, najefikasniji za predviđanje stope konverzije posjetilaca *online* prodavnice, dok su *Chaudhuri et al.* (2021) najbolje rezultate ostvarili korišćenjem algoritama dubokog učenja (eng. *deep learning*) i neuronskih mreža. S obzirom na to da se odnose na *online* kupovinu, ovi radovi će detaljnije biti opisani u okviru sekcije 3.3.

Sažetak relevantnih radova koji tretiraju modeliranje odgovora kupaca predstavljenih u ovom dijelu rada dat je u Tabeli 2[2].

Tabela 2. Performanse najefikasnijih modela u prethodnim studijama modeliranja odgovora na kampanju

Autor(i)	Stopa odgovora	Metoda	Tačnost	Senzitivnost	AUC
Kang et al. (2012)	9,42% (DMEF4)	CUE sa k-NN klasifikatorom	84,5%	BCR - 83,7%	-
Farquad i *Bose* (2012)	6%	25% poduzorkovanje u kombinaciji sa RF klasifikatorom	40,28%	71,01%	0, 5467
Kim et al. (2013)	2,46%	2:1 poduzorkovanje u kombinaciji sa SVM klasifikatorom	95,2%	23,8%	-
Daneshmandi i *Ahmadzadeh* (2013)	19,81%	*K-means* klasterizacija u kombinaciji sa BNN	96,5%	89%	0,985
Asare-Frempong i *Jayabalan* (2017)	11,63%	Balansirani (poduzorkovani) RF	86,8%	90,2%	0,927

[2] U slučaju da je u radu prikazano više metoda i rezultata, odabran je model s najboljim performansama.

Migueis et al. (2017)	11,2%	RF metoda na poduzorkovanom skupu podataka (*EasyEnsemble*)	-	-	0,989
Marinakos i *Daskalaki* (2017)	11,7%	Poduzorkovanje bazirano na klasterizaciji i k-NN	-	88%	0,90
Lee et al. (2021)	2,29% (stopa konverzije)	XGB sa SMOTE preuzorkovanjem	74,17%	73,92%	0,791
Chaudhuri et al. (2021)	-	*Deep learning* neuronske mreže	89%	96%	0,89

Najniža stopa odgovora u skupovima podataka u prethodnim studijama bila je 2,29% (stopa konverzije) (Lee et al., 2021), dok je najviša bila 27,42% (Kim et al., 2013), što je značajno više od stope odgovora u bazi podataka koja je korišćena u empirijskom dijelu ovog istraživanja, čiji će rezultati biti predstavljeni u sekcijama 5.4 i 5.5.

Važno je napomenuti da visoka vjerovatnoća odgovora na kampanju ne mora podrazumijevati i visoku profitabilnost kupca. Stoga, donosioci odluka u direktnom marketingu prilikom targetiranja potencijalnih kupaca, osim vjerovatnoće odgovora na kampanju, moraju uzeti u obzir i stepen profitabilnosti kupaca. Dakle, osim maksimizacije stope odgovora, neophodno je maksimizirati i profitabilnost, pa u kampanji kompanija teži da targetira potencijalne kupce s visokom vjerovatnoćom odgovora na kampanju, koji spadaju u grupu visokoprofitabilnih kupaca. Ovaj problem posebno dolazi do izražaja u tradicionalnom (ili *offline*) direktnom marketingu, kada postoje značajnija budžetska ograničenja za sprovođenje kampanje (slanje flajera, brošura ili slično). Međutim, ovaj argument je aktuelan i u *online* direktnom marketingu, kako se poruka ne bi rasipala i kako bi se kampanja mogla ocijeniti kao efikasna.

Adekvatno upotrijebljeni *data mining* modeli u direktnom marketingu mogu da povećaju efikasnost i efektivnost budućih aktivnosti, da snize troškove i povećaju stopu zadržavanja kupaca, kao i stopu konverzije (Barwise & Farley, 2005; Bose &

Chen, 2009; Watjatrakul & Drennan, 2005). Uzimajući sve ovo u obzir, možemo istači da je razumijevanje tehnika koje rezultiraju najtačnijom predikcijom profitabilnosti od presudne važnosti za donosioce odluka u direktnom marketingu.

3.1.3 Modeli bazirani na profitabilnosti kupaca

S obzirom na to da je u nauci i praksi rasprostranjen stav da je zadržavanje potrošača profitabilnije za kompaniju od privlačenja novih, marketing odnosa s potrošačima zadobija sve veću pažnju menadžera i istraživača u ovoj oblasti. *Lejeune* (2001) navodi da nove kompanije teže privlačenju kupaca, dok zrele kompanije teže da zadrže svoje postojeće potrošače u cilju razvijanja odnosa s njima, stvaranja baze lojalnih i profitabilnih kupaca, gdje se mogu realizovati određene *cross-selling* strategije. Ovu tezu potvrđuju i *Lin et al.* (2012), koji ističu da kompanije svoju pažnju pri stvaranju održivog profita usmjeravaju ka stvaranju mogućnosti za širenje obima prodaje postojećim kupcima. *Larivie`re* i *Van den Poel* (2005) potvrđuju da je zadržavanje potrošača centralna tema marketing i menadžment odluka, što se objašnjava pretpostavkom da postoji značajna veza između zadržavanja potrošača i njihove profitabilnosti. Naime, u literaturi se navodi da dugoročni potrošači kupuju više, a zahtijevaju manje ulaganja u opsluživanje (Ganesh et al., 2000; Hwang et al., 2004). Organizacije danas koriste različite strategije kako bi na osnovu toga adekvatno organizovale podatke o svojim kupcima i izgradile bazu lojalnih potrošača. U istraživanjima se ističe važnost sposobnosti identifikovanja određenih segmenata kupaca unutar opštih segmenata i definisanja različitih marketing strategija na osnovu njihovih obrazaca ponašanja (Sheikh et al., 2019). Slični rezultati, osim na opštem nivou kupaca, mogu se pronaći i za specifične sektore. Na primjer, u svom istraživanju, *Ansari* i *Riasi* (2016) su naglasili važnost određivanja klastera klijenata u bankarskom sektoru, objašnjavajući da upotreba klastera omogućava bankama da identifikuju svoje najprofitabilnije klijente i da za svaku grupu kupaca na osnovu njihovih atributa dizajniraju marketing strategije.

Kotler i *Armstrong* (1996) definišu profitabilnog kupca kao *"osobu, domaćinstvo ili kompaniju, čiji prihodi vremenom premašuju, za prihvatljiv iznos, troškove kompanije za privlačenje, prodaju i servisiranje tog kupca."*

U direktnom marketingu, pored stope odgovora kupaca na kampanju, koja je opisana u prethodnom poglavlju, važno je tačno identifikovati visokoprofitabilne kupce. Obično ovakvih kupaca nije mnogo, tako da nemaju veliki uticaj na stopu odgovora, ali imaju veliki uticaj na prihod koji se otvaruje kroz kampanju. Stopa odgovora kupaca s niskim profitom može biti visoka, dok se istovremeno zadržava nizak prihod od kampanje.

Iz perspektive poslovanja orijentisanog ka kupcima, važno je uočiti razlike između kupaca, čak i u situacijama kada oni kupuju iste proizvode ili usluge, što uslovljava potrebu za objektivnom segmentacijom. Dodatno, za kompaniju svi segmenti nemaju istu vrijednost, te marketing napori usmjereni ka određenim grupama ne moraju donijeti benefite. S tim u vezi, kao jedna od najznačajnijih strategija segmentacije potrošača ističe se analiza profitabilnosti (Fang et al., 2016), kojom se može napraviti razlika između potrošača vrijednih za kompaniju, kao i onih niže vrijednosti. Veliki broj donosilaca odluka u kompanijama svjestan je da se, u određenoj mjeri, njihovi kupci razlikuju po vrijednosti i profitabilnosti i prepoznaju „pravilo 80/20" – 20% kupaca stvara 80% profita za kompaniju, što je i potvrđeno u velikom broju empirijskih studija (Pitta et al., 2006). Longitudinalna studija koja je sprovedena u bankarskom sektoru pokazala je da 20% najvrednijih potrošača kreira 82% profita banke (Zeithaml et al., 2001), što savršeno odslikava „pravilo 80/20".

Slične rezultate prikazao je i *Mulhern* (1999) na primjeru *business-to-business* (B2B) prodaje farmaceutskih proizvoda, gdje je istakao da je analiza profitabilnosti krucijalna za otkrivanje veličine grupe kupaca koja se smatra najprofitabilnijom. Analiza prodaje farmaceutskih proizvoda pokazala je da 20% kupaca donosi 65,5% ukupnog profita, dok polovina baze kupaca generiše 95,5% profita (Mulhern, 1999). Ovaj rezultat, kao i prethodni, podržava tezu da najveći dio profita generiše relativno mala grupa kupaca. Autor je naveo da se informacije dobijene iz analize distribucije

profitabinosti mogu efikasno primijeniti tokom targetiranja kupaca u budućim marketing aktivnostima s obzirom na to da svi kupci nisu jednako profitabilni i ne mogu se tretirati na isti način.

Zeithaml et al. (2001) su kreirali konceptualni okvir za kategorizaciju potrošača, koji je baziran na njihovoj očekivanoj profitabilnosti. Ovaj sistem uključuje piramidalni prikaz od četiri nivoa, koje su nazvali: platina, zlato, gvožđe i olovo, pri čemu prva dva sloja predstavljaju atraktivne grupe kupaca, dok posljednja dva čine manje atraktivne grupe prema vrijednosti za kompaniju. Autori navode da je ovakav sistem za diferencijaciju kupaca značajan u situacijama kada kompanija pruža isti nivo usluge svim kupcima, uprkos činjenici da ima kupce koji se razlikuju prema profitabilnosti, što će dovesti do toga da se ne posvećuje značajna pažnja segmentu najvrednijih kupaca. Prilagođavanje usluge prema nivou profitabilnosti kupca može učiniti bazu klijenata kompanije profitabilnijom, povećavajući joj tako šanse za uspjeh na tržištu. Autori navode da proces ispitivanja kupaca prema profitabilnosti može doprinijeti povećanju trenutne i buduće profitabilnosti svih kupaca u portfoliju kompanije (Zeithaml et al., 2001).

U skladu sa ovim saznanjima o profitabilnosti kupaca, mogu se kreirati odgovarajuće strategije i unaprijediti sistemi za donošenje odluka u marketingu. Analiza profitabilnosti kupaca pruža važne informacije i matrike koje su ključne za alokaciju marketing resursa, kako pojedinačnim potrošačima, tako i definisanim segmentima. S obzirom na to da su savremeni potrošači izuzetno zahtjevni i informatički pismeni, kompanijama se nameće potreba kreiranja prilagođenih ponuda, kao i personalizovanih „jedan na jedan" strategija. U cilju kreiranja ovih strategija, potreban je permanentni monitoring definisanih potrošačkih grupa, kao i njihova precizna klasifikacija.

Profitabilnost potrošača (eng. *customer profitability*) izučava se pod različitim terminima: cjeloživotna vrijednost (eng. *lifetime value – LTV*), potrošački kapital (eng. *customer equilty*) i vrijednost potrošača (eng. *customer value*). Profitabilnost potrošača predstavlja varijablu koja, na nivou pojedinačnih kupaca, opisuje prihode koje potrošač generiše u određenom vremenskom periodu. Dakle, cilj analize

profitabilnosti kupaca pri segmentaciji tržišta je da alocira marketing resurse i prilagodi strategije na način da obezbijedi visoku profitabilnost.

U marketing literaturi, ovaj pojam se javlja u dvije forme (Donio et al., 2006). Prva forma odnosi se na istorijske zapise – a analiza potrošačke profitabilnosti slična je analizi profita i gubitka kompanije, sa osnovnom razlikom što se ova analiza sprovodi za pojedinačne potrošače, dok se bilans uspjeha odnosi na sve kupce. Druga tačka gledišta u literaturi odnosi se na budući period. Naime, profitabilnost potrošača, u ovom smislu, često ima oblik rezultata iz analize neto sadašnje vrijednosti. U tim situacijama, često se koristi upravo pojam cjeloživotne vrijednosti kupaca (Khajvand et al., 2011; Rust et al., 1995; Shen & Chuang, 2009). *Peppers* i *Rogers* (1997) su definisali ovaj pojam kao *„tok očekivane buduće dobiti od transakcija klijenata, umanjene za troškove, diskontovane po nekoj odgovarajućoj stopi na trenutnu neto sadašnju vrijednost"*.

Budući da baze podataka kupaca sadrže velike količine podataka, teško je ili gotovo nemoguće njima ručno manipulisati. Stoga je korisno primijeniti *data mining* metode, koje omogućavaju automatizaciju ovog procesa, što u savremenom tržišnom okruženju postaje neophodno za stvaranje adekvatnih marketing strategija. Dakle, predikcija profitabilnosti u direktnom marketingu omogućava da donosioci odluka dobiju informaciju ne samo o broju i strukturi kupaca koji će odgovoriti na kampanju, već i o njihovoj potencijalnoj profitabilnosti. Uzimajući navedeno u obzir, ove dvije predikcije se često zajedno sprovode radi dobijanja kompletne slike o potencijalnim kupcima pri planiranju budućih aktivnosti direktnog marketinga.

U ranijim radovima za predikciju profitabilnosti u direktnom marketingu najzastupljeniji su linearna regresija i *Tobit* model. Linearna regresija generiše kontinuirani rezultat, poput procijenjenog iznosa potrošnje (Malthouse, 1999). Model linearne regresije selektuje potrošače na bazi unaprijed definisanih granica, što znači da će potrošač biti selektovan ukoliko ima rezultat veći od definisane granice (Bose & Chen, 2009). Za razliku od ovog modela, *Tobit* model predstavlja regresioni model baziran na cenzorisanju distribucije zavisne varijable (Tobin,

1958). Stoga, u marketing analizi, *Tobit* model bi konvertovao negativne vrijednosti u nulu, a zadržao bi pozitivne vrijednosti. Primjer neprekidne vrijednosti za *Tobit* model je iznos novca koji potrošač troši, a izbor potrošača vrši se na isti način kao kod linearne regresije – u odnosu na definisanu granicu.

Najčešće primjenjivana tehnika u okviru linearne regresije za procjenu profitabilnosti potrošača u direktnom marketingu je regresija zasnovana na običnim najmanjim kvadratima (eng. *Ordinary Least Squares* - OLS). Standardna OLS regresija je uobičajeni pristup bodovanja koji se koristi za dobijanje procjena uslovne sredine neke promjenljive, s obzirom na neki skup kovarijacija. Iako je ova tehnika važna i često se primjenjuje, nedostatak ovog pristupa je taj što je dobijena procjena samo jedan broj koji se koristi za sumiranje odnosa između zavisne promjenljive i nezavisnih promjenljivih. Međutim, profit kupaca može da varira u velikom rasponu među kupcima, što podrazumijeva heterogenost u ponašanju potrošača (Zhang, 2009). Dakle, tradicionalna OLS regresija ne može izmjeriti disparitet profita među kupcima, a osim toga, na OLS regresiju utiče vrlo iskrivljena raspodjela profita, te ova metoda ne daje adekvatne rezultate u slučaju postojanja značajnih odstupanja (eng. *outliers*). Konkretno, u analizi profitabilnosti kupaca, donosioci odluka su posebno zainteresovani za kupce koji kompaniji donose visok profit. Obično su sredstva koja su predviđena marketing budžetima ograničena, pa marketing menadžeri ne mogu targetirati svakog kupca sa određenom vjerovatnoćom odgovora. U tom slučaju, oni će odabrati najprofitabilnije kupce kako bi maksimizirali profit. S tim u vezi, OLS regresija, koja se fokusira samo na centralnu lokaciju, nije adekvatna metoda za primjenu u planiranju kampanje, jer ne može odgovoriti na pitanje izbora ciljnih kupaca kod iskrivljene distribucije profitabilnosti.

Neki od ranijih radova koji su tretirali ovu problematiku koristili su logističke modele. Na primjer, *Bult* i *Wansbeek* (1995) predstavili su strategiju maksimizacije profita za logističke modele, koja uključuje komponente profita i troškova u funkciji procjene. Sličan pristup primijenio je i *Van der Sheer* (1998), gdje je, takođe u funkciji procjene profitabilnosti, za predloženi pristup maksimizacije profita zasnovan na

logici, uključen faktor profita i troškova. Međutim, strategija maksimizacije profita za odabir klijenata, koja uključuje sve kupce s pozitivnim predviđenim marginalnim profitom, nije izvodljiva u situacijama kada se kompanije susriječu sa ograničenim budžetom.

U poređenju s prethodno predstavljenim pristupom procjene profitabilnosti kupaca, istraživači iz ove oblasti sve više primjenjuju pristup cjeloživotne vrijednosti kupaca, a prvi CLV modeli pojavili su se još tokom osamdesetih godina prošlog vijeka. Autori iz ove oblasti ističu da primjena ovog pristupa može ostvariti veliki potencijal u segmentu poboljšanja kvaliteta marketing odluka i poslovnih strategija (Gupta & Lehmann, 2003; Peppers & Rogers, 1997). Dakle, analiza cjeloživotne vrijednosti kupca omogućava izbor klijenata visoke vrijednosti, procijenjene na osnovu njihovog potencijalnog budućeg ekonomskog doprinosa kompaniji (Venkatesan & Kumar, 2004).

U svom istraživanju, *Malthouse* (1999) je prikazao upotrebu *Ridge* regresije za poboljšanje efikasnosti bodovnog modela direktnog marketinga. Naime, u ovom radu je dokazano da predložena metoda daje procjene koeficijenta nagiba s nižom srednjom kvadratnom greškom od regularnih modela najmanjih kvadrata, kao i da se *Ridge* regresija može koristiti kao alternativa pristupima odabira podskupa varijabli za predikciju, kao što je postepena regresija (eng. *stepwise regression*). S tim u vezi, u ovom radu je implicirano da donosioci odluka u direktnom marketingu mogu uključiti više faktora u bodovni model, bez rizika od prekomjernog prilagođavanja.

Donkers et al. (2007) su predložili metodu selekcije ciljne grupe kupaca za slanje promotivnih materijala direktnom poštom, koja je bazirana na *Probit* modelu za procjenu vjerovatnoće odgovora i loglinearnoj regresiji za predikciju iznosa, odnosno vrijednosti transakcije. Da bi izvršili predikciju buduće profitabilnost kupaca, *Malthouse* i *Blattberg* (2005) su koristili linearnu regresiju, linearnu regresiju izračunatu sa iterativno reponderisanim najmanjim kvadratima, kao i neuronske mreže. Na osnovu najboljih nalaza iz tri primijenjena modela, autori su na pet skupova podataka formirali pravilo 20–55, što sugeriše da je među 20%

najprofitabilnijih potrošača 55% pogrešno klasifikovano. Pored toga, formirali su pravilo 80–15, koje kaže da je 15% od preostalih 80% potrošača pogrešno klasifikovano u kategoriju najprofitabilnijih kupaca. Kao rezultat toga, gubi se značajan broj najznačajnijih klijenata.

Dakle, osnovni izazov predikcije profitabilnosti kupaca je asimetrija, tj. zakrivljenost distribucije profitabilnosti. Do ovog problema dolazi zbog toga što je broj visokoprofitabilnih kupaca najčešće značajno manji u poređenju sa ostalim, manje profitabilnim grupama. U tradicionalnim regresionim modelima (Donkers et al., 2007; Glady et al., 2008; Malthouse & Blattberg, 2005; Rust et al., 2011; Verhoef & Donkers, 2001), kao što je prethodno istaknuto, fokus je na prosječnom kupcu, dok se heterogenost kupaca ne uzima u obzir. Na ovaj način se ne prave dovoljno precizne razlike između vrijednih i manje vrijednih kupaca. Uz navedene nedostatke klasičnih regresionih metoda, prethodna istraživanja su pokazala da se ove metode teško mogu uspješno izboriti sa asimetričnom distribucijom i dati precizne predikcije, posebno za veoma visoke profitabilnosti, kojih je značajno manje od ostalih, ukoliko se prethodno ne izvrši segmentacija ili klasterizacija podataka (Christmann, 2004; Kašćelan et al., 2016). S tim u vezi, u sekciji 4.5.5 predstavljena je nova metoda za predikciju profitabilnosti kupaca, zasnovana na *Support Vector* regresiji, koja je empirijski testirana u sekciji 5.6.

U poređenju sa statističkim modelima, kao što su logistička regresija i konvencionalna regresija najmanjih kvadrata, *data mining* modeli pokazuju značajno bolje performanse, što podstiče njihovu primjenu u oblasti predikcije profitabilnosti u direktnom marketingu.

U svom istraživanju, *Hwang, Jung* i *Suh* (2004) razvili su model za procjenu cjeloživotne vrijednosti kupaca, koji je uključio tri aspekta: prethodni doprinos, potencijalna vrijednost i vjerovatnoća napuštanja. Marketing menadžeri mogu koristiti ovaj model za proračun profitabilnosti kupaca, kao i za klasifikaciju različitih segmenata. Naime, autori navode da korišćenjem ova tri aspekta, donosioci odluka mogu dobiti uvid u trenutno finansijsko stanje, tj. trenutnu

vrijednost kupaca, kao i mogućnosti za *cross-selling*, što predstavlja potencijalnu vrijednost.

Kim et al. (2008) su koristili SVR u modeliranju odgovora da bi predvidjeli ukupan iznos novca koji će svaki od kupaca potrošiti u kampanjama direktnog marketinga. Zbog vremenske složenosti, autori ukazuju na poteškoće u obučavanju SVR modela na velikim uzorcima. Oni preporučuju korišćenje ove tehnike za predviđanje profitabilnosti na manjem uzorku kako bi se skratilo vrijeme obuke. Dodatno, autori su preporučili redukovanje uzorka uklanjanjem ekstremnih vrijednosti. Međutim, s obzirom na to da se iz modela isključuju upravo kupci sa ekstremno visokim vrijednostima profita, tj. najprofitabilniji segment, ovo dovodi do gubitka važnih informacija za obučavanje modela.

D'Haen je sa grupom autora (2013) analizirao efikasnost različitih *data mining* tehnika u predviđanju profitabilnosti potrošača, a osim poređenja samih tehnika, ispitivali su ih u kombinaciji s različitim izvorima podataka o kupcima. *Data mining* tehnike koje su upoređivali bile su logistička regresija, DT i *Bagging DT*, a dva korišćena izvora podataka su bili podaci dobijeni *web mining* procesom (besplatni podaci koji su dostupni svima sa internet pristupom) i podaci koji se kupuju od specijalizovanih dobavljača. Efikasnost *data mining* tehnika autori su procjenjivali koristeći AUC vrijednost. Rezultati ovog istraživanja su pokazali da je, bez obzira na izvor podataka, *Bagging* DT tehnika dala najbolje rezultate (osim u slučaju komercijalnih podataka, kada i logistička regresija pokazuje slične rezultate), pa autori preporučuju ovu tehniku za kreiranje modela, u odnosu na logističku regresiju i klasični DT (D'Haen et al., 2013). Dodatno, autori ističu da su besplatni podaci dobijeni kroz *web mining* proces idealan input za ovakve modele, te da je uključivanje eksternih podataka za poboljšanje prediktivnih performansi modela opravdano isključivo ukoliko je na raspolaganju odgovarajući budžet. S tim u vezi, autori predlažu sprovođenje *cost-benefit* analize, koja bi kao rezultat pokazala da li je ekonomski opravdano ulaganje u kupovinu skupih eksternih podataka, u odnosu na relativno malo poboljšanje prediktivnih performansi modela.

U cilju predviđanja profitabilnosti kupaca, *Lam* (2018) je predložio dvostepeni pristup zasnovan na dvije *data mining* metode – *Gradient Boosting* i neuronske mreže. Prvi korak je izgradnja dihotomnog modela koji predviđa vjerovatnoću budućih kupovina od strane potrošača. Drugi korak sastoji se od dizajniranja modela s kontinualnom ciljnom promjenljivom, koja može predvidjeti privremenu buduću dobit koju će kupac generisati ako je obavio kupovinu. Vještačke neuronske mreže su primijenili i *Vassiljeva et al.* (2017), koji su kreirali model za procjenu profitabilnosti klijenata automobilskog osiguranja.

Ponašanje potrošača predstavlja rezultat kombinacije različitih faktora, koji uključuju: nivo marketing aktivnosti, konkurentsko okruženje, percepciju brenda, uticaj novih tehnologija, kao i individualne želje i potrebe (Wang & Hong, 2006). U skladu s tim, promjena fokusa u marketingu sa agregatnog i masovnog pristupa ka analizi pojedinačnih potrošača, važan je aspekt „jedan na jedan" marketinga ili „mikro marketinga". Savremene kompanije teže kreiranju strategija zasnovanih na objektivnoj segmentaciji tržišta, pa iz ove perspektive, profitabilnost potrošača postaje važna dimenzija analize svakog pojedinačnog kupca. Prilagođavanje marketing napora segmentima koji se razlikuju u trenutnoj i/ili budućoj profitabilnosti, strategiju kompanije čini efikasnijom, identifikujući profitabilnosti različitih grupa kupaca, što omogućava prilagođavanje ponude proizvoda i usluga svakom niovu i prisvajanje njihove potencijalne vrijednosti (Donio et al., 2006).

3.2 Prednosti DM metoda za selekciju i targetiranje kupaca

Segmentacija je postala jedna od ključnih aktivnosti savremenog marketinga kako u teoriji, tako i u praksi. Različiti autori su u literaturi tokom godina predlagali različite tehnike segmentacije, međutim, odgovarajuća segmentacija tržišta predstavlja goruči problem među istraživačima tržišta i marketing menadžerima (Dutta et al., 2015). Primjena različitih tehnika zavisi od karakteristika kompanije, kao i podataka s kojima ona raspolaže. Koncept segmentacije tržišta prvi je predložio Smith (1956), a kasnije su ga razvijali brojni istraživači. Međutim, novija literatura sve više se

fokusira na primjenu *data mining* metoda, koje su se pokazale kao efikasne tehnike za selekciju i targetiranje kupaca.

Data mining predstavlja proces otkrivanja novog, implicitnog, korisnog i sveobuhvatnog znanja iz velikih baza podataka. Primjena *data mining* metoda omogućava uvid u obrasce i znanja koja su skriveni u podacima i nisu očigledni, te se ne mogu dobiti jednostavnijom statističkom analizom. Kao efikasne metode u sferama klasifikacije i regresije, *data mining* metode imaju značajnu primjenu u istraživanjima iz oblasti marketinga i cjelokupnog procesa selekcije kupaca za targetiranje u kampanjama. Iz analize profitabilnosti kupaca, predstavljene u prethodnom poglavlju, jasno je da se kupci ne mogu tretirati na isti način, te da kompanije moraju biti selektivne po pitanju izbora ciljne publike. U tržišnom okruženju koje karakterišu previranja i oskudni resursi, donosioci odluka moraju pažljivo razmotriti i analizirati tržište koje žele da opsluže (Pitta et al., 2006).

Kompanije generišu i čuvaju velike količine podataka o svojim kupcima. Ovi podaci, najčešće iz kategorije prethodnog kupovnog ponašanja, mogu se analizirati korišćenjem *data mining* tehnika, koje mogu dati uvid u interesovanja i želje potrošača. *Data mining* procesi se u marketingu koriste u cilju generisanja važnih zaključaka i znanja iz naizgled nepovezanog skupa podataka, što može značajno olakšati i unaprijediti proces donošenja odluka. Dakle, *data mining* tehnike fokusirane su na ekstrakciju znanja i razumijevanje velike količine podataka, kroz transformaciju raspoloživih sirovih podataka u znanje i informacije, koje mogu biti osnov za donošenje odluka (Tsiptsis & Chorianopoulos, 2010). Cilj ovog procesa je pronalaženje neočekivanih karakteristika, skrivenih osobina i naizgled nejasnih veza u podacima (Elsalamony, 2014).

Uzimajući u obzir napredak u objektivnom odlučivanju koji se može postići korišćenjem mašinskog učenja, *big data* analitike i *data mining* tehnika, u literaturi se često navodi da ove tehnike mijenjanju način na koje kompanije komuniciraju sa svojim potrošačima (Chagas et al., 2020).

Prethodne studije su ukazale na prednosti koje *data mining* metode stvaraju u procesu segmentacije i targetiranja potrošača, kao i procjene njihove profitabilnosti. Kao osnovni razlog korišćenja *data mining* metoda, u odnosu na tradicionalne statističke i ekonometrijske metode, autori navode veličinu skupova podataka koji se koriste u analizi. Osim toga, u situacijama kada je u skupu podataka dat veći broj prediktora nego što je primjenljivo za analizu, *data mining* metode omogućavaju dizajniranje procesa izbora varijabli, što je njihova dodatna prednost, a pored toga, veliki skupovi podataka mogu omogućiti fleksibilnije veze od jednostavnih linearnih modela (Varian, 2014). *Data mining* metode efikasno rješavaju probleme klasifikacije i regresije, a u literaturi iz oblasti marketinga primjenjivane su za segmentaciju tržišta, modeliranje odgovora na kampanju, procjenu profitabilnosti i cjeloživotne vrijednosti kupaca, predviđanje stope napuštanja potrošača, analizu potrošačke korpe i sličnih problema. U domenu direktnog marketinga, *data mining* metode tokom posljednje decenije dobijaju sve više na značaju, posebno u pogledu kreiranja strategija za targetirani marketing.

Uobičajene statističke i ekonometrijske tehnike poput regresije često dobro funkcionišu, ali mogu se javiti prethodno pomenuti problemi koji su jedinstveni za velike skupove podataka i koji mogu zahtijevati različite alate. Tehnike mašinskog učenja, kao što su drvo odlučivanja, *Support Vector Machine*, neuronske mreže i druge mogu omogućiti efikasnije načine za modeliranje složenih odnosa u bazama podataka (Varian, 2014). S tim u vezi, tražnja za tehnikama prediktivnog modeliranja u mnogim industrijama raste, uključujući i direktni marketing (Bose & Chen, 2009). Dakle, marketing menadžeri mogu donositi odluke zasnovane na rezultatima prediktivnih tehnika, što će povećati njihovu efikasnost u odnosu na tradicionalno odlučivanje na osnovu deskriptivnih podataka. Ovi napredni prediktivni modeli veoma su korisni u praksi direktnog marketinga, jer mogu da koriste više nezavisnih promjenljivih, te da generišu rezultate znatno bolje od tradicionalnih RFM modela i drugih jednostavnih statističkih tehnika.

Među najčešće korišćenim *data mining* metodama u direktnom marketingu ističu se: drvo odlučivanja, neuronske mreže i evolucioni algoritmi, čija efikasnost zavisi od

specifičnih karakteristika podataka na kojim se primjenjuju. Dodatno, kao jedna od metoda koja sve više dobija na značaju ističe se *Support Vector Machine*, koja će biti primijenjena i u ovom radu. U nastavku ovog poglavlja biće ukratko opisana primjena pomenutih metoda u procesu segmentacije i targetiranja kupaca i njihove prednosti, dok će algoritmi na kojima se zasnivaju ove metode detaljnije biti opisani u četvrtom poglavlju.

Drvo odlučivanja je metoda koja u cilju definisanja određenih odluka i njihovih direktnih posljedica koristi strukturu u obliku drveta (EMC Education Services, 2015), pri čemu je cilj da se, na osnovu inputa predvidi autput, tj. vrijednost zavisne varijable. Drvo odlučivanja koristi pristup „od vrha ka dnu" (eng. *top-down approach*) i sastoji se od grana i čvorova. Grana predstavlja ishod odluke i predstavljena je linijom koja spaja dva čvora. Čvorovi predstavljaju testove, a čvor na vrhu naziva se korijen. Na dnu drveta nalaze se listovi, koji predstavljaju pojedinačne klase, odnosno ishod svih prethodnih odluka.

Interesantno je da je koautor jednog od najranijih radova iz oblasti automatskog kreiranja stabla odlučivanja (Morgan & Sonquist, 1963) bio ekonomista. Međutim, tek nakon dvadeset godina, ovu tehniku su oživjeli u svom radu autori *Breiman, Friedman, Olsen* i Stone (1984), a danas je ona poznata kao „klasifikaciona i regresiona stabla" (eng. *classification and regression trees* – CART). S tim u vezi, stabla odlučivanja mogu biti klasifikaciona i regresiona – klasifikaciona za izlaz imaju kategoričke varijable, često binarne, poput da ili ne, dok regresiona mogu biti primijenjena i za numeričke ili neprekidne varijable, kao što je, na primjer, vjerovatnoća da kupac obavi kupovinu.

Jedan od prvih DT algoritama je ID3, koji je kreirao *Quinlan* (1986). Ovaj algoritam radi s kategoričkim varijablama s višestrukom podjelom i koristi *information gain* kao mjeru za kvalitet podjele. Pored ovog algoritma, kreiran je i CART (Breiman et al., 1984), sa specifičnošću da radi i sa kategoričkim i numeričkim varijablama, koji podržava samo binarnu podjelu, a za kvalitet podjele koristi *Gini indeks*. Pored ovih, kreirani su i C4.5 algoritam (Quinlan, 1992), CHAID (Kass, 1980), QUEST (Loh &

Shin, 1997) i drugi, sa svojim specifičnostima. Detaljniji opis različitih DT algoritama biće predstavljen u poglavlju 4.1.2.

Neuronske mreže, kao tip modela za mašinsko učenje, tokom posljednjih par decenija dobile su na popularnosti kao efikasan model za klasifikaciju, klasterizaciju, prepoznavanje obrazaca i predikciju u različitim disciplinama (Dave & Dutta, 2014). Neuronske mreže su mreže jednostavnih elemenata procesiranja (nazvanih „neuroni"), koji djeluju na njihovim lokalnim podacima i komuniciraju s drugim elementima. Dizajn neuronskih mreža motivisan je strukturom ljudskog mozga, ali su elementi obrade i arhitekture korišćeni u ovim modelima otišli daleko od svoje biološke inspiracije (Svozil et al., 1997).

U cilju pojednostavljenja procesa identifikacije potencijalnih potrošača, *Kim* i *Street* (2004) su dizajnirali hibridni prediktivni model koji kombinuje vještačke neuronske mreže i genetske algoritme. U razvoju ovog modela, autori su se fokusirali na dva cilja: interpretabilnost modela i tačnost predikcije. Dodatno, *Lee* i *Park* (2005) su u svom radu uzastopnim istraživanjem zadovoljstva kupaca i sociodemografskim podacima definisali profitabilne kupce iz cjelokupne baze anketiranih kupaca. U tom istraživanju, autori su primijenili sljedeće *data mining* alate: samoorganizujuće mape (SOM), neuronske mreže i C4.5 algoritam za formiranje stabla odlučivanja, koji su omogućili efikasnu i objektivnu segmentaciju profitabilnih kupaca kompanije u pogledu njihove profitabilnosti. Takođe, *Valero-Fernandez et al.* (2017) koristili su različite algoritme za segmentaciju kupaca na osnovu istorijskih transakcionih podataka iz *online* prodavnice iz Velike Britanije. Ciljevi ovog rada bili su definisanje osnovnih metrika, poput stope napuštanja potrošača i njihove cjeloživotne vrijednosti, kao i utvrđivanje prediktora za ponašanje segmentiranih grupa kupaca, kako bi se efikasnije sprovodilo targetiranje u budućim promotivnim kampanjama i na kraju - povećala profitabilnost. Pored neuronskih mreža, testirali su prediktivne sposobnosti logističke regresije, *SVM* (SVM sa linearnim kernelom i RBF SVM), DT, *ensemble* modele – *Random Forest* i *AdaBoost* i druge, pri čemu su linearni SVM, neuronske mreže, *AdaBoost* i logistička regresija postigli najbolju tačnost.

Za definisanje optimalnog broja segmenata koje treba odrediti u okviru tržišta, autori *Boone* i *Roehm* (2002) koristili su analitičku tehniku zasnovanu na vještačkim neuronskim mrežama. Naime, u slučaju prekomjernog broja segmenata, kompanije bi odvojeno tretirale segmente koji se mogu tretirati inkluzivno, što bi izazvalo veće troškove. S druge strane, u slučaju nedovoljnog broja segmenata, donosioci odluka bi podsegmentirali tržište, te ne bi uspjeli da identifikuju različite, održive segmente koje treba zasebno posmatrati i tretirati. Autori navode da predloženi algoritam - *Membership Clustering Criterion* (MOC) tačnije definiše stvarni broj segmenata, što potencijalno omogućava efikasniju upotrebu marketing resursa kroz odgovarajuću segmentaciju i targetiranje potrošača (Boone & Roehm, 2002).

Neuronske mreže su se u različitim oblicima primjenjivale u cilju segmentacije, selekcije i targetiranja potrošača u velikom broju radova. *Cuadros* i *Domínguez* (2014) primijenili su samoorganizujuće mape za proračun cjeloživotne vrijednosti kupaca, njihove lojalnosti i definisanje segmenata kupaca, dok je *Chan* (2005) primijenio istu tehniku za segmentaciju kupaca sa *online* aukcija u homogene grupe (definisao je tri segmenta: strpljivi kupci, impulsivni kupci i analitični kupci). Takođe, samoorganizujuće mape su koristili i *Vellido et al.* (1999) za segmentaciju potrošača iz *online* maloprodaje i predstavili su dva različita seta rezultata. U prvom, SOM je primijenjena u interpretaciji rezultata proizvedenim od nadgledane neuronske mreže u vezi s predviđanjem usvajanja online kupovine. U drugom, SOM je djelovala u potpuno nenadgledanom režimu za klasterizaciju podataka (Vellido et al., 1999). Dodatno, neuronske mreže s genetskim algoritmom primijenili su *Kim et al.* (2005) za predviđanje i targetiranje domaćinstava zainteresovanih za kupovinu polise osiguranja za vozila za rekreaciju, dok su *Salehinejad* i *Rahnamayan* (2016) predložili model predikcije ponašanja potrošača, koji je zasnovan na RFM atributima i rekurentnim neuronskim mrežama (eng. *Recurrent Neural Network* - RNN), a na osnovu koga se potrošači mogu segmentirati i precizno targetirati u prilagođenim promotivnim akcijama.

Na osnovu pregleda navedenih radova, možemo istaći da će upravo korišćenje prediktivnih *data mining* tehnika zamijeniti klasične statističke analize kupaca, kao

i subjektivne procjene donosilaca odluka. Klasični *data mining* algoritmi, poput DT, logističke regresije i neuronskih mreža, često se smatraju *benchmark* prediktivnim tehnikama (Bose & Chen, 2009), uzimajući u obzir da u svakom pogledu nadmašuju jednostavne statističke modele, posebno u pogledu preciznosti i tačnosti.

Konačno, tokom posljednjih godina SVM se ističe kao jedna od najefikasnijih tehnika za klasifikaciju i selekciju kupaca u direktnom marketingu (Govidarajan, 2013; Kurnia & Kusuma, 2018; Lawi et al., 2018; Moro et al., 2011; Panigrahi & Patnaik, 2020; Sagala & Permai, 2021). Ovaj pristup odnosi se na savremenu tehniku klasifikacije, koja koristi matematičku teoriju učenja (Vapnik, 2010). U kontekstu binarne klasifikacije, SVM pokušava da pronađe optimalnu hiperravan, tako da je margina razdvajanja između pozitivnih i negativnih primjera maksimalna. Ovo je ekvivalentno rješavanju kvadratnog problema optimizacije, u kome presudnu ulogu igraju vektori oslonca (eng. *support vectors*), tj. tačke najbliže optimalnoj hiperravni (Coussement & Van den Poel, 2008). Međutim, u praksi podaci često nisu linearno odvojivi. Da bi se povećala mogućnost linearnog razdvajanja podataka, ulazni prostor može se transformisati nelinearnim preslikavanjem u prostor veće dimenzije. Ova transformacija se vrši upotrebom kernel funkcije. Detaljniji opis SVM klasifikatora predstavljen je u poglavlju 4.1.3.

Uzimajući sve ovo navedeno u obzir, može se zaključiti da u odnosu na klasične *data mining* metode, kao i neuronske mreže, SVM metoda ima značajne prednosti, koje se ogledaju u rješavanju problema preklapanja klasa i poteškoća koje za klasične algoritme izaziva nelinearnost podataka. Dodatno, ova metoda se može uspješno koristiti za prečišćavanje podataka, odnosno njihovo pretprocesiranje, pri čemu se manja klasa dopunjava relevantnim primjerima iz veće klase. Pored toga, u regresionim problemima, SVR se uspješno bori sa asimetričnošću distribucije ciljne varijable, te za razliku od neuronskih mreža, ne zapada u lokalni minimum, već pronalazi globalni minimum greške. Stoga, imajući u vidu navedene prednosti, u ovom radu će za definisanje prediktivnih modela za selekciju i targetiranje kupaca biti primijenjena SVM metoda.

3.3 Metode za online targetiranje kupaca

Dinamičan razvoj elektronske trgovine, komunikacije putem interneta, društvenih mreža i direktnog marketinga, stvorio je mogućnosti za kreiranje nove vrijednosti u *online* tržišnom okruženju. Dodatno, ovi faktori su uticali na promjene u sferi direktnog marketinga, koji tokom posljednje decenije sve više postaje digitalni direktni marketing. Savremeni potrošač prisutan je na mreži, pretražuje *online* prodavnice i posjećuje stranice na društvenim mrežama svojih omiljenih brendova, te samim tim ostavlja za sobom digitalni trag i podatke o svojim interesovanjima i potencijalnim željama. Za razliku od tradicionalne trgovine, digitalni marketing i e-trgovina generišu ogromnu količinu vrijednih podataka o proizvodima, namjerama i željama kupaca, kao i njihovom ponašanju, poput: podataka o tome odakle kupci dolaze, koje uređaje koriste, koje proizvode pregledaju i kupuju, koliko dugo se zadržavaju na pojedinačnim stranicama, te da li odgovaraju na promotivne sadržaje i poruke (Esmeli et al., 2020).

Podaci o ponašanju potrošača na internetu predstavljaju dragocjeni izvor informacija za donosioce odluka u direktnom marketingu. S obzirom na to da se targetiranje kupaca za e-tržišta vrši *online* putem, u ovom dijelu rada biće opisane metode za *online* targetiranje kupaca i primjena metoda u prethodnim istraživanjima.

Kao što je opisano u prethodnim poglavljima, industrija elektronske trgovine, kao i sam marketing sve više se kreću ka personalizovanim i prilagođenim porukama. Skorija istraživanja su pokazala da generičke ponude usmjerene ka potrošačima predstavljaju veoma neefikasnu strategiju oglašavanja (Behera et al., 2020; Stewart-Knox et al., 2016). U *online* okruženju, uz mogućnost targetiranja većeg broja potencijalnih kupaca uz niže troškove, kao jedan od osnovnih izazova javlja se stopa konverzije – broj sesija ili posjeta sajtu koji se završi kupovinom neznatan je u odnosu na ukupan broj pristupa (Behera et al., 2020; Liu et al., 2019).

Aktivnosti koje se odnose na akviziciju i zadržavanje kupaca u okviru *online* maloprodaja predstavljaju dio sistema *web* marketinga (Schafer et al., 2001), a

najčešće kao inpute ove aktivnosti koriste *web log* podatke, kao i *clickstream* podatke. Za potrebe marketing sektora u kompanijama, ovi podaci se najčešće posmatraju i ispituju iz perspektive klijenta. Dakle, kako bi se pružila bolja usluga i unaprijedilo potrošačko iskustvo na *web* sajtovima, kompanije orijentisane na kupce sve više se fokusiraju na podatke iz pomenutih izvora, kako bi bolje razumjele njihove potrebe i želje. *Clickstream* podaci detaljno prikazuju navigaciju posjetilaca kroz *web* sajt i uključuju podatke, kao što su: broj posjećenih stranica, vrijeme provedeno na svakoj od njih, kako su došli na tu stranicu (organskom pretragom, preko društvenih mreža, klikom na oglas ili slično), kao i gdje su otišli nakon posjete sajtu. Agregatno posmatrano, podaci koji se prikupljaju kroz *web log* mogu obezbijediti veoma značajan uvid u način korišćenja *web* sajta od strane posjetilaca, otkriti stranice na kojima se posjetioci najviše zadržavaju, kao i one koje se često ignorišu. Dakle, ovaj tip podataka bilježi interaciju korisnika i *web* sajta (ili aplikacije), tako što se svaki naredni klik bilježi i kreira tok (eng. *stream*). Zbog pristupačnosti i niskih troškova prikupljanja, ovi podaci su našli primjenu u analizi e-trgovine, društvenih mreža, e-učenja, kao i analizi korišćenja internet portala i pretraživača, u cilju definisanja karakteristika korisnika, njihove klasterizacije i modeliranja njihovog ponašanja (Jiang et al., 2018). Ovaj tip podataka izuzetno je koristan resurs za marketing praktičare, e-trgovine i istraživače, kako bi razumjeli ponašanje posjetilaca sajta pri donošenju odluka o kupovini ili suprotno - napuštanju e-prodavnice.

Van den Poel i *Buckinx* (2005) su istakli da detaljni *clickstream* podaci predstavljaju najznačajniji skup atributa za segmentaciju, odnosno klasifikaciju kupaca na osnovu njihovog prethodnog ponašanja u kupovini. U svom istraživanju, oni su procijenili prediktivnu moć za čak 92 varijable, što je mnogo više u poređenju s prethodnim studijama koje su tretirale program *online* kupovine. Ovi autori su varijable grupisali u četiri kategorije: opšti *clickstream* podaci na nivou posjete sajtu, detaljni *clickstream* podaci, demografski podaci i podaci o prethodnom kupovnom ponašanju. Na osnovu sveobuhvatne analize omogućili su dublje razumijevanje karakteristika koje utiču na odluku posjetioca sajta - da li da izvrši kupovinu ili ne. Na osnovu predstavljenih rezultata, marketing menadžeri mogu da definišu ko će od

kupaca posjetiti njihovu stranicu s namjerom kupovine, a u skladu sa CRM mogućnostima kompanije, oni mogu prilagoditi ponudu i precizno targetirati upravo te zainteresovane kupce (Van den Poel & Buckinx, 2005). Dakle, detaljniji podaci koji su na raspolaganju *online* trgovcima mogu im pomoći u dizajniranju strategija za segmentaciju i targetiranje kupaca. S obzirom na to da se uklanjaju geografske i vremenske barijere posjete, ovo predstavlja dodatnu prednost u odnosu na tradicionalne maloprodaje. Zadovoljavajuće korisničko iskustvo i prepoznavanje želja potrošača pozitivno utiču na jednostavnost korišćenja platforme i pretrage proizvoda, štede vrijeme potrošača kroz precizno targetiranje i sistem preporuka, što doprinosi rastu stope konverzije. Stoga je jasno i opravdano interesovanje kompanija za kreiranje strategija koje mogu povećati stepen interakcija s kupcima, čiji je cilj konverzija, odnosno realizovanje transakcije.

U prethodnim istraživanjima, autori su koristili određene *web* metrike za analizu ponašanja potrošača na *online* tržištu. S tim u vezi, *Raphaeli et al.* (2017) su u svom radu uporedili proces pretraživanja kod korisnika koji e-maloprodaji pristupaju s mobilnog telefona, u odnosu na one koji pristupaju sa PC računara. Njihovi rezultati su pokazali da je stepen interakcije (eng. *user engagement*) značajno veći u PC sesijama, u poređenju s pristupima s mobilnog uređaja, te da PC sesije traju duže. Analizom *clickstream* podataka, autori su zaključili da je pretraživanje korisnika PC uređaja dominantno orijentisano na istraživanje, dok su korisnici mobilnih uređaja više fokusirani na specifičan zadatak, tj. imaju jasnije definisan cilj pristupa sajtu e-trgovine. Slično njima, *Kaatz et al.* (2019) ispitivali su uticaj korišćenja različitih uređaja – desktop kompjutera, tablet računara i pametnih telefona na stopu konverzije, kombinujući *clickstream* i podatke dobijene anketiranjem kupaca. Njihovi rezultati anketnog istraživanja otkrili su da su odluke o kupovini korisnika računara uglavnom vođene kognitivnim komponentama korisničkog iskustva kupovine, dok se korisnici pametnih telefona uglavnom oslanjaju na afektivna iskustva i iskustva u ponašanju. Suprotno tome, veća je vjerovatnoća da će korisnici tablet računara odluku o kupovini odrediti na osnovu sveukupnog iskustva. Dodatno, autori navode da korisnicima računara kompanije treba da se obraćaju putem marketing kanala koji izazivaju slučajno ponašanje prilikom pretraživanja,

poput *newslettera* ili društvenih mreža, dok korisnici mobilnih telefona imaju veću vjerovatnoću korišćenja marketing kanala koji ne zahtijevaju proširenu pretragu informacija, poput oglašavanja u pretraživačima (eng. *search engine advertising*) ili čak direktnih posjeta njima poznatim *online* prodavnicama (Kaatz et al., 2019).

Web log podatke su u prethodnim istraživanjima koristili i *Lee et al.* (2002), koji su predložili *fuzzy cognitive map* pristup za analizu obrazaca kupovine. Cilj ovog rada je bio da autori, korišćenjem asocijativnih pravila, dizajniraju marketing strategiju na osnovu obrazaca u kupovnom ponašanju, sakrivenom u podacima, što potvrđuje značaj i vrijednost ovog tipa podataka za planiranje budućih poslovnih aktivnosti. Pored pomenutih istraživanja, *web log* podatke su koristili i *Rho et al.* (2011), koji su, u cilju segmentacije kupaca i procjene njihove vrijednosti, ove podatke kombinovali s demografskim. Autori su kroz klasterizaciju i primjenu DT metode pokazali značajne razlike između segmenata kupaca *online* prodavnice za kućne ljubimce, na osnovu kojih su predstavili strategijske implikacije za donosioce odluka u marketingu. Za potencijalne potrošače sa izraženom namjerom za kupovinu, utvrđenom na osnovu prethodnog ponašanja, personalizovane strategije, poput targetiranih popusta, personalizovanih preporuka i elektronske pošte čiji je cilj da podsjeti korisnika da obavi kupovinu, mogu biti veoma efikasne. Ove taktike usmjerene su na povećanje stope konverzije, kroz precizno identifikovanje i targetiranje kupaca i unapređenje obima prodaje.

Najveći broj istraživanja u oblasti utvrđivanja namjere za kupovinu i targetiranja kupaca s prilagođenim ili personalizovanim ponudama koristi podatke o prethodnom ponašanju, tj. analizira ih nakon završenih sesija, te procjenjuje mogućnosti za uticaj na odluku o kupovini tokom naredne sesije, odnosno naredne posjete sajtu (Kytö et al., 2019; Martínez et al., 2020; Park & Park, 2016). Međutim, *Esmeli et al.* (2020) su dizajnirali EPP okvir (rano predviđanje namjere kupovine – eng. *early purchase intention prediction*), koji ima mogućnost da predvidi namjeru kupovine za tekuću sesiju na *web* sajtu e-trgovine. Takođe, autori su u ovom radu predložili i metodu bodovanja, koja je testirala kako modeli mašinskog učenja mogu otkriti moguću kupovinu u toku trajanja sesije, i to prije nego što se kupovina inicira

i realizuje. Rezultati ovog istraživanja pokazali su da je DT model bio najuspješniji za predviđanje kupovine, u poređenju s drugim modelima koji su testirani, i to: RF, *Bagging, k-NN* i *Naive Bayes*. Dodatno, autori su istakli da je vrijeme trajanja sesije bio najznačajniji atribut za predviđanje namjere za kupovinu.

Osim utvrđivanja obrazaca kupovine i segmentacije kupaca, *clickstream* podaci su se u prethodnim istraživanjima koristili za modeliranje *online* pretrage i analize ponašanja potrošača, kao i procjene i predviđanja odgovora na kampanju. *Hofgesang* i *Kowalczyk* (2006) su, na osnovu metrika poput posjete određenim stranicama, vremenu i frekvenciji posjete, klasterizovali korisnike na one koji posjećuju bez konkretnog cilja pretrage (eng. *browsers*), one koji sajt pretražuju namjenski (eng. *searchers*) i one koji obavljaju kupovinu (eng. *purchasers*). Ovakva podjela potencijalnih potrošača je, sa aspekta marketinga, veoma značajna, uzimajući u obzir da se za svaku definisanu grupu može kreirati specifična strategija targetiranja u cilju maksimizacije vrijednosti svakog pojedinačnog kupca ili klastera kupaca. Ova marketing aktivnost posebno dobija na značaju u *online* okruženju, gdje je konkurentska e-maloprodaja samo „klik" daleko, pa akvizicija i zadržavanje potrošača predstavljaju jedan od najznačajnijih izazova u e-trgovini. S tim u vezi, *Villanueva et al.* (2007) su u svom istraživanju naveli da kupci stečeni kroz marketing i promotivne aktivnosti generišu veću vrijednost u kratkom roku. Ovaj rezultat dokazuje značaj „poziva na akciju", što je jedna od osnovnih karatkeristika direktnog marketinga. Tokom kampanje, kupcima se prezentuje prilagođena ponuda, a ukoliko je segmentacija kvalitetno odrađena, mogu se očekivati značajni kratkoročni rezultati, a rasipanje poruke svodi se na minimum.

Empirijski je potvrđeno da su *data mining* tehnike korisne za analiziranje velikih i kompleksnih baza podataka (Mrzic & Zaimovic, 2020), u koje spadaju i *clickstream* podaci, odnosno *web log* podaci, kao i podaci o *online* transakcijama i karakteristikama kupaca uopšte. Da bi se ove tehnike mogle adekvatno primijeniti, potrebno je da se integrišu procesi koji podrazumijevaju prepoznavanje potrebe i ciljeva za koje se analiza sprovodi, da se pripremi i pretprocesira sirova baza podataka na kojoj je zasnovana analiza, te da se kreiraju i implementiraju modeli s

konačnim ciljem pružanja preporuka za donošenje odluka i dizajniranje strategija (Noviantoro & Huang, 2021).

U cilju procjene uticaja korišćenja različitih platformi (mobilni telefon, tablet ili PC) i operativnih sistema za pristup *web* sajtovima na proces odlučivanja o kupovini, *Esmeli et al.* (2020) su primijenili tri prediktivna modela mašinskog učenja, i to: k-NN, DT, i *Bagging* klasifikator. Evaluaciju predloženih modela vršili su upotrebom *10-fold* kros-validacije. Atributi čiji je uticaj bio razmatran su: dan u nedjelji, vrijeme pristupa, vrijeme trajanja sesije, lokacija korisnika, broj prethodnih posjeta sajtu i broj prethodnih kupovina, kao i broj pregledanih stranica. Rezultati navedenih eksperimenata su pokazali da DT i *Bagging* imaju superiorne performanse pri predviđanju kupovine u odnosu na k-NN klasifikator, kao i da postoji pozitivna korelacija performansi predviđanja kupovine u sesijama kada se uključe podaci o operativnim sistemima korisnika i tipovima platformi s kojih pristupaju, što potvrđuje da su ovi atributi izuzetno značajni za proces predviđanja odluke o kupovini (Esmeli et al., 2020). Slično istraživanje sproveli su *Noviantoro* i *Huang* (2021), u kome su za istraživanje ponašanja potrošača i predviđanja kupovine na osnovu *clickstream* podataka iz *online* prodavnica, primijenili DT, RF, NN, *Deep Learning*, *Naïve Bayes*, k-NN, LR i *Rule Induction*. Model neuronskih mreža je ostvario najbolji rezultat tačnosti, a odmah zatim je, sa neznatno slabijom tačnošću, rezultat RF modela (Noviantoro & Huang, 2021).

Kada je riječ o *online* direktnim marketing kampanjama i ukupnom predviđanju kupovine preko interneta pomoću podataka *web* evidencije, postoji niz novijih radova koji tretiraju ovaj problem. Na primjer, *Lee et al.* (2021) istraživali su modele mašinskog učenja, kao i potencijalno efikasne metode uzorkovanja podataka za predviđanje ponašanja potrošača na mreži, za posjetioce *Google Merchandise Store* prodavnice. Autori su istakli da je algoritam *eXstreme Gradient Boosting* (XGB) najefikasniji za predviđanje konverzije kupovine *online* potrošača, dok se pokazalo da je preuzorkovanje (SMOTE) najbolja metoda za rješavanje problema neravnoteže podataka. Rezultati njihovog istraživanja su sljedeći: tačnost - 74,17%, senzitivnost - 73,92% i AUC - 0,791. Međutim, važno je navesti da je korišćeni skup podataka

sadržao podatke za sve posjete *web* sajtu, a ne samo pristup posredstvom direktnih marketing kampanja. Stoga se ne može govoriti o stopi odgovora, već o stopi konverzije, koja je u ovom skupu podataka iznosila 2,29%.

Slično njima, *Chaudhuri et al.* (Chaudhuri et al., 2021) su koristili skup podataka s platforme za elektronsku trgovinu da predvide ponašanje pri kupovini i uporedili su performanse algoritama mašinskog učenja s dubokim učenjem (eng. *deep learning* - DL). Njihovi rezultati pokazuju da DL tehnike, koje su uključivale razvijene napredne varijante vještačkih neuronskih mreža, pokazuju bolje performanse od ML algoritama – najbolji model je postigao tačnost od 89%, senzitivnost 96% i AUC 0,89. Međutim, autori su naveli da je DL algoritam znatno intenzivniji od ML algoritama, te da zahtijeva značajno više resursa.

Osim procjene odgovora na kampanju, slični modeli se mogu koristiti i u druge svrhe, poput predviđanja da li će potencijalni potrošači dodati u korpu određene proizvode u *online* kupovini (Migueis & Teixeira, 2020). Ovi autori su primjenom logističke regresije i metoda *Random Forest* predložili klasifikacioni model, koji rezultate predviđanja dobija na osnovu navigacionih obrazaca, tj. *clickstream* podataka koji su prikupljeni iz *online* maloprodaje. *Migueis* i *Teixeira* (2020) navode da korišćenje ovog modela može omogućiti predviđanje strukture potrošačke korpe korisnika i pravovremeno reagovanje na potrebe i namjere kupaca.

Važno je istaći da *online* targetiranje kupaca ima ekstremno nisku stopu odgovora, u određenim slučajevima čak i manju od 0,5%, što utiče na to da standardne prediktivne metode mogu djelimično ili potpuno pogrešno klasifikovati klasu respondenata. Dodatno, po saznanju autora, izuzetno mali broj prethodno objavljenih radova tretira značaj *web* metrika u odnosu na RFM atribute pri predikciji odgovora na kampanju. Stoga će u ovom radu biće predstavljen model za *online* targetiranje kupaca i predikciju njihovog odgovora, koji pokazuje dobre performanse čak i sa ekstremno niskim stopama odgovora zabilježenim iz prethodnog kupovnog ponašanja, kao i procjena značaja web metrika pri toj predikciji.

Na osnovu rezultata navedenih i sličnih modela, donosioci odluka u marketingu mogu koristiti neke od alata za *online* targetiranje kupaca s najvećom vjerovatnoćom kupovine, poput: marketinga u pretraživačima (eng. *search engine marketing - SEM*), oglašavanja na društvenim mrežama (eng. *social media marketing - SMM*), *newslettera* i drugih vidova targetiranog *e-mail* marketinga. Važno je napomenuti da je za posljednju kategoriju, targetirani *e-mail* marketing, potrebno da kompanija posjeduje svoju internu bazu korisnika i njihovih kontakata, dok za prve dvije kategorije to nije slučaj.

Nakon ovog dijela rada, u kome su predstavljene metode za segmentaciju i targetiranje potrošača, u narednom, četvrtom poglavlju, biće predstavljen predlog *data mining* koncepta sistema za efikasno targetiranje kupaca. Osim toga, biće detaljnije opisane i analizirane tehnike, metode i algoritmi na kojima se bazira predloženi sistem, a koji su već pominjani, kao što su: DT, SVM, *ensemble* metode i slično.

3.4 Identifikovanje istraživačkog jaza

Sumirajući istraživanja navedena u prethodnim sekcijama, mogu se uočiti sljedeće praznine, odnosno problemi kada su u pitanju metode za selekciju i targetiranje kupaca, kojima se bavi ovo istraživanje:

- subjektivnost i manuelni (neautomatizovani) pristup pri RFM segmentaciji postojećih kupaca;
- nemogućnost klasifikovanja i targetiranja unaprijed nepoznatih kupaca;
- visoka nebalansiranost klasa, tj. veoma mala stopa odgovora na kampanju direktnog marketinga, usljed čega dolazi do pristrasnosti klasifikatora;
- iskošenost distribucije profitabilnosti usljed malog broja najvrednijih kupaca, zbog čega prediktivne metode propuštaju njihovo identifikovanje;
- linarno neseparabilne klase, preklapanje klasa i prekomjerno prilagođavanje klasičnih DM metoda postojećim podacima;
- nemogućnost postizanja globalnog minimuma kod neuronskih mreža;

- ekstremno niska stopa odgovora kod *online* targetiranja kupaca, što često dovodi do potpuno pogrešne klasifikacije minorne klase;
- nedostatak kombinovanja RFM atributa sa *web* metrikama i nepoznat značaj jedne i druge grupe atributa za predikciju odgovora i predikciju profitabilnosti kupaca.

Uzimajući sve dosad navedeno u obzir, u sekcijama 4.5.1-4.5.5 biće predstavljena konceptuelna metoda za efikasno targetiranje kupaca, zasnovana na *data mining* tehnikama, kojom se teži prevazići postojeći istraživački jaz.

4. DATA MINING KONCEPT SISTEMA ZA EFIKASNO TARGETIRANJE KUPACA BAZIRAN NA SVM METODI

Postoje ubjedljivi dokazi da donošenje odluka zasnovano na podacima i tehnologiji velikih podataka značajno poboljšava poslovne performanse (Provost & Fawcett, 2013).

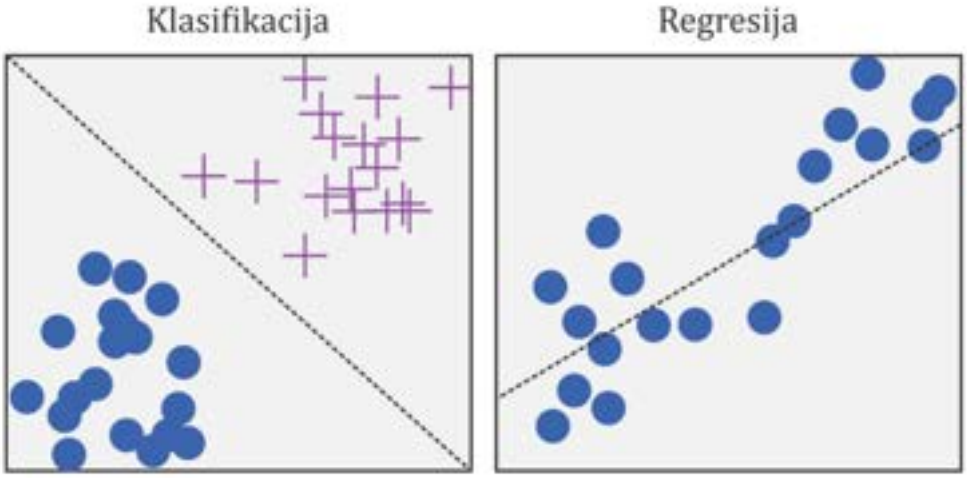
U oblasti nauke o podacima postoje dvije osnovne vrste zadataka: nadgledani (eng. *supervised*) i nenadgledani (eng. *unsupervised*), pri čemu je terminologija naslijeđena iz mašinskog učenja (Provost & Fawcett, 2013). Glavna razlika između učenja pod nadzorom i učenja bez nadzora je korišćenje označenih ili neoznačenih podataka (eng. *labeled/unlabeled data*).

Generalno gledano, za primjenu *data mining* tehnika, koristi se skup podataka primjera, od kojih svaki sadrži vrijednosti određenog broja promjenljivih koje se nazivaju atributima. U osnovi, postoje dvije vrste podataka, koji se tretiraju na radikalno različite načine (Bramer, 2020). Za prvi tip postoji posebno određen atribut i cilj je da se pomoću datih podataka predvidi vrijednost tog atributa za instance koje još uvijek nisu viđene. Podaci ove vrste se nazivaju označenim. S druge strane, podaci koji nemaju takav posebno naznačen atribut nazivaju se neoznačenim.

Na primjeru u vezi s kategorijama potrošača, mogu se uporediti pitanja koja predodređuju da li će se koristiti nadgledano ili nenadgledano učenje. Prvo pitanje može glasiti: *Da li kupci kompanije prirodno pripadaju različitim segmentima?* U ovom slučaju nema eksplicitnog cilja ili pravila za grupisanje kupaca i njihovu podjelu u segmente. Za *data mining* problem koji nema takav cilj, odnosno koristi neoznačene podatke, kaže se da je nenadgledan. S druge stane, naredno pitanje može glasiti: *Možemo li identifikovati grupe potrošača za koje je veća vjerovatnoća da će odgovoriti na kampanju direktnog marketinga?* Ovdje se pominje jasan cilj - da li bi kupac kupio proizvod na osnovu ponude kojom je targetiran? U ovom scenariju,

klasifikacija kupaca služi specifičnoj svrsi: preduzimanje radnji u zavisnosti od vjerovatnoće odgovora, tako da se ovaj problem može označiti kao problem nadgledanog *data mininga*. Nadgledano učenje se najčešće odnosi na dva tipa problema u *data miningu* - klasifikacija i regresija.

Problemi sa klasifikacijom koriste algoritam za preciznu podjelu podataka iz testnog skupa određene kategorije, kao što je odvajanje respondenata od nerespondenata. S druge strane, regresija koristi algoritam za razumijevanje odnosa između zavisne i nezavisnih varijabli. Regresioni modeli su korisni za predviđanje numeričkih vrijednosti na osnovu različitih podataka, kao što je, na primjer, apsolutni iznos profita koji se očekuje od kupca tokom njegovog životnog ciklusa u kompaniji. Slika 3 ilustruje razliku između klasifikacije i regresije.



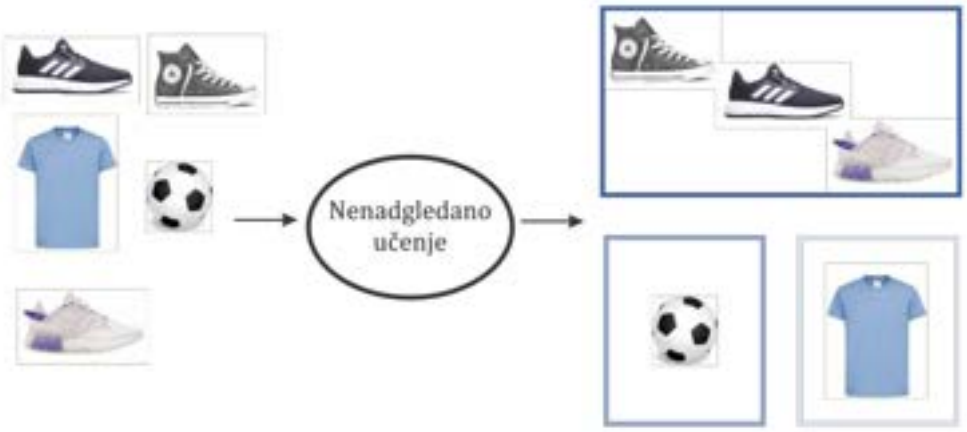
Slika 3. Poređenje klasifikacije i regresije (Soni, 2018)

Uobičajeni algoritmi u nadgledanom učenju su: logistička regresija, drvo odlučivanja, *Support Vector Machine*, vještačke neuronske mreže i *Random Forest*. Kako kod regresije, tako i kod klasifikacije, cilj je pronaći specifične odnose ili strukturu u ulaznim podacima koji omogućavaju da se efikasno proizvedu tačni izlazni podaci (Soni, 2018).

Učenje bez nadzora koristi algoritme mašinskog učenja za analizu i grupisanje neoznačenih skupova podataka. Ovi algoritmi otkrivaju skrivene obrasce u podacima, bez potrebe za ljudskom intervencijom (Delua, 2021). Učenje bez nadzora je veoma korisno u istraživačkoj analizi jer može automatski identifikovati strukturu u podacima (Soni, 2018). Modeli učenja bez nadzora se koriste za tri glavna zadatka: klasterizaciju, asocijaciju i smanjenje dimenzionalnosti.

Klasterizacija je *data mining* tehnika za grupisanje neoznačenih podataka na osnovu njihovih sličnosti ili razlika. Asocijacija koristi različita pravila za pronalaženje odnosa između varijabli u datom skupu podataka. Ova tehnika se često koristi za analizu tržišne korpe i razvijanje sistema za preporuke. Konačno, smanjenje dimenzionalnosti je tehnika učenja koja se koristi kada je broj karakteristika (ili dimenzija) u datom skupu podataka preveliki. Smanjuje broj ulaznih podataka na veličinu kojom se može upravljati, uz istovremeno očuvanje integriteta podataka (Delua, 2021).

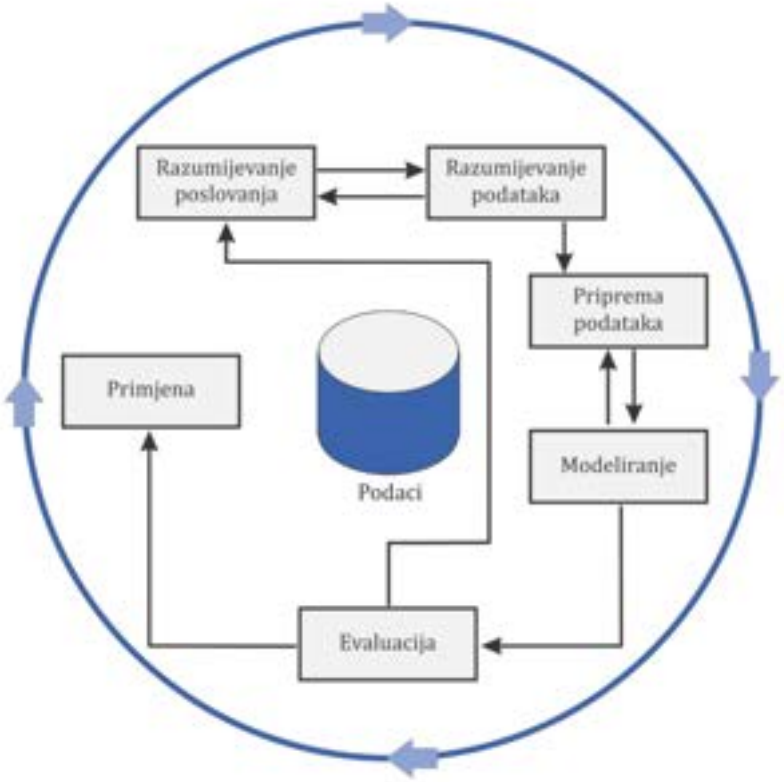
Naredna slika (Slika 4) prikazuje logiku nenadgledanog učenja.



Slika 4. Input i autput nenadgledanog učenja

Neki uobičajeni algoritmi za nenadgledano učenje uključuju k-means klasterizaciju (eng. *k-means clustering*), analizu glavnih komponenti (eng. *principal component analysis*) i autoenkodere (eng. *autoencoders*). Pošto nisu obezbijeđene oznake podataka, u većini metoda učenja bez nadzora ne može se definisati poseban način za upoređivanje performansi modela.

Postoji više usvojenih modela procesa DM-a, a jedan od najčešće korišćenih je CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Kao metodologija, on uključuje opise tipičnih faza projekta, zadataka uključenih u svaku fazu i objašnjenje odnosa između ovih zadataka. Kao model procesa, CRISP-DM pruža pregled životnog ciklusa *data mininga* (IBM, 2021).



Slika 5. *Data mining* životni ciklus (Binu & Rajakumar, 2021; IBM, 2021; Provost & Fawcett, 2013)

Kao što je prikazano na Slici 5, model životnog ciklusa sastoji se od šest faza sa strelicama koje ukazuju na najvažnije i najčešće zavisnosti između faza. Redosljed faza nije striktan, već, u stvari, omogućava kretanje naprijed-nazad između faza u okviru potreba pojedinačnih projekata (IBM, 2021). Za marketing menadžera, *data mining* proces je koristan kao okvir za analizu projekta ili predloga *data mininga*. Proces obezbjeđuje sistematsku organizaciju, uključujući skup pitanja koja se mogu

postaviti o predloženom projektu, kako bi se razumjelo da li je projekat dobro zamišljen ili je suštinski pogrešan (Provost & Fawcett, 2013).

U okviru sljedeće sekcije biće detaljno opisane sve DM tehnike i metode koje će biti korišćene za izgradnju konceptualnih modela za selekciju i targetiranje kupaca.

4.1 Primijenjene data mining metode

U ovom dijelu rada biće predstavljene sve DM metode koje su primijenjene za kreiranje konceptualnih modela za segmentaciju, selekciju i targetiranje kupaca.

4.1.1 K-means klasterizacija

Jedna od osnovnih sposobnosti živih bića je prepoznavanje sličnosti i zakonitosti u objektima, pojavama i podacima. Jedan od načina za izražavanje tih sličnosti je grupisanje određenog broja objekata koji imaju iste karakteristike. Ideja sortiranja i grupisanja sličnih objekata u kategorije ujedno predstavlja i jednu od najranije razvijenih ljudskih sposobnosti. U svakodnevnom životu, ovo predstavlja dio procesa učenja, te, na primjer, dijete uči da razlikuje sto od stolice, plavu od crvene boje, bananu od jabuke i slično, tako što u kontinuitetu unapređuje svoje podsvjesne klasifikacione šeme. Slično tome, klasifikacija je i u nauci takođe imala od davnina zapaženo mjesto. *Aristotle* je razvio kompleksan sistem klasifikacije životinjskih vrsta, počevši od podjele u dvije grupe – onih koje imaju crvenu krv (kičmenjaci) i onih koji nemaju ovu karakteristiku (beskičmenjaci), a klasifikacija elemenata periodnog sistema, koju je kreirao *Mendeleyev* tokom 1860-ih godina, značajno je uticala na razumijevanje strukture atoma (Everitt et al., 2011). Zatim, od 1911. godine, *Hertzsprung* i *Russell* grupisali su zvijezde na osnovu dvije varijable: njihovog intenziteta svjetlosti i temperature površine (Strömgren, 1956).

Naravno, klasterizacija ima značajno mjesto i u društvenim naukama – često se pojedinci grupišu u odnosu na njihovo ponašanje ili preferencije, posebno u marketingu, gdje je cilj identifikovanje tržišnih segmenata, tj. grupa potrošača sa sličnim potrebama (Kaufman & Rousseuw, 1991). Klasterizacija omogućava segmentaciju kupaca, formiranje homogenih grupa ili kategorija kupaca, koje

karakteriše mala varijansa unutar grupa i velika varijansa između definisanih grupa. Na osnovu podataka, ova *data mining* metoda ima mogućnost da otkrije zajedničke osobine pojedinaca koji pripadaju istom klasteru (Lejeune, 2001). Upravo klasterizacija predstavlja jednu od tehnika neohodnih za efikasno obavljanje procesa profilisanja kupaca, koji može da omogući detaljniju analizu manjih grupa koje predstavljaju tržišne segmente (Bose & Chen, 2009). Grupisanje kupaca može se postići korišćenjem efikasnih algoritama klasterizacije, koji se oslanjaju na nenadgledane klasifikatore i na podatke o kupcima, što može uključivati demografske podatke, kao i podatke o prethodnom ponašanju u kupovini.

Iako postoje različite definicije klastera, jednu od najprihvaćenijih je dao *Gordon* (1999), koji je klastere objasnio pomoću termina interne kohezije – homogenosti, kao i eksterne izolacije – odvojenosti.

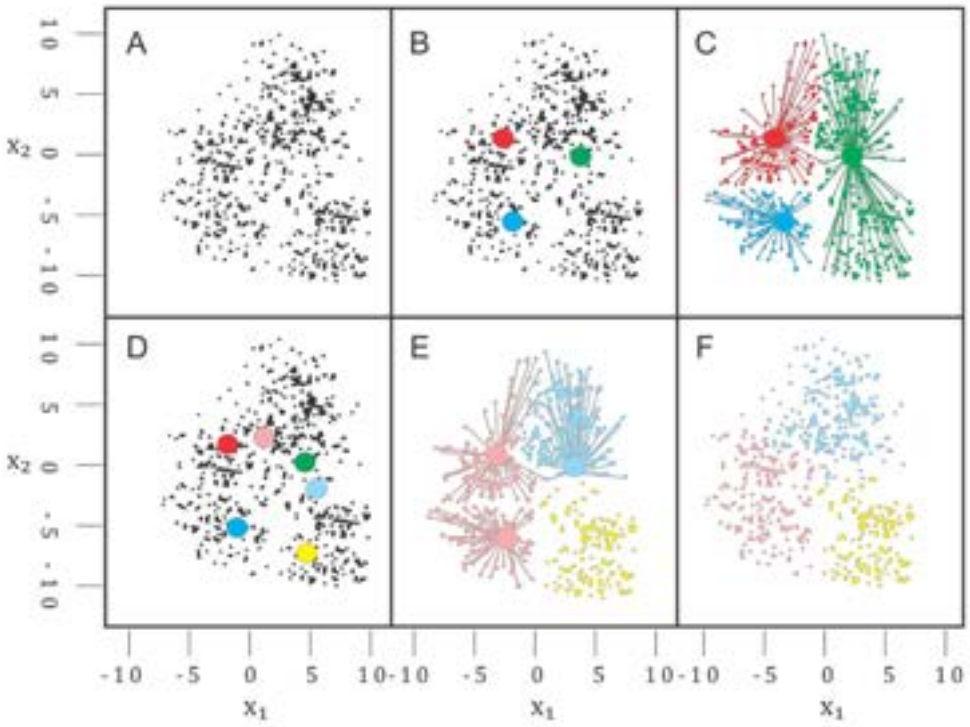
Naučnici su, uz napredak informacionih tehnologija, počeli da traže sistematične načine za pronalaženje sličnosti i grupa u podacima. Danas se metode klasterzacije primjenjuju u mnogim domenima, uključujući vještačku inteligenciju i prepoznavanje obrazaca, ekologiju, ekonomiju, marketing, medicinska istraživanja, političke nauke, psihometriju i mnoge druge. To je dovelo do razvoja različitih metoda, a članci o ovom procesu pojavili su se ne samo u statističkim časopisima, već i u periodičnim naučnim publikacijama svih navedenih domena (Kaufman & Rousseuw, 1991).

Klasterizacija je metoda koja se često koristi za eksploratornu analizu podataka, te se u ovoj analizi ne sprovode predviđanja. Umjesto toga, metode klasterizacije pronalaze sličnosti između objekata prema njihovim atributima, a zatim grupišu slične objekte u klastere (EMC Education Services, 2015). Jedna od najčešće korišćenih metoda klasterizacije je *k-means* klasterizacija, koja će biti primijenjena u empirijskom dijelu ovog rada.

K-means klasterizacija predstavlja jednu od najjednostavnijih i najpopularnijih algoritama mašinskog učenja bez nadzora. Generalno gledano, algoritmi bez nadzora donose zaključke iz skupova podataka, koristeći samo ulazne vektore bez

pozivanja na poznate ili označene ishode. Klaster se odnosi na kolekciju tačaka podataka ili instanci, koje su objedinjene zbog određenih sličnosti. Ovaj tip klasterske analize razvijen je tokom pedesetih godina prošlog vijeka (MacQueen, 1967).

Slika 6 pokazuje korake koje sprovodi algoritam u cilju podjele skupa podataka u klastere koji se ne preklapaju. Na prvom dijelu slike (dio A) prikazani su neoznačeni podaci (eng. *unlabeled data*). Prvi korak *k-means* klasterizacije uključuje izbor broja klastera, odnosno vrijednosti "*k*" u *k-means* algoritmu. Drugi korak predstavlja definisanje inicijalnih centroida za *k* klastera (dio B) i naziva se inicijalizacija algoritma. Treći korak je dodjela svake instance iz uzorka najsličnijem klasteru, u skladu sa odabranom metrikom sličnosti, poput Euklidskog rastojanja (dio C). Na slici je prikazano mjerenje Euklidskog rastojanja od centroida klastera do svake tačke i dodjeljivanje instance najbližem centru. Konvergencija algoritma postiže se ponavljanjem posljednjeg koraka, tj. redefinisanjem centara klastera na osnovu centroida svih instanci dodijeljenih svakom klasteru, kao što je prikazano na dijelu slike D, gdje je ilustrovano preraspoređivanje centroida u svakom klasteru, pri čemu su prethodni centroidi prikazani u tamnijim bojama, a novi u svjetlijim bojama. Dakle, da bi obradio podatke u procesu učenja, *k-means* algoritam započinje prvom grupom slučajno izabranih centroida, koji se koriste kao početne tačke za svaki klaster, a zatim izvodi iterativne proračune za optimizaciju položaja centroida. Ponavljanje trećeg i četvrtog koraka predstavljeno je na dijelu slike E, dok se ne ispuni kriterijum konvergencije. Najčešće se kao kriterijum konvergencije uzima ili prag u varijansi unutar klastera ili minimalan broj preraspoređivanja instanci po klasterima između dvije uzastopne iteracije. Posljednji dio slike (dio F) predstavlja rezultat algoritma sa tri definisana klastera (Garcia-Dias et al., 2020).



Slika 6. Ilustracija procesa *k-means* algoritma (Garcia-Dias et al., 2020)

Drugim riječima, *k-means* algoritam identifikuje *k* centroida, a zatim dodjeljuje svaku tačku podataka najbližem klasteru, istovremeno zadržavajući centroide što je moguće manjim. „*Means*" u nazivu algoritma odnosi se na traženje prosjeka u podacima (eng. *means*), odnosno pronalaženje centroida ili težišta.

Kvalitet klasterizacije može se procijeniti korišćenjem različitih mjera. Klasterizacija se uglavnom smatra dobrom ako ima maksimalnu udaljenost između tačaka različitih klastera (eng. *intercluster distance*) i minimalnu udaljenost između tačaka u okviru jednog klastera (eng. *intracluster distance*). U okviru ovog poglavlja, biće predstavljene neke od najkorišćenijih mjera za procjenu kvaliteta klasterizacije, poput: sume kvadratne greške, *Dunn* indeksa, "lakat" krive (eng. *Elbow Method*), krive siluete (eng. *Silhouette Curve*), kao i *Davies-Bouldin* (DB) indeksa, koji će, kao mjera kvaliteta klasterizacije, biti korišćen u empirijskom dijelu ovog rada.

Jedna od metoda za procjenjivanje kvaliteta klasterizacije je izračunavanje sume kvadratne greške (eng. *sum of the squared error* – SSE). Dakle, ovom mjerom se određuje greška za sve tačke, tj. njihovo Euklidsko rastojanje do najbližeg centroida, nakon čega se računa ukupna suma kvadratne greške (Tan et al., 2019). U slučaju da postoje dva seta klastera koja su rezultat dva *k-means* algoritma, kao onaj s boljim performansama prepoznaje se onaj s manjom kvadratnom greškom. Ova mjera ukazuje na to da su centroidi definisani tako da što bolje predstavljaju tačke iz svog klastera. Formula za izračunavanje SSE data je u nastavku:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$$

gdje je *dist* standardno Euklidsko rastojanje između dva objekta u Euklidskom prostoru.

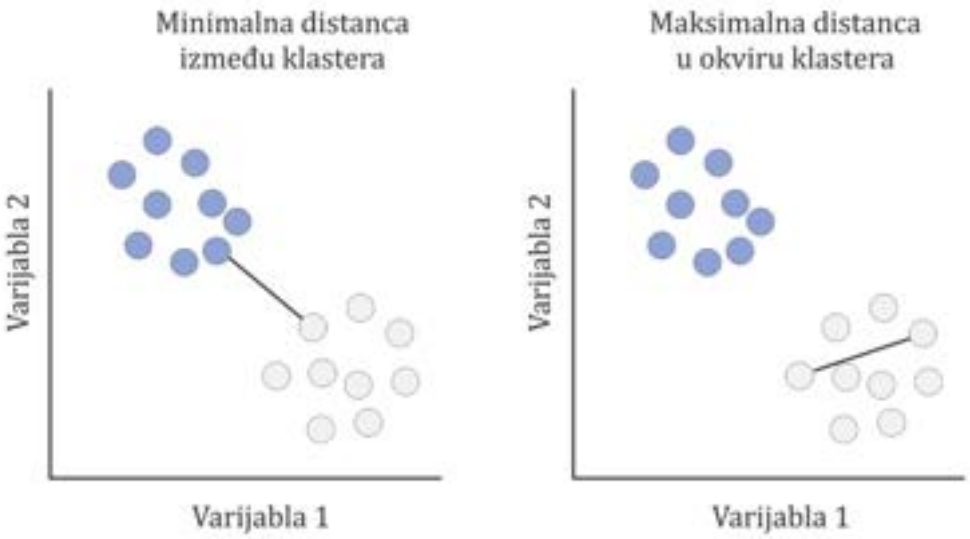
Vraćajući se na proceduru predstavljenu na Slici 6, možemo istaći da treći korak formira klastere dodjeljivanjem tačaka njihovom najbližem centroidu, što minimizira SSE za dati skup centroida. Nakon toga, u četvrtom koraku se preračunavaju centroidi, kako bi dodatno bio smanjen SSE. Međutim, važno je naglasiti da se u koracima 3 i 4 garantovano pronalazi samo lokalni minimum u odnosu na SSE, imajući u vidu da se proces zasniva na optimizaciji SSE za određene izbore centroida i klastera, a ne za sve moguće izbore.

Druga metrika za evaluaciju kvaliteta klasterizacije je *Dunn* indeks (Dunn, 1973). Formula za izračunavanje ovog indeksa data je u nastavku:

$$Dunn = \min_{1 \le i \le k} \left\{ \min_k \left(\frac{\delta(c_i, c_j)}{\max \Delta(c_k)_{1 \le i \ne i \le k}} \right) \right\}$$

gdje $\delta(c_i,c_j)$ predstavlja sve razlike u parovima između slučajeva u klasterima i i j, a $\Delta(c_i)$ predstavlja sve razlike u parovima između slučajeva u klasteru k. Drugim riječima, u odnos se stavlja udaljenost između klastera i i j sa distancom između tačaka u okviru klastera k. Dakle, *Dunn* indeks je odnos minimalne udaljenosti među

klasterima i maksimuma razdaljine u okviru klastera. Na Slici 7 prikazana je udaljenost između klastera (lijevo) i udaljenost tačaka u okviru klastera (desno).



Slika 7. Grafički prikaz distance između i unutar klastera (Rhys, 2020)

Najbolji kvalitet klasterizacije odraziće se kroz najveću vrijednost *Dunn* indeksa.

Još jedna metoda kojom se može procijeniti kvalitet klasterizacije i koja može uputiti na optimalan broj klastera je tzv. "lakat" metoda. Kako je osnovna ideja klasterizacije upravo definisanje klastera s minimalnom ukupnom varijacijom unutar klastera, ova metoda izbora optimalnog broja klastera se zasniva na minimiziranju ukupnog zbira kvadrata unutar klastera:

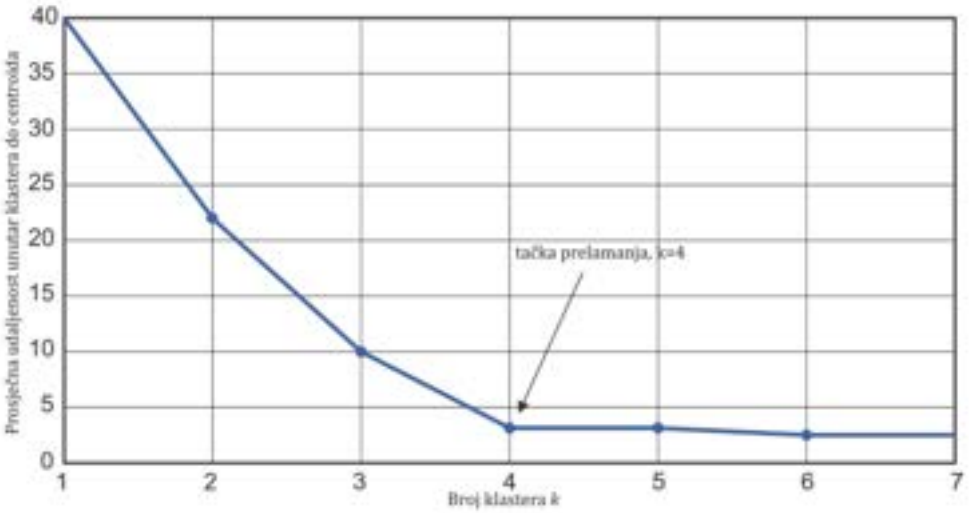
$$minimize\left(\sum_{k=1}^{k} W(C_k)\right)$$

gdje C_k predstavlja k-ti klaster, a $W(C_k)$ varijaciju unutar klastera. Ukupna suma kvadrata unutar klastera (eng. *Within-cluster Sum of Square - WSS*) mjeri kompaktnost klastera i težnja je da ovaj koeficijent bude što manji. S tim u vezi,

procedura za definisanje optimalnog broja klastera upotrebom „lakat" metode obuhvata sljedeće korake (University of Cincinatti, n.d.):

1. Pokrenuti algoritam za klasterizaciju (npr. *k-means*) za različite vrijednosti *k* (na primjer, uzeti vrijednosti od 1 do 10 za *k*);
2. Za svako k izračunati ukupni zbir kvadrata unutar klastera (*WSS*);
3. Nacrtati krivu *WSS* prema broju klastera *k*;
4. Mjesto savijanja (lakat) na grafiku se generalno smatra pokazateljem optimalnog broja klastera.

Na Slici 8 prikazana je *wss* kriva, gdje se na osnovu mjesta prelamanja može utvrditi odgovarajući broj klastera za analizu.



Slika 8. Grafički prikaz "lakat" metode (prilagođeno prema: Gandhi, 2018)

Na osnovu Slike 8 može se zaključiti da je optimalan broj klastera k=4. Iako se rastojanje unutar klastera smanjuje nakon ove vrijednosti, veći broj klastera izazvao bi kreiranje većeg broja proračuna, što je analogno zakonu opadajućeg prinosa (Gandhi, 2018). Stoga se sa mjesta preloma linije (k=4) bira optimalan broj klastera.

Pored "lakat" krive, još jedna grafička metoda se koristi za evaluaciju klasterizacije – kriva siluete. Kriva siluete generiše se na osnovu koeficijenta analize siluete. Ovaj

koeficijent izračunava gustinu klastera, generišući rezultat za svaki uzorak na osnovu razlike između prosječne udaljenosti unutar klastera i srednje udaljenosti najbližeg klastera za taj uzorak, koja je normalizovana maksimalnom vrijednošću (Goel, 2020). Na osnovu dobijenog koeficijenta, grafički se može predstaviti stepen odvojenosti između klastera, te se na taj način može utvrditi njihov optimalan broj. Koeficijent analize siluete (SA) se izračunava na sljedeći način:

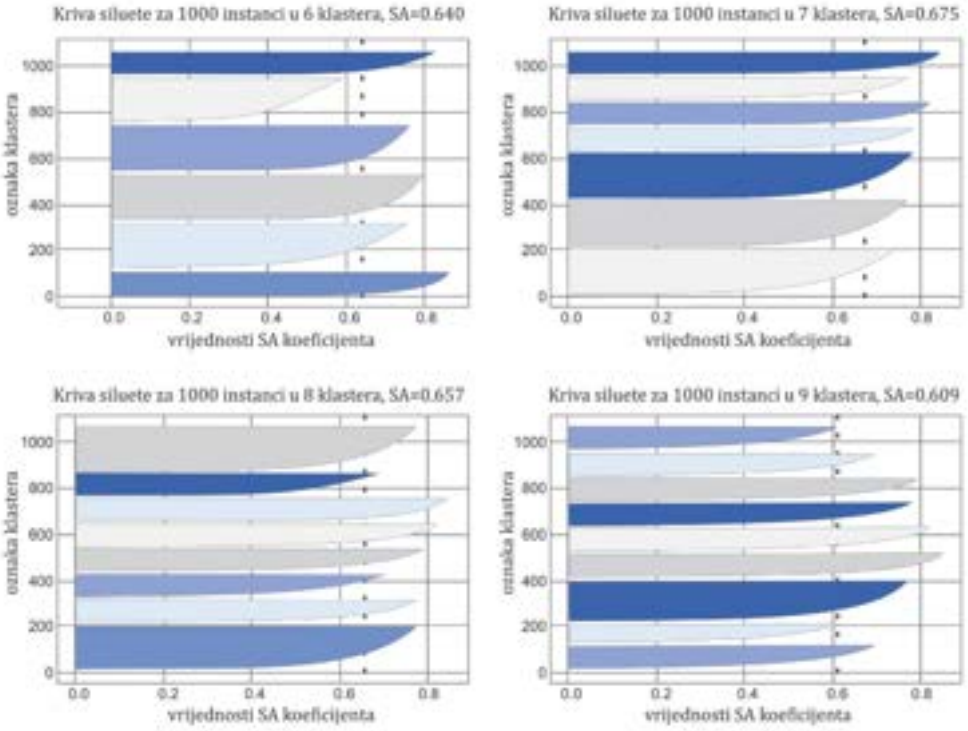
$$SA = \frac{b^i - a^i}{max(b^i, a^i)}$$

Dakle, analizom siluete utvrđuje se stepen odvojenosti između klastera. U navedenoj formuli, za svaki uzorak, a_i predstavlja prosječnu distancu od svih tačaka u istom klasteru, dok b_i predstavlja prosječnu udaljenost od svih tačaka u najbližem klasteru (Kumar, 2020). Koeficijent analize siluete ima raspon vrijednosti od -1 do 1, sa sljedećim značenjem:

- SA = 0: uzorak je vrlo blizu susjednih klastera;
- SA = 1: uzorak je daleko od susjednih klastera;
- SA = -1: uzorak je dodijeljen pogrešnim klasterima.

S tim u vezi, najbolja vrijednost koeficijenta je kada je ona što veća i bliža jedinici.

Na Slici 9 predstavljena je kriva siluete za k = 6, 7, 8 i 9. Iz navedenog prikaza može se uočiti da je optimalna vrijednost k = 7, tj. da je potrebno kreirati sedam klastera, uzimajući u obzir maksimalnu vrijednost koeficijenta od 0,675.



Slika 9. Grafički prikaz analize siluete (prilagođeno prema: Goel, 2020)

Metoda za evaluaciju klasterizacije, tj. izbor optimalnog broja klastera, koja će biti primijenjena u empirijskom dijelu ovog rada je *Davies-Bouldin* indeks. Naime, *Davies-Bouldin* indeks kvantifikuje prosječnu odvojenost svakog klastera od sljedećeg najbližeg klastera, tako što izračunava varijansu unutar klastera i rastojanje između centroida svakog klastera (Davies & Bouldin, 1979). Za svaki klaster identifikuje se njegov najbliži susjedni klaster, a zbir varijacija unutar klastera dijeli se razlikom između njihovih centroida. Ova vrijednost se izračunava za svaki klaster, a *Davies-Bouldin* indeks predstavlja srednju vrijednost ovih vrijednosti (Rhys, 2020). Što je prosječna sličnost niža, to su klasteri bolje odvojeni i rezultat izvršene klasterizacije je bolji.

Procedura izračunavanja *Davies-Bouldin* indeksa obavlja se kroz pet koraka (Sidyakov, n.d.):

1. Izračunavanje disperzije unutar klastera,

2. Izračunavanje mjere razdvajanja,
3. Izračunavanje sličnosti između klastera,
4. Pronalažanje najsličnijeg klastera za svaki klaster,
5. Izračunavanje *Davies-Bouldin* indeksa.

Za izračunavanje disperzije unutar klastera, *Davies* i *Bouldin* (1979) dali su sljedeću formulu:

$$scatter_k = \left(\frac{1}{n_k} \sum_{i \in k}^{n_k} (x_i - c_k)^2 \right)^{1/2}$$

gdje je $scatter_k$ mjera disperzije unutar klastera k, n_k je broj primjera u klasteru k, x_i je i-ti primjer u klasteru k, a c_k je centroid klastera k. Dakle, da bi se dobila mjera disperzije, računa se prosječna udaljenost između svakog primjera unutar klastera i njegovog težišta, tj. centroida.

Za naredni korak, izračunavanje mjere odvojenosti između klastera (j i k), autori su dali sljedeću formulu:

$$separation_{j,k} = \left(\sum_{1 \le j \le k}^{N} (c_j - c_k)^2 \right)^{1/2}$$

gdje je $separation_{j,k}$ mjera razdvajanja između klastera j i k, c_j i c_k su njihovi odgovarajući centroidi, a N ukupan broj klastera.

Zatim se izračunava sličnost između klastera na osnovu formule:

$$ratio_{j,k} = \frac{scatter_j + scatter_k}{separation_{j,k}}$$

U osnovi, u okviru ovog koraka računa se sličnost klastera kao zbir dvije disperzije unutar klastera podijeljene mjerom njihovog razdvajanja. Što je veći $ratio_{ij}$, to su

sličniji klasteri *i* i *j*. Dakle, najbolji rezultat se dobija kada je ovaj broj što manji (Sidyakov, n.d.).

Ovaj odnos se izračunava za sve parove klastera, a za svaki klaster definisan je najveći odnos između njega i ostalih klastera - R_k. Nakon toga, *Davies-Bouldin* indeks se dobija kao prosjek ovih najvećih odnosa:

$$DB = \frac{1}{N}\sum_{k=1}^{N} R_k$$

Iako predstavlja jedan od najčešće korišćenih algoritama za klasterizaciju zbog jednostavnosti primjene, *k-means* ima određene nedostatke (Garcia-Dias et al., 2020).

Kao što je prethodno pomenuto, jedan od nedostataka ovog algoritma odnosi se na nemogućnost garancije da se konvergencija izvrši do globalnog minimuma – ponekad *k-means* konvergira na lokalni minimum. Međutim, ovaj problem je rješiv ponavljenjem procesa klasterizacije više puta i upoređivanjem njihovih rezultata. S obzirom na to da ovaj algoritam ne zahtijeva značajnu kompjutersku moć, ponavljanje procesa *k-means* algoritma je vremenski efikasnije od korišćenja drugih algoritama klasterizacije koji nemaju ovaj nedostatak.

Drugi nedostatak ovog algoritma je činjenica da se broj klastera mora upisati kao input, tj. broj klastera mora biti definisan *a priori*. Slično kao i kod prethodno pomenutog nedostatka, i ovaj karakteriše većinu algoritama za klasterizaciju. Postoje određeni algoritmi gdje se broj klastera ne mora unaprijed definisati, ali, oni zavise od definisanja drugih parametara. Iako se u literaturi ovo navodi kao nedostatak, on se može prevazići korišćenjem metoda za odabir optimalnog broja klastera, od kojih su neki pomenuti u ovom dijelu rada – *„Lakat kriva"*, *Dunn* indeks ili *Davies-Bouldin* indeks, koji će, kao mjera, biti korišćen u empirijskom dijelu ovog rada.

Treći nedostatak *k-means* algoritma je da on kao rezultat uvijek ima grupe, čak i kada u distribuciji podataka nema stvarnih grupa, slično kao što linearna regresija kao rezultat ima liniju, čak iako se primjenjuje na podacima sa eksponencijalnom raspodjelom. Ovo takođe važi za većinu algoritama klasterizacije i ne treba ga smatrati nedostatkom, već karakteristikom metode koju treba uzeti u obzir. Međutim, ovo nije nedostatak isključivo *k-means* algoritma, već generalno metoda klasterizacije.

Četvrti nedostatak se odnosi na neefikasnost algoritma u situacijama linearne neodvojivosti podataka. Kao što je prethodno navedeno, *k-means* generiše klastere koji se ne preklapaju na osnovu kriterijuma sličnosti, najčešće Euklidskog rastojanja, i u odnosu na centar klastera. S tim u vezi, granice za odlučivanje mogu biti isključivo linearne, te algoritam neće biti efikasan ukoliko se primijeni na podacima koji nisu linearno odvojivi. S obzirom na to da ovaj nedostatak važi za ovaj kriterijum sličnosti, može se izbjeći korišćenjem drugog kriterijuma ili transformacijom podataka kako bi se mogli linearno odvojiti.

Dodatni nedostaci *k-means* algoritma koji se navode u literaturi odnose se neefikasnost algoritma kada klasteri imaju različite razmjere, različite oblike ili neuravnotežen broj primjera (Garcia-Dias et al., 2020). Konkretno, kada su prisutna značajna odstupanja (eng. *outliers*), definisani klaster centroidi obično nisu toliko reprezentativni kao što bi inače bili, pa će zbir kvadratne greške biti veći (Tan et al., 2019).

Kao što je prethodno navedeno, uprkos nedostacima, *k-means* metoda predstavlja jednu od najkorišćenijih za klasterizaciju i ima značajnu primjenu u istraživanjima koja tretiraju segmentaciju kupaca. Na primjer, *Gončarovs* (2018) je primijenio *k-means* u cilju segmentacije korisnika iz finansijske institucije; *Martínez et al.* (2019) koristili su ovu metodu uz RFM za segmentaciju kupaca i upravljanje kampanjama; *Liao et al.* (2011) su na osnovu *k-means* metode izvršili klasterizaciju kupaca, a zatim generisali asocijativna pravila za svaki od definisanih klastera. Na ovaj način, autori su kreirali prilagođene predloge i rješenja za kompanije koje sprovode direktni marketing. *Khajvand et al.* (2011) su ovaj algoritam primijenili u cilju segmentacije

kupaca kompanije koja se bavi prodajom kozmetičkih proizvoda i procjene njihove cjeloživotne vrijednosti. Rezultati sprovedene klasterizacije omogućavaju kreiranje marketing strategija za svaki pojedinačni segment, što može rezultirati većom stopom zadržavanja kupaca. *Abdi* i *Abolmakarem* (2019) su, takođe, primjenom *k-means* klasterizacije izvršili segmentaciju kupaca s ciljem predikcije stope napuštanja. Na osnovu dobijenih rezultata, oni su razvili nekoliko strategija za osnose s kupcima. Pored ovih, mnogi drugi autori su koristili *k-means* klasterizaciju za segmentaciju kupaca, što potvrđuje popularnost ovog modela u akademskim krugovima.

U narednom dijelu rada biće predstavljena DT metoda (drvo odlučivanja), kao sljedeća *data mining* metoda koja će biti primijenjena u ovom istraživanju.

4.1.2 Decision Tree metoda

U sekciji 3.2 opisane su osnove DT metode, kao jedne od *data mining* metoda koje se najčešće koriste za prediktivnu segmentaciju kupaca. S obzirom na to da konvencionalne statističke i ekonometrijske metode, u slučaju primjene na velikim podacima i uz veći broj prediktora, ne mogu obezbijediti nivo efikasnosti kao što to mogu *data mining* tehnike, često se upravo metode, kao što su DT, SVM, NN i druge koriste u cilju modeliranja kompleksnijih veza u podacima (Varian, 2014). Jedna od najkorišćenijih tehnika upravo je drvo odlučivanja, koje može biti klasifikaciono i regresiono. Klasifikaciona DT metoda kao rezultat ima kategoričke varijable poput „da" i „ne" u pogledu obavljanja kupovine, dok regresiona za rezultat ima numeričku ili neprekidnu vrijednost varijable, poput vjerovatnoće da selektovani kupac obavi kupovinu.

Metoda DT dijeli skup podataka po vrijednostima atributa, tako da podgrupe sadrže što više primjera jedne klase, tj. da se njihova nečistoća svede na minimum, kako bi se osigurala homogenost po pitanju klase koju predstavljaju. Tokom induktivne podjele formira se model u obliku drveta, na osnovu čega je i sama metoda dobila naziv. Djelovi stabla odlučivanja koji se razlažu po određenim kriterijumima nazivaju se čvorovi (eng. *node*), a prvi čvor se naziva korijenski čvor (eng. *root node*).

Čvorovi imaju jednu granu koja vodi do njih i dvije ili više grana koje vode od njih. Čvorovi na dnu stabla se nazivaju čvorovi listova ili listovi (eng. *leaf nodes/leaves*). Oni predstavljaju oznake klase (eng. *label*), tj. rezultat svih prethodnih odluka. S tim u vezi, listovi imaju jednu granu koja vodi do njih, ali nema grana koje vode dalje od njih. Kada primjer pronađe svoj put od korijena do lista, on dalje ne napreduje i klasifikuje se kao većinska klasa u tom listu (Rhys, 2020).

U svakoj fazi procesa kreiranja stabla, algoritam uzima u obzir sve prediktore i bira onaj koji najbolje može da izvrši diskriminaciju klasa, tj. onaj koji daje podjelu na najčistije skupove. Ovaj proces počinje u korijenu, a zatim se na svakoj grani ponovo traži naredna osobina (atribut, tj. prediktor), koja će najbolje razlikovati klase primjera koji se nalaze na toj grani (Rhys, 2020). Dakle, biraju se atributi koji pružaju najbolju podjelu prema datom kriterijumu. Kriterijum po kome se vrši podjela, odnosno mjera kvaliteta podjele može biti *information gain* (Quinlan, 1986), *gain ratio* (Quinlan, 1992), *Gini index* (Breiman, 1984) ili tačnost cijelog drveta (eng. *accuracy of the whole tree*), koji će biti opisani u nastavku ovog poglavlja.

Putanja od korijena do lišća definiše pravila klasifikacije ako-onda (eng. *if-then*) u terminima prediktivnih atributa (čvorova). Složenost i tačnost generisanog modela zavisi od dubine stabla (eng. *depth of the tree*), minimalne veličine čvora pomoću kojeg se može izvršiti podjela, tj. broja primjera u njegovoj podgrupi (eng. *minimum size for split*), veličine lista (eng. *leaf size*) i definisanog minimalnog dobitka koji je postignut podjelom čvora (eng. *minimum gain*). Što je dubina manja, veća minimalna veličina čvora za podjelu, veća veličina lista i veći minimalni dobitak, drvo odlučivanja će biti manje složeno, ali će zato imati i manju tačnost (S. Rogić & Kašćelan, 2020).

Najznačajniji algoritmi za konstruisanje DT modela su: *Iterative Dichotomiser 3* – ID3, C4.5 i *Classification And Regression Trees* - CART.

Jedan od prvih DT algoritama, ID3, razvio je *John Ross Quinlan* (1986). Ovaj algoritam radi isključivo s kategoričkim varijablama, rezultat, tj. *label* je nominalni (kategorički), a kao mjeru za evaluaciju koristi *information gain*. Isti autor je razvio

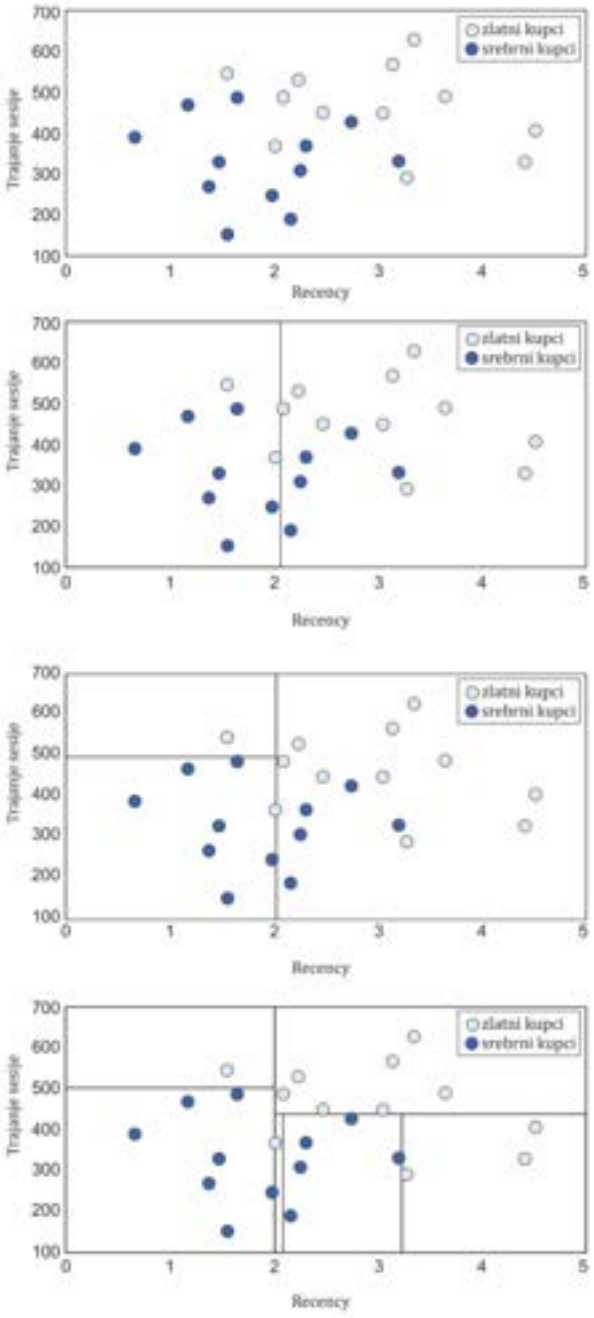
i napredniji DT algoritam – C4.5 (Quinlan, 1992). Za razliku od prethodnog, algoritam C4.5 može da radi s nedostajućim vrijednostima i podržava i kategoričke i numeričke prediktorske varijable, dok je rezultat, tj. *label* takođe nominalni. Za vrijednosti kontinuirane (numeričke) ulazne varijable - prediktora, algoritam obavlja sortiranje i podjelu. Dobitak se računa (eng. *gain*) za odgovarajuće zapise svake particije, a za podjelu se bira ona koja maksimizira dobitak (EMC Education Services, 2015). Još jedan nedostatak algoritma ID3, koji je riješen razvojem C4.5 algoritma je izbjegavanje prevelike adaptacije prema podacima za obučavanje modela – (eng. *overfitting*). Previše prilagođeni model dobro opisuje podatke na kojima uči, ali ima lošu prediktivnu moć na neviđenim podacima. Naime, ID3 može da generiše duboko i složeno stablo, koje zbog toga može biti sklono prekomjernom prilagođavanju podacima, dok C4.5 koristi tehniku podrezivanja drveta (eng. *pruning*) u cilju smanjenja kompleksnosti. Kao kriterijum podjele, ovaj algoritam koristi *gain index*. Konačno, CART algoritam (Breiman, 1984) radi i s kategoričkim i numeričkim prediktivnim varijablama, dok, za razliku od prethodnih, kod ovog algoritma ciljna varijabla može biti kategorička - u slučaju klasifikacije, ili numerička - u slučaju regresije. Kao kriterijum podjele koristi *gini index*.

Kada je riječ o DT metodi, važno je opisati dvije osnovne ideje za njenu realizaciju: rekurzivna particija (eng. *recursive partitioning*) za konstruisanje drveta i podrezivanje drveta, za smanjenje njegove dubine ili kompleksnosti (Shmueli et al., 2018). U procesu konstruisanja stabla odlučivanja, kao što je prethodno pomenuto, rekurzivnom particijom se uzimaju u obzir sve promjenljive prediktora i bira se prediktor koji najbolje vrši diskriminaciju klasa.

Za potrebe opisivanja procesa rekurzivne particije, označimo varijablu ishoda sa Y, a ulazne (prediktorske) promjenljive sa $X_1, X_2, X_3 .. X_p$, na osnovu primjera koji su u svojoj knjizi predstavili autori *Shmueli et al.* (2018). U klasifikaciji će izlazna varijabla biti kategorička. Rekurzivna particija dijeli p-dimenzionalni prostor X prediktorskih varijabli na višedimenzionalne pravougaonike koji se ne preklapaju. Ovdje prediktorske promjenljive mogu biti kontinualne, binarne ili nominalne. Ova podjela se vrši rekurzivno, tj. proces se nastavlja na osnovu rezultata prethodnih

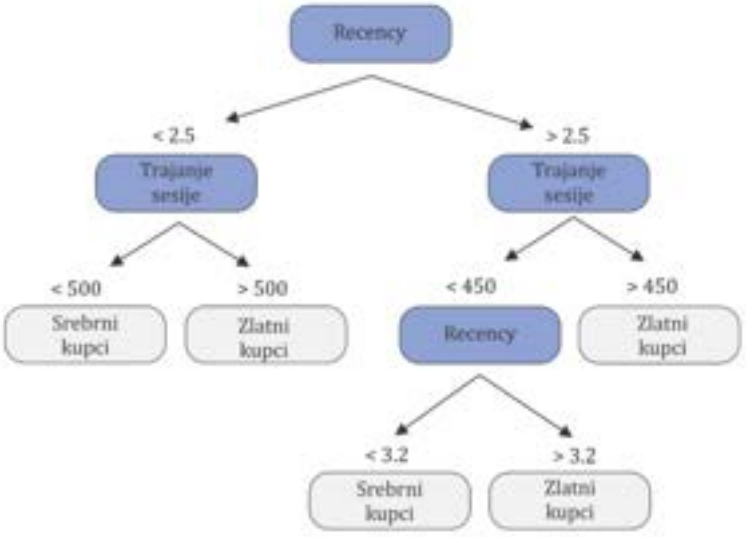
podjela. Prvo se odabere jedna od prediktorskih promjenljivih, recimo X_i, a vrijednost X_i-a, koju možemo označiti sa s_i, odabere se za podjelu p-dimenzionalnog prostora na dva dijela: jedan dio koji sadrži sve tačke sa $X_i < s_i$, a drugi sve tačke za koje važi $X_i \geq s_i$.

Nakon ovog koraka, jedan od ova dva pravougaonika dijeli na sličan način ponovnim odabirom prediktorske promjenljive, koja opet može biti ista - X_i ili neka druga promjenljiva, kao i vrijednosti podjele za tu promjenljivu. Rezultat su tri višedimenzionalna pravougaona područja. Ovaj proces se nastavlja tako da se dobijaju sve manji i manji pravougaoni regioni. Ideja je da se cjelokupan X-prostor podijeli na pravougaonike tako da je svaki pravougaonik što je moguće homogeniji ili „čistiji". Pod tim podrazumijevamo zapise koji pripadaju samo jednoj klasi. Naravno, to nije uvijek moguće, jer se može desiti da postoje zapisi koji pripadaju različitim klasama, ali imaju potpuno iste vrijednosti za svaki od promjenljivih prediktora (Shmueli et al., 2018). Na Slici 10 grafički je predstavljen proces podjele prostora.



Slika 10. Grafički prikaz procesa rekurzivne particije

Na osnovu kompletirane podjele (posljednji segment Slike 10), može se konstruisati drvo odlučivanja, predstavljeno na Slici 11.



Slika 11. Ilustracija stabla odlučivanja

Sveobuhvatni i detaljni opisi procesa rekurzivne particije predstavljeni su u radovima autora *Landau* i *Barthel* (2010), kao i *Strobl et al.* (2009).

Kao što je prethodno navedeno, kriterijumi za podjelu, tj. mjere kvaliteta podjele zavise od odabranog DT algoritma. U ovom dijelu rada biće ukratko objašnjeni kriterijumi: *information gain, gain ratio* i *Gini index*. Kriterijumi za podjelu se nazivaju i mjere nečistoće.

Information gain je kriterijum za podjelu, koji koristi informacionu entropiju za mjeru nečistoće, kao funkciju distribucije vjerovatnoće. Neka je *D* polazni skup podataka, a *D1, … , Dn* podjela polaznog skupa podataka na osnovu *n* klasa zavisne varijable. U ovom slučaju, *n* klasa zavisne varijable označava se sa *i = 1, 2, 3… n*, a distribucija vjerovatnoće sa $P = (p_1, p_2, ..., p_n)$, gdje je p_i vjerovatnoća da se tačka nalazi u podskupu *Di* (od skupa *D*), koja se izračunava kao frekvencija (tj. udio) instanci iz podskupa *Di* u odnosu na potpuni skup instanci *D*. Entropija se računa pomoću sljedeće formule (Dubinets, n.d.):

$$Entropy(P) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

Ova mjera se kreće između 0 (najčistiji, svi zapisi pripadaju istoj klasi) i *log2 (p_i)* (kada su svih *n* klasa predstavljene podjednako). U slučaju dvije klase, mjera entropije je maksimizirana pri $p_i = 0{,}5$. Na osnovu izračunate entropije, *information gain* se dobija korišćenjem sljedeće formule (Dubinets, n.d.):

$$Information\ Gain_Y(X_i, D) = Entropy\left(P_Y(D)\right) - \sum_{j=1}^{m} \frac{\left|\sigma_{X_i=v_j}(D)\right|}{|D|}\ Entropy\left(P_Y\left(\sigma_{X_i=v_j}(D)\right)\right)$$

$$Information\ Gain(X_i, D) = Entropy\ Before\ Split - Entropy\ After\ Split$$

Information gain određuje smanjenje nesigurnosti nakon podjele skupa podataka po određenom atributu, što znači, ako se vrijednost ovog kriterijuma poveća, taj atribut je najkorisniji za klasifikaciju. Dakle, atribut koji ima najveću vrijednost kriterijuma smatra se najboljom varijablom koja se bira za podjelu.

Kao jedan od nedostataka ovog kriterijuma za podjelu navodi se pristrasnost prema testovima koji sadrže veći broj ishoda (Quinlan, 1992). Na primjer, ukoliko se konstruiše DT u cilju predviđanja budućeg ponašanja potrošača, a jedna od ulaznih varijabli predstavlja nešto jedinstveno za svakog kupca, poput broja kreditne kartice ili broja telefona, onda će *information gain* za taj atribut biti veoma visok. Kao rezultat toga, ovaj čvor će biti postavljen visoko, pri korijenu DT modela. Međutim, testiranje na nepoznatim podacima neće dati dobar rezultat s obzirom na to da su podaci za nove kupce nepoznati i njihov jedinstveni identifikator neće biti među rezultatima iz skupa za obučavanje (Dubinets, n.d.). Kako bi se uklonio ovaj nedostatak, razvijen je drugi kriterijum za podjelu – *gain ratio* (Quinlan, 1992), koji normalizuje *information gain* za atribut X_i, u odnosu na vrijednost entropije tog atributa. *Gain ratio* se računa pomoću sljedeće formule (Dubinets, n.d.):

$$Gain\ Ratio_Y(X_i, D) = \frac{Information\ Gain_Y\ (X_i, D)}{Entropy\ \left(P_{X_i}(D)\right)}$$

Iz formule se može konstatovati – ako je entropija vrlo mala, onda će *gain ratio* biti visok i obrnuto.

Autor ovog kriterijuma predlaže sljedeću proceduru:

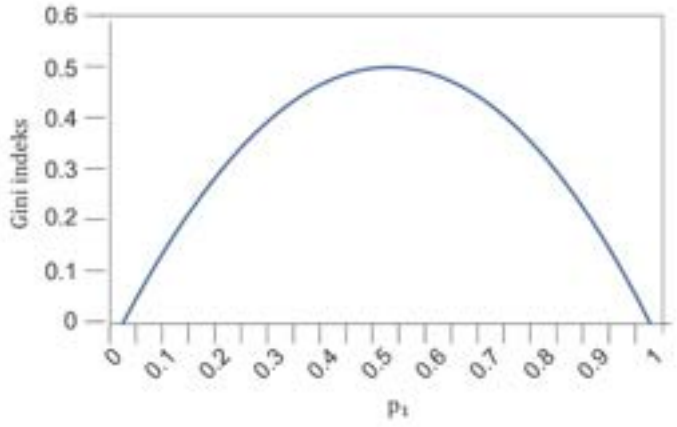
1. izračunavanje vrijednosti *information gain* za sve atribute i dobijanje prosječne vrijednosti za *information gain;*
2. izračunavanje vrijednosti *gain ratio* za sve atribute za koje je vrijednost *information gain* kriterijuma veća ili jednaka prosječnoj vrijednosti za *information gain*, te selektovati atribute koji imaju najveću vrijednost za *gain ratio*.

Posljednja mjera koja će u ovom radu biti predstavljena jeste *Gini indeks*, koji se takođe može koristiti kao mjera nečistoće. *Gini* indeks predstavlja varijaciju *Gini* koeficijenta, koji se u ekonomiji koristi kao mjera disperzije bogatstva populacije (González Abril et al., 2010). Ovaj indeks računa se pomoću sljedeće formule:

$$Gini(P) = \sum_{i=1}^{n} p_i(1 - p_i) = 1 - \sum_{i=1}^{n} (p_i)^2$$

Dakle, n klasa izlazne varijable označene su sa $i = 1, 2, \ldots, n$, a *gini indeks* se računa za pravougaonik, odnosno podskup P. U ovoj formuli, sa p_i označena je proporcija primjera u pravougaoniku D, koji pripadaju klasi i ili vjerovatnoća da primjer pripada klasi i. Ova mjera uzima vrijednosti između 0 (kada svi zapisi pripadaju istoj klasi ili kada postoji samo jedna klasa) i $1 - (1 / n)$ (kada su svi elementi nasumično distribuirani kroz različite klase).

Na Slici 12 prikazane su vrijednosti *Gini indeksa* za slučaj sa dvije klase, kao funkcija p_i. Na osnovu slike se može uočiti da je mjera nečistoće na vrhuncu kada je $p_i = 0{,}5$, tj. kada pravougaonik sadrži 50% svake od dvije klase.



Slika 12. Vrijednosti *Gini* indeksa za slučaj sa dvije klase kao funkcija proporcije zapisa u klasi 1 (p_1) (Shmueli et al., 2018)

Za konstriusanje DT modela korišćenjem *Gini indeksa*, za korijenski čvor se bira onaj atribut koji ima najmanju vrijednost ovog kriterijuma.

U konačnom, izbor kriterijuma za podjelu i dalju konstrukciju DT modela zavisi od strukture podataka, kao i potencijalne pristrasnosti različitih algoritama.

Proces podjele *X*-prostora, odnosno rasta stabla nastavlja se sve dok se ne postigne kriterijum zaustavljanja. Neki od najprihvaćenijih kriterijuma za zaustavljanje procesa rekurzivne particije su (Rokach & Maimon, 2015):

1. Svi primjeri u skupu podataka za obučavanje pripadaju jednoj vrijednosti izlazne varijable;
2. Dostignuta je maksimalna dubina stabla;
3. Broj primjera u listu manji je od minimalnog broja primjera za roditeljske čvorove (eng. *parent nodes*);
4. Ako je čvor podijeljen, broj primjera u jednom ili više podčvorova (eng. *child nodes*) bio bi manji od minimalnog broja primjera za podčvorove;
5. Najbolji kriterijum podjele nije veći od definisanog praga.

Međutim, za razliku od definisanja kriterijuma koji zaustavljaju rast stabla odlučivanja, drugo rješenje predstavlja potkresivanje stabla. Ovaj proces pokreće se

kada se konstruiše kompletno stablo odlučivanja velike dubine i kompleksnosti, koje može biti sklono pretjeranom prilagođavanju podacima iz skupa za obučavanje modela. U tom slučaju, da bi se izbjegla takva prilagođenost, koja dovodi do slabih performansi primjenom modela na novim i nepoznatim podacima, stablo odlučivanja se smanjuje uklanjanjem najslabijih grana, koje ne smanjuju značajno stopu greške, tj. ne doprinose tačnosti generalizacije.

Podrezivanje se sastoji od sukcesivnog odabira čvora odluke i njegovog redizajniranja kao lista. Dakle, uklanjaju se grane koje se šire izvan tog čvora i na taj način se smanjuje veličina stabla. Ovaj proces smanjuje grešku klasifikacije u testnom skupu, tj. na nepoznatim podacima, jer se podrezivanjem stabla smanjuje prilagođavanje modela izuzecima u podacima za obučavanje. Na taj način, model u podacima za obučavanje bilježi obrasce, a ne šumove (izuzetke) (Shmueli et al., 2018).

Postoje različite tehnike za potkresivanje stabla odlučivanja, kao što su: podrezivanje na osnovu složenosti troškova (eng. *cost complexity pruning),* podrezivanje na osnovu smanjene greške (eng. *reduced error pruning),* podrezivanje na osnovu minimalne greške *(eng. minimum error pruning),* pesimističko podrezivanje *(eng. pessimistic pruning),* podrezivanje na osnovu greške *(eng. error-based pruning),* podrezivanje na osnovu najmanje dužine opisa *(eng. minimum description length pruning)* i druge. S obzirom da prevazilaze okvir ovog rada, tehnike za podrezivanje stabla neće biti pojedinačno predstavljene. Pomenute tehnike su detaljno u svojoj knjizi opisali autori *Rokach* i *Maimon* (2015).

Drvo odlučivanja je zbog jednostavnosti upotrebe i mogućnosti interpretacije rezultata postala veoma popularna metoda za klasifikaciju. Ova metoda ima značajne prednosti, kao što su: visok stepen automatizacije, robustnost prema izuzecima u podacima i sposobnost da se nosi s nedostajućim vrijednostima.

Kao neke od osnovnih prednosti klasifikacione DT metode u literaturi se navođe (Rokach & Maimon, 2015):

1. Stabla odlučivanja su sama po sebi objašnjena i lako ih je pratiti. S tim u vezi, ako DT model ima razuman broj listova, i neprofesionalni korisnici mogu ga razumjeti. Uz to, model se može predstaviti i skupom *if-then* pravila, što dodatno olakšava razumijevanje rezultata;
2. Stabla odlučivanja mogu raditi i s nominalnim i numeričkim nezavisnim varijablama;
3. Zastupljenost stabla odlučivanja je dovoljno bogata da predstavlja bilo koji klasifikator diskretnih vrijednosti;
4. Stabla odlučivanja mogu raditi i sa skupovima podataka koji mogu imati greške;
5. Stabla odlučivanja mogu raditi i sa skupovima podataka koji mogu imati nedostajuće vrijednosti;
6. Stabla odlučivanja smatraju se neparametarskom metodom, tj. odluke ne uključuju pretpostavke o raspodjeli prostora prediktorskih varijabli u procesu rekurzivne particije, kao ni o strukturi klasifikatora;
7. Kada su troškovi klasifikacije visoki zbog velikog broja podataka, stabla odlučivanja mogu biti privlačna, jer traže samo vrijednosti atributa za podskup podataka (jednu particiju - jedan put od korijena ka listu).

Međutim, autori iz ove oblasti navode i neke nedostatke ovog modela, poput nedostatka robustnosti i neoptimalnih performansi (Larivière & Van Den Poel, 2005), kao i nestabilnost modela – male promjene u skupu podataka za obučavanje modela mogu izazvati velike promjene u rezultatima, tj. predikciji (Breiman, 1996). Osim toga, DT nije dobar izbor ako skup podataka sadrži veliki broj irelevantnih varijabli. Neki od nedostataka ove metode uklonjeni su sa razvojem *ensemble* metoda, tj. optimizacijom DT tehnike, kao što su *Random Forest* (Breiman, 2001) i *Bootstrap aggregating – Bagging* (Breiman, 1996). Pomenute *ensemble* tehnike, koje će biti korišćene u empirijskom dijelu ovog rada, biće opisane u sekciji 4.1.4. U prethodnim istraživanjima potvrđeno je da ove metode poboljšavaju prediktivnu moć u poređenju s jednim stablom odlučivanja (EMC Education Services, 2015).

Kao i *k-means* metoda za klasterizaciju, DT je jedna od najkorišćenijih metoda za klasifikaciju i segmentaciju kupaca i često se primjenjuje u studijama iz oblasti direktnog marketinga. *Khalili-Damghani et al.* (2018) su predložili pristup za problem segmentacije kupaca, koji se sastoji od klasterizacije, generisanja pravila i DT metode. Cilj ovog istraživanja, koje je sprovedeno u formi dvije studije slučaja u oblasti telekomunikacija i osiguranja bio je predikcija profitabilnosti potencijalnih kupaca, te njihovog zadržavanja kroz marketing aktivnosti kompanija. Dodatno, *Kim et al.* (2006) su koristili DT metodu za otkrivanje karakteristika segmentiranih kupaca, kako bi se kreirale specifične marketing strategije za svaku pojedinačnu grupu. Za predviđanje vrijednosti potrošača u direktnom marketingu, baziranoj na atributima o izvršenim transakcijama, *Rogić* i *Kašćelan* (2019) su generisali DT model, koji je ujedno imao funkciju opisivanja SVM rezultata, tj. ekstrakcije pravila iz SVM modela. Slično, *Rogić* i *Kašćelan* (2021) su, nakon definisanja tržišnih segmenata u na osnovu RFM atributa, generisali DT model za predikciju pripadnosti kupaca nekom od segmenata, na osnovu atributa o kupovnim transakcijama i raspoloživim demografskim atributima, poput pola i regiona. U cilju opisivanja individualnih segmentata pomoću DT metode, generisan je set pravila za opisivanje kupaca u okviru segmenata, što može unaprijediti aktivnosti direktnog marketinga u kompanijama.

Xiahou et al. (2016) su primijenili *k-means* klasterizaciju i DT metodu za klasifikaciju kupaca na osnovu njihove profitabilnosti, pri čemu su definisali četiri nivoa: ekstremno niska, niska, srednja i visoka profitabilnost. Dodatno, *Olson* i *Chae* (2012) su, uz LR, RFM metodu i NN, koristili i DT metodu za prediktivno modeliranje odgovora na kampanju od strane potrošača. Autori su potvrdili da je korišćenje *data mining* tehnika za podršku odlučivanju u marketingu veoma značajno, posebno ukoliko se rezultati ovih prediktivnih tehnika uporede s prostijim - tradicionalnim deskriptivnim modelima.

Han et al. (2012) su u svom radu predložili metodu za segmentaciju korisnika telekomunikacione kompanije, koja je zasnovana na životnom ciklusu kupaca. Autori su naveli da su, zbog poteškoća u kvantitativnom izračunavanju dugoročne

vrijednosti kupaca, kredita i lojalnosti, primijenili DT metodu za ekstrakciju važnih parametara koji se odnose na ove karakteristike. Konačno, interesantnu primjenu DT metode izvršili su *Lipyanina et al.* (2020), koji su predložili model za evaluaciju targetiranja kupaca u kampanji plasiranoj putem društvene mreže *Facebook*, a u cilju formiranja strategije oglašavanja.

Ovi, kao i mnogi drugi radovi potvrđuju značaj i popularnost primjene DT metode u akademskim istraživanjima iz sfere direktnog marketinga.

U narednom dijelu rada biće predstavljen *Support Vector Machine* metoda, koja će takođe biti primijenjena u empirijskom dijelu ovog rada.

4.1.3 Support Vector Machine metoda

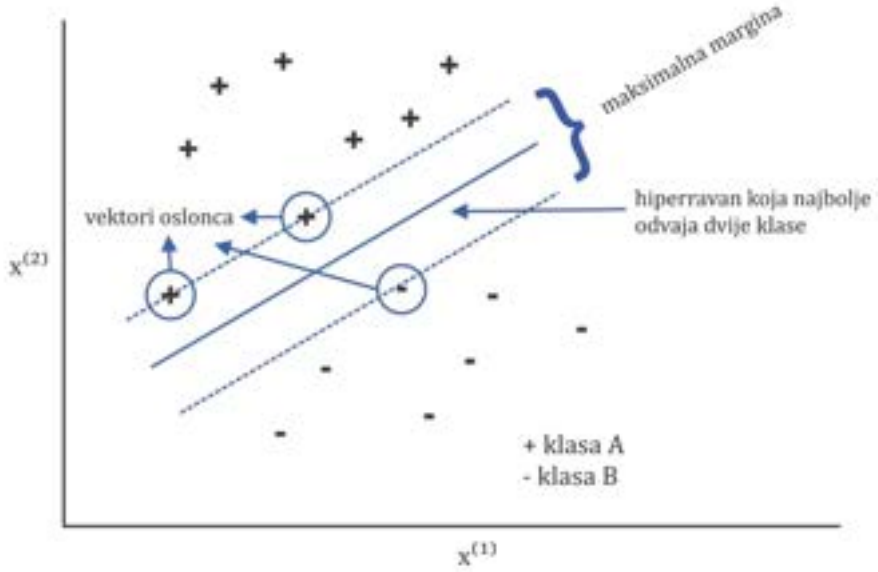
Vapnik (2010) je osmislio SVM kao nadgledanu tehniku učenja, koja sprovodi analize podataka u cilju identifikacije obrazaca. Naime, uz dati skup podataka koji uključuju i oznaku klase, ova tehnika opservacije, odnosno primjere, predstavlja tačkama u n-dimenzionom prostoru, gdje teži da identifikuje hiperravan koja će na najbolji način odvojiti date tačke. Novi, nepoznati podaci predstavljaju se u istom prostoru i klasifikuju se na osnovu njihove blizine margini koja odvaja klase (Sabbeh, 2018). Karakteristike koje izdvajaju SVM, u odnosu na ostale algoritme mašinskog učenja koji se koriste za klasifikaciju, odnose se na mogućnost ove metode da nauči obrasce klasifikacije podataka sa uravnoteženom tačnošću i ponovljivošću. Važno je napomenuti da se SVM može koristiti i za regresiju – *Support Vector Regression*, o čemu će biti riječi u sekciji 4.1.7.

SVM, kao tehnika klasifikacije, koristi statističku teoriju učenja (Vapnik, 2010). U kontekstu binarne klasifikacije, SVM pokušava da pronađe optimalnu hiperravan, tako da se maksimizira margina razdvajanja između pozitivnih i negativnih primjera. Ovaj proces ekvivalentan je rješavanju kvadratnog problema optimizacije, u kome presudnu ulogu igraju samo vektori oslonca (eng. *support vectors*), tj. tačke podataka koje su najbliže optimalnoj hiperravni (Coussement & Van den Poel, 2008). Dakle, funkcija odlučivanja SVM upravo predstavlja određivanje optimalne

hiperravni, koja služi za odvajanje, odnosno klasifikovanje primjera koji pripadaju jednoj klasi od primjera koji pripadaju drugoj klasi, pri čemu se klasifikacija vrši na osnovu obrazaca informacija o atributima. Na taj način, definisana hiperravan se u narednom koraku koristi za određivanje oznake klase za nepoznate primjere.

Efikasna upotreba SVM tehnike u marketingu ne zahtijeva detaljno razumijevanje njene matematičke osnove, ali zahtijeva jasno konceptualno razumijevanje u pogledu primjene. S tim u vezi, važno je razumjeti ideju iza procesa obuke funkcije odlučivanja SVM modela, koji obuhvata identifikovanje hiperravni koja maksimizira rastojanje, tj. marginu između vektora oslonca koji pripadaju različitim klasama.

Za klasifikacione probleme gdje su primjeri u potpunosti linearno odvojivi, može postojati veliki broj različitih hiperravni, koje podjednako dobro vrše diferencijaciju između klasa u skupu podataka za obučavanje modela. Međutim, optimizacija hiperravni, odnosno pronalaženje one koja maksimizira marginu oko sebe, vrši se s ciljem omogućavanja bolje generalizacije na neviđenim podacima, tj. podacima za testiranje ili validaciju modela. U tom smislu, margina se definiše kao udaljenost od hiperravni koja dodiruje najmanje primjera iz skupa podataka za obučavanje. Ovi primjeri, nazvani vektorima oslonca, podržavaju položaj hiperravni, a po njima je i sam algoritam dobio naziv. Upravo su te tačke najznačajniji slučajevi u skupu podataka za obučavanje, s obzirom na to da oni određuju granicu između klasa, dok ostale tačke nemaju uticaj na položaj i veličinu margine (Rhys, 2020). Na Slici 13 prikazana je ilustracija jedne hiperravni, koja odvaja primjere iz dvije različite klase.



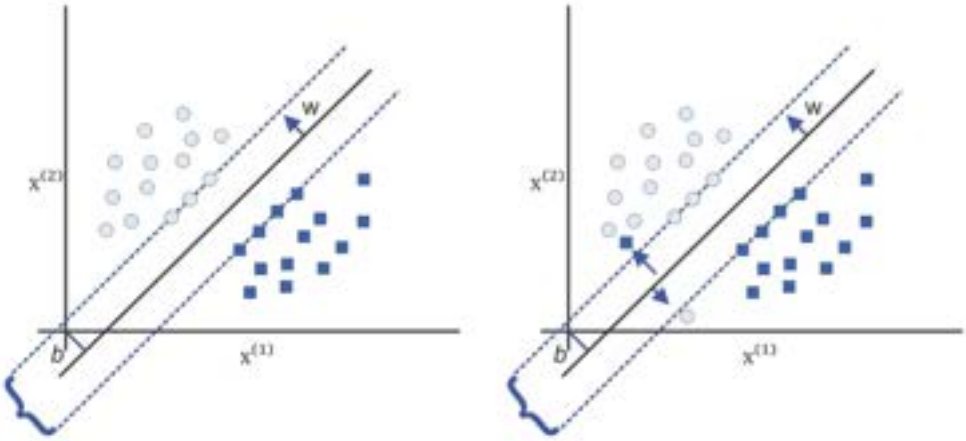
Slika 13. Ilustracija hiperravni koja maksimalno odvaja vektore oslonca

Na Slici 13 predstavljen je slučaj linearnog SVM problema s dvije dimenzije (dva atributa), gdje hiperravan odgovara pravoj liniji. U slučaju postojanja tri atributa, hiperravan bi odgovarala dvodimenzionalnoj ravni. Dakle, tražena hiperravan uvijek predstavlja površinu koja ima jednu dimenziju manje nego što je broj promjenljivih u skupu podataka. Bez obzira na nivo dimenzionalnosti, odnosno kompleksnost SVM modela, klasifikacioni problemi su najčešće linearni – hiperravan je ravna, a ne zakrivljena (Pisner & Schnyer, 2019).

Na prethodnoj slici prikazana je tzv. "čvrsta" margina (eng. *hard margin*), odnosno situacija kada nisu dozvoljene greške u obučavanju. Kada SVM koristi čvrstu marginu, nijedan primjer ne smije da upadne u definisani prostor margine. U tom slučaju, ako klase nisu u potpunosti odvojive, algoritam neće uspjeti (Rhys, 2020), što predstavlja veliki problem, jer klasifikacioni problemi često nisu toliko jednostavni. Stoga, veću primjenu za obučavanje klasifikatora ima definisanje "meke" margine (eng. *soft margin*). Na ovaj način se kreira veća margina, koja omogućava veći stepen generalizacije novih podataka, a klasifikatoru se zauzvrat dozvoljava da određene primjere pogrešno klasifikuje. U SVM modelu s mekom marginom, algoritam i dalje definiše hiperravan koja najbolje razdvaja klase,

međutim, postoji mogućnost da neki primjeri upadnu u marginu i budu pogrešno klasifikovani.

Što je margina čvršća, to će unutar nje biti manje instanci, a hiperravan će zavisiti od manjeg broja vektora oslonca. Što je margina mekša, to će unutar nje biti više instanci, dok će hiperravan zavisiti od većeg broja vektora oslonca (Rhys, 2020). Na Slici 14 predstavljena je ilustracija hiperravni sa tvrdom (lijevi segment) i s mekom marginom (desni segment).



Slika 14. Ilustracija tvrde i meke margine za definisanu SVM hiperravan

Osim slučaja linearno odvojivih klasa, postoje i situacije odslikane u skupovima podataka kada klase nisu linearno odvojive. Kod nelinearnih problema, SVM mapira podatke iz originalnog prostora (eng. *input space*) u prostor veće dimenzije (eng. *feature space*), gdje klasifikacija opet postaje linearna i klase je tada moguće odvojiti korišćenjem hiperravni (Vapnik, 2010). Umjesto eksplicitne funkcije preslikavanja u prostor veće dimenzije, ova transformacija se vrši korišćenjem kernel funkcije, koja omogućava izračunavanje skalarnog proizvoda vektora u originalnom prostoru (kernel trik). Dakle, maksimizacija margine u prostoru veće dimenzije se svodi na optimizacioni problem kvadratnog konveksnog programiranja u originalnom prostoru, uz upotrebu kernel funkcije. U literaturi su, u ove svrhe, prepoznate različite kernel funkcije, ali je često najefikasnija, a samim tim i najkorišćenija -

Radial Basis Function (RBF) (Sanderson, 2010), koja će biti primijenjena i u empirijskom dijelu ovog rada.

$$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right)$$

Dakle, SVM algoritam nastoji da maksimizira marginu u prostoru veće dimenzije, što se svodi na konveksnu optimizaciju, odnosno problem kvadratnog programiranja u originalnom prostoru:

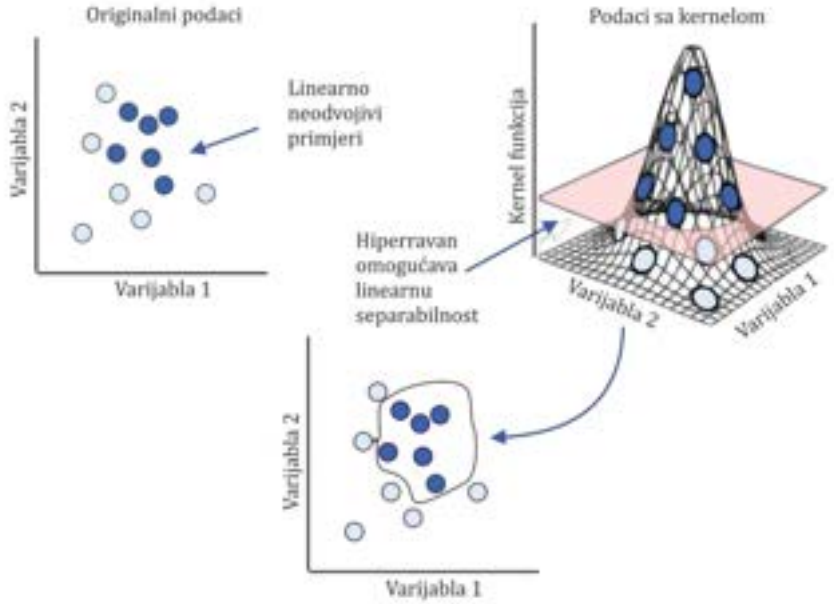
$$max_{a_i} \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i,j=1}^{n} a_i a_j y_i y_j K(x_i, x_j)$$

$$\sum_{i=1}^{n} y_i a_i = 0$$

$$0 \leq a_i \leq C, i = 1, \dots, n$$

gdje je K – kernel funkcija, α_i su Lagranžovi multiplikatori (eng. *Lagrange multipliers*), n je broj primjera u skupu podataka za obučavanje modela, a C je parametar koji vrši *trade-off* između maksimizacije margine, u odnosu na minimizaciju greške u klasifikaciji. Dakle, parametar C će definisati koliko će margina koja razdvaja klase biti tvrda odnosno meka. S tim u vezi, obučavanje SVM klasifikatora se realizuje izborom optimalnih vrijednosti parametra γ za RBF kernel i parametra C, koji predstavlja granicu za marginu, tj. prazan prostor između klasa ili, preciznije - kaznu za grešku kod meke margine. Izbor manjih vrijednosti za parametar C, tj. manje kazne za grešku na podacima za obučavanje, smanjuje prekomjerno prilagođavanje modela podacima i povećava generalnost SVM modela, tj. njegove prediktivne performanse (Rogić & Kaščelan, 2020).

Na Slici 15 predstavljena je ilustracija linearno neodvojivih primjera i način na koji SVM algoritam kreira dodatnu dimenziju, kako bi omogućio linearnu separabilnost.



Slika 15. Ilustracija linearno neodvojivih klasa u dvodimenzionalnom prostoru

U osnovi, SVM se sastoji od dvije komponente – kernela i algoritma za optimizaciju. Kernel dijeli nelinearne podatke u višedimenzionalnom prostoru i čini ih linearno odvojivim. Obučavanje modela se vrši u prostoru više dimenzije. Druga komponenta, optimizacioni algoritam, primjenjuje se za rješavanje problema optimizacije. Budući da se SVM zasniva na induktivnom principu strukturne minimizacije rizika, algoritam teži da minimizira gornju granicu generalizacione greške, koja se sastoji od zbira greške u obučavanju (eng. *training error*) i nivoa pouzdanosti na nepoznatim podacima (eng. *confidence level*). Ovo čini SVM superiornim u odnosu na uobičajeni princip empirijskog minimiziranja rizika (eng. *empirical risk minimization*), koji samo minimalizuje grešku u obučavanju. Na osnovu takvog principa indukcije, u literaturi se navodi da SVM obično postiže bolje performanse generalizacije od ostalih tehnika mašinskog učenja (Zhang et al., 2016).

Prednost SVM metode u odnosu, recimo, na neuronske mreže kod nelinearnih problema, potiče od geometrijske interpretacije i činjenice da je u suštini SVM metode problem konveksnog matematičkog programiranja koji garantuje globalni minimum. Naime, neuronske mreže imaju empirijsku minimizaciju greške, tj. greška

se koriguje svaki put kada neki primjer prođe kroz mrežu, što može dovesti do zapadanja u lokalni minimum. Za razliku od neuronskih mreža, SVM ima strukturnu minimizaciju (minimizacija rastojanja između margine i vektora oslonca), koja se svodi na minimum kvadratne konveksne forme uz odgovarajuće uslove, tako da SVM uvijek pronalazi globalni minimum. Modeli neuronske mreže su često previše prilagođeni podacima za obučavanje i imaju lošije rezultate na nepoznatom skupu podataka. Za razliku od njih, SVM ima izvanrednu moć generalizacije zahvaljujući mogućnosti širenja margine. Osim toga, neuronska mreža zahtijeva podešavanje velikog broja parametara, dok složenost modela u velikoj mjeri zavisi od broja prediktora. Složenost SVM modela zavisi isključivo od broja vektora oslonca, pa njegovo obučavanje zahtijeva manje kompjuterskog vremena i prostora bez obzira na broj prediktora. Dodatno, klasifikaciona hiperravan u SVM metodi zavisi isključivo od nekoliko tačaka podataka – vektora oslonca i samim tim je njegova robustnost jača od mnogih drugih metoda mašinskog učenja (Liu & Zio, 2019).

Rad sa SVM klasifikatorom uključuje proces balansiranja dva komplementarna cilja, što je često slučaj i s drugim algoritmima mašinskog učenja (Pisner & Schnyer, 2019):

1. Optimizacija tačnosti, odnosno maksimizacija procenta tačnih oznaka klase dodijeljenih novim instancama od strane klasifikatora;
2. Obezbjeđivanje da se klasifikator može generalizovati za nove podatke, tj. optimizacija njegove ponovljivosti.

Imajući sve navedeno u vidu, prednosti SVM algoritma koje se mogu istaći su (Coussement & Van den Poel, 2008; Kim et al., 2005; Rhys, 2020):

1. Efikasan je u pronalaženju načina razdvajanja linearno neodvojivih klasa;
2. Ima tendenciju da dobro radi u širokom spektru zadataka;
3. Potrebno je izabrati samo dva parametra – gornju granicu kažnjavanja greške na primjerima za obuku i kernel parametar;
4. Rješenje koje SVM daje je jedinstveno, optimalno i globalno, jer se njegova obuka vrši rješavanjem lenearno ograničenog kvadratnog problema;

5. Ne zahtijeva pretpostavke o distribuciji prediktorskih promjenljivih;
6. S obzirom na to da se zasniva na principu strukturnog minimiziranja rizika, ovaj tip klasifikatora minimizira gornju granicu stvarnog rizika u poređenju s drugim klasifikatorima, koji minimiziraju empirijski rizik.

U literaturi se, s druge strane, navode i određeni nedostaci:

1. Ovaj algoritam je jedan od računski najskupljih i najkompleksnijih za obuku;
2. Parametri koje je potrebno podesiti, moraju se podešavati istovremeno;
3. Može da obrađuje samo kontinuirane prediktorske promjenljive (međutim, u određenim slučajevima može se sprovesti kodiranje kategoričke varijable kao numeričke);
4. SVM je metoda "crne kutije" (eng. *black box method*), tj. karakteriše je nedostatak u pogledu nemogućnosti interpretacije rezultata.

Uzimajući u obzir posljednji nedostatak, a zbog značaja interpretiranja rezultata, u ovom empirijskom istraživanju će biti testirane mogućnosti hibridnog pristupa primjene SVM metode u kombinaciji sa DT metodom.

Tokom posljednjih godina, SVM metoda sve više dobija na značaju i ima sve veću primjenu u ekonomskim i marketing istraživanjima. Iako se često u literaturi porede performanse SVM metode i neuronskih mreža, određeni broj istraživanja naglašava prednost prve metode. Naime, *Shin* i *Cho* (2006) su primijenili SVM za predikciju odgovora na kampanju, kako bi prevazišli nedostatke neuronskih mreža. Slično istraživanje sproveli su i *Mandapaka et al.* (2014), koji su razvili nekoliko prediktivnih modela koristeći DT, neuronske mreže, linearnu regresiju, kao i SVM, u cilju predviđanja odgovora na ponudu od strane potrošača i utvrdili superiornost SVM metode. Dodatno, *Gordini* i *Veglio* (2017) su u svom radu testirali sposobnost predviđanja SVM modela u cilju predikcije napuštanja kupaca. Autori su koristili bazu podataka od 80.000 kupaca italijanske kompanije za B2B trgovinu, a rezultate svog modela uporedili su s performansama neuronskih mreža i logističke regresije. Rezultati njihovog istraživanja su pokazali da je SVM nadmašila ostale metode u

pogledu stope tačnosti predviđanja, i to +4,57% u odnosu na neuronske mreže i +5,87% u odnosu na logističku regresiju.

U studijama iz oblasti marketinga, neki autori su SVM klasifikator koristili i u kombinaciji s drugim algoritmima mašinskog učenja. Na primjer, *Lawi et al.* (2018) primijenili su SVM u kombinaciji sa *AdaBoost* algoritmom u cilju klasifikacije potencijalnih klijenata za direktni marketing banke, koristeći bazu podataka iz repozitorijuma Univerziteta u Kaliforniji, *Irvine* (UCI). Njihovi rezultati su pokazali da je ovom kombinacijom ostvaren bolji rezultat u odnosu na standardni SVM klasifikator – senzitivnost je povećana sa 83,8% na 91,65%.

Baesens et al. (2003) su istražili performanse različitih klasifikacionih tehnika za procjenu kreditnog rejtinga i na osnovu rezultata su istakli da je, uz model neuronskih mreža, *Radial Basis Function Least Square SVM (RBF LS-SVM)* pokazao veoma dobre performanse. Dodatno, u cilju unapređenja modela odgovora na kampanju, i *Govindarajan* (2016) je u svom radu primijenio RBF-SVM klasifikator. Rezultati ovog istraživanja ukazuju da je hibridni RBF-SVM pokazao bolju prediktivnu tačnost u odnosu na individualni SVM klasifikator.

Koristeći RFM varijable, *Kim et al.* (2013) su primijenili SVM metodu za kreiranje modela odgovora na kampanju. Svoj model su primijenili na tri skupa podataka iz direktnog marketinga, koje je karakterisala različita stopa nebalansiranosti klasa. Dodatno, za smanjenje stepena nebalansiranosti u podacima, autori su primijenili i metodu slučajnog poduzorkovanja. Rezultati ovog rada su pokazali da nivo poduzorkovanja može ili pozitivno ili negativno uticati na performanse modela, pri čemu se 30% poduzorkovanje pokazalo kao bolja opcija od 50% poduzorkovanja. Naime, 50% poduzorkovanje bi dovelo do uklanjanja potencijalno važnih podataka o velikoj klasi u veoma nebalansiranim skupovima podataka. Ovo istraživanje je od posebnog značaja za empirijski dio ovog rada, s obzirom na to da se odnosi na primjenu SVM metode u direktnom marketingu, i to primjenom na podacima s visokom stopom nebalansiranosti klasa. Kao što je pomenuto u sekciji 3.1.2, ovaj problem karakteriše najveći broj baza podataka u direktnom marketingu, s obzirom na činjenicu da je broj lica koja odgovore na kampanju značajno manji od ukupnog

broja plasiranih ponuda. Upravo su autori *Kim et al.* (2013) u jednoj od svojih preporuka naglasili važnost uključivanja većeg broja varijabli (pored RFM), što će biti sprovedeno u empirijskom dijelu ovog rada. U okviru sekcije 5.4 biće testiran predloženi konceptualni model za rješavanje problema nebalansiranosti klasa, takođe primjenom SVM metode, dok će u poglavlju 4.1.5 biti detaljnije opisan ovaj problem, kao i prethodna istraživanja koja su ga tretirala.

U narednom dijelu rada biće opisane osnovne ideje *ensemble* metoda, koje će biti korišćene u empirijskom dijelu ovog rada, uz pregled dosadašnjih istraživanja iz oblasti marketinga u kojima su ove metode primijenjene.

4.1.4 Ensemble metode

Koncept simultane upotrebe više modela za predikciju (eng. *ensemble* modela) razvio se tokom devedesetih godina prošlog vijeka, kada je posebno dobio na značaju u oblastima mašinskog učenja, *data mininga* i statistike. Ove discipline su s različitih aspekata počele da istražuju *ensemble* metode, koje će biti opisane u ovom dijelu rada. U sekciji 4.1.2, pri opisivanju DT metode, pomenuto je da ona ima određene nedostatke, poput nedostatka robustnosti i suboptimalnih performansi (Larivière & Van Den Poel, 2005), što je motivisalo kreiranje tehnika za optimizaciju ove metode, između ostalog – *ensemble* tehnika.

Ensemble metode koriste više modela kako bi obezbijedile bolje prediktivne performanse u odnosu na one koje se mogu dobiti iz bilo kojeg od pojedinačnih modela koji konstituišu konačni model. Drugim riječima, *ensemble* tehnike kombinuju više slabijih pojedinačnih algoritama u cilju stvaranja jakog algoritma. Kao osnovna ideja *ensemble* metoda ističe se generisanje većeg broja modela na podskupovima podataka koji su slučajni. Na ovaj način, kreiraju se uzorci sa vraćanjem - ponavljanjem (eng. *sampling with replacement*), što znači da isti podatak može da se uključi u podskup i kod narednog uzorkovanja. Smisao uzorkovanja s vraćanjem je učiniti ponovno uzorkovanje zaista slučajnim. Ako se izvrši bez ponavljanja, generisani uzorci će zavisiti od prethodnih i stoga neće biti slučajni.

Nakon kreiranja uzoraka, vrši se agregacija rezultata, tako što modeli glasaju, a kao konačan rezultat uzima se onaj s najvećim brojem glasova od strane modela.

Dakle, ključna ideja *ensemble* metoda je da se umjesto obučavanja jednog modela, obučava više modela, nakon čega se traži evaluacija od strane svih njih po pitanju predikcije za nove podatke. S tim u vezi, predikcija zasnova na glasovima većeg broja modela imaće manju varijansu u odnosu na predikciju jednog individualnog modela. U literaturi se *ensemble* metode grupišu u tri kategorije (Rhys, 2020):

1. *Bootstrap aggregating (Bagging),*
2. *Boosting,*
3. *Stacking.*

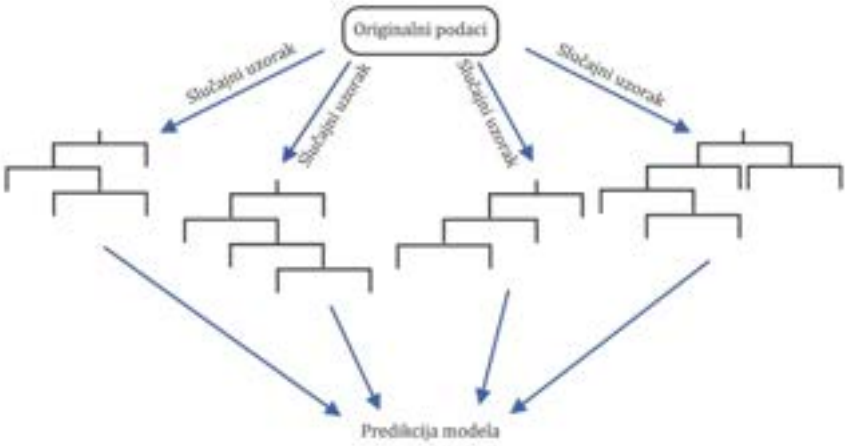
U cilju prevazilaženja problema s kojim se suočavaju algoritmi mašinskog učenja, poput osjetljivosti na šumove (eng. *noise*), koji su posljedica odstupanja (eng. *outliers*) i greške mjerenja, *Bootstrap Aggregating*, odnosno *Bagging* omogućava obučavanje modela koji koristi sve dostupne podatke u više iteracija. Na ovaj način se smanjuje vjerovatnoća da će model imati velika odstupanja prilikom predikcije na novim podacima. Dvije ključne ideje *bagging* algoritma su *bootstrapping* i agregacija, odakle je i dobio naziv. *Bootstrapping* (Efron & Tibshirani, 1993) predstavlja tehniku uzorkovanja, koja podrazumijeva da se podskupovi skupa podataka za obučavanje modela slučajno biraju korišćenjem modela s vraćanjem (isti primjer se može birati u novom uzorkovanju). Naime, s obzirom na skup podataka za obučavanje koji sadrži n broj primjera obuke, generisaće se uzorak od m primjera obuke uzorkovanjem s vraćanjem. Neki originalni primjeri će se pojaviti više puta, dok neki originalni primjeri uopšte neće biti zastupljeni u uzorku. Primjenom ovog procesa T puta, dobija se T poduzoraka sa m primjera obuke, a na svakom od T podskupova generiše se po jedan model. Kao što je prethodno navedeno, uzorkovanje s ponavljanjem učiniće ga zaista slučajnim uzorkovanjem. Agregacija rezultata vrši se sakupljanjem predviđanja iz pojedinačnih T modela, te se na taj način dobija finalno kombinovano predviđanje. Korišćenjem kombinacije predviđanja osnovnih, odnosno pojedinačnih modela, smanjuje se varijansa i izbjegava prekomjerno prilagođavanje modela podacima na kojima se obučava.

Optimizacija *bagging* modela uključuje izbor optimalnog broja pojedinačnih *T* modela.

Prediktivna procedura *bagging* modela odvija se na sljedeći način (Rhys, 2020):

1. Definisati broj pojedinačnih modela koje je potrebno obučiti;
2. Za svaki sub-model, korišćenjem uzorkovanja s ponavljanjem, izabrati uzorak skupa za obučavanje;
3. Obučiti pojedinačne sub-modele na svim kreiranim uzorcima;
4. Pustiti nove podatke kroz svaki pojedinačni model, na osnovu čega će modeli glasati o predikciji;
5. Kao konačnu predikciju uzeti onu za koju je glasalo najviše sub-modela.

Jedna od osnovnih prednosti i motiva za korišćenje *bagging* tehnike je upravo njena mogućnost da poboljša stabilnost i tačnost algoritama mašinskog učenja, koji se primjenjuju kako za klasifikaciju, tako i za regresiju. Najčešće se ova metoda primjenjuje u kombinaciji sa DT metodom, iako se može koristiti i sa drugim metodama. Na Slici 16 prikazana je ilustracija *bagging* DT modela.



Slika 16. Ilustracija *bagging* modela sa stablom odlučivanja (Rhys, 2020)

Jedna od najpopularnijih implementacija *bagging* metode sa DT metodom naziva se *Random Forest*, koja predstavlja *state-of-the-art ensemble* metodu. Ovu tehniku

razvio je *Breiman* (2001), a u njenoj osnovi je kombinacija *bootstraping* skupa podataka za obučavanje, slučajnog izbora prediktora i DT metode. *Random Forest* generiše „šumu" DT modela za svaki slučajno izabrani podskup podataka, a optimizacija podrazumijeva izbor optimalnog broja pojedinačnih DT modela. Ova tehnika može podržati i klasifikaciju i regresiju. Autori *Howard* i *Bowles* (2012) su istakli da RF predstavlja jedan od najuspješnijih algoritama opšte namjene u savremenom dobu.

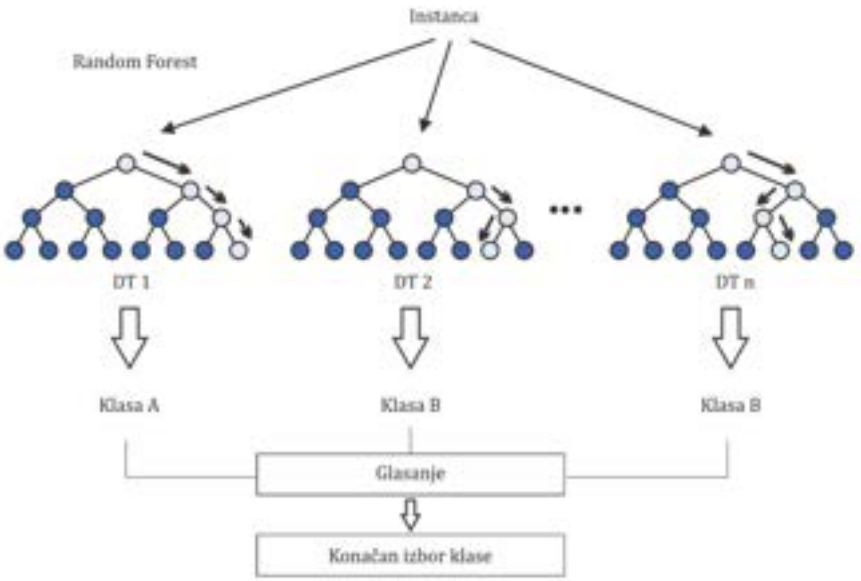
Dakle, RF algoritam koristi *bagging* za kreiranje većeg broja stabala odlučivanja u fazi obučavanja. S tim u vezi, on predstavlja ekstenziju bazične ideje o individualnim klasifikatorima tako što kreira veći broj klasifikacionih stabala odlučivanja. Ova stabla predstavljaju dio modela, a uključivanjem novih podataka u model, svako stablo kreira svoju predikciju. Na osnovu pojedinačnih predikcija, kao i u samom *bagging* procesu, bira se ona s najvećim brojem glasova. S tim u vezi, za klasifikaciju jednog primjera, svako stablo u šumi generiše odgovor, odnosno glasa za određenu klasu. Na osnovu prikupljenih glasova, model bira onu klasu koja je ostvarila najveći broj glasova od strane svih stabala. Tako će RF model, za razliku od individualnog DT modela, biti zaštićen od problema prekomjernog prilagođavanja modela podacima iz skupa za obučavanje i pružiće bolje performanse (Larivière & Van Den Poel, 2005). Međutim, za razliku od *bagging* metode, ovaj algoritam, pored uzorkovanja, uključuje još jednu dimenziju slučajnog izbora – slučajni izbor prediktora. Naime, na svakom čvoru određenog stabla, algoritam nasumično bira dio prediktorskih varijabli, koje će uzeti u obzir za tu podjelu. Na sljedećem čvoru, algoritam vrši još jedan slučajni izbor prediktorskih varijabli, koje će uzeti u obzir za narednu podjelu i tako dalje (Rhys, 2020). Na ovaj način, kroz slučajni izbor primjera za pojedinačne uzorke, kao i slučajni izbor varijabli kreiraju se stabla odlučivanja koja su nepovezana, odnosno nekorelirana. *Rhys* (2020) objašnjava značaj nepovezanih stabala činjenicom da će se, u slučaju postojanja varijabli koje u velikoj mjeri predviđaju ishod, upravo te varijable iskoristiti kao kriterijum za podjelu velikog broja stabala. S tim u vezi, stabla koja sadrže iste kriterijume za podjelu neće doprinijeti obezbjeđivanju dodatnih informacija, te je od posebnog

značaja slučajan izbor varijabli, koji će stvoriti različita stabla odlučivanja, a samim tim i različite prediktivne informacije.

Procedura razvijanja RF modela obuhvata sljedeće faze (Varian, 2014):

1. Bira se *bootstrap* uzorak primjera i počinje izgradnja stabla;
2. Na svakom čvoru se bira slučajan uzorak prediktora za podjelu (ne vrši se potkresivanje stabla);
3. Ovaj proces se ponavlja više puta - sve dok se ne kreira šuma stabala;
4. U cilju klasifikovanja novog primjera, vrši se predikcija od strane svakog stabla i koristi se izbor klase s najvećim brojem glasova.

Na Slici 17 predstavljena je ilustracija RF modela.



Slika 17. Ilustracija *Random Forest* modela

Kao jedna od savremenih *ensemble* metoda, *Random Forest* ima značajne prednosti u odnosu na pojedinačne klasifikatore (Breiman, 2001; Buckinx & Van den Poel, 2005; Coussement & Van den Poel, 2008):

1. Prediktivne performanse RF metode su među najboljim od svih raspoloživih prediktivnih tehnika;
2. Rezultat ovog klasifikatora je robustan i otporan na odstupanja i šum u podacima;
3. Ovaj klasifikator daje korisne interne procjene grešaka, snage, korelacije i značaja promjenljivih;
4. Ne zahtijeva značajno vrijeme za sprovođenje i značajnu računarsku snagu;
5. Jednostavan je za implementaciju, jer zahtijeva podešavanje samo dva parametra – broj slučajno izabranih prediktora i ukupan broj stabala koji se kreira.

Međutim, ovaj algoritam ima i jedan značajan nedostatak. Naime, kao i SVM algoritam, i RF se može okarakterisati kao „crna kutija" – ovaj algoritam ne omogućava uvid u veze koje postoje u podacima (Varian, 2014). Za razliku od individualnog stabla odlučivanja, koje može da pruži informacije o odnosu među prediktorima, cijela šuma stabala odlučivanja prilično je komplikovana za interpretaciju. S druge strane, RF nudi informaciju o skupu varijabli koje su važni prediktori, u smislu njihovog uticaja na poboljšanje prediktivne tačnosti.

Kao što je prethodno opisano, za kreiranje konačnog modela primjenom *bagging* metode, svi individualni modeli se obučavaju paralelno. S druge strane, druga kategorija *ensemble* metoda – *boosting*, obučava veći broj individualnih modela tako što ih kreira sekvencijalno. S tim u vezi, svaki naredni model teži da ispravi greške prethodne grupe modela. U okviru boosting tehnike razlikuju se dva algoritma: *Adaptive Boosting* (*AdaBoost)* i *Gradient Boosting*, pri čemu će prvi biti detaljnije opisan zbog značajnije primjene u istraživanjima iz oblasti marketinga. Obje tehnike teže da konvertuju slabi algoritam u jači algoritam, vodeći se idejom da će slabi algoritam dati bar malo bolje performanse od slučajnog pogađanja (Sabbeh, 2018). *AdaBoost* i *Gradient Boosting* razlikuju se u pogledu iterativne procedure u kojoj se kreiraju pomenuti slabi algoritmi.

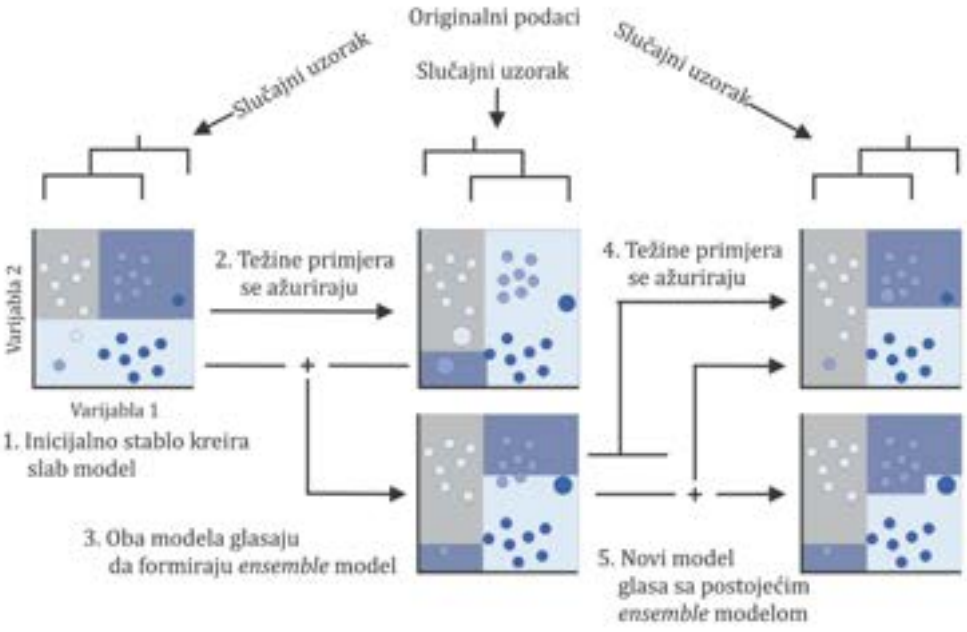
AdaBoost algoritam se može koristiti i za probleme klasifikacije i regresije. Ovaj klasifikator kombinuje više slabih klasifikatora da bi se dobio jaki klasifikator. Jedan

klasifikator može loše klasifikovati primjere, međutim, ako se kreira kombinacija većeg broja klasifikatora, uz odabir različitog skupa podataka za obučavanje pri svakoj iteraciji i dodjeljivanjem odgovarajuće težine podacima za konačno glasanje, može se dobiti dobra ocjena tačnosti za ukupni klasifikator.

Inicijalno, svi primjeri u skupu podataka za obučavanje imaju isti značaj, odnosno težinu. Početni model se obučava na *bootstrap* uzorku seta za obuku, gdje je vjerovatnoća da se primjer uzme u obzir proporcionalna njegovoj težini (u ovom trenutku svi jednaki). Primjerima koje ovaj početni model pogrešno klasifikuje dodeljuje se veća težina, tj. važnost, dok se slučajevima koji su pravilno klasifikovani, dodjeljuje manja težina. Dodatno, algoritam se ponavlja tako što se biraju podaci za obučavanje modela na osnovu tačnosti prethodnog skupa za obučavanje, a težina svakog obučenog klasifikatora u bilo kojoj iteraciji zavisi od postignute tačnosti u prethodnoj. Dodjeljivanje veće težine pogrešno klasifikovanom primjeru omogućava veću vjerovatnoću njegovog ponavljanja u narednom podskupu za obučavanje klasifikatora. Optimizacija *AdaBoost* modela podrazumijeva izbor optimalnog broja iteracija.

Dakle, *AdaBoost* filtrira primjere dodjeljivanjem veće težine problematičnim primjerima, kao i onim koje slabi algoritam nije tačno klasifikovao. Primarni fokus je razvoj novih slabih algoritama koji mogu da rade s primjerima koji su prethodno pogrešno klasifikovani. Ovaj proces se nastavlja tako što se dodaje novi model u *ensemble*, nakon čega slijedi glasanje svih modela, ažuriranje težina i izbor narednog uzorka podataka na osnovu novih težina (Rhys, 2020). Kada se dostigne maksimalan broj definisanih stabala, ovaj proces se završava i rezultira konačnim *ensemble* modelom. Nakon obučavanja, slabi algoritmi se dodaju jačem algoritmu na osnovu njihove alfa-težine, odnosno tačnosti. Pri tome, veća alfa-težina označava veći doprinos finalnom jakom algoritmu. Oznaka klase kojoj će primjer biti pridružen bazira se na kombinaciji izlaza svih slabih algoritama, ocijenjenih na osnovu njihove tačnosti (Sabbeh, 2018). Dakle, procedura konačne klasifikacije primjera biće gotovo identična kao pri korišćenju *bagging* metode – agregiranjem glasova.

Važno je istaći da u okviru *AdaBoost* algoritma i sami modeli takođe imaju težinu. Težina modela zavisi od broja grešaka koje model napravi – više grešaka uticaće na dobijanje manje težine. Ovdje se može uočiti razlika u odnosu na glasanje u okviru *bagging* metode. Naime, iako svaki pojedinačni slabi algoritam glasa individualno, kao i u *bagging* proceduri, u slučaju *AdaBoost* algoritma, težina svakog glasa zavisi od težine modela, odnosno njegove prethodne tačnosti. Na Slici 18 predstavljena je ilustracija *AdaBoost* DT modela.



Slika 18. Ilustracija *AdaBoost* modela sa stablom odlučivanja (Rhys, 2020)

Na prethodnoj slici se vidi da se inicijalni model obučava na slučajnom uzorku iz skupa podataka za obučavanje. Tačno klasifikovanim primjerima dodjeljuje se manja težina, koja je ilustrovana veličinom tačke podataka. Nadalje, vjerovatnoća narednog biranja uzorka direktno zavisi od težine primjera. Kako se dodaju nova stabla odlučivanja, vrši se glasanje u cilju formiranja *ensemble* modela, a na osnovu tih predikcija se ažuriraju težine instanci u svakoj iteraciji (Rhys, 2020).

Za razliku od *Adaptive Boosting* algoritma, gdje se primjeri različito ponderišu u zavisnosti od tačnosti njihove klasifikacije, *Gradient Boosting* modeli pokušavaju da

predvide reziduale prethodnog *ensemble* modela (Sabbeh, 2018). Dakle, ovaj algoritam daje značaj pogrešno klasifikovanim primjerima korišćenjem reziduala, odnosno greške reziduala, koja predstavlja razliku između stvarne vrijednosti i vrijednosti predviđene modelom. U svakoj iteraciji se računaju reziduali, na osnovu čega se prilagođava slabi algoritam, a doprinos slabog algoritma jakom predstavlja umanjivanje ukupne greške jakog algoritma. Konačno, *Gradient Boosting* algoritam bira onaj naredni model koji minimizira grešku reziduala prethodne grupe modela. Na ovaj način, minimiziranjem greške reziduala, sljedeći modeli će, zapravo, favorizovati ispravnu klasifikaciju slučajeva koji su prethodno bili pogrešno klasifikovani (Rhys, 2020).

Posljednja iz kategorije *ensemble* metoda, *Stacking* metoda, ne koristi se tako često kao *Bagging* i *Boosting*. Za razliku od *Bagging* i *Boosting* tehnika, gdje su korišćeni algoritmi za sve modele najčešće homogeni (kao npr. DT), *Stacking* metoda isključivo koristi različite algoritme za obučavanje sub-modela. Na primjer, ona može koristiti logističku regresiju, neuronske mreže i SVM algoritam za kreiranje tri nezavisna bazna modela. Osnovna ideja ove tehnike je kreiranje baznih modela koji će biti efikasni pri učenju različitih obrazaca u prostoru više dimenzije – jedan model može izvršiti efikasno predviđanje u jednom dijelu prostora, ali isto tako može praviti greške u drugom. Stoga, zadatak ostalih modela je da izvrše dobru predikciju u dijelu prostora u kome drugi modeli griješe. Na taj način, predikcije kreirane od strane baznih modela koriste se, uz originalne prediktore, kao prediktorske varijable *stacked* modela (Rhys, 2020). Dakle, *stacking* model uči iz predikcija kreiranih od strane baznih modela, pa je na taj način u mogućnosti da kreira svoju tačniju predikciju. Zbog svoje kompleksnosti, *Stacking* metoda još uvijek nema značajnu primjenu u ekonomskim i marketing studijama[3].

Bose i *Chen* (2009) navode da se u direktnom marketingu koriste prve dvije kategorije *ensemble* tehnika – *bagging* i *boosting*, s obzirom na to da najčešće pokazuju dobre performanse na podacima za obučavanje modela jer kombinuju više

[3] *Boosting* i *Stacking* tehnike prevazilaze okvir ovog rada i neće biti primijenjeni u empirijskom dijelu istraživanja. Detaljni opisi i matematičke formulacije svih *ensemble* metoda dati su u knjigama (Rhys, 2020; Zhou, 2012).

rješenja koja pojedinačno mogu pokazati određenu pristrasnost. Dodatno, *Lemmens* i *Croux* (2006) su pokazali da *bagging* i *boosting* unapređuju performanse klasičnih i tradicionalnih klasifikacionih modela za predikciju napuštanja potrošača smanjenjem klasifikacione greške i poboljšanjem prediktivne tačnosti, što potvrđuje upotrebu i značaj ovih tehnika u studijama iz oblasti marketinga. U empirijskom dijelu ovog rada biće primijenjena *Bagging* metoda, kao i *Random Forest*, u kombinaciji sa SVM klasifikatorom (sekcija 5.5).

Uzimajući u obzir da je nebalansiranost klasa jedan od istaknutih problema u skupovima podataka iz direktnog marketinga, *ensemble* tehnike sve više dobijaju na značaju u prevazilaženju ovog problema. Naime, prilikom balansiranja klasa u procesu obuke, broj primjera iz veće klase koji se uzima slučajnim izborom jednak je broju primjera iz minorne klase. S obzirom na to da se podaci iz veće klase biraju slučajnim uzorkovanjem u više iteracija, *ensemble* metode smanjuju potencijalni problem gubitka informacija, koje mogu biti veoma značajne za diferencijaciju između definisanih klasa. U prethodnim istraživanjima potvrđeno je da ove tehnike daju bolje rezultate od standardnih metoda poduzorkovanja ili preuzorkovanja (Galar et al., 2012; Miguéis et al., 2017).

Koristeći *Random Forest* tehniku, *Larivière* i *Van den Poel* (2005) su, u cilju procjene profitabilnosti i mogućnosti zadržavanja potrošača, istraživali ponašanje potrošača, kao i uočene sličnosti između njih, analizirajući bazu podataka od 100.000 kupaca, preuzetu iz skladišta podataka velike evropske kompanije koja se bavi finansijskim uslugama. Autori su primijenili dvije RF tehnike za analizu ovih podataka. Naime, za binarnu klasifikaciju koristili su klasičnu tehniku, dok su za modele s linearno zavisnim varijablama primijenili regresionu tehniku. Rezultati ovog istraživanja pokazali su da obje RF tehnike, u poređenju s tradicionalnim metodama linearne regresije i logističke regresije, pružaju bolje rezultate u uzorcima za testiranje i validaciju. Uz to, autori su istakli i jedan interesantan nalaz rada – isti skup varijabli imao je različit uticaj na kupovinu u poređenju sa uticajem na mogućnost napuštanja ili profitabilnost potrošača. Analizom varijabli, ustanovili su da upravo prethodno

ponašanje u kupovini ima najznačajniji uticaj na generisanje ponovljene kupovine, kao i na povoljan rezultat u pogledu profitabilnosti.

U cilju razvoja konceptualnog modela za targetiranje klijenata banke putem direktnog marketinga, *Ładyżyński et al.* (2019) su uporedili tri algoritma mašinskog učenja: klasifikacioni DT (CART), RF i *Deep Belief Network* (forma neuronskih mreža). Nakon testiranja svih modela, autori su odabrali klasifikacioni DT kao najbolji, sa odličnim rezultatom odziva klasa, srednjom preciznošću i najvećom efikasnošću u domenu računarskog procesiranja. Predloženi model uključio je podatke za analizu ponašanja potrošača u cilju procjene tipa personalizovanih marketing strategija koje se mogu usmjeriti konkretnom kupcu, kao i najpogodniji momenat za targetiranje odabranih kupaca. Iako je pristup testiran i validiran u finansijskoj instituciji, njegova primjena nije ograničena na banke, već ima opštu primjenu za upravljanje kampanjama direktnog marketinga.

U cilju predviđanja odgovora na kampanju direktnog marketinga, *Asare-Frempong* i *Jayabalan* (2017) su primijenili četiri klasifikatora: DT, RF, LR i MLPNN. Za ovu studiju, autori su koristi bazu podataka za direktni marketing banke. Rezultati njihovog istraživanja su ukazali na superiornost RF klasifikatora za predikciju, koji je ostvario tačnost od 87%. Pored procjene odgovora na kampanju, kao drugi cilj ove studije autori navode identifikovanje najznačajnijih atributa potrošača koji su prijavljeni i koji će s najvećom vjerovatnoćom pozitivno odgovoriti na buduće ponude banke. U tom cilju, primijenjena je klaster analiza, koja je pokazala da su neki od najznačajnijih pokazatelja za predikciju dužina telefonskog razgovora s potencijalnim klijentom, kao i stepen obrazovanja klijenta. Dakle, potencijalni klijenti s kojima je vođen duži telefonski razgovor, kao i oni s minimalno srednjim stepenom obrazovanja istakli su se kao najznačajniji prospekti za targetiranje u narednim kampanjama direktnog marketinga.

U cilju predikcije profitabilnosti potrošača, *D'Haen et al.* (2013) su istraživali koje *data mining* tehnike pokazuju najbolje performanse u kombinaciji s različitim izvorima podataka. *Data mining* tehnike koje su primijenili su logistička regresija, DT i *bagging* DT, dok je jedna baza podataka prikupljena putem *web mininga*,

besplatno, a druga je kupljena od specijalizovanog prodavca. Rezultati ovog istraživanja su pokazali da, nezavisno od izvora podataka, *bagging* u kombinaciji sa DT metodom, pruža najbolje performanse, koje se ogledaju kroz najveći AUC. Iako su *web* podaci obezbijedili bolje prediktivne performanse u odnosu na komercijalne podatke, najbolji rezultat je ostvaren kombinacijom oba skupa podataka.

Migueis i *Teixeira* (2020) su predložili model za predikciju dodavanja proizvoda u virtuelnu korpu, koji je baziran na *Random Forest* i tehnici logističke regresije. Naime, koristeći *clickstream* podatke iz kompanije koja se bavi e-trgovinom, izvršili su predviđanje da li će potencijalni kupac dodati određeni proizvod u svoju korpu, na osnovu podataka o prethodnim obrascima navigacije po e-prodavnici. Rezultati su pokazali da je *Random Forest* rezultirao boljim performansama u odnosu na logističku regresiju u pogledu svih metrika, a posebno u pogledu vrijednosti AUC, koji je za 2,6% veći korišćenjem ove tehnike. Ostvareni rezultati ukazuju na efikasnost korišćenja navedenih prediktivnih modela u ovom domenu, čime se može uticati na stvaranje personalizovanih strategija targetiranja kupaca određenim sistemima preporuke proizvoda na osnovu njihovog prethodnog *web* ponašanja.

Fang et al. (2016) su razvili model za predikciju profitabilnosti korisnika osiguranja, koristeći podatke iz osiguravajuće kuće iz Tajvana i RF regresiju, poredeći je sa drugim *data mining* tehnikama. Njihovi rezultati su pokazali da RF model može efikasno predvidjeti profitabilnost korisnika korišćenjem svega nekoliko varijabli. Na ovaj način, autori su podijelili klijente na klijente visoke, srednje i male vrijednosti. Poredeći RF sa ostalim modelima, autori su istakli da je ovaj model pokazao bolje performanse u odnosu na linearnu regresiju, DT, SVM i generalizovani *boosting* model. U pogledu podataka koji su bili najznačajniji za predikciju profitabilnosti, istakli su se: region klijenta, starost, status osiguranja i pol.

U cilju prevazilaženja problema nebalansiranosti klasa u direktnom marketingu, *Kang et al.* (2012) su razvili novu metodu za balansiranje podataka - CUE (*clustering, undersampling, ensemble*). Svojom metodom za predikciju odgovora na kampanju direktnog marketinga težili su da unaprijede performanse prethodnih metoda, koje karakteriše nebalansiranost klasa. Autori su u ovom radu primijenili četiri

klasifikatora: LR, MLP, k-NN i SVM, iskoristili su kombinaciju šest metoda za balansiranje podataka (bez reuzorkovanja, slučajno poduzorkovanje, jednostrana selekcija, slučajno preuzorkovanje, SMOTE i CUE) i kreirali su 24 modela za predikciju odgovora na kampanju, koje su procjenjivali na osnovu ostvarene stope odgovora i profitabilnosti. Na osnovu rezultata razvijenog pristupa, autori su istakli da predložena metoda balansiranja podataka (CUE) poboljšava prediktivnu tačnost, stabilizuje model i povećava profit kroz značajno povećanje prihoda ostvarenih u kampanji.

U prethodnim istraživanjima iz domena direktnog marketinga, *ensemble* metode su se pokazale kao efikasne za predikciju profitabilnosti kupaca, kao i za predikciju odgovora na kampanju, što u velikoj mjeri može unaprijediti sistem za selekciju i targetiranje kupaca u budućim aktivnostima direktnog marketinga.

Nakon opisa *ensemble* metoda, u narednom dijelu rada biće detaljnije opisan prethodno pomenuti problem i jedan od najčešćih problema u skupovima podataka iz direktnog marketinga – nebalansiranost klasa. Pored toga, biće predstavljena i dosadašnja istraživanja koja su tretirala ovaj problem, s posebnim naglaskom na radove u kojima je primijenjena SVM metoda, koja će u empirijskom dijelu ovog rada biti korišćena u cilju prevazilaženja nebalansiranosti klasa.

4.1.5 Problem nebalansiranosti klasa u direktnom marketingu i SVM balansiranje

Napredak u tehnologiji i nauci doprinio je stvaranju mogućnosti za otkrivanje znanja iz podataka i napredne analitike, što je pronašlo značajnu primjenu u različitim oblastima, od nacionalne i sajber (eng. *cyber*) bezbjednosti, medicinske dijagnostike, do olakšavanja i unapređivanja donošenja odluka u kompanijama i vladama. Međutim, tokom posljednjih godina, uz rast obima raspoloživih podataka, istakao se i jedan novi problem koji je karakterisao veliki broj baza podataka – neravnoteža klasa. Ovaj problem privukao je pažnju i akademske zajednice i privrede, uzimajući u obzir da značajno kompromituje performanse standardnih algoritama.

Naime, problem koji je postojao u tradicionalnoj direktnoj pošti – stopa konverzije, zapravo je i dalje prisutan i u eri digitalnog marketinga. Broj sesija koje uključuju obavljenu kupovinu je znatno manji od ukupnog broja sesija u vezi s kampanjom (Behera et al., 2020), što doprinosi stvaranju problema minorne klase. Problem neravnoteže klasa dovodi do pristrasnih rezultata prediktivne procedure s obzirom na to da nema dovoljno pozitivnih primjera za proces obučavanja modela. Ovo obično dovodi do loših performansi klasifikacije, jer bi model sve testne primjere klasifikovao kao negativne (Wang & Pineau, 2016). S druge strane, razvoj prediktivne analitike i društvenih medija, kao i sama količina dostupnih podataka, čini proces modeliranja odgovora kupaca preciznijim. Umjesto čistog menadžerskog prosuđivanja u odabiru targetiranih segmenata, donosioci odluka mogu da koriste podatke i analitiku kako bi mnogo efikasnije identifikovali svoje ispitanike, dok, ujedno, tretiraju pitanje neravnoteže klasa (Rogić, Kašćelan, & Pejić Bach, 2022).

Kao što je istaknuto u prethodnom poglavlju, kao i u sekciji 3.1.2, jedan od najvećih izazova kod razvoja prediktivnih modela u direktnom marketingu je upravo problem neravnoteže klasa (problem minorne klase). Prema *Pareto* principu, segment najvrednijih kupaca je najmanji i čini oko 20% kupaca, ali je i najvažniji za uspjeh kampanje. Stopa odgovora u direktnoj kampanji često je manja od 5%, dok oni koji ne odgovore čine čak 95% (ili više) od ukupnog broja kupaca. To dovodi do vrlo neuravnoteženih baza podataka za obuku prediktivnih klasifikatora u direktnom marketingu (Bose & Chen, 2009; Kang et al., 2012; Kim et al., 2013). Problem balansiranja klasa u ovoj oblasti veoma je aktuelan i u skladu s tim, veliki broj istraživanja tretira probleme i metode koje efikasno rješavaju ovaj problem. Aktuelnost ovog problema u akademskoj zajednici zasnovana je na činjenici da algoritmi mašinskog učenja obično imaju poteškoće u učenju iz podataka s neuravnoteženim klasama zbog težnje da minimiziraju ukupnu stopu greške (Vassiljeva et al., 2017). Kao rezultat, to bi moglo dovesti do pogrešne klasifikacije svih instanci, što rezultira lošim performansama klasifikacije (He & Garcia, 2009; Maldonado & López, 2014; Wang & Pineau, 2016). Kako bi se skup podataka okarakterisao kao nebalansiran, najčešće se uzima prag odnosa od 5:1 (ili veći). Pri opisivanju problema neravnoteže klasa, potrebno je formalno opisati koncept

nadgledane klasifikacije (eng. *supervised classification*). Naime, cilj klasifikacije u mašinskom učenju je da obuči sistem koji je sposoban za predikciju nepoznate klase prethodno nepoznatih primjera, uz dobru sposobnost generalizacije (Galar et al., 2012). S tim u vezi, klasa s najmanjim brojem instanci je obično klasa interesovanja sa stanovišta obučavanja modela, pa se u tom smislu ističe problem neravnoteže primjera.

Weiss (2004) navodi šest kategorija problema koji se javljaju kada se primjenjuju *data mining* algoritmi na podacima koje karakteriše nebalansiranost klasa:

1. Neadekvatne metrike za evaluaciju: često dolazi do problema korišćenja neadekvatnih metrika, što onemogućava objektivnu evaluaciju modela;
2. Nedostatak podataka – apsolutna rijetkost: broj podataka koji pripada minornoj klasi često je veoma mali u apsolutnom smislu, što otežava pronalaženje pravilnosti u okviru rijetke klase;
3. Nedostatak podataka – relativna rijetkost: primjeri iz minorne klase nisu rijetki u apsolutnom smislu, ali su rijetki u odnosu na primjere iz drugih klasa, što onemogućava primjenu određenih metoda;
4. Fragmentacija podataka: mnogi *data mining* algoritmi (poput DT algoritma) koriste pristup *podijeli pa osvoji* (eng. *divide-and-conquer*), gdje se izvorni problem razlaže na sve manje probleme, što rezultira dijeljenjem prostora podataka na sve manje djelove. S tim u vezi, nastaje problem, jer se pravilnosti tada mogu pronaći samo unutar svake pojedinačne particije, koja će sadržavati manje podataka;
5. Neodgovarajuća induktivna pristrasnost (eng. *inappropriate inductive bias*): generalizovanje iz konkretnih primjera ili indukcija zahtijeva vandokaznu pristrasnost (eng. *extra-evidentiary bias*). Bez takve pristrasnosti, „induktivni skokovi" nisu mogući i obučavanje se ne može sprovesti. Stoga je pristrasnost *data mining* sistema ključna za njegove performanse. Autor navodi da mnogi algoritmi koriste opštu pristrasnost, kako bi podstakli generalizaciju i izbjegli pretjerano prilagođavanje. Ova pristrasnost može negativno uticati na sposobnost obučavanja za rijetke slučajeve i rijetke klase;

6. Šum (eng. *noise*): šum u podacima će uticati na ponašanje bilo kog *data mining* sistema, međutim, važno je istaći da on ima veći uticaj na rijetke slučajeve nego na uobičajene slučajeve iz većih klasa.

Osim toga, nebalansirane skupove podataka obično karakteriše i preklapanje klasa, mala veličina uzorka ili nizak stepen razdvajanja, što dodatno otežava obučavanje klasifikatora (Galar et al., 2012; Japkowicz & Stephen, 2002). Ovi problemi mogu dovesti do zanemarivanja primjera manjinske klase tako što će ih tretirati kao šum u podacima. Na primjer, ukoliko je u skupu podataka odnos klasa 100:1, odnosno kada za svakih 100 primjera negativne klase postoji samo jedan primjer pozitivne klase, klasifikator može težiti maksimizaciji tačnosti pravila klasifikacije i ostvariti tačnost od 99%, tako što će jednostavno ignorisati pozitivne primjere i klasifikovati sve primjere kao negativne (Galar et al., 2012). Dakle, standardni algoritmi očekuju balansiranu distribuciju podataka i jednake troškove pogrešne klasifikacije, a kada su primijenjeni na kompleksnim i nebalansiranim podacima, oni rezultiraju nepovoljnom tačnošću za sve klase podataka (He & Garcia, 2009).

Važno je istaći da problem nebalansiranosti klasa nije ekskluzivan samo u direktnom marketingu, već generalno u mnogim klasifikacionim problemima iz stvarnog života, poput: dijagnostike greške (Zhu & Song, 2010), otkrivanja anomalija (Khreich et al., 2010), medicinske dijagnoze (Mazurowski et al., 2008), klasifikacije e-pošte (Bermejo et al., 2011), prepoznavanja lica (Huang et al., 2020), upravljanja rizikom (Huang et al., 2006) i slično.

Navedenim problemom su se bavili *Sun et al.* (2009), koji su istakli da su za tretiranje problema nebalansiranosti klasa u klasifikacionim modelima razvijena rješenja na nivou podataka, na nivou algoritama, kao i rješenja senzitivna na troškove (eng. *cost-sensitive)*. Na nivou podataka, cilj je uravnotežiti klase sa uzorkovanjem, koje može biti slučajno ili ciljano poduzorkovanje i preuzorkovanje. Na nivou algoritma, rješenjima se teži prilagoditi algoritam u cilju unapređenja obuke u malim klasama. Rješenja osjetljiva na troškove, kako na nivou podataka, tako i na nivou algoritma, pridružiće veće troškove pogrešne klasifikacije primjerima iz minorne klase. Tokom

posljednjih nekoliko godina, ovaj problem je obrađivan u nekoliko studija (Liu & Zio, 2019; Lopez-Garcia et al., 2019; Wong et al., 2020).

Osnovne metode uzorkovanja uključuju poduzorkovanje i preuzorkovanje. Poduzorkovanje eliminiše primjere većinske klase, dok preuzorkovanje, u svom najjednostavnijem obliku, duplira primjere manjinske klase. Obje tehnike uzorkovanja smanjuju ukupan nivo neravnoteže klasa, te na taj način rezultiraju manjom rijetkošću minorne klase. Međutim, ove metode uzorkovanja imaju određene nedostatke (Weiss, 2004). Poduzorkovanje odbacuje potencijalno korisne primjere većinske klase, što može potencijalno dovesti do pogoršanja performansi klasifikatora. S druge strane, budući da preuzorkovanje uvodi dodatne instance u skup podataka za obučavanje, ono može povećati vrijeme potrebno za izgradnju klasifikatora. Preuzorkovanje često uključuje kreiranje dodatnih kopija već postojećih instanci, te na taj način može dovesti do *overfitting* problema. Dakle, uprkos eliminaciji problema nebalansiranosti klasa, ovaj pristup ima značajna ograničenja i nedostatke, među kojima se, pored navedenih, mogu istaći i sljedeći: nepoznata distribucija optimalne klase, nedefinisan kriterijum za izbor primjera koji će se ukloniti, kao i rizik gubljenja informacija relevatnih za diferenciranje klasa pri poduzorkovanju većih klasa.

Pristupi koji ovaj problem tretiraju na nivou algoritama zahtijevaju ekstenzivno poznavanje algoritama i oblasti njihove primjene, dok troškovno osjetljiva rješenja uključuju dodatne troškove za obučavanje u cilju istraživanja efikasnih postavki troškova, kada stvarne vrijednosti troškova nisu dostupne.

Međutim, uprkos navedenim nedostacima, ova rješenja se koriste u skorašnjim istraživanjima iz oblasti direktnog marketinga. S tim u vezi, *Kim et al.* (2013) su poredili efikasnost SVM klasifikatora sa drvetom odlučivanja i neuronskim mrežama na veoma nebalansiranim setovima podataka u direktnom marketingu. Njihovi rezultati su pokazali da jedino SVM u potpunosti ne griješi pri klasifikaciji minorne klase, ali da je senzitivnost veoma niska, što znači da je problem nebalansiranosti klasa prisutan i kod primjene SVM metode. Korišćenjem slučajnog poduzorkovanja velike klase od 33% (odnos klasa 2:1), svi klasifikatori su poboljšali

svoje performanse, dok je SVM i dalje bio bolji od ostalih. Međutim, sa odnosom klasa 1:1, performanse SVM modela su oslabile, što sugeriše da se uklanjanjem velikog broja primjera iz velike klase gube relevantni podaci za proces obučavanja modela.

Pored pomenuta tri osnovna rješenja, važno je istaći i dodatno – korišćenje *ensemble* metoda. Određene kombinacije *ensemble* metoda s drugim tehnikama za prevazilaženje problema nebalansiranosti klasa već su predložene u dosadašnjim istraživanjima, koja su rezultirala veoma pozitivnim rezultatima (Galar et al., 2012). Na primjer, poduzorkovanje se može kombinovati sa *ensemble* tehnikama, tako da se slučajni odabir podataka iz veće klase ponovi nekoliko puta i smanji vjerovatnoća da se značajni podaci potpuno izuzmu, pa neki radovi koji se bave problemom neravnoteže klasa u direktnom marketingu idu u tom pravcu.

Na primjer, *Kang et al.* (2012) su u svom radu sugerisali da se modeli za predviđanje odgovora na kampanju mogu unaprijediti balansiranjem klasa korišćenjem klasterzacije, poduzorkovanja i *ensemble* metodama. Prvo se klasterizuju instance koje pripadaju klasi nerespondenata. U narednom koraku sprovodi se poduzorkovanje, kao dio *ensemble* procedure, tako što se slučajnim izborom selektuje određeni broj predstavnika svih klastera, proporcionalno veličini klastera, tako da ukupan broj selektovanih instanci bude jednak minornoj klasi (eng. *balanced ensemble*). Na taj način se postiže uzimanje određenog broja reprezentativnih članova veće klase i smanjuje gubitak informacija relevantnih za diferencijaciju klasa. Izvođenjem *ensemble* postupka u k iteracija, na k takvih uravnoteženih uzoraka generiše se k klasifikatora i kombinuju se njihova predviđanja. Rezultati su pokazali da u poređenju s metodama slučajnog uzorkovanja, ovaj pristup ima stabilnije prediktivne performanse kojima donosioci odluka mogu više vjerovati (Kang et al., 2012).

Migueis et al. (2017) uporedili su *ensemble* uravnoteženo poduzorkovanje (algoritam *EasyEnsemble* koji koristi uzorkovanje bez vraćanja) sa SMOTE metodom preuzorkovanja (*Synthetic Minority Oversampling Technique*) za predviđanje odgovora na kampanju direktnog marketiga u bankarstvu. Njihovi rezultati su

pokazali da je metoda *EasyEnsemble* dala bolje rezultate. Međutim, model uzorkovanja bez vraćanja može ugroziti nezavisnost klasifikatora u *ensemble* postupku, jer uzorkovanje u sljedećem koraku zavisi od onog napravljenog u prethodnom koraku. *Marinakos* i *Daskalaki* (2017) su testirali tehnike poduzorkovanja zasnovane na klasterima i reuzorkovanje na osnovu udaljenosti (eng. *distance-based resampling*), za model odgovora na kampanju od strane klijenata banke (sa 12% respondenata i 88% nerespondenata) s nekoliko različitih klasifikatora, kao što je linearna diskriminantna analiza (eng. *Linear Discriminant Analysis* - LDA), LR, k-NN, DT, NN i SVM. Najveća tačnost klasifikacije manjinske klase postignuta je kombinacijom poduzorkovanja klastera i k-NN.

U cilju predikcije namjere za kupovinu, *Kurniawan et al.* (2020) su u svom radu predložili metodu, koja uključuje *Random Under-Sampling* (RUS) i *Synthetic Minority Over-sampling Technique* (SMOTE). Dakle, ovaj pristup predikciji namjere za kupovinu uključuje i rješavanje problema nebalansiranosti klasa za klasifikaciju, što su autori obrazložili navođenjem činjenice da su podaci generalno veoma nebalansirani, što značajno utiče na smanjenje efikasnosti algoritama mašinskog učenja. Autori su koristili javno dostupan skup podataka sa Univerziteta u Kaliforniji, *Irvine* (UCI) repozitorijuma za mašinsko učenje. Od sprovedenih eksperimenata, kombinacija *SMOTE+AdaBoost+Random Forest* je pokazala najbolje performanse, sa AUC rezultatom od 0,960.

Peng et al. (2010) predložili su rješenje zasnovano na adaptaciji algoritma u obliku troškovno osjetljivog obučavanja SVM modela za segmentaciju korisnika kreditnih kartica. Autori su pokazali da ovo rješenje daje bolje rezultate za najmanju klasu korisnika, tj. korisnika najveće vrijednosti, od osnovnog SVM modela sa slučajnim poduzorkovanjem. Ovaj pristup zahtijeva opsežno znanje o SVM metodi kako bi se uključili pogrešno klasifikovani troškovi.

Na osnovu sprovedenih eksperimenata, *Japkowicz* i *Stephen* (2002) donijeli su nekoliko zaključaka o problemu neravnoteže klasa. Okarakterisali su ga kao relativni problem, koji zavisi od nekoliko faktora: stepena neravnoteže klasa, složenosti koncepta predstavljenog u podacima, ukupne veličine skupa podataka za

obučavanje klasifikatora i izbora samog klasifikatora. Preciznije, autori su istakli da što je veći stepen neravnoteže klasa, što je koncept složeniji i što je manja ukupna veličina skupa za obučavanje modela, veći će biti efekat problema nebalansiranosti klasa. Ovo se objašnjava činjenicom da navedeni faktori dovode do pojave izuzetno malih klastera, koji se ne mogu tačno klasifikovati. Međutim, problem neravnoteže klasa ne nanosi nikakvu štetu kada svi (minorni) klasteri imaju razumnu veličinu, čime se odbacuje vjerovanje da će se greške klasifikacije nužno pojaviti ako je jedna klasa predstavljena velikim skupom podataka, a druga malim (Japkowicz & Stephen, 2002). Dodatno, u pogledu osjetljivosti klasifikatora na problem neravnoteže u podacima, autori su zaključili da je C5.0 DT najosjetljiviji, MLP je u sredini, dok je SVM pokazao najbolji rezultat i najmanju osjetljivost na problem. Rezultati eksperimenata koji su kombinovali preuzorkovanje/poduzorkovanje i SVM pokazali su da slučajno preuzorkovanje ne poboljšava performanse SVM modela, dok poduzorkovanje čak pogoršava njegove performanse. Uzimajući sve ovo u obzir, autori predlažu SVM klasifikator za tretiranje problema nebalansiranosti klasa.

Raskutti i *Kowalczyk* (2004) su primijenili nekoliko strategija za rješenje navedenog problema korišćenjem SVM metode, pri čemu su poseban fokus stavili na situaciju kada se jedna klasa (negativnih primjera) u potpunosti ignoriše, a model se obučava isključivo koristeći pozitivne primjere. Ovaj pristup primijenili su na skupu podataka koji karakteriše značajna neravnoteža klasa, s obzirom na to da su pozitivni primjeri bili zastupljeni sa svega 3%. Njihovi rezultati su ukazali na to da obučavanje s negativnim primjerima u određenim situacijama urušava performanse SVM modela. Autori su ovaj pristup (eng. *one-class learning*) preporučili za primjenu na veoma nebalansiranim setovima podataka, sa značajnim prisustvom šuma i visoke dimenzionalnosti.

Izbor SVM klasifikatora za rješavanje problema neravnoteže klasa može se opisati i činjenicom da ovaj algoritam u procesu obučavanja modela uzima u obzir one primjere koji se nalaze najbliže margini, tj. vektore oslonca. S tim u vezi, na performanse SVM algoritma neće uticati negativni primjeri koji se nalaze daleko od margine, čak i ako postoje u velikom broju (Akbani et al., 2004).

I *Farquad* i *Bose* (2012) su testirali SVM kao pretprocesor za balansiranje klasa podataka klijenata osiguranja. Ukazali su na to da se dobija mnogo veća senzitivnost, tj. broj stvarnih primjera manje klase koje model tačno klasifikuje, kada se klasifikatori primjenjuju na prečišćenom skupu. Takođe su otkrili da je balansiranje podataka SVM metodom efikasnije od ostalih tehnika balansiranja, kao što je 100% i 200% SMOTE preuzorkovanje ili 25% i 50% poduzorkovanje.

U preliminarnom istraživanju, *Rogić* i *Kaščelan* (2019) testirali su s kojom stopom uspješnosti hibridni model koji kombinuje SVM i drvo odlučivanja za ekstrakciju pravila (SVM-DT) rješava problem minorne klase najvrednijih kupaca. Rezultati su pokazali da se ovim pristupom segment najvrednijih kupaca može predvidjeti s tačnošću od 77%, što je za 44% bolje od samostalnog DT modela. Dakle, SVM je, kao pretprocesor, efikasno povećao preciznost predviđanja minorne klase. Poboljšanje je još značajnije za procenat postojećih kupaca koji su prepoznati kao članovi najvrednijeg klastera. Samostalni DT identifikovao ih je samo 4%, dok je SVM-DT model uspio da identifikuje 63% takvih kupaca. Iako se model dobro pokazao na skupu podataka za obučavanje, on nije testiran na nepoznatom skupu podataka, pa njegova stvarna prediktivna snaga nije bila potvrđena u ovoj preliminarnoj studiji. Dodatno, u svom radu, *Đurišić et al.* (2020) testirali su kako SVM obavlja pretprocesiranje podataka i optimizuje CRM u bankama. Rezultati su pokazali da pri segmentaciji korisnika kreditnih kartica, ova metoda uspješno rješava problem preklapanja i nebalansiranosti klasa. Međutim, kao i u prethodno pomenutom radu, ni u ovom radu model nije bio testiran, odnosno validiran na potpuno nepoznatom skupu podataka.

U prethodnim istraživanjima, da bi se prevazišao problem nebalansiranosti klasa, uglavnom su testirane balansirane *ensemble* metode u kombinaciji s različitim klasifikatorima ili samostalni SVM, kao pretprocesor koji rješava problem preklapanja klasa i na taj način balansira podatke. U ovoj disertaciji će se testirati

kombinovanje *ensemble* pristupa i SVM metode za poboljšanje performansi prediktivnih modela odlučivanja u direktnom marketingu[4].

4.1.6 SVM Rule Extraction metoda

Tokom posljednje dvije decenije, veliki broj autora je, korišćenjem SVM klasifikatora, demonstrirao superiorne performanse generalizacije u odnosu na bilo koju drugi tehniku klasifikacije. Međutim, SVM nema mogućnost interpretacije i objašnjenja rezultata. Upravo taj nedostatak, koji karakteriše i SVM i neuronske mreže kao metode „crne kutije", predstavlja jednu od osnovnih prepreka koja sprečava njihovu praktičnu primjenu. Opisivanje i interpretacija rezultata posebno su značajne kada je u pitanju poslovna primjena ovih algoritama, kao i primjena u medicini. Zbog toga su uvedene tehnike za ekstrakciju pravila prvo iz neuronskih mreža, a kasnije i iz SVM modela, kako bi se ublažio ovaj problem i kako bi se pomoglo u objašnjenju klasifikacionih odluka modela (Barakat & Bradley, 2010; Diederich, 2008). Konačni cilj ekstrakcije pravila je generisanje pravila iz modela, a ne direktno iz podataka.

Autori *Martens et al.* (2008) navode dva najznačajnija razloga za generisanje pravila iz zatvorenih modela:

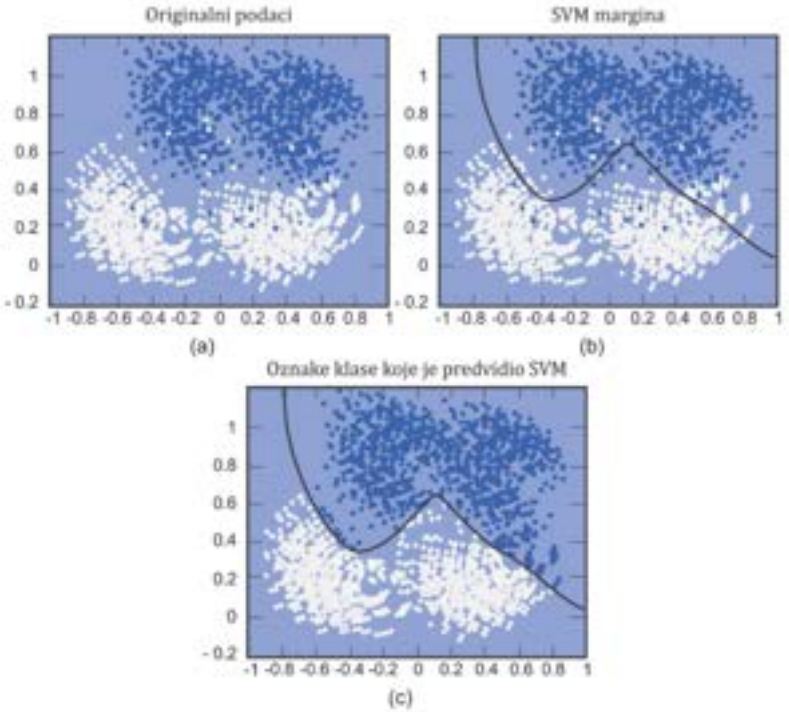
1. Razumijevanje klasifikacije obavljene u okviru modela (otvaranje „crne kutije");
2. Poboljšanje performansi tehnika za generisanje pravila, kroz uklanjanje specifičnih razlika u podacima.

Prvi navedeni razlog ujedno predstavlja i najčešći motiv za korišćenje tehnika za ekstrakciju pravila - dobijanje skupa pravila koja mogu objasniti model „crne kutije". Naime, dobijanjem skupa pravila koja oponašaju predviđanja SVM modela, stiče se određeni uvid u njegovo logičko funkcionisanje. Procjena kvaliteta generisanih

[4] Opisivanje različitih pristupa za rješavanje problema nebalansiranosti klasa prevazilazi okvir ovog rada. Detaljniji opis i kritički osvrt na prirodu problema, kao najsavremenije tehnologije, dali su *He* i *García* (2009), koji su u svom preglednom istraživanju dali sveobuhvatan prikaz razvoja istraživanja o učenju iz neuravnoteženih podataka

pravila vrši se metrikom vjernosti (eng. *fidelity*), koja pokazuje procenat primjera iz skupa podataka za testiranje modela za koje se SVM i skup pravila podudaraju u pogledu oznake klase (Martens et al., 2008). S tim u vezi, nakon dobijanja informacija o generisanim pravilima i mjeri vjernosti, donosioci odluka mogu procijeniti da li će prihvatiti i koristiti SVM kao model za podršku odlučivanju.

Kada je u pitanju drugi navedeni razlog za sprovođenje ekstrakcije pravila, autori navode zapažanje da se, generalno, nelinearni modeli sa odličnim performansama mogu koristiti uz pretprocesiranje podataka, odnosno njihovu prethodnu obradu i čišćenje. Na ovaj način, može se izvršiti zamjena oznake klase u podacima onom koja je predviđena od strane SVM modela, čime se uklanja šum i preklapanje klasa iz podataka. Na Slici 19 predstavljen je sintetički skup podataka (Ripley, 1994) s dvije varijable i dvije klase, s visokim stepenom preklapanja.



Slika 19. Ilustracija zamjene oznake klase u skupu podataka pomoću SVM pretprocesiranja (Martens et al., 2008)

U prvom dijelu slike (a) predstavljen je originalni skup podataka, gdje se može uočiti preklapanje sivih i plavih oznaka klase. Drugi dio slike ilustruje SVM granicu odlučivanja sa osnovnim skupom podataka. Konačno, u trećem dijelu slike (c), oznake klase podataka su zamijenjene sa oznakama klase predviđenim od strane SVM modela – svi primjeri iznad SVM granice postaju plave tačke, a svi primjeri koji se nalaze ispod postaju sive. Iz ove ilustracije može se zaključiti da u podacima nema više šuma i preklapanja klasa, te da je konkretan SVM model riješio ovaj problem.

Barakat i *Bradley* (2010) su metode za ekstrakciju pravila iz SVM modela grupisali u dvije kategorije: one zasnovane na komponentama modela SVM i one koje ne koriste unutrašnju strukturu SVM modela, već crpe pravila iz SVM izlaza. Kada se tumači SVM model ili se SVM koristi kao pretprocesor podataka, autori preporučuju tehnike iz druge grupe jer pružaju razumljivija pravila. Dakle, tehnike iz ove kategorije pristupaju SVM algoritmu kao „crnoj kutiji" i generišu pravila koja opisuju vezu između ulaza i izlaza modela (Barakat & Bradley, 2010; He et al., 2006; Martens et al., 2008). Osnovna ideja ove tehnike je kreiranje vještačkih primjera sa oznakom klase, pri čemu se ciljana oznaka klase u skupu podataka za obučavanje zamijeni sa oznakom klase predviđenom od strane SVM modela. Zatim, vještački skup podataka se koristi s drugom metodom mašinskog učenja, koja ima eksplanatorne mogućnosti, poput DT metode. Dakle, cilj ovog procesa je da se modelira izlaz originalnog klasifikatora korišćenjem drugog klasifikatora, koji ima mogućnosti interpretacije i razumljivosti. Slijedeći preporuke autora (Barakat & Bradley, 2010), u empirijskom dijelu ovog rada biće primijenjena ekstrakcija pravila iz SVM izlaza korišćenjem DT metode (Rogić & Kaščelan, 2020).

Pomenuti pristup je do sada imao primjenu u oblastima poput bioinformatike (He et al., 2006) i procjene kreditnog rizika (Martens et al., 2007), međutim, prema saznanjima autora ove disertacije, još uvijek nije primijenjen za balansiranje klasa u studijama iz oblasti marketinga. Autori *He et al.* (2006) su predstavili pristup koji kombinuje SVM sa stablom odlučivanja u novi algoritam, nazvan SVM_DT, koji se sastoji od tri osnovna koraka. Prvo, ovaj algoritam obučava SVM, a zatim se, u drugom koraku, novi skup podataka za obuku generiše odabirom iz SVM izlaza. Ovaj

novi skup podataka za obučavanje DT modela biće kvalitetniji od originalnog, imajući u vidu ostvarene koristi od SVM modela. Konačno, dobijeni skup se koristi za obuku sistema učenja za stablo odlučivanja i za izdvajanje odgovarajućih skupova pravila. Autori su istakli da je razumljivost SVM_DT modela mnogo bolja od samostalnog SVM modela. Osim toga, sposobnost generalizacije SVM_DT je bolja od sposobnosti stabala odlučivanja C4.5 i slična je onoj kod SVM modela. I autori *Martens et al.* (2007) su za procjenu kreditnog rizika predložili sličan model, ističući da je SVM klasifikator, kao kompleksna matematička funkcija, prilično nerazumljiva, što sprečava primjenu ovog efikasnog algoritma u stvarnom životu. U cilju prevazilaženja tog problema, predložili su generisanje pravila iz SVM modela, uz zadržavanje što je moguće većeg stepena tačnosti, koji je obezbijeđen kroz SVM. Korišćenje SVM izlaza, umjesto originalnih skupova, stvara čistiji skup podataka. Na osnovu sprovedenih eksperimenata, autori navode da pravila koja generiše DT za podatke sa oznakama klase predviđenim od strane SVM modela čak nadmašuju DT pravila, koja su rezultat skupa podataka sa stvarnim oznakama klasa. S tim u vezi, autori predlažu primjenu ovog pristupa u situacijama kada su donosiocima odluka potrebni i tačnost i razumljivost (Martens et al., 2007), uzimajući u obzir da se upravo taj *trade-off* u literaturi navodi kao jedan od izazova prilikom izbora tehnika za ekstrakciju pravila iz SVM modela (Barakat & Bradley, 2010; Huysmans et al., 2006).

Pomenutom procedurom najčešće se generišu *if-then* pravila[5]. Dio generisanih pravila koji se odnosi na uslov predstavlja logičku kombinaciju uslova vezanih za nezavisne promjenljive. Ovaj dio uslova može sadržati konjukcije, disjunkcije i negacije, ali većina algoritama će generisati pravila koja sadrže samo konjukcije. Primjer takvog pravila je: ako je $X=a$ i $Y=b$, onda je klasa=1, pri čemu su nezavisne promjenljive X i Y, s mogućim vrijednostima a i b. Za neprekidne ulazne promjenljive, uslovi se obično navode kao ograničenja dozvoljenih vrijednosti, poput: „$X \in [c_1, c_2]$" ili „$X > c_3$", pri čemu $c_1, c_2, c_3 \in R$. Većina algoritama će osigurati da djelovi svakog pravila koji se odnose na uslove razgraniče zasebna područja u

[5] Ostale, manje korišćene tipove generisanih pravila su u svom radu opisali *Huysmans et al.* (2006).

ulaznom prostoru, tj. da pravila budu međusobno isključiva. Prema tome, samo jedno pravilo može biti zadovoljeno kada se predstavi novi primjer i to pravilo biće jedino koje će se koristiti za donošenje odluke o klasifikaciji ili regresiji (Huysmans et al., 2006).

Za evaluaciju performansi generisanih pravila, u literaturi se navodi pet mjera (Craven & Shavlik, 1995):

1. Razumljivost (eng. *comprehensibility*) - stepen razumljivosti generisanih pravila;
2. Vjernost (eng. *fidelity*) - mjera u kojoj izdvojena pravila modeliraju model „crne kutije" iz kojeg su generisani (procenat primjera za koje se primjeri predviđeni pravilima poklapaju sa SVM oznakom (eng. *label*));
3. Tačnost (eng. *accuracy*) - sposobnost izdvojenih pravila da daju tačna predviđanja za ranije neviđene slučajeve (procenat primjera za koje se primjeri predviđeni pravilima poklapaju sa originalnom oznakom (eng. *label*));
4. Skalabilnost (eng. *scalability*) - sposobnost metode da se proširi na druge modele s velikim ulaznim prostorom i velikim brojem podataka;
5. Generalnost (eng. *generality*) - u kojoj mjeri metoda zahtijeva posebne režime obuke ili ograničenja u arhitekturi modela.

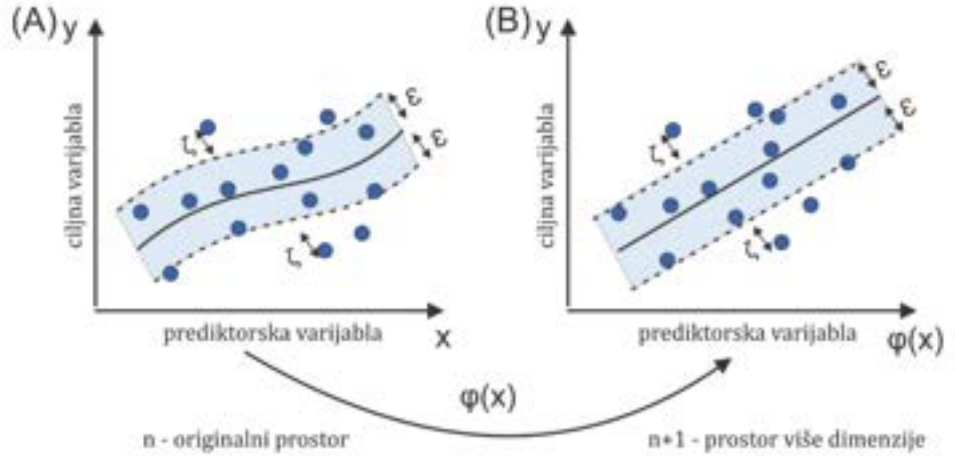
Autori ističu da se posljednje dvije mjere performansi rijetko koriste, uzimajući u obzir nemogućnost njihovog kvantifikovanja. Međutim, u slučaju ekstrakcije pravila iz SVM modela, skalabilnost ostaje važan aspekt, uzimajući u obzir da SVM daje dobre rezultate kada se primijeni na bazama podataka velikih dimenzija.

U prethodnim djelovima ovog poglavlja istaknute su karakteristike SVM metoda, s posebnim fokusom na klasifikaciju. Međutim, kako je i prethodno istaknuto, SVM se može koristiti i u regresionim analizama. U narednom dijelu rada biće predstavljena ideja *Support Vector* regresije, kao i jedan od najznačajnijih izazova u regresionim analizama – problem asimetrične distribucije podataka, koji ova metoda efikasno rješava.

4.1.7 Support Vector Regression metoda

Osnovna ideja *Support Vector* Regression (SVR) metode, koju su uveli *Vapnik et al.* (1997), kao i klasifikacione SVM metode, je da se podaci smatraju vektorima u *n*-dimenzionalnom prostoru (eng. *input space*). Kao i kod klasifikacione SVM metode, u slučaju da je veza između regresora i zavisne varijable nelinearna, vektori se preslikavaju u prostor veće dimenzije (eng. *feature space*), gdje je moguće naći optimalnu hiperravan koja linearno modelira ovu vezu. Metoda SVR, s jedne strane, teži da minimizira grešku u ocjeni zavisne varijable, dok s druge strane nastoji da model u prostoru veće dimenzije bude što ravniji kako bi se povećala njegova generalnost, tj. tačnost predviđanja na nepoznatom skupu podataka. Da bi model bio ravniji potrebno je minimizirati intezitet vektora normalnosti hiperravni, ali tako da odstupanja zavisne varijable, dobijena modelom od njenih aktuelnih vrijednosti, budu najviše ε. Drugim riječima, prilikom minimizacije, greške manje od ε se ne uzimaju u obzir (ε *insensitive – loss function*). Dakle, SVR, kao i SVM, pokušava da riješi konveksni optimizacioni problem.

Nesenzitivna ε zona (koja sadrži tačke za koje je greška u ocjeni manja od ε) može se malo proširiti uvođenjem dozvoljene devijacije za ε. Elastičnost ε zone kontorliše parametar C (eng. *error penalty)*, kojim se postiže kompromis između ravnosti modela i ε tačnosti. Veće vrijednosti parametra C dozvoljavaju da model postane neravniji, tj. da se oblikuje prema podacima u skupu podataka za obučavanje modela (eng. *overfitting*), smanjujući na taj način njegovu generalnost (pri čemu devijacije ε ostaju male, tj. postiže se ε tačnost). Manje vrijednosti parametra C smanjuju broj vektora oslonca, tj. složenost modela i povećavaju njegovu ravnost, a samim tim i generalnost (pri čemu se dopuštaju veće devijacije za ε, tj. smanjuje se ε tačnost). Na Slici 20 ilustrovan je prikaz *Support Vector* regresije i transformacije u prostor veće dimenzije.



Slika 20. Ilustracija SVR sa $\mathcal{E}$ tubom – originalni i prostor više dimenzije (Zhang & O'Donnell, 2019)

Ovdje je važno istaći određene razlike između klasifikacionog SVM algoritma i regresionog. Naime, za razliku od SVM klasifikacije koja proizvodi diskretni izlaz (tj. oznaku klase, poput definisanog segmenta kojem kupac pripada), SVR rješava problem regresije koji omogućava procjenu stvarne vrijednosti (neprekidna vrijednost, poput apsolutnih vrijednosti profita koji se očekuje od kupca). Za regresiju, umjesto pronalaženja hiperravni koja u velikoj mjeri može odvojiti primjere u skupu podataka za obučavanje modela, SVR uvodi $\mathcal{E}$-nesenzitivnu funkciju gubitka za izračunavanje hiperravni, tako da predviđene vrijednosti primjera iz uzorka za obučavanje definisane tom hiperravni, imaju najviše $\mathcal{E}$ odstupanja od posmatranih, odnosno stvarnih vrijednosti. Hiperravan plus $\mathcal{E}$ definišu $\mathcal{E}$-neosjetljivu tubu (ili opseg) za izračunavanje granica generalizacije za regresiju. Optimizacija se vrši minimiziranjem $\mathcal{E}$-neosjetljive tube, tako da bude što je moguće ravnija (uža), a da sadrži većinu uzoraka za obuku. U ovom slučaju, hiperravan je predstavljena kroz nekoliko vektora oslonca, odnosno primjera za obuku koji leže izvan, ali blizu granice tube. Kao rezultat SVR obuke, obučava se regresioni model za predviđanje izlaznog odgovora za novi uzorak (Zhang & O'Donnell, 2019).

S tim u vezi, obučavanje SVR modela sastoji se od izbora optimalne kombinacije parametara C, ε i γ (iz RBF kernela, koji se koristi za transformaciju iz originalnog u prostor veće dimenzije). Veće vrijednosti parametra C i manje vrijednosti parametra ε dovode do manje greške u ocjeni zavisne varijable na podacima za obučavanje modela, ali smanjuju generalnost modela, tj. njegovu prediktivnu moć na nepoznatom skupu podataka. Parametar γ zavisi od distribucije podataka za obučavanje modela. Naime, kada je γ previše malo, model ne obuhavata dobro oblik podataka, tj. previše je ravan, a kada je ovaj parametar preveliki, nijedan izbor C neće spriječiti *overfitting* modela.

Dakle, prednost SVR metode u odnosu na tradicionalne regresione metode ogleda se u tome što SVR pruža fleksibilnost definisanja stepena dozvoljene greške u modelu, te pronalazi odgovarajuću liniju ili hiperravan, koja će odgovarati podacima. Generalno gledano, SVR je prilično sličan SVM algoritmu, uz nekoliko važnih razlika (Dobilas, 2020):

1. SVR ima dodatni parametar za podešavanje - ε, čija vrijednost definiše širinu tube oko ocijenjene funkcije – hiperravni. Tačke, odnosno primjeri, koji se nađu u okviru determinisane cijevi smatraju se tačnim predikcijama;
2. Vektori oslonca su u ovom slučaju tačke koje se nalaze izvan tube, u odnosu na SVM klasifikaciju, gdje su to bile tačke koje se nalaze na margini;
3. Podešavanjem parametra C, može se kontrolisati značaj udaljenosti do tačaka koje se nalaze izvan tube.

Tokom posljednjih godina, SVR metoda dobija sve više na značaju, a nakon značajne primjene u oblastima poput medicine (Goli et al., 2016; Schnack et al., 2016; Zhang et al., 2014), ova metoda našla je primjenu i u studijama iz oblasti ekonomije i marketinga.

Za procjenu profitabilnosti klijenata u osiguranju, *Fang et al.* (2016) su testirali SVR metodu i utvrdili da ni ona nije, kao ni ostale metode, efikasna kada je u pitanju izuzetno heterogeni uzorak, specifičan za osiguranje. Naime, u ovom slučaju, pored malog broja najvrednijih kupaca, postoji i problem negativne profitabilnosti za

kupce s velikim štetama, što dovodi do pristrasnosti modela. U tom slučaju, uz prethodnu klasifikaciju, moguće je postići dobru tačnost (Rogić et al., 2022).

Empirijska potvrda efikasnosti SVR metode predstavljena je u radu *Rogić et al.* (2022), gdje je ova metoda upravo primijenjena za predikciju profitabilnosti kupaca u osiguranju. Ova studija, sprovedena za validaciju SVR metode na skupu podataka koji nije iz direktnog marketinga, predviđa profitabilnost postojećih kupaca za koje već postoji iskustvo kupovine registrovano u bazi podataka korisnika osiguranja. Ovo predviđanje profitabilnosti se koristi za automatski odabir najvrednijih kupaca u velikim CRM bazama podataka, za slanje ponude u direktnoj marketing kampanji ili za izgradnju boljih odnosa s kupcima.

U skladu s navedenim razmatranjima, aktivnosti direktnog marketinga kompanija mogu se unaprijediti i ovom metodom poslovne inteligencije, kao što je odabir visokoprofitabilnih potrošača u strategijama akvizicije, omogućavanje interakcije s profitabilnim kupcima kroz kreiranje personalizovanih iskustava kako bi se oni zadržali, kao i povećanje vrijednosti i profitabilnosti odabranih kupaca kroz različite prodajne strategije i preporuke dodatnih proizvoda i usluga (Stone & Woodcock, 2014).

Još jedno istraživanje iz oblasti osiguranja, u kome je primijenjena SVR metoda, sproveli su *Kaščelan et al.* (2016). Uz SVR, autori su koristili *k-means* klasterizaciju za predviđanje rizika u osiguranju. Korišćenjem metoda klasterizacije, oni su svrstali polise u grupe na osnovu stepena rizika, nakon čega su vršili predikciju očekivane vrijednosti potraživanja upotrebom SVR metoda unutar takvih klastera, kao i očekivani nastanak štete upotrebom kernel logističke regresije (KLR). Autori su ostvarili 80% tačnosti predikcije na malom skupu podataka, gdje je 30% polisa imalo nepoznatu vrijednost naknade. S tim u vezi, zaključili su da SVR može biti efikasno primijenjen za predikciju u malim skupovima podataka.

I *Christmann* (2004) je koristio neparametarski pristup zasnovan na kombinaciji KLR i SVR, kako bi izvukao dodatne informacije koje je potencijalno teško otkriti ili modelirati klasičnim metodama, a što bi moglo pomoći kompanijama za osiguranje

motornih vozila u razvoju tarifa osiguranja. U ovom radu je ukazano na prednosti neparametarskih metoda, kao što je SVR, u odnosu na metode linearne regresije u industriji osiguranja. Zbog velikog broja kategoričkih prediktora koji moraju biti kodirani dihotomnim (eng. *dummy*) varijablama, specifikacija funkcionalnog oblika može biti složena, posebno ako su uključeni interaktivni članovi. Zbog izuzetno iskošene distribucije zavisne promjenljive (iznos potraživanja), autor predlaže stratifikaciju promjenljive, definišući mali broj klasa za njihove vrijednosti (ekstremni iznos potraživanja, visok iznos potraživanja, srednji iznos potraživanja i nizak iznos potraživanja), kao i generisanje zasebnih SVR modela za svaku klasu. Iznos potraživanja koji je dao SVR unutar klase množi se sa vjerovatnoćom da korisnik pripada toj klasi. Konačno, dodavanjem ovih proizvoda za sve klase, dobija se očekivani iznos potraživanja. Prednost ovog pristupa je smanjenje potrebnog računskog vremena, zbog prilagođavanja regresionih modela malom podskupu. Dodatno, u ovom radu se posebno ističu prednosti ovog pristupa za senzitivnu analizu segmenata kupaca (Christmann, 2004).

U cilju predikcije ukupne novčane vrijednosti transakcije za svakog respondenta u kampanji, *Kim et al.* (2008) su primijenili SVR za modeliranje odgovora na kampanju direktnog marketinga. Autori su ukazali na problem obuke SVR modela na velikim uzorcima zbog vremenske složenosti. S tim u vezi, oni predlažu primjenu ove metode za predviđanje profitabilnosti, ali na smanjenom uzorku radi smanjenja vremena potrebnog za obučavanje modela. Smanjivanje se vrši uklanjanjem tačaka izvan ε-tube, odnosno ekstremnih vrijednosti. Međutim, na ovaj način se uklanjaju i najvažniji kupci, tj. oni sa ekstremnim vrijednostima profitabilnosti.

Chen i *Wang* (2007) primijenili su SVR za predviđanje tražnje u turizmu i uporedili su performanse ove metode s neuronskim mrežama (eng. *back-propagation neural networks* - *BPNN*) i modelom autoregresivnog integrisanog pokretnog prosjeka (eng. *autoregressive integrated moving average* - ARIMA), koristeći podatke o turistima koji su posjetili Kinu od 1985. do 2001. godine. Oni su osmislili novu proceduru - GA-SVR, koja istražuje opcione parametre za SVM koristeći genetske algoritme, nakon čega slijedi prihvatanje optimalnih parametara za konstrukciju

SVR modela. Njihov GA-SVR model premašio je BPNN i ARIMA modele na osnovu normalizovane srednje kvadratne greške (eng. *normalized mean square error* - NMSE) i srednje apsolutne procentualne greške (eng. *mean absolute percentage error* - MAPE). Takođe, *Wu et al.* (2004) formulisali su SVR model za predviđanje vremena putovanja za korisnike auto-puta i zaključili da ovaj model, s predloženim skupom parametara, nadmašuje ostale osnovne standardne regresione modele.

Kao osnovne prednosti SVR metode ističu se (Raj, 2020):

1. Robustan je za ekstremne vrijednosti;
2. Model odlučivanja može se jednostavno ažurirati;
3. Ima odlične mogućnosti generalizacije, s visokom tačnošću predviđanja;
4. Implementacija je relativno jednostavna.

Zbog istaknutih prednosti, ova metoda će biti korišćena u empirijskom dijelu ove disertacije, za predikciju profitabilnosti kupaca u direktnom marketingu.

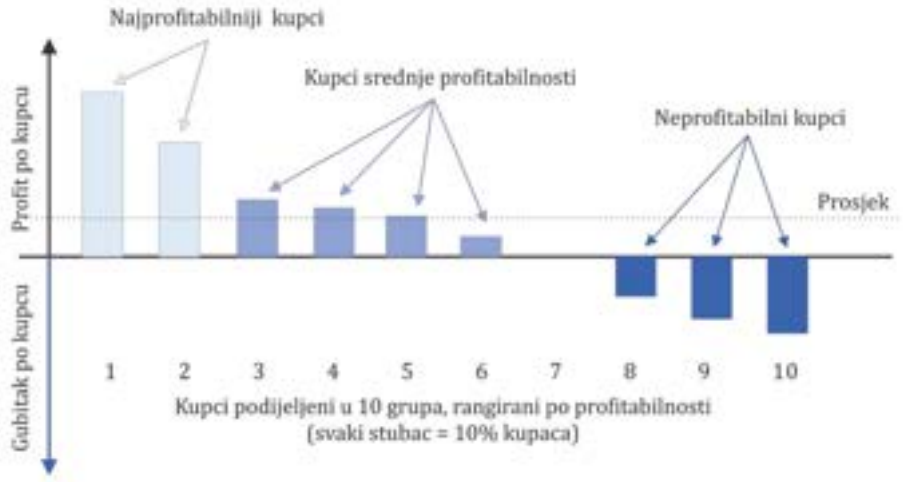
4.1.8 Problem asimetrične distribucije profitabilnosti kupca

U direktnom marketingu, pored predikcije odgovora kupaca na kampanju, važno je precizno identifikovati visokoprofitabilne kupce. Obično ih je malo, zbog čega nemaju veliki uticaj na stopu odgovora, ali imaju veliki uticaj na prihod kampanje. Stopa odziva korisnika s niskim profitom može biti visoka, a da prihodi od kampanje budu niski. *Wang et al.* (2005) naglašavaju da postoji obrnuta korelacija između vjerovatnoće kupovine i iznosa novca za potrošnju. Takođe, autori *Verhoef* i *Donkers* (2001) navode da je u ugovornim transakcijama, kao što je na primjer osiguranje, vjerovatnoća kupovine je stabilna, pa se modeliranje odgovora uglavnom oslanja na predviđanje profitabilnosti kupaca. Zbog toga je za efikasan izbor kupaca potrebna najtačnija procjena njihove očekivane profitabilnosti na osnovu prethodnih podataka. Budući da baze podataka o korisnicima sadrže velike količine podataka, teško je, ili gotovo nemoguće, obrađivati ručno – analizirajući slučaj po slučaj. Zbog toga su u ovom pogledu značajne i neophodne *data mining* metode, koje

omogućavaju automatizaciju ovog procesa i kreiranje adekvatnih marketing strategija.

Glavni izazov u predviđanju profitabilnosti kupaca je asimetrija, odnosno iskošenost distribucije profitabilnosti, jer je broj visokoprofitabilnih kupaca veoma mali u poređenju s drugima. U tradicionalnim metodama zasnovanim na regresiji (Donkers et al., 2007; Glady et al., 2008; Malthouse & Blattberg, 2005; Rust et al., 2011; Verhoef & Donkers, 2001), fokus je na prosječnom kupcu, dok se njihova heterogenost ne uzima u obzir. Na ovaj način se ne prave dovoljno precizne razlike između vrijednih i manje vrijednih kupaca.

Slika 21 ilustruje iskošenost distribucije probitabilnosti kupaca.



Slika 21. Ilustracija iskošenosti distribucije profitabilnosti kupaca

Tradicionalna, odnosno statistička regresiona procedura opisuje se kao proces koji izvodi funkciju *f(x)* koja ima najmanje odstupanje između predviđenih i eksperimentalno posmatranih odgovora za sve primjere obuke. Jedna od glavnih karakteristika SVR metode je da, umjesto da minimizira uočenu grešku u obučavanju (eng. *observed training error*), pokušava da minimizira granicu generalizovane greške (eng. *generalized error bound*), kako bi se postigle generalizovane performanse (Basak et al., 2007).

Kada je u pitanju *data mining* pristup za procjenu profitabilnosti kupaca, generalno se primjenjuju metode klasifikacije, koje mogu predvidjeti kojoj kategoriji kupac pripada: kupci visoke, srednje ili manje vrijednosti (Rogic & Kascelan, 2020; Sujah & Rathnayaka, 2019; Xiahou et al., 2016). Međutim, ove metode ne mogu predvidjeti apsolutni iznos dobiti, pa je teško napraviti fine razlike između vrijednih kupaca. S tim u vezi, nekada je važnije identifikovati svega nekoliko najvrednijih kupca u odnosu na veliki broj kupaca prosječne vrijednosti, posebno uzimajući u obzir njihovu sadašnju i potencijalnu buduću profitabilnost. Da bi to bilo moguće, profitabilnost se mora tretirati kao neprekidna varijabla. Zbog iskošenosti distribucije ove promjenljive, s povećanjem vrijednosti dobiti, tačnost predviđanja se često smanjuje (Fang et al., 2016; Lam, 2018). Kao rezultat toga, ponovo se može desiti da se ne identifikuju najvredniji kupci.

Uzimajući u obzir izvanredne mogućnosti generalizacije i robustnosti prema ekstremnim vrijednostima, može se zaključiti da primjena SVR regresije kod iskošene distribucije zavisne varijable može biti uspješna. Kada je u pitanju profitabilnost kupaca, to potvrđuju i prethodna istraživanja. Osnovne ideje i pristupi za predviđanje profitabilnosti kupaca iz prethodnih istraživanja sumirani su u sljedećoj tabeli.

Tabela 3. Sumarni pregled literature iz oblasti predikcije profitabilnosti kupaca

Autor(i)	Predloženi pristup	Prednosti/nedostaci
Malthouse i *Blattberg* (2005)	Linearna regresija, linearna regresija procijenjena sa iterativno ponderisanim najmanjim kvadratima i metode *feedforward* neuronske mreže primijenjene su da bi se predvidjela buduća profitabilnost kupaca	Na osnovu rezultata sve tri metode za više od 50% najvrednijih kupaca profitabilnost nije precizno predviđena i oni ne bi bili targetirani u budućnosti.
Xiahou et al. (2016)	*k-means* klasterizacija i DT metoda za dobijanje	Segmentacija zasnovana na profitabilnosti bez predviđanja apsolutnog iznosa profita

	klasifikacionih pravila za profitabilnost kupaca	
Verhoef i *Donkers* (2001)	*Probit* model za klasifikaciju klijenata osiguranja	Klasifikacioni pristup bez predviđanja apsolutnog iznosa profita i manja tačnost klasifikacije
Ballestar et al. (2019)	Prediktivni model baziran na MLP vještačkim neuronskim mrežama za predviđanje kvaliteta kupaca	Klasifikacioni pristup bez predviđanja apsolutnog iznosa profita i manja tačnost klasifikacije
Lam (2018)	*Ensemble* koji čine *gradient boosting* i neuronske mreže za predviđanje apsolutnog iznosa profitabilnosti kupaca	Greška za prvi poludecil najvrednijih kupaca (najvažniji segment) je najveća.
Fang et al. (2016)	RF u poređenju sa generalizovanom linearnom regresijom, DT, SVR i *generalized boosted regression* za predviđanje profitabilnosti kupaca	Tačnost predviđanja apsolutnog iznosa profitabilnosti opada kako se vrijednost profita povećava
Kim et al. (2008)	SVR za modeliranje odgovora na kampanju u cilju predviđanja ukupnog iznosa koji bi potrošio svaki respondent	Zbog vremenske složenosti, autori predlažu smanjenje uzorka uklanjanjem *outliersa*, odnosno kupaca sa ekstremnim vrijednostima profitabilnosti. Pored toga, ovaj postupak može dovesti do gubitka važnih informacija za klasifikaciju.
Kašćelan et al. (2016)	*K-means*, SVR i KLR za predviđanje rizika u osiguranju automobila	SVR je uspješno primijenjen za predviđanje nepoznatih šteta na malim skupovima podataka, dobijenim klasterizacijom polaznih podataka. Dodatna prednost je u rješavanju problema asimetrične distribucije ciljne varijable.
Christmann (2004)	KLR i SVR za podršku odlučivanju u industriji osiguranja	Smanjenje potrebnih računskih resursa prilagođavanjem

	regresionih modela malom podskupu – stratifikacija varijable definisanjem malog broja klasa i generisanjem SVR modela za svaku klasu. Takođe, riješen je problem asimetrične distribucije ciljne varijable.

Analiza istraživanja iz prethodne tabele ukazuje na to da su *data mining* metode fleksibilnije i povoljnije za predviđanje profitabilnosti kupaca u direktnom marketingu, u poređenju sa statističkim, ali i da postoji potreba za poboljšanjem tačnosti predviđanja, posebno za najvrednije i najrizičnije kupce. Velike baze podataka o korisnicima zahtijevaju mnogo vremena za obuku modela zasnovanih na podacima, pa prethodna literatura sugeriše smanjenje uzorka. Međutim, ova procedura može učiniti da se izgube značajne informacije za donošenje odluka (Rogić et al., 2022).

Cilj empirijskog dijela ove disertacije, čiji će rezultati biti predstavljeni u sekciji 5.6, jeste da prevaziđe uočene nedostatke primjenom SVR metode. Dodatno, dosadašnja istraživanja opravdavaju primjenu SVR metode, koja ima sposobnost da riješi problem heterogenosti uzorka i potvrđuju da se dobre prediktivne performanse SVR metode mogu očekivati i za kategoriju najvrednijih kupaca.

Nakon predstavljanja najznačajnijih ideja SVM metoda za klasifikaciju i regresiju, u narednom dijelu rada biće opisani pokazatelji prediktivnih performansi klasifikacionih i regresionih modela.

4.2 Pokazatelji prediktivnih performansi za klasifikaciju i regresiju

Veliki broj razvijenih tehnika prediktivnog modeliranja stvorio je mogućnost za kreirenje velikog broja modela, koji mogu omogućiti efikasnu predikciju za zadati problem. Međutim, u tim uslovima, odabir pravog modela nameće se kao izazov. U

cilju pronalaženja najboljeg modela za pojedinačne svrhe, potrebno je razumjeti njihove performanse, analizirajući skup metrika koje mjere kvalitet izgrađenih modela. Dakle, u skladu s poslovnim ciljevima koji motivišu upotrebu prediktivnih modela, potrebno je izvršiti njihovu evaluaciju. Evaluacija modela je od posebnog značaja ukoliko se uzme u obzir potreba da prediktivni model ima sličnu efikasnost na različitim skupovima podataka, a ne samo na onom na kojem je obučen (Shchutskaya, 2021). S tim u vezi, rezultati prediktivnih modela moraju biti uporedivi, mjerljivi i ponovljivi.

Faza evaluacije, dakle, pomaže da se osigura da model zadovoljava prvobitno definisane poslovne ciljeve. Primarni cilj nauke o podacima za kompanije je upravo podrška u donošenju odluka, te je fokus ovog procesa upravo na problemu koji kompanija teži da riješi, poput adekvatnog odabira kupaca koji će biti targetirani u narednoj kampanji direktnog marketinga. Rezultati *data mining* modela najčešće predstavljaju samo dio sveobuhvatnog rješenja, zbog čega je neophodno izvršiti njihovu evaluaciju. Evaluacija rezultata prediktivnih *data mining* modela uključuje kako kvalitativne, tako i kvantitativne procjene (Provost & Fawcett, 2013). Različiti stejkholderi imaju interese u donošenju poslovnih odluka, koje će biti postignute ili u nekoj mjeri podržane rezultatima kreiranih modela. Kako bi se olakšala kvalitativna procjena modela, analitičari zaduženi za njihov razvoj dužni su osigurati razumljivost modela i rezultata svim zainteresovanim stranama. Kada je u pitanju kvantitativna evaluacija prediktivnih modela, analitičari mogu koristiti kombinacije metrika, odnosno pokazatelja prediktivnih performansi modela, koji će biti opisani u nastavku poglavlja.

U zavisnosti od toga da li je inicijalni problem koji se teži riješiti klasifikacione ili regresione prirode, analitičari će izvršiti odabir različitih metrika za evaluaciju prediktivnih modela. Postoji fundamentalna razlika između metoda za vrednovanje regresionog i klasifikacionog modela. Naime, regresija se bavi neprekidnim vrijednostima, gdje se može identifikovati greška između stvarnog i predviđenog izlaza. Međutim, prilikom ocjenjivanja klasifikacionog modela, fokus je na broju predviđanja koja možemo pravilno klasifikovati. Da bismo pravilno ocijenili model

klasifikacije, moramo uzeti u obzir i podatke koji su pogrešno klasifikovani (Singh, 2019).

U Tabeli 4 navedene su najznačajnije metrike za evaluaciju prediktivnih modela za klasifikaciju i regresiju.

Tabela 4. Metrike za evaluaciju prediktivnih modela za klasifikaciju i regresiju

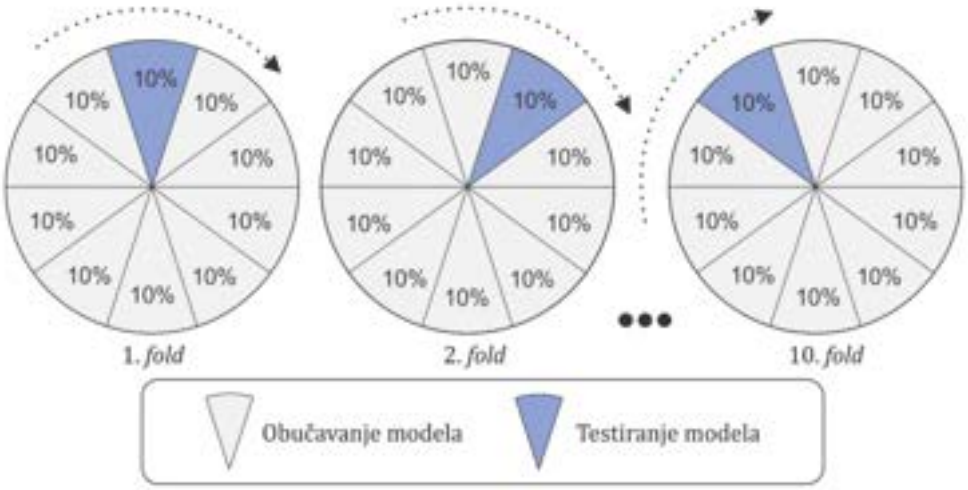
Tip prediktivnog modela: klasifikacioni / regresioni	**Metrika za evaluaciju prediktivnih performansi modela**
Klasifikacioni modeli	Konfuziona matrica (eng. Confusion matrix)
	Tačnost (eng. *Accuracy*)
	Stopa greške (eng. *Error Rate; Overall Error Rate; Estimated Misclassification Rate*)
	Senzitivnost (eng. *Sensitivity; Recall; True Positive Rate – TPR*)
	Specifičnost (eng. *Specificity; Selectivity; True Negative Rate – TNR*)
	Preciznost (eng. *Precision; Positive Predictive Value – PPV*)
	Stopa ispadanja / stopa lažno pozitivnih primjera (*eng. Fall-out; False Positive Rate – FPR*)
	F-metrika (eng. *F-measure; F1 Score*)
	ROC kriva (eng. *Receiver Operating Characteristic – ROC curve*) i AUC metrika (eng. *Area Under the Curve – AUC*)
	Lift kriva (eng. *lift curve; lift chart*)
Regresioni modeli	Srednja apsolutna greška (eng. *Mean Absolute Error/Deviation - MAE*)
	Srednja kvadratna greška (eng. *Mean Square Error*)
	Procenat srednje greške (eng. *Mean Percentage Error*)
	Srednji apsolutni procenat greške (eng. *Mean Absolute Percentage Error – MAPE*)
	Korijen srednje kvadratne greške (eng. *Root Mean Squared Error – RMSE*)
	Koeficijent determinacije – R^2 (eng. *Coefficient of Determination*)

Evaluacija modela podrazumijeva sastavni dio samog procesa razvoja modela i pomaže u pronalaženju najboljeg modela koji može predstaviti raspoložive podatke.

Takođe, važan aspekt ovog procesa odnosi se na procjenu efikasnosti funkcionisanja kreiranog modela u budućnosti.

Evaluacija performansi modela korišćenjem podataka za njegovo obučavanje, u nauci o podacima uglavnom nije prihvatljiva. Ovakva praksa dovodi do generisanja previše optimističnih modela, koji su zavisni od podataka. Umjesto ovog pristupa, nauka o podacima podržava korišćenje sljedeće dvije metode: zadržavanje (eng. *hold-out*) i kros-validacija (eng. *cross-validation*) (Singh, 2019), koje obuhvataju podjelu inicijalnog skupa podataka tako što se kreira podskup za testiranje modela, koji će do momenta validacije modelu biti nepoznat. Na ovaj način, izbjegava se pretjerano prilagođavanje modela skupu podataka za njegovo obučavanje.

Metoda zadržavanja primjenjuje se s većom učestalošću u slučaju korišćenja većih baza podataka. Ovaj pristup obuhvata podjelu cjelokupne baze podataka na tri dijela, odnosno podskupa: skup za obučavanje, set za validaciju i set za testiranje. Prvi – skup za obučavanje, koristi se za izgradnju i obučavanje prediktivnog modela. Drugi – skup za validaciju, koristi se za procjenu performansi modela izgrađenog u prvom koraku. Ovaj podskup podataka pruža platformu za fino podešavanje parametara modela i odabir modela s najboljim performansama. Konačno, podskup za testiranje modela predstavlja dio originalne baze podataka, koji je nepoznat modelu. Posljednji skup podataka koristi se za procjenu budućih performansi modela. Ako se model mnogo bolje uklapa u skup za obuku nego što se uklapa u testni skup, vjerovatno je uzrok preveliko prilagođavanje. S druge strane, kada su analitičarima na raspolaganju manji i ograničeni skupovi podataka, da bi se postigla nepristrasna procjena performansi modela, koristi se *k-fold* kros-validacija. Ovaj pristup uključuje podjelu podataka u *k* podskupova jednake veličine. Modeli se kreiraju *k* puta, pri čemu se u svakoj iteraciji isključuje jedan podskup iz skupa za obučavanje i koristi se za testiranje (Singh, 2019). Na Slici 22 prikazan je proces *10-fold* kros-validacije.



Slika 22. Proces *10-fold* kros-validacije

Kao što se vidi sa Slike 22, cjelokupni skup podataka se dijeli na 10 (približno) istih djelova. U slučaju *10-fold* kros-validacije, u svakoj od 10 iteracija, jedan dio podataka se koristi za testiranje, dok se preostalih devet koristi za obučavanje modela. Detaljan opis procesa kros-validacije biće predstavljen u sekciji 4.3.

Za procjenu prediktivnih performansi modela može se koristiti kombinacija različitih metrika. U svakom slučaju, mjere se zasnivaju na testnom skupu, koji služi kao objektivniji osnov od skupa obuke za procjenu tačnosti predviđanja – primjeri u testnom skupu po pravilu su sličniji primjerima koje je potrebno predvidjeti, u smislu da se ne koriste za izbor prediktora ili za procjenu parametara modela, a nisu poznati modelu (Shmueli et al., 2018). Dakle, modeli se izgrađuju korišćenjem podataka za obučavanje, a zatim se primjenjuju na testnim podacima, čije se performanse procjenjuju korišćenjem različitih metrika. U nastavku poglavlja će biti opisani načini za evaluaciju prediktivnih klasifikacionih i regresionih modela, respektivno.

4.2.1 Evaluacija klasifikacionih modela

U ovom dijelu rada biće opisane najznačajnije metrike za evaluaciju prediktivnih klasifikacionih modela.

1. Konfuziona matrica (eng. *confusion matrix*)

Konfuziona matrica, koja se u literaturi naziva i klasifikaciona matrica ili matrica greške, predstavlja sumarni prikaz tačne i netačne klasifikacije, koje je model proizveo za određeni skup podataka. Redovi i kolone matrice odgovaraju predviđenim i tačnim (stvarnim) klasama, respektivno (Shmueli et al., 2018). Dakle, za svaku instancu u testnom skupu podataka, konfuziona matrica upoređuje stvarnu klasu sa klasom kojoj je instanca dodijeljena od strane obučenog klasifikatora.

Slika koja slijedi ilustruje konfuzionu matricu 2x2 za dvije klase (pozitivna i negativna).

		Stvarna vrijednost: pozitivna	Stvarna vrijednost: negativna
Predviđena vrijednost	pozitivna	TP True Positive	FP False Positive
	negativna	FN False Negative	TN True Negative

Slika 23. Ilustracija 2x2 konfuzione matrice

Dakle, na osnovu predstavljene tabele, može se uočiti da svaki red matrice predstavlja instance u predviđenoj klasi, a svaka kolona predstavlja instance u stvarnoj klasi. Ova matrica izvještava o broju tačno pozitivnih (eng. *true positive - TP*), lažno pozitivnih (eng. *false positive – FP*), lažno negativnih (eng. *false negative - FN*) i tačno negativnih (eng. *true negative – TN*) predikcija, odnosno rezultata. Glavna dijagonala obuhvata broj instanci koje su tačno klasifikovane, dok suprotno važi za elemente van dijagonale. Navedene metrike omogućavaju sveobuhvatniju evaluaciju modela u odnosu na korišćenje tačnosti kao mjerila performansi modela.

Naime, u slučaju izražene nebalansiranosti klasa, preciznost kao mjera neće biti odraz objektivnih rezultata[6].

Konfuziona matrica, iako nije metrika sama po sebi, jasno pokazuje da li model adekvatno razdvaja dvije klase, tj. da li jednu klasu pogrešno označava kao drugu. Dodatno, na osnovu elemenata konfuzione matrice – TP, FP, FN, TN, izračunava se značajan broj metrika koje se koriste za evaluaciju prediktivnih modela klasifikacije, od kojih će najkorišćenije biti opisane u nastavku.

2. **Tačnost (eng. *accuracy*)**

Tačnost klasifikacije predstavlja jednu od najjednostavnijih metrika koje se koriste u evaluaciji modela. Definiše se kao udio ukupnog broja predviđanja koja su bila tačna. S tim u vezi, tačnost klasifikacije računa se kao broj tačnih predikcija u odnosu na ukupan broj predikcija. Formula za izračunavanje tačnosti klasifikacije je:

$$Accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$

3. **Stopa greške (eng. *error rate; overall error rate; estimated misclassification rate*)**

Stopa greške opisuje drugu stranu novčića u odnosu na tačnost – udio netačnih predikcija u odnosu na ukupan broj predikcija. Računa se na sljedeći način:

$$Error\ rate = 1 - Accuracy = \frac{FN + FP}{TN + FN + FP + TP} = \frac{FN + FP}{P + N}$$

4. **Senzitivnost (eng. *sensitivity; recall; true positive rate – TPR*)**

Kao što je navedeno pri opisivanju metrike tačnosti, u situacijama kada klase u bazi podataka nisu približno jednake, isključivo posmatranje tačnosti klasifikacije može

[6] Formula za izračunavanje preciznosti biće predstavljena u nastavku poglavlja.

dati lažnu sliku o performansama modela. U tim situacijama, češće se koriste alternativne metrike – senzitivnost i specifičnost.

Senzitivnost mjeri sposobnost modela da pozitivno klasifikuje zapis (Larose & Larose, 2015).

$$Sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN}$$

U slučaju da je zadatak klasifikatora da predvidi pripadnost kupca visoko ili nisko profitabilnoj klasi, dobar klasifikacioni model trebalo bi da bude senzitivan, tj. sposoban da identifikuje značajan udio kupaca koji su pozitivni (pripadaju profitabilnom segmentu). Dakle, senzitivnost pokazuje koji je procenat klase od interesa model zaista obuhvatio.

U teoriji bi savršeni klasifikator imao senzitivnost = 1 (100%). Međutim, nulti model koji bi sve korisnike jednostavno klasifikovao kao pozitivne, takođe bi imao senzitivnost = 1 (Larose & Larose, 2015). S tim u vezi, za adekvatnu evaluaciju modela, senzitivnost, kao izolovana metrika, nije dovoljna.

5. **Specifičnost (eng. *specificity; selectivity; true negative rate – TNR*)**

Klasifikacioni model treba da karakteriše i specifičnost, odnosno tačna klasifikacija negativnih primjera. Koristeći isti primjer kao za senzitivnost, dobar prediktivni model trebalo bi da identifikuje značajan udio kupaca koji pripadaju negativnoj klasi (segmentu niskoprofitabilnih kupaca). Dakle, specifičnost predstavlja procenat stvarnih negativnih slučajeva koji su pravilno identifikovani, odnosno mjeri sposobnost modela da prepozna negativne instance. Računa se na sljedeći način:

$$Specificity = \frac{TN}{N} = \frac{TN}{TN + FP}$$

Kao i u slučaju senzitivnosti, savršeni klasifikacioni model imao bi specifičnost = 1. S tim u vezi, i model koji bi sve instance predvidio kao negativne, imao bi ovu vrijednost metrike specifičnosti (Larose & Larose, 2015). Stoga, dobar klasifikacioni

model treba da pokaže zadovoljavajuće vrijednosti i senzitivnosti i specifičnosti, pri čemu zadovoljavajuće vrijednosti variraju od slučaja do slučaja.

Senzitivnost i specifičnost daju sliku o sposobnostima modela za pravu distinkciju između klasa i često se koriste za evaluaciju prediktivnih performansi modela za segmentaciju kupaca.

6. **Preciznost (eng. *precision; positive predictive value – PPV*)**

Preciznost pokazuje koliko primjera, klasifikovanih kao „pozitivni", zapravo pripada predviđenoj pozitivnoj klasi. Dakle, preciznost mjeri udio tačno predviđenih pozitivnih vrijednosti i računa se na sljedeći način:

$$Precision = \frac{TP}{TP + FP}$$

Veći broj lažno pozitivnih primjera ukazuje na nižu stopu preciznosti klasifikatora.

7. **Stopa ispadanja / stopa lažno pozitivnih primjera (*eng. fallout; false positive rate – FPR*)**

Stopa ispadanja ili stopa lažno pozitivnih primjera računa se kao odnos broja lažno pozitivnih primjera i ukupnog broja stvarno negativnih primjera:

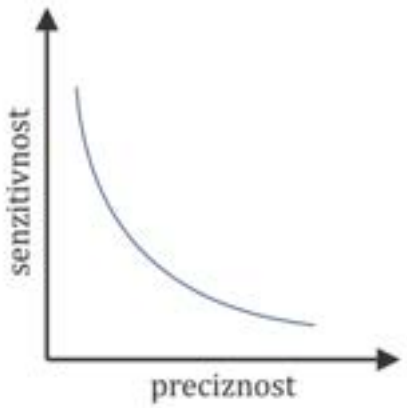
$$Fallout = \frac{FP}{N} = \frac{FP}{FP + TN}$$

Ova metrika se dovodi u vezu s preciznošću. Dok preciznost mjeri vjerovatnoću da je primjer klasifikovan kao pozitivan zapravo pozitivan, stopa lažno pozitivnih mjeri odnos lažno pozitivnih u negativnim primjerima. Ako je broj negativnih primjera veoma veliki, odnosno ako skup podataka karakteriše izražena nebalansiranost klasa, stopa ispadanja raste sporije, s obzirom na to da bi broj tačno negativnih (FP+TN) bio veoma visok, što bi ovu metriku činilo manjom. S druge strane, na preciznost ne utiče veliki broj negativnih primjera, što se odslikava kroz zbir u imeniocu (TP+FP). Dakle, preciznost je dominantno fokusirana na pozitivnu klasu,

dok je stopa ispadanja jedna od mjera koja pokazuje sposobnost klasifikatora da, uz senzitivnost, pravi razliku između klasa (Lador, 2017). Ovdje je važno je istaći značaj ove mjere za direktni marketing, uzimajući u obzir da kod predviđanja odgovora kupca ukazuje upravo na kupce koji su lažno proglašeni respondentima, te za koje će se neefikasno potrošiti sredstva u kampanji.

8. F-metrika (eng. *F-measure; F1 score*)

U modelima klasifikacije uglavnom postoji *trade-off* između mjera preciznosti i senzitivnosti. Težnja da se jedna mjera popravi najčešće rezultira u pogoršanju druge (Rokach & Maimon, 2015). Slika 24 ilustruje *trade-off* između preciznosti i senzitivnosti.



Slika 24. Trade off između preciznosti i senzitivnosti

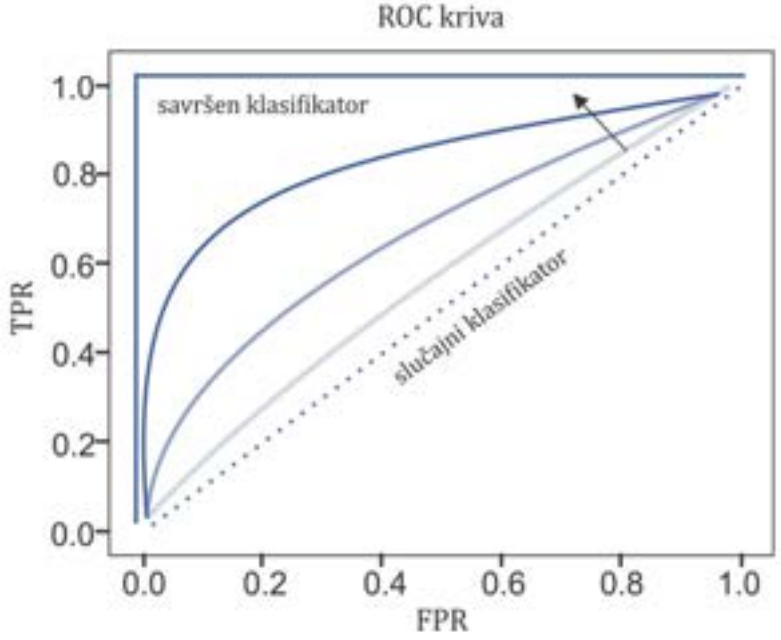
Metrike preciznosti i senzitivnosti, ako se izolovano posmatraju, ne mogu dati pravu sliku o performansama modela. Stoga je kreirana dodatna metrika, kojom se teži riješiti ovaj problem - F-metrika. Nakon što se izračunaju senzitivnost i preciznost individualno, njihove vrijednosti se kombinuju za dobijanje F vrijednosti, koja predstavlja jednu od najčešće korišćenih metrika za probleme neuravnotežene klasifikacije (He & Ma, 2013). Računa se na sljedeći način:

$$F - measure = 2 \times \frac{PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

Dakle, F-rezultat je metrika koja uzima u obzir i preciznost i senzitivnost i može se tumačiti kao ponderisani harmonijski prosjek vrijednosti preciznosti i senzitivnosti. Kada preciznost i senzitivnost nisu visoke, ni ovaj rezultat takođe ne može biti visok, što ukazuje na dobre performanse modela.

9. **ROC kriva (eng. *Receiver Operating Characteristic – ROC curve*) i AUC metrika (eng. *Area Under the Curve – AUC*)**

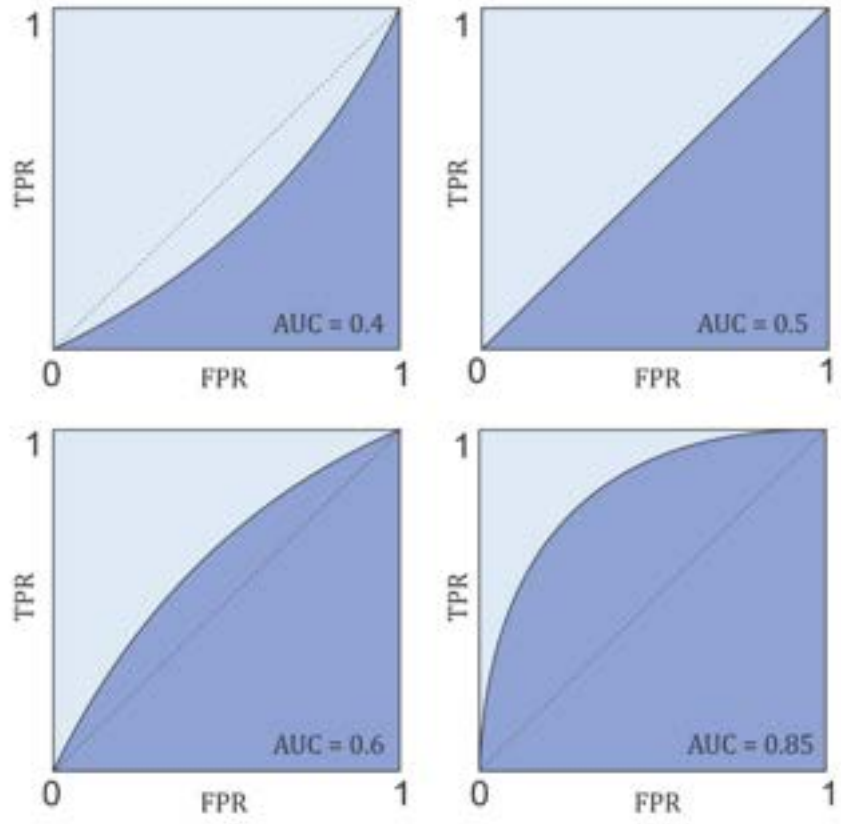
ROC kriva grafički predstavlja odnos između dvije prethodno pomenute metrike – senzitivnosti (TPR) i stope ispadanja (FPR), tj. stope tačno pozitivnih i stope lažnih negativnih primjera. Slika 25 prikazuje ilustraciju ROC krive, pri čemu se na *X* osi odslikava FPR, a na *Y* osi TPR.



Slika 25. Ilustracija ROC krive

Pri evaluaciji klasifikatora, težnja je da je ROC kriva što bliža lijevoj strani grafika, odnosno da je TRP veće u odnosu na FPR.

Prostor ispod ROC krive koristi se kao mjera kvaliteta klasifikacionih modela – AUC (eng. *Area Under the Curve*). Na Slici 26 prikazane su ilustracije različitih ROC krivih, s različitim AUC vrijednostima. Slučajni klasifikator ima AUC vrijednost 0,5, savršeni klasifikator ima vrijednost 1, dok u praksi, većina klasifikacionih modela ima AUC između 0,5 i 1.



Slika 26. Ilustracija različitih vrijednosti AUC metrike

Kada klasifikator ne može razlikovati dva segmenta, AUC će biti 0,5 (poklapaće se sa dijagonalom). S druge strane, kada postoji savršeno razdvajanje segmenata, odnosno kada nema preklapanja u podacima, površina ispod ROC krive dostiže vrijednost 1 (ROC kriva će doći do gornjeg lijevog ugla grafikona). U tom slučaju, svi pozitivni primjeri biće tačno klasifikovani i neće biti pogrešno klasifikovanih negativnih primjera (FP). Dakle, što je veća vrijednost AUC metrike, bolje su

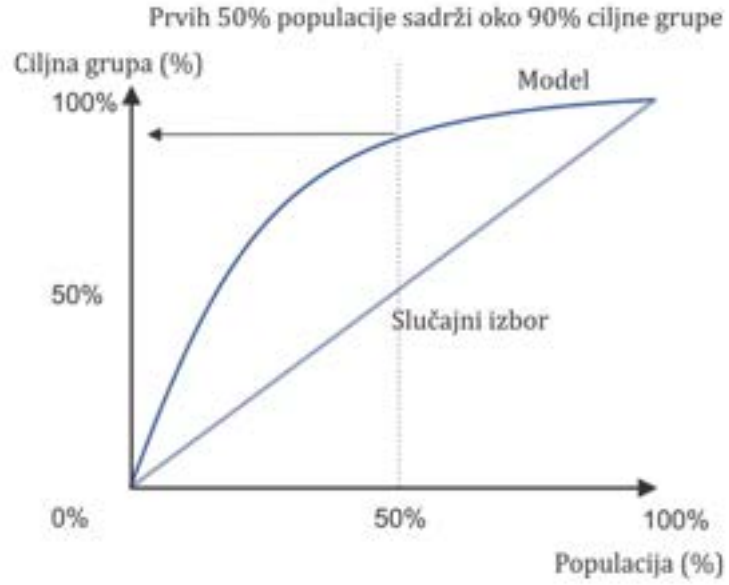
performanse modela. Ova metrika se može tumačiti kao vjerovatnoća da model bolje rangira slučajne pozitivne u odnosu na slučajne negativne primjere.

10. Lift kriva (eng. *lift curve; lift chart*)

Pored ROC krive, *Lift* je još jedna vizuelna tehnika za evaluaciju prediktivnih modela. *Lift* kriva predstavlja koncept, kreiran u oblasti marketinga, koji nastoji da uporedi stope odgovora sa i bez korišćenja klasifikacionog modela.

Lift predstavlja mjeru efikasnosti klasifikacionog modela, koja se računa kao odnos između rezultata dobijenih s modelom i bez njega. Dakle, kao grafička metoda za evaluaciju prediktivnih klasifikacionih modela, *Lift* kriva procjenjuje i upoređuje korisnost modela (Larose & Larose, 2015). Međutim, za razliku od konfuizone matrice, koja procjenjuje modele na osnovnu cjelokupne populacije, *Lift* vrši evaluaciju modela na osnovu jednog segmenta populacije (Singh, 2019). Lift se najčešće koristi za procjenu kvaliteta klasifikatora u odnosu na slučajni izbor.

Ova tehnika se često koristi u direktnom marketingu (Prati et al., 2011), posebno u cilju identifikovanja potencijalnih respondenata, koji se targetiraju u kampanjama. Svakom potencijalnom kupcu dodijeli se vrijednost koja odslikava vjerovatnoću odgovora, a *Lift* kriva pomaže u otkrivanju dijela uzorka koji je potrebno targetirati, a koji će doprinijeti velikom udjelu odgovora. Na narednoj slici prikazan je primjer *Lift* krive.



Slika 27. Prikaz *Lift* krive

Na Slici 27 ilustrovan je primjer *Lift* krive. Slučajni izbor je, kao i u analizi ROC krive, predstavljen dijagonalom koja kreće iz koordinatnog početka. U ovom primjeru, uzorak od 50% populacije sadrži oko 90% respondenata. Povećanjem uzorka, odnosno segmenta kupaca koji će biti targetirani dobija se zanemarljivo povećanje broja odgovora na kampanju.

U narednom dijelu rada biće opisane metrike za evaluaciju performansi regresionih modela.

4.2.2 Evaluacija regresionih modela

Za evaluaciju regresionih model najčešće se koristi kombinacija metrika predstavljenih u Tabeli 4, od kojih će najvažnije biti ukratko opisane u ovom dijelu rada. U poređenju s klasifikacionim modelima, tačnost regresionog modela je teže ilustrovati, s obzirom na to da je praktično nemoguće predvidjeti tačnu vrijednost, već koliko je blizu predviđena u odnosu na stvarnu vrijednost.

9

1. **Srednja apsolutna greška (eng. *Mean Absolute Error/Deviation - MAE*)**

Srednja apsolutna greška (MAE) pronalazi prosječnu apsolutnu distancu između predviđenih i stvarnih vrijednosti. Ova metrika uzima u obzir zbir apsolutnih vrijednosti greške i računa se na sljedeći način:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

gdje je N – broj instanci, y_i stvarna vrijednost, a $\hat{y}_i$ predviđena vrijednost zavisne varijable.

Sredna apsolutna greška izražena je u istim jedinicama mjera kao i originalni podaci, pa se može porediti samo s modelima čije su greške izražene istim mjernim jednicama.

2. **Srednja greška (eng. Mean Error – ME)**

Mjera srednje greške (ME) slična je prethodnoj (MAE), s tom razlikom što ME zadržava znak greške, tako da negativne greške poništavaju pozitivne greške iste veličine. S tim u vezi, ova metrika daje informaciju o tome da li model vrši precjenjivanje ili potcjenjivanje vrijednosti zavisne varijable (Shmueli et al., 2018).

Formula za izračunavanje srednje greške je:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

3. **Srednja kvadratna greška (eng. *Mean Square Error - MSE*)**

Srednja kvadratna greška (MSE) se izračunava kao suma kvadrata greške predviđanja $(y_i - \hat{y}_i)$, koja se zatim dijeli sa brojem instanci. Formula za izračunavanje MSE je:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

gdje je N – broj instanci, y_i stvarna vrijednost, a $\hat{y}_i$ predviđena vrijednost zavisne varijable.

Ova metrika takođe daje informaciju o tome koliko predviđeni rezultati odstupaju od stvarnih. Međutim, evaluacija cjelokupnog modela ne može se izvršiti tumačenjem individualne MSE metrike, već se po pravilu koristi skup metrika koje će pomoći da se izabere najbolji regresioni model.

Kvadriranjem grešaka, ekstremne vrijednosti u modelu, koje obično imaju veće greške u odnosu na druge instance, dobijaju na značaju i dominiraju u konačnom rezultatu MSE. Stoga je za evaluaciju regresionih modela, MAE metrika, u poređenju sa MSE, robustnija u odnosu na ekstremne vrijednosti (Minaee, 2019). Srednja kvadratna greška uzima isključivo ne-negativne vrijednosti, pri čemu model boljih performansi ima vrijednost metrike bližu nuli.

4. Srednja procentualna greška (eng. *Mean Percentage Error - MPE*)

Srednja procentualna greška se računa kao prosjek procentualnih grešaka po kojima se predviđene vrijednosti u okviru modela razlikuju od stvarnih vrijednosti. Formula za MPE je:

$$MPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)}{y_i}$$

gdje je y_i stvarna vrijednost promjenljive, $\hat{y}_i$ predviđena vrijednost, a n je broj instanci.

Uzimajući u obzir da se za računanje MPE koriste stvarne, a ne apsolutne vrijednosti grešaka predviđanja, pozitivne i negativne greške se mogu međusobno kompenzovati. S tim u vezi, formula se može koristiti kao mjera pristrasnosti u

prognozama. S druge strane, osnovni nedostatak ove metrike je što ostaje nedefinisana u situacijama kada je jedna stvarna vrijednost nula.

5. **Srednji apsolutni procenat greške (eng. *Mean Absolute Percentage Error – MAPE*)**

Srednja apsolutna procentualna greška (MAPE) takođe predstavlja mjeru tačnosti predviđanja i računa se pomoću sljedeće formule:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

gde je y_i stvarna vrijednost, $\hat{y}_i$ predviđena vrijednost, a n broj instanci. Kao i u slučaju MPE metrike, ne može se koristiti ako postoje nulte vrijednosti, jer bi u tom slučaju došlo do dijeljenja s nulom.

6. **Korijen srednje kvadratne greške (eng. *Root Mean Squared Error – RMSE*)**

Korijen srednje kvadratne greške (RMSE) može imati samo ne-negativne vrijednosti, pri čemu bi RMSE=0 bila vrijednost metrike koja ukazuje na idealno uklapanje prema podacima. Dakle, savršena vrijednost ove metrike bila bi nula i označavala bi jednakost svih predviđenih vrijednosti u odnosu na stvarne. Generalno, nultu vrijednost u praksi gotovo nije moguće postići (osim u slučaju trivijalnih modela), te se kao indikator dobrih performansi modela uzima niža vrijednost metrike - ako su predviđene vrijednosti veoma blizu stvarnim, vrijednosti RMSE metrike biće manja.

Formula za izračunavanje ove metrike data je u nastavku:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

Dodatno, korijen srednje kvadratne greške može se računati i na sljedeći način:

$$RMSE = \sqrt{MSE}$$

Ovdje je važno naglasiti da se prilikom računanja kvadratnog korijena jedinice mjere RMSE metrike vraćaju na originalne jedinice mjere zavisne varijable koja je predmet predikcije. Dakle, ako je zavisna varijabla – profitabilnost kupaca, izražena u eurima, tada će vrijednost RMSE također biti izražena u eurima, za razliku od MSE metrike, na primjer, gdje bi bila izražena u „eurima na kvadrat".

7. **Koeficijent determinacije – R^2 (eng. *Coefficient of Determination; Squared Correlation*)**

Koeficijent determinacije R^2 mjeri koliko se varijabilnost u zavisnoj promjenljivoj može objasniti modelom. Ova metrika predstavlja kvadrat koeficijenta korelacije (R). Dok korelacija objašnjava jačinu veze između nezavisne i zavisne varijable, R^2 objašnjava u kojoj mjeri varijansa jedne varijable objašnjava varijansu druge varijable. Dakle, ako je R^2 modela 0,50, onda se otprilike polovina uočene varijacije može objasniti inputima modela.

Dakle, R^2 se izračunava kao suma kvadrata greške predviđanja podijeljena sa ukupnom sumom kvadrata, koja predviđenu vrijednost zamjenjuje srednjom vrijednošću. Vrijednost ove metrike kreće se između 0 i 1, a veća vrijednost ukazuje na bolje uklapanje između predviđanja i stvarne vrijednosti.

$$R^2 = 1 - \frac{SS_{regression}}{SS_{total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

R^2 je korisna mjera za utvrđivanje koliko dobro model determiniše zavisnu promjenljivu. Međutim, ona ne uzima u obzir problem prekomjernog prilagođavanja podacima za obučavanje modela. Ako regresioni model ima mnogo nezavisnih varijabli, što čini model komplikovanim, može se vrlo dobro uklopiti u podatke za obučavanje, međutim, performanse na podacima za testiranje mogu biti nezadovoljavajuće.

U narednom dijelu rada biće detaljnije opisana kros-validacija, kao značajna tehnika testiranja prediktivnih performansi modela.

4.3 Testiranje prediktivnih performansi kros-validacijom

Početkom tridesetih godina prethodnog vijeka, *Larson* (1931) je zaključio da obučavanje algoritma i evaluacija njegovih statističkih performansi na istim podacima daje previše optimističan rezultat. S tim u vezi, a u cilju prevazilaženja ovog problema, razvijen je koncept kros-validacije. Naime, kros-validacija podrazumijeva testiranje razvijenog i obučenog modela na novim, nepoznatim podacima, kako bi se adekvatno procijenile njegove performanse (Geisser, 1975; Mosteller & Tukey, 1968).

Kros-validacija posebno dobija na značaju kada je raspoloživi skup podataka ograničen, s obzirom na to da omogućava podjelu tog skupa na podskup za obučavanje (skup za obučavanje) i podskup za validaciju (skup za testiranje). U tom slučaju, drugi podskup za model predstavlja neviđene – nove podatke, što će doprinijeti objektivnoj evaluaciji performansi modela.

Kros-validacija se može definisati kao tehnika koja obezbjeđuje da se rezultati otkriveni u analizi kroz model mogu generalizovati, korišćenjem nezavisnog, neviđenog skupa podataka (Larose & Larose, 2015). U primjeni *data mining* metoda, najčešće korišćene forme kros-validacije su dvostruka kros-validacija i *k-fold* kros-validacija. Prva, jednostavnija forma, nazvana i "kros-validacija sa zadržavanjem" (eng. *holdout cross-validation*), obuhvata nasumičnu podjelu kompletnog skupa podataka na setove za obučavanje i za testiranje, pri čemu je u testnom skupu potrebno izostaviti zavisnu promjenljivu. Dakle, jedina razlika između skupa za obučavanje i testiranje modela ogleda se u tome što je u drugom skupu izostavljena zavisna, odnosno ciljna varijabla. Na primjer, ukoliko je zadatak modela klasifikacija potrošača prema nivou profitabilnosti, model će se obučavati korišćenjem skupa podataka koji sadrži veliki broj zapisa o karakteristikama kupaca, uključujući i nivo njihove profitabilnosti. Ovo podrazumijeva da je klasa svakog kupca već poznata u

skupu za obučavanje. Međutim, ovaj skup podataka ne sadrži informacije o novim, odnosno potencijalnim kupcima, za čiju je klasifikaciju kompanija zainteresovana.

S tim u vezi, *Larose* i *Larose* (2015) navode da je potrebno da se model zaštiti od memorisanja skupa podataka za obučavanje modela, u cilju izbjegavanja primjene obrazaca uočenih u ovom skupu na skupu novih i nepoznatih podataka. Stoga se nakon obučavanja modela, on primjenjuje na testnom skupu s nepoznatom klasom kupaca, na osnovu čega se procjenjuje njegova tačnost. S obzirom na to da su podaci o zavisnoj varijabli sakriveni u testnom skupu, njegova evaluacija vrši se upoređivanjem predviđenih vrijednosti u testnom skupu i stvarno zabilježenih vrijednosti zavisne varijable. Posljednji korak u ovom procesu obuhvata prilagođavanje modela u cilju minimizacije stope greške na testnom skupu (Larose & Larose, 2015).

Pri sprovođenju dvostruke kros-validacije potrebno je donijeti odluku koji će se dio podataka koristiti za skup za obučavanje, a koji za testiranje modela. Odabir većeg skupa za obučavanje dovešće do manjeg skupa za testiranje i obratno. Ukoliko je testni skup previše mali, tada će procjena performansi imati veliku varijaciju. S druge strane, ako je skup za obučavanje modela previše mali, tada će evaluaciju karakterisati velika pristrasnost. S tim u vezi, u literaturi se navodi preporuka da se za obučavanje koriste dvije trećine podataka, a preostala trećina za validaciju modela (Rhys, 2020). Međutim, važno je istaći da i ova proporcija zavisi od broja primjera u cjelokupnom skupu podataka.

Jedan od nedostataka primjene dvostruke kros-validacije, odnosno slučajnog izbora skupa podataka za obučavanje i testiranje za procjenu prediktivnih performansi modela, odnosi se na pristrasnost. Minimizacija pristrasnosti ovog tipa realizuje se kroz *k-fold* kros-validaciju, koja je ukratko opisana u sekciji 4.3. U ovom procesu, cjelokupni skup podataka se nasumično dijeli na k međusobno isključivih podskupova približno jednake veličine, a zatim se model obučava i testira k puta (Olson & Delen, 2008; Rhys, 2020; Rokach & Maimon, 2015). U svakoj se iteraciji obučavanje modela vrši korišćenjem $k-1$ podskupova, pri čemu se jedan preostali podskup koristi za testiranje modela, a procjena sveukupne tačnosti modela računa

se kao prosjek individualnih mjera tačnosti dobijenih iz k modela. Formula za računanje kros-validacione tačnosti data je u nastavku:

$$CVA = \frac{1}{k}\sum_{i=1}^{k} A_i$$

gdje je CVA kros-validaciona tačnost, k broj korišćenih iteracija, A mjera tačnosti pojedinačnih iteracija (poput senzitivnosti ili specifičnosti – detaljnije u sekciji 4.2.1).

Kros-validaciona tačnost direktno bi zavisila od slučajnog nasumičnog raspoređivanja pojedinačnih slučajeva u k različitih podskupova, pa je uobičajena praksa da se sami podskupovi stratifikuju. U slučaju stratifikovane *k-folds* procedure kros-validacije, svaki podskup u svakoj iteraciji sadrži približno isti udio klasa kao u originalnom skupu podataka. Dakle, proces stratifikacije se često primjenjuje u kros-validaciji, kako bi se osiguralo da je distribucija klasa iz cjelokupne baze podataka sačuvana i preslikana na skupove za obučavanje i testiranje. Ovaj proces pokazao se značajnim za smanjenje varijanse procijenjene greške u bazama podataka s velikim brojem klasa (Rokach & Maimon, 2015). Osim ove prednosti stratifikovane procedure, empirijska istraživanja su potvrdila i generisanje rezultata s nižom stopom pristrasnosti u odnosu na klasični proces kros-validacije (Kohavi, 1995; Olson & Delen, 2008).

Ukoliko se kao primjer uzme skup podataka sa 100 instanci i dvije klase, pri čemu pozitivna klasa sadrži 80 primjera, a negativna 20, slučajno uzorkovanje bez stratifikacije moglo bi da dovede do situacije da skup podataka za validaciju modela sadrži samo pozitivne primjere ili samo negativne. S druge strane, uz uključivanje stratifikacije, u slučaju *10-fold* kros-validacije, svaki validacioni podskup sadržaće (približno) osam pozitivnih i dva negativna primjera, odslikavajući odnos klasa u cijelom skupu (Berrar, 2018). Dakle, kako skup podataka za obučavanje modela predstavlja uzorak iz interesne populacije i odslikava odnos klasa u cjelokupnom skupu, tako i podskupovi koji se koriste za evaluaciju modela treba da odražavaju

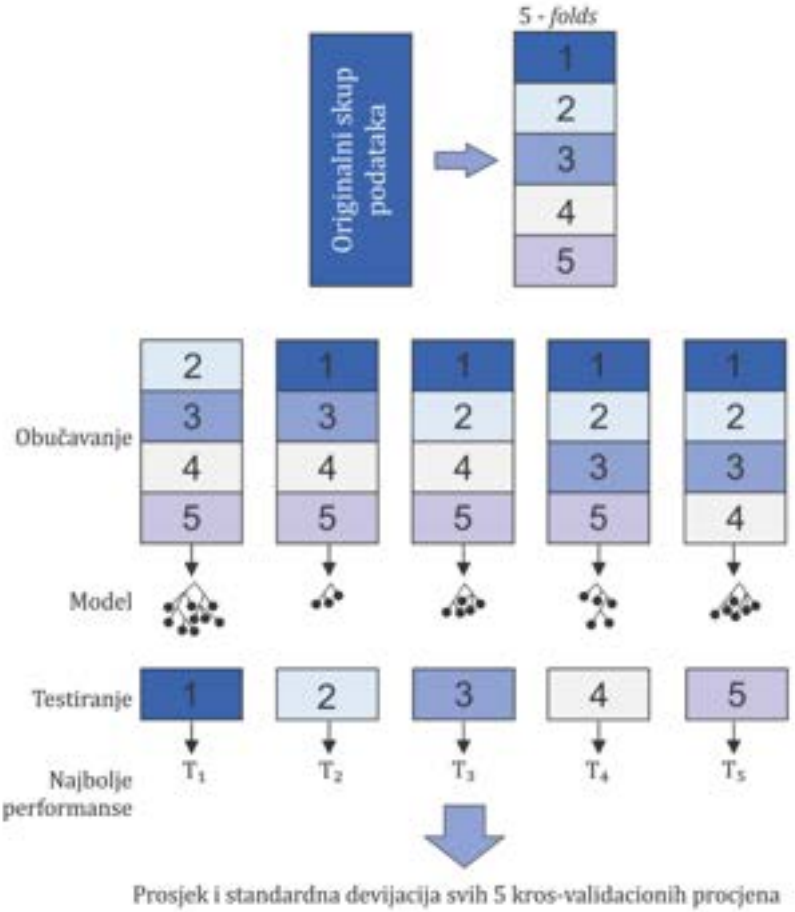
ovaj odnos. Na ovaj način se izbjegava pristrasna evaluacija i potvrđuje se potreba za uključivanjem stratifikacije u kros-validacioni proces.

Praksa je pokazala da se za *k* najčešće uzimaju vrijednosti pet ili 10. Autori navode *10-fold* kros-validaciju kao najčešće korišćenu, navodeći da je 10 optimalan broj iteracija koji optimizuje vrijeme potrebno za izvršenje testa, kao i za rezultirajući stepen pristrasnosti i varijanse u validacionom procesu (Breiman et al., 1984; Olson & Delen, 2008). *Arlot* i *Celise* (2010) preporučuju definisanje vrijednosti za *k* između pet i 10, navodeći da se statističke performanse ne mogu značajno unaprijediti za veće vrijednosti *k*, dok izračunavanje prosjeka za manje od 10 podjela ostaje i dalje računski izvodljivo.

Proces *k-fold* kros-validacije obavlja se kroz tri faze (Olson & Delen, 2008):

1. Kompletan skup podataka nasumično se dijeli na *k* nepovezanih podskupova, od kojih svaki sadrži približno isti broj zapisa. Kao što je prethodno istaknuto, uzorkovanje se obavlja uz stratifikaciju, kako bi se osiguralo da je proporcionalna zastupljenost klasa otprilike ista kao u izvornom skupu podataka;
2. Za svaki podskup se konstruiše klasifikator, koji koristi sve zapise iz ostalih *k-1* podskupova za obuku. Zatim se klasifikator testira na preostalom podskupu, kako bi se dobila kros-validaciona procjena njegove stope greške. Ovaj rezultat se bilježi;
3. Nakon ponavljanja drugog koraka za svih *k* podskupova, traži se prosjek svih *k* kros-validacionih procjena, kako bi se utvrdila agregirana kros-validaciona tačnost.

Na narednoj slici predstavljen je proces za *k*=5.



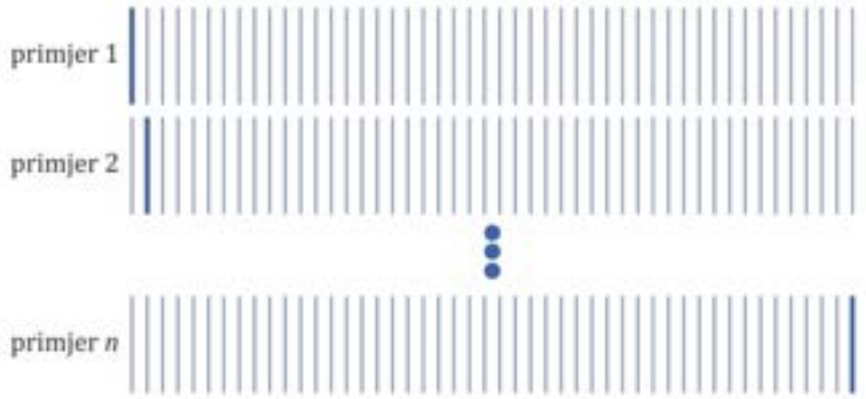
Slika 28. Ilustracija *5-fold* kros-validacije (prilagođeno prema: Provost & Fawcett, 2013)

Prateći prethodno opisane faze sprovođenja kros-validacije, na ilustraciji je predstavljen proces za pet definisanih podskupova. Dakle, svaki od podskupova se kroz iteracije koristi kao skup podataka za testiranje, pri čemu se ostala četiri koriste u procesu obučavanja. Konačan rezultat tačnosti i varijanse dobija se na osvnou prosjeka svih ostvarenih rezultata tačnosti u svakoj pojedinačnoj iteraciji (Provost & Fawcett, 2013).

Kros-validacija je pogodna za korišćenje u ograničenim skupovima podataka. Naime, za razliku od razdvajanja podataka u skup za obučavanje i skup za testiranje,

kros-validacija izračunava svoje procjene za sve podatke izvođenjem više podjela i sistematskom zamjenom uzoraka za testiranje.

Dodatno, pored dvostruke kros-validacije i *k-fold* kros-validacije, *Rhys* (2020) navodi i kros-validaciju sa izostavljanjem jedne instance (eng. *leave-one-out cross-validation*), kao ekstremnu formu ove tehnike. U ovom slučaju, umjesto podjele podataka u podskupove, samo jedna instanca se izdvaja iz skupa podataka za testiranje, a model se obučava na ostatku skupa. Nakon obučavanja, model se testira na izdvojenom primjeru i bilježe se relevantne metrike performansi. Ovaj proces se ponavlja sve dok se ne iskoristi i posljednja instanca za testiranje modela, nakon čega se traži prosjek metrika performansi, kao i u *k-fold* kros-validaciji. Ilustracija procesa kros-validacije sa izostavljanjem jedne instance prikazana je na Slici 29.



Slika 29. Ilustracija "*leave-one-out*" kros-validacije (prilagođeno prema: Rhys, 2020)

Kros-validacija sa isključivanjem jedne instance ima tendenciju da daje promjenljive procjene performansi modela, s obzirom na to da procjena performansi u svakoj iteraciji direktno zavisi od tačnog označavanja jednog testnog primjera. Međutim, ova procedura može pružiti bolje performanse od *k-fold* kros-validacije u slučajevima primjene na veoma malim bazama podataka. Kada se *k-fold* kros-validacija primjenjuje na malim skupovima podataka, podjela na *k* podskupova generiše veoma male skupove za obučavanje. S tim u vezi, varijansa modela obučenog na malom skupu podataka ima tendenciju da bude veća, jer će na nju više uticati greška uzorkovanja i neobični slučajevi. Stoga je ova specifična forma kros-

validacije korisna za male skupove podataka, gdje se kao dodatna prednost ističe i manja potreba računskog vremena (Rhys, 2020).

Konačno, prednost kros-validacije, generalno posmatrano, ogleda se u mogućnosti primjene u različitim algoritmima i namjenama, kako u regresiji, tako i u klasifikaciji.

4.4 Izbor optimalne kombinacije parametara – Grid Search tehnika

Grid-Search (mrežno pretraživanje) tehnika predstavlja proces skeniranja podataka u cilju konfiguracije optimalnih parametara za dati model. Ova tehnika optimizacije parametara ne odnosi se na specifičan tip modela, već se može primijeniti u mašinskom učenju za izračunavanje najboljih parametara za bilo koji dati model, a parametri koji će se tražiti zavise od konkretnog tipa modela. Optimizacija parametara može oduzeti mnogo vremena ako se obavlja ručno, posebno ako algoritam ima mnogo parametara. Dodatno, većina algoritama mašinskog učenja neće postići optimalne rezultate ako se njihovi parametri ne podese pravilno. U ovom dijelu rada biće opisan proces *Grid-Search* tehnike, koja izgrađuje model na svakoj zadatoj kombinaciji parametara koji su predmet optimizacije.

Parametri modela predstavljaju eksterne karakteristike modela, čija se vrijednost ne može procijeniti iz samih podataka (Joseph, 2018). Dakle, vrijednosti parametara moraju se podesiti prije otpočinjanja procesa obučavanja modela. Na primjer, u slučaju razvijanja SVM modela, to može biti parametar *C* ili, u slučaju DT modela – maksimalna dubina stabla, čija se optimalna vrijednost može tražiti kroz *Grid-Search*, a koja rezultira predikcijama najveće tačnosti. Vrijednost parametara definiše tačnost modela. Kod SVM parametra, na primjer, parametar *C* određuje *trade-off* između minimiziranja greške i maksimiziranja margine klasifikacije (Tharwat, 2019), pa samim tim utiče na vrijednost greške, broj vektora oslonca i samu marginu.

Grid-Search koristi mrežu parametara (eng. *grid*) koja se pretražuje i iz koje se bira najbolja kombinacija parametara u skladu s raspoloživim podacima. Dakle, ovaj alat radi pod pretpostavkom da postoji specifična kombinacija vrijednosti različitih parametara, koji će minimizirati grešku prediktivnog modela (Korstanje, 2020). *Grid-Search* vrši testiranje različitih kombinacija parametara i bira najbolju. Važno je istaći da nije moguće testirati svaku moguću kombinaciju, jer bi za kontinualnu skalu bilo potrebno beskonačno mnogo kombinacija za testiranje. U tom smislu, ova tehnika bazirana je na kreiranju mreže, koja definiše određene vrijednosti koje treba testirati. Na narednoj slici dat je šematski prikaz *Grid-Search* pretrage za dva parametra – Alfa i Beta.

	Alfa 0.1	0.01	0.001	0.0001
Beta 0.1	Tačnost 0.716	Tačnost 0.76	Tačnost 0.81	Tačnost 0.78
0.01	Tačnost 0.721	Tačnost 0.81	Tačnost 0.94	Tačnost 0.86
0.001	Tačnost 0.709	Tačnost 0.88	Tačnost 0.80	Tačnost 0.73
0.0001	Tačnost 0.72	Tačnost 0.69	Tačnost 0.71	Tačnost 0.70

Slika 30. Šematski pregled *Grid-Search* tehnike za dva parametra *Alfa* i *Beta* (prilagođeno prema: Korstanje, 2020)

Na Slici 30 predstavljen je primjera *Grid-Search* testiranja parametara *Alfa* i *Beta*. Definisane vrijednosti za testiranje u okviru mreže su 0,1, 0,01, 0,001, 0,0001, a rezultirajuće vrijednosti svih kombinacija predstavljene su u odgovarajućim poljima

matrice. Nakon testiranja svih polja u okviru predviđene mreže, model će izabrati kombinaciju koja daje (u ovom slučaju) najveću tačnost.

Ova tehnika često se koristi u kombinaciji sa kros-validacijom. Za slučaj SVM optimizacije parametara, izbor C i γ parametara, korišćenjem *k-fold* kros-validacije, vrši se na sljedeći način (Syarif et al., 2016):

1. Raspoloživi skup podataka se dijeli na k podskupova, pri čemu se jedan podskup čuva za testiranje modela, dok se model obučava na preostalih $k-1$ podskupova;
2. Kros-validaciona tačnost za SVM klasifikator računa se korišćenjem različitih vrijednosti parametara C i γ – različite kombinacije parametara se testiraju i bira se ona s najboljom kros-validacionom tačnošću i koristi za treniranje modela na cijelom skupu podataka.

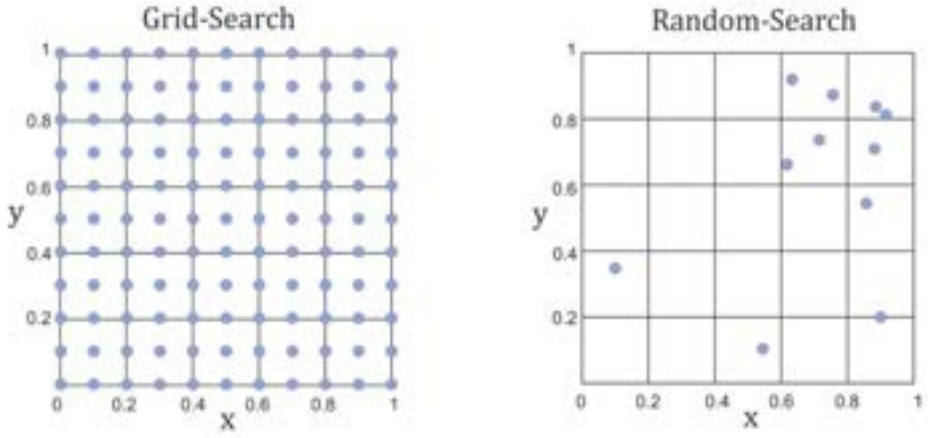
Kao najznačajniji nedostatak *Grid-Search* tehnike navodi se sporo obavljanje procesa, posebno u situacijama kada se model obučava i testira na velikim skupovima podataka i velikim rangovima definisanih parametara (Liashchynskyi & Liashchynskyi, 2019; Syarif et al., 2016). Na primjer, ukoliko se optimizuje pet parametara sa 25 iteracija za svaki od njih, ukupan broj kombinacija bi bio 25^5, odnosno 9.765.625, što zahtijeva značajno računsko vrijeme.

Kako bi se prevazišao problem značajnog računskog vremena potrebnog za optimizaciju parametara *Grid-Search* tehnikom, *Lameski et al.* (2015) su predložili sprovođenje *Grid-Search* procedure u dvije faze. U prvoj fazi, koristeći eksponencijalno rastuće korake za pretraživanje mreže, pronašli su odgovarajuće vrijednosti za C i γ parametre za SVM klasifikaciju. Zatim, koristeći manje rangove, tražili su vrijednosti za C i γ blizu vrijednosti utvrđenih u prethodnom koraku. Rezultati su pokazali unapređenje prediktivnih performansi modela nakon odgovarajućeg podešavanja parametara, a uz to, pokazali su stabilnost modela i izbjegavanje problema prekomjernog prilagođavanja podacima.

Pored *Grid-Search* tehnike, proces optimizacije parametara može se sprovesti i ručno, kao i *Random-Search* tehnikom (Zahedi et al., 2021).

Kako je navedeno u uvodnom dijelu ovog poglavlja, ručno podešavanje i pretraga optimalnih parametara predstavlja najbazičniju metodu pronalaženja parametara za model mašinskog učenja. Ovaj pristup obuhvata testiranje različitih vrijednosti parametara, pri čemu se izbor vrijednosti za testiranje bazira na pukom pogađanju ili, s druge strane, poznavanju domena. Ovaj proces se ponavlja dok se ne pronađe kombinacija parametara koja omogućava zadovoljavajuće poboljšanje modela. Međutim, kompleksna priroda modela mašinskog učenja i potencijalni broj kombinacija za testiranje čine ovaj proces nepraktičnim (Steinholtz, 2018).

Random-Search predstavlja još jednu često korišćenu metodu za podešavanje parametara. Za razliku od *Grid-Search* tehnike, *Random-Search* bira konfiguracije parametara slučajnim izborom, bez ispitivanja svake moguće kombinacije, te ponavlja proces dok se ne iscrpe svi resursi (Zahedi et al., 2021). Ova tehnika je brža od *Grid-Search* tehnike, samim tim što ne testira sve kombinacije. Međutim, zbog ovakvog principa rada, moguća je situacija da se ne pronađe optimalna kombinacija parametara. Na Slici 31 dat je uporedni prikaz *Grid-Search* i *Random-Search* tehnika.



Slika 31. Ilustracija *Grid-Search* i *Random-Search* pristupa

S obzirom na to da različiti izbori parametara mogu rezultirati izrazito različitim performansama, pronalaženje optimalnih vrijednosti parametara za modele obično je zahtjevan zadatak, a ključan s tačke gledišta performansi modela. U tom smislu, poželjan je automatizovani pristup, poput *Grid-Search* tehnike, koja omogućava efikasno pronalaženje optimalnih parametara na bazi kojih će se model adekvatno obučavati.

Nakon teorijskog pregleda metoda koje će biti primijenjene u empirijskom dijelu distertacije, kao i pregleda prethodnih istraživanja koja su koristila navedene metode, u narednom dijelu rada biće predstavljen koncept tri prediktivna modela za targetiranje kupaca: prediktivna segmentacija kupaca, predikcija odgovora kupca i predikcija profitabilnosti kupca. Cilj predloženog koncepta je prevazilaženje problema minorne klase najvrednijih kupaca i respondenata, kao i asimetričnosti profitabilnosti. Na ovom konceptu će biti zasnovan i empirijski dio istraživanja. Naredno poglavlje najprije opisuje prvi od tri navedena modela, tj. konceptualni model prediktivne RFM segmentacije baziran na *k-means* klasterizaciji i SVM-RE metodu.

4.5 Koncept predloženih prediktivnih metoda

U ovom dijelu rada biće predstavljen koncept tri prediktivna modela odlučivanja u direktnom marketingu bazirana na SVM metodi: model RFM segmentacije, tj. model za predikciju stepena vrijednosti kupca, model odgovora na kampanju, tj. model za predikciju odgovora kupca i model targetiranja na osnovu profitabilnosti, tj. model za predikciju profitabilnosti kupca.

4.5.1 Koncept modela prediktivne RFM segmentacije baziran na SVM metodi

U ovoj sekciji predložen je model za prediktivnu segmentaciju na osnovu vrijednosti kupaca. U kampanji će biti targetirani kupci s najvećim stepenom vrijednosti. Polazi se od skupa podataka o kupovnim transakcijama iz direktnih kampanja, koji može sadržati podatke o kupcima, podatke o proizvodima, kao i podatke o kupovnom

ponašanju u vidu RFM atributa. RFM atributi su definisani na sljedeći način: R - datum posljednje porudžbine, F - ukupan broj transakcija u razmatranom periodu i M - novčani iznos koji je potrošen u razmatranom periodu. Atribut R je kodiran tako što je za 20% najskorijih datuma dodijeljen skor 5, sljedećih 20% datuma dobija skor 4 i tako dalje do skora 1. Atributi F i M su zadržani u originalnom obliku.

20

Koristeći RFM atribute, pomoću *k-means* klasterizacije kupci se dijele u klastere, pri čemu pripadnici klastera imaju slično kupovno ponašanje, a zatim se kupcima pojedinih klastera dodjeljuje odgovarajući stepen vrijednosti (eng. *customer value* - CV).

CV-nivo kupca (tj. pripadnost odgovarajućem klasteru) se zatim predviđa pomoću prediktivne klasifikacione metode. Na ovaj način je moguće, ako postoje podaci o kupcu i proizvodima koji mu se nude u kampanji, predvidjeti njegov CV-nivo, te na osnovu njega donijeti odluku da li ga targetirati ili ne. Prilikom klasifikacije kupaca na osnovu vrijednosti, prisutan je problem male, a najvažnije klase (klasa najvrednijih kupaca kojih je po pravilu najmanje). Za potrebe direktnog marketinga je korisno otkriti i pravila koja opisuju klastere kupaca s većim CV-nivoom u terminima podataka o kupcima i proizvodima koje preferiraju. Ta pravila se mogu koristiti za grupno targetiranje kupaca, kao i za targetiranje novih potencijalnih kupaca. Stoga je od velikog značaja da prediktivna klasifikaciona metoda bude interpretabilna.

U prethodnoj literaturi već postoje modeli prediktivne segmentcije, međutim, prisutni su određeni nedostaci koje ovdje predloženi koncept nastoji da prevaziđe. U svom istraživanju, *Cheng* i *Chen* (2009) su koristili *k-means* klasterizaciju za segmentaciju kupaca pomoću skaliranih RFM atributa, koristeći jedinstven pristup skaliranja (podjelom podataka na segmente od po 20%). S obzirom na to da *k-means* klasterizacija funkcioniše s numeričkim atributima, u ovoj disertaciji se predlaže da samo atribut R (tj. datumi) treba da se skalira. Na ovaj način se izbjegava gubitak važnih informacija i subjektivnost u procjeni da li najviše ocjene za F i M treba dodijeliti prvih 20% ili više/manje kupaca. Osim toga, navedeni autori su testirali pristup kreiranjem tri, pet i sedam klastera, dok ova disteracija sugeriše ocjenu

optimalnog broja klastera zasnovanu na *Davies-Bouldin* indeksu, koji garantuje maksimalnu homogenost unutar klastera i maksimalnu heterogenost između klastera.

Koristeći karakteristike kupaca (region i kreditni dug, u ovom slučaju), *Cheng* i *Chen* (2009) koriste *rough set* i metodu ekstrakcije pravila LEM2 da bi generisali skup eksplicitnih pravila koja se mogu koristiti za targetiranje kupaca. Da bi se postigla visoka stopa tačnosti, prediktivni atributi uključuju i RFM atribute koji imaju najveći uticaj na klasifikaciju, jer se već na ovoj osnovi formiraju klasteri. Zbog toga, pravila možda neće pokazati neke karakteristike kupaca koje su veoma važne u targetiranju (mogu se apsorbovati efektom RFM atributa). U ovoj disertaciji, karakteristike kupaca i podaci o proizvodima se koriste kao prediktivni atributi, koji mogu da obezbijede pravila predviđanja s korisnijim informacijama za selekciju i targetiranje kupaca (Tsai & Chiu, 2004).

Da bi procijenili performanse klasifikacije *rough set* LEM2 metode, *Cheng* i *Chen* (2009) su koristili isključivo stopu tačnosti. Klasteri ne sadrže isti broj kupaca, jer je obično najmanji broj kupaca s najvišim CV-nivoom. Dakle, kod prediktivne klasifikacije postoji problem neuravnoteženosti klasa, što može dovesti do niske preciznosti klase (procenat precizno predviđenih primjera u okviru predviđene klase) i/ili odziva klase (procenat tačno klasifikovanih primjera u okviru trenutne klase) za najmanji segment koji je najvažniji za ovaj istraživani problem. S tim u vezi, s obzirom na to da SVM metoda uspješno rješava problem neravnoteže klasa, ona se i u ovoj disertaciji predlaže za takvo predviđanje. Uzimajući u obzir da SVM ne generiše pravila neophodna za grupno targetiranje kupaca, predložena je hibridna metoda ekstrakcije SVM pravila pomoću DT metode, koja tretira problem neinterpretabilnosti, tj. SVM-RE metoda. Detaljan postupak predložene prediktivne procedure izložen je u nastavku.

Procedura prediktivne klasifikacije

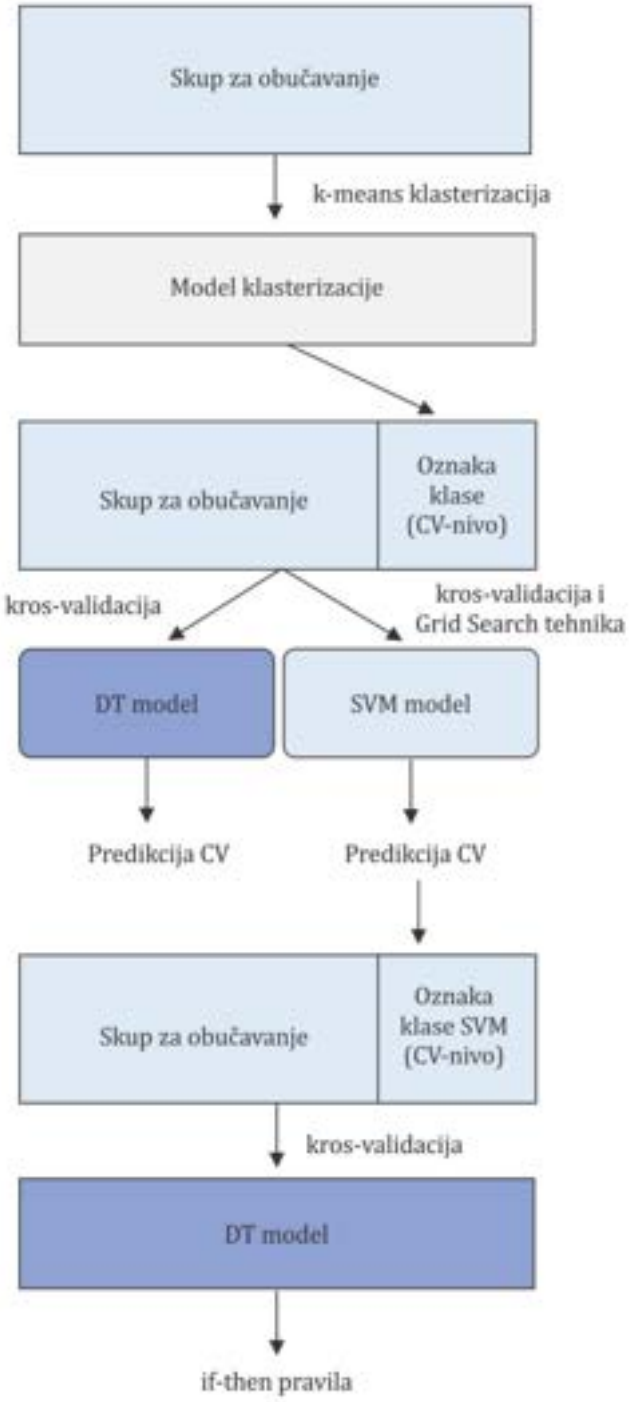
Za ocjenu performansi prediktivne klasifikacije biće korišćena stopa tačnosti (*mean accuracy rate*), *class precision* i *class recall*, dobijeni postupkom *k-fold* kros-validacije sa stratifikovanim uzorkovanjem. Osim generisanja prediktivnih pravila s visokom tačnošću, za targetiranje kupaca je važno i da se postojeći kupci što tačnije klasifikuju, pa je *class recall* važan pokazatelj performansi modela. Na svakoj iteraciji kros-validacije se izračunavaju parametri za klasifikacione performanse (tačnost, *class precision* i *class recall*) i na kraju nalazi njihova prosječna vrijednost.

Predložena procedura prediktivne klasifikacije kupaca po CV-nivou sastoji se iz sljedećih koraka:

1. Priprema podataka (izračunavanje RFM atributa);
2. Segmentacija kupaca, tj. transakcija po normalizovanim RFM atributima pomoću *k-means* klasterizacije, uz ocjenu *centroid cluster* modela na osnovu *Davies-Bouldin* indeksa;
3. Opis klastera putem RFM atributa i dodjela CV-nivoa kupcima na osnovu pripadnosti odgovarajućem klasteru (polaznom skupu podataka se dodaje oznaka klase *Cluster*);
4. Generisanje DT modela za predviđanje CV-nivoa na osnovu atributa o kupovnim transakcijama i oznake klase *Cluster*. Ovaj korak podrazumijeva pronalaženje optimalnih parametara za DT model (kriterijum za evaluaciju podjele, *minimal size for split, minimal leaf size, minimal gain, maximal depth*) u postupku *k-fold* kros-validacije, tako da se dobije najveća tačnost predikcije;
5. Generisanje SVM modela za predviđanje CV-nivoa na osnovu atributa o kupovnim transakcijama. Treniranje SVM-a zahtijeva da se utvrdi optimalna kombinacija parametara γ iz RBF kernel funkcije i širine margine C, koja će dati najveću prosječnu tačnost klasifikacije u toku *k-fold* kros-validacije. Za izbor optimalne kombinacije parametara koristi se *Grid-Search* pristup;

6. Generisanje CV-nivoa na osnovu SVM predikcije (polaznom skupu podataka se dodaje oznaka klase SVM). Na ovom koraku se kupcima dodjeljuje CV-nivo, koji je predvidio SVM klasifikator;
7. Generisanje DT modela za predviđanje CV-nivoa na osnovu atributa o kupovnim transakcijama i SVM oznake klase. DT koji se generiše na ovom koraku je interpretator SVM modela, tj. on obavlja ekstrakciju pravila iz SVM modela. Slično kao u četvrtom koraku, na osnovu *k-fold* kros-validacije se biraju optimalni DT parametri za maksimalnu tačnost. U poređenju sa DT modelom iz tačke 4, ovaj DT model bi trebalo da ima značajno veću prediktivnu tačnost;
8. Generisanje *if-then* pravila iz DT modela dobijenog u prethodnom koraku. Ova pravila eksplicitno ukazuju na atribute iz kupovnih transakcija koji predviđaju odgovarajući CV-nivo i na taj način mogu pomoći da se kupci lakše targetiraju.

Ova prediktivna procedura se može šematski prikazati kao na Slici 32.



Slika 32. Šematski prikaz prediktivne procedure

Kako je već navedeno, *ensemble* metode mogu poboljšati tačnost klasifikatora, a samim tim i performanse predloženog modela za prediktivnu segmentaciju. U sljedećoj sekciji se predlaže koncept koji uključuje ovo poboljšanje.

4.5.2 Koncept modela prediktivne RFM segmentacije sa ensemble metodama

U ovom dijelu rada biće opisan koncept modela prediktivne RFM segmentacije, koji je baziran na SVM-RE metodi u kombinaciji sa *ensemble* metodama u cilju poboljšanja njegovih performansi.

Kao što je već istaknuto u prethodnim istraživanjima, da bi se prevazišao problem neravnoteže klasa, uglavnom su testirane balansirane *ensemble* metode u kombinaciji s različitim klasifikatorima ili samostalni SVM, kao pretprocesor koji eliminiše preklapanje klasa i time balansira podatke. Ovdje se predlaže kombinacija *ensemble* pristupa i SVM metode za poboljšanje performansi pretprocesiranja, kao i balansirani *ensemble* u kombinaciji sa DT metodom na prethodno obrađenom skupu podataka, za poboljšanje performansi ekstrakcije pravila iz SVM izlaza, što bi na kraju trebalo da dovede do poboljšanja performansi klasifikacije segmenata korisnika.

Prediktivna procedura se odvija na sljedeći način: priprema podataka i CV stepen kupca se određuje na isti način kao u prethodnoj sekciji (klasterizacijom na osnovu RFM atributa). Zatim se *Bagging* SVM obučava kao pretprocesor podataka. Na skupu za obučavanje pretprocesor predviđa nivo CV. Da bi se dobili što čistiji klasteri, s manje preklapanja i postigao bolji balans klasa, uzimaju se samo oni rezultati za koje je u *Bagging* postupku glasalo više od 90% SVM modela, tj. rezultati za koje je *Confidence* > 0,9. Na ovaj način se dobija poduzorkovani skup podataka za obučavanje modela s novom oznakom klase (eng. *class label*) koju predviđa *Bagging SVM*. Nova oznaka klase definiše klase koje se manje preklapaju i uravnoteženije su.

Na ovom *Bagging SVM* izlazu obučen je balansirani *Bagging DT* model. Model je sada obučen na mnogo balansiranijim podacima, a balansirani *ensemble* meta-algoritam dodatno pomaže u rješavanju problema minorne klase (najvrednije klase kupaca) i poboljšava performanse klasifikacije segmenata kupaca. Pored toga, balansirani *Bagging DT* model generiše pravila koja precizno opisuju segmente kupaca, što je od posebnog značaja za minornu klasu. Naime, rješavajući problem manje klase, dobija se znatno više pravila za najvažniji i najvredniji segment kupaca.

Dakle, konačni model za klasifikaciju segmenata kupaca stvoren je kombinovanjem *ensemble* SVM klasifikatora i *ensemble* DT ekstraktora pravila, tako da se može nazvati *ensemble* SVM-RE modelom.

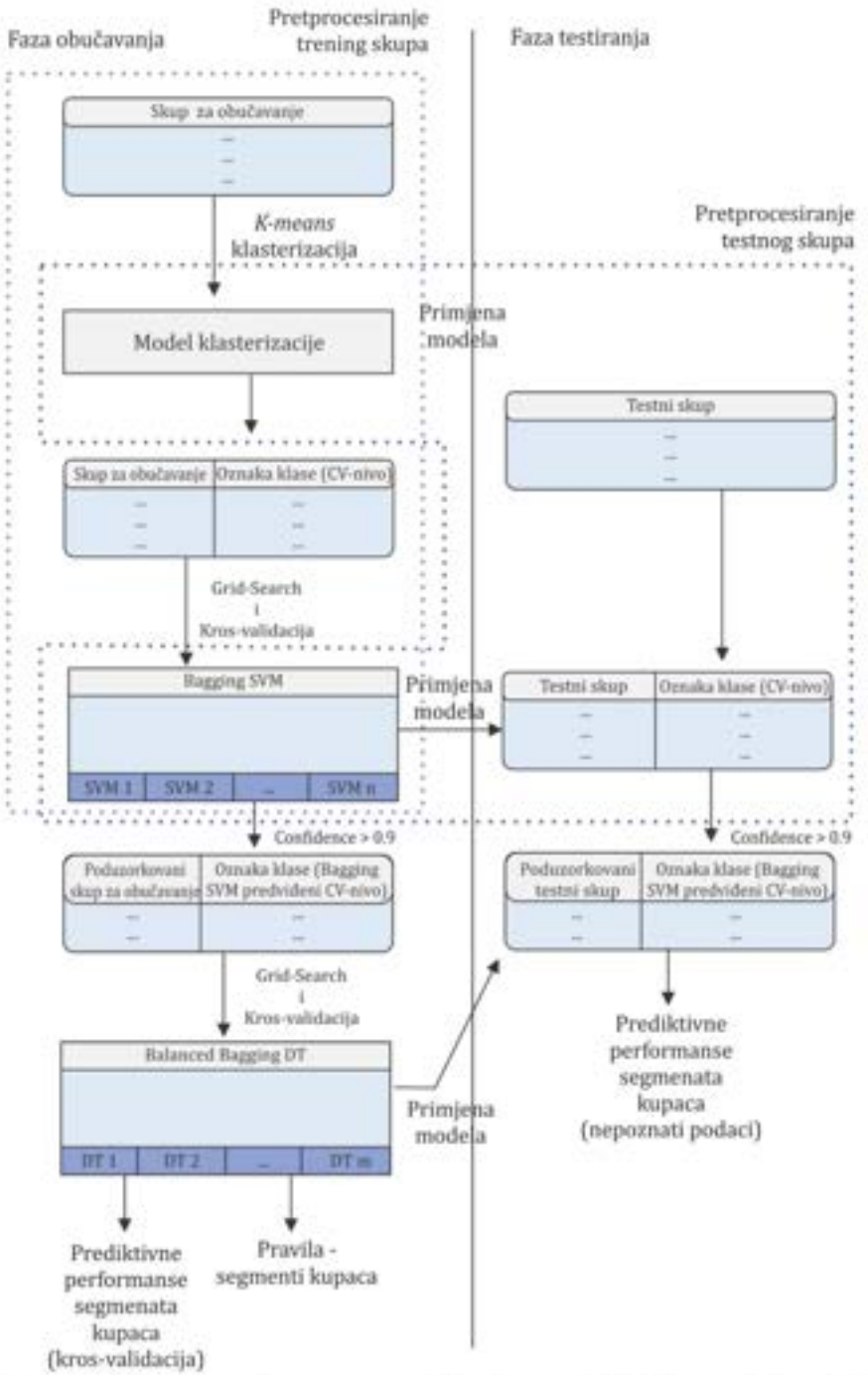
Obučavanjem modela, pronalazi se optimalna kombinacija parametara modela, koja postiže maksimalne prediktivne performanse. To se postiže kombinovanjem tehnologije *Grid-Search* sa *k-fold* kros-validacijom.

Za procjenu prediktivnih performansi, koriste se stepen tačnosti (eng. *accuracy rate*), preciznost klase (eng. *class precision)* i odziv klase *(*eng. *class recall)*. Pored velike preciznosti predviđanja segmenta kupaca, za targetiranje kupaca važno je tačnije klasifikovati postojeće potrošače, pa je *class recall* važan pokazatelj performansi modela.

U fazi testiranja, kako bi se procijenila tačnost modela na testnom skupu, stvarni CV-nivo se određuje pomoću klaster modela generisanog u fazi obučavanja modela. Zatim se CV-nivo predviđa pomoću obučenog *Bagging* SVM modela, a testni skup zadržava primjere čija predviđanja imaju *Confidence* > 0,9. Predviđeni CV-nivo tada se proglašava oznakom klase. Obučeni balansirani *Bagging* DT model se zatim primjenjuje na ovaj prethodno obrađeni testni skup i tako se dobijaju predviđanja za CV-nivo.

U fazi testiranja, prediktivne performanse modela određuju se upoređivanjem CV-nivoa, koji je dobijen predviđanjem pomoću obučenih modela, sa stvarnim vrijednostima CV-nivoa u testnom skupu.

Slika 33 prikazuje dijagram toka za faze obuke i testiranja prediktivnog postupka.



Slika 33. Prediktivna procedura segmentacije bazirana na SVM-RE i *ensemble* metodama

Postupak će biti sproveden pomoću *Rapid Miner* alata i realizovaće se kao proces spreman za upotrebu od strane korisnika. Rezultati ovog postupka biće predstavljeni u empirijskom dijelu rada, u sekciji 5.3.

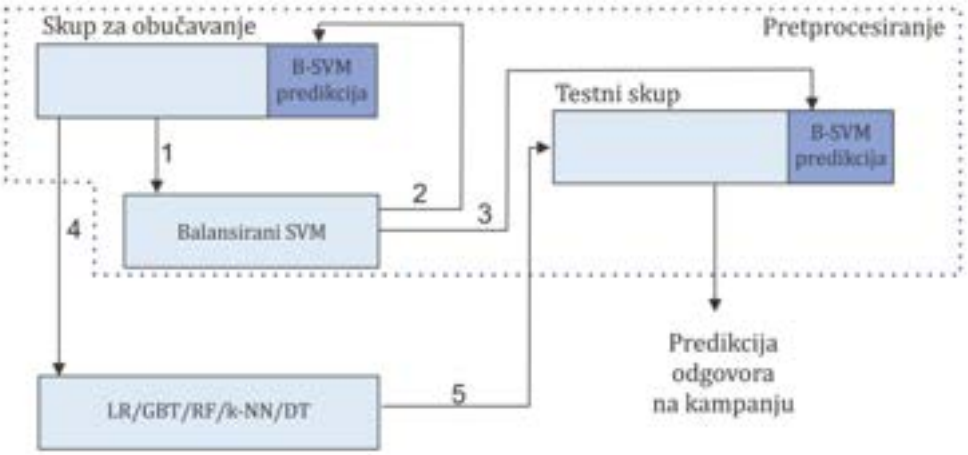
4.5.3 Koncept modela za predikciju odgovora kupca baziran na SVM metodi

Da bi se tretirao problem neravnoteže klasa u modeliranju odgovora kupca na kampanju, koji onemogućava klasifikacione algoritme da tačno prepoznaju primjere pozitivne (manje) klase respondenata, za pretprocesiranje podataka primijenjena je kombinacija slučajnog poduzorkovanja i SVM klasifikacije. Ovo poduzorkovanje, u svom najosnovnijem obliku, nasumično uklanja primjere veće klase iz baze podataka. U ovom slučaju, poduzorkovanje 1:1 (isti broj primjera kao u manjoj klasi) je sprovedeno na skupu podataka za obučavanje modela, čime je generisan balansirani SVM (B-SVM) pretprocesor. Za razliku od modela iz prethodne sekcije, koji je takođe suočen s problemom minorne klase, kod modela ogovora na kampanju neravnoteža klasa je mnogo veća, jer je minorna klasa ekstremno mala (nekada procenat respondenata može biti ispod 0,5%, posebno u bazama *online* kupaca). Stoga je kod SVM metode, kao pretprocesora, nužno i balansirano poduzorkovanje.

Naime, u slučaju ekstremne klasne neravnoteže, kao što je slučaj s bazom podataka koja će biti korišćena u empirijskom dijelu ovog istraživanja, SVM je takođe pristrasan prema dominantnoj klasi (Kim et aL, 2013).

U prethodnoj sekciji je kao klasifikator na pretprocesiranim podacima predložen DT, zbog njegove interpretabilnosti, koja je u modelu prediktivne segmentacije veoma važna. S obzirom na to da je kod predikcije odgovora kupaca od najvećeg značaja tačnost, postavlja se pitanje da li neki drugi klasifikator može biti bolji u tom smislu. Stoga će na pretprocesiranim podacima, osim DT metode, biti testirani sljedeći klasifikatori: logistička regresija (Berkson, 1944), *Gradient Boosted Trees* (Friedman, 2001), *Random Forest* (Breiman, 2001) i *k-Nearest Neighbor* (Fix &

Hodges Jr., 1989). Također, treba napomenuti da klasifikatori nisu poboljšavani iterativnim slučajnim poduzorkovanjem s vraćanjem (kao kod modela iz prethodne sekcije), jer je balansiranje obavljeno u fazi pretprocesiranja. Predložena prediktivna procedura prikazana je na Slici 34.



Slika 34. Ilustracija prediktivne procedure modela odgovora na kampanju

Predložena prediktivna procedura sastoji se od sljedećih koraka (koraci 1, 2 i 3 uključuju pretprocesiranje podataka, dok je predikcija odgovora na kampanju realizovana u koracima 4 i 5):

- Korak 1 – B-SVM se obučava na polaznom skupu podataka za obučavanje koji je balansiran slučajnim poduzorkovanjem. Model s najboljim prediktivnim performansama dobija se u proceduri *k-fold* kros-validacije;
- Korak 2 - Obučeni B-SVM je primijenjen na cjelokupnom skupu podataka za obučavanje modela i njegova oznaka klase (eng. *class label*) je zamijenjena sa B-SVM predikcijom, čime je minorna klasa respondenata dopunjena najsličnijim primjerima iz klase nerespondenata i postignuto balansiranje klasa;
- Korak 3 - Obučeni B-SVM je primijenjen na originalnom neuravnoteženom skupu podataka za testiranje modela i njegova oznaka klase je zamijenjena sa B-SVM predikcijom, čime su u testnom skupu kupci najsličniji

respondentima proglašeni vjerovatnim respondentima i na taj način su balansirane klase;

- Korak 4 - Na modifikovanim (balansiranim) podacima za obučavanje modela iz Koraka 2, trenirani su različiti klasifikatori, kao što su: DT, LR, GBT, RF i k-NN. Modeli s najboljim prediktivnim performansama izabrani su pomoću *k–fold* kros-validacije;
- Korak 5 - Obučeni klasifikatori su primijenjeni na skupu podataka za testiranje modela. Prilikom ocjene prediktivnih performansi, umjesto originalne oznake klase, korišćena je B-SVM predikcija.

U empirijskom dijelu ovog istraživanja, za ocjenu performansi modela biće korišćene sljedeće metrike: AUC, tačnost, senzitivnost i *fallout*. Kako je od izuzetnog značaja procjena mogućnosti modela da razlikuje respondente od nerespondenata, prilikom odgovora na kampanju biće korišćena AUC metrika, koja je, kako je već istaknuto, u literaturi često korišćena da pokaže stepen odvajanja između klasa (Asare-Frempong & Jayabalan, 2017; Chen et al., 2011; D'Haen et al., 2013).

Proces koji realizuje predloženi prediktivni postupak biće generisan u *Rapid Miner* alatu i kao gotov proces biće spreman za upotrebu od strane krajnjih korisnika.

Kao što je ranije istaknuto, u literaturi su za prevazilaženje problema neravnoteže klasa prepoznate i *ensemble* metode u kombinaciji s poduzorkovanjem. U tom smislu, postavlja se pitanje da li ove metode mogu dati bolje rezultate u slučaju predikcije odgovora na kampanju od pristupa sa SVM pretprocesiranjem. Stoga je u narednoj sekciji predložen model baziran na njima, da bi se napravila komparacija s modelom baziranim na B-SVM pretprocesiranju. Takođe, s obzirom na to da skup podataka koji se koristi u ovom radu sadrži *online* transakcije s *web* metrikama, postavlja se pitanje koliko su one važne za predikciju odgovora kupca, te da li su važnije od klasičnih metrika kao što su RFM. Da bi se dobio odgovor na ovo pitanje, predložena je i procedura za testiranje uticaja ovih metrika na predikciju kod modela zasnovanih na *ensemble* metodama.

4.5.4 Koncept modela za predikciju odgovora kupca baziran na ensemble metodama

U *online* okruženju, uz mogućnost targetiranja većeg broja potencijalnih kupaca s nižim troškovima, jedan od glavnih izazova je stopa konverzije, uzimajući u obzir da je broj sesija ili posjeta sajtu koje se završavaju kupovinom zanemarljiv u odnosu na ukupan broj pristupa (Behera et al., 2020; Liu et al., 2019). Ovo stvara problem male klase, što predstavlja izazov za *data mining* modele. U ovom dijelu rada će biti predstavljene *ensemble* tehnike balansiranja podataka u cilju prevazilaženja ovog problema, kako bi se poboljšala efikasnost modeliranja odgovora kupaca. Pored toga, pitanje koje je manje istraženo, a koje je veoma relevantno u kontekstu rasta *online* trgovine, jeste da li je i u kojoj mjeri ponašanje korisnika na internetu važno za predviđanje odgovora korišćenjem metoda mašinskog učenja.

S obzirom na izazove u radu sa nebalansiranim podacima, pomoću ovog modela biće testirano koliko efikasno *ensemble* tehnike u kombinaciji s balansiranim klasifikatorima (u budućem tekstu *ensemble* balansiranih klasifikatora) mogu da riješe problem ekstremne neravnoteže klasa u predviđanju odgovora na kampanju kupaca u e-trgovini i da li to daje bolje rezultate od pristupa sa B-SVM pretprocesiranjem polaznog skupa podataka (iz prethodne sekcije), kao i u kojoj mjeri prediktivni učinak zavisi od *web* metrike, tj. ponašanja.

Preciznije, osnovni ciljevi predloženog koncepta su sljedeći:

1. Rješavanje problema klasne neravnoteže korišćenjem različitih *ensemble* balansiranih klasifikatora;
2. Komparativna analiza modela odgovora kupaca sa i bez *web* ponašanja;
3. Analiza modela odgovora kupaca sa i bez podataka o ponašanju prilikom kupovine i atributa o proizvodima;
4. Upoređivanje modela iz tačaka 3 i 4.

Aktivnosti koje se odnose na akviziciju i zadržavanje kupaca u okviru *online* trgovine su dio sistema internet marketinga (Schafer et al., 2001), a najčešće koriste *web log*

podatke i *clickstream* podatke, kao inpute za ovu aktivnost. Podaci o klikovima detaljno prikazuju navigaciju posjetilaca kroz *web* sajt, što može pružiti veoma značajan uvid u način na koji posjetioci koriste *web* sajt, otkriti stranice na kojima se najviše zadržavaju, kao i one koje se često ignorišu. Dakle, ova vrsta podataka bilježi interakciju korisnika i *web* sajta (ili aplikacije), tako da se svaki naredni klik bilježi i kreira niz podataka. Zbog pristupačnosti i niskih troškova prikupljanja, ovi podaci su našli primjenu u analizi elektronske trgovine, društvenih mreža, e-učenja, kao i analizi korišćenja internet portala i pretraživača, u cilju definisanja karakteristika korisnika, grupisanja i modeliranja njihovog ponašanja (Jiang et al., 2018).

U sekciji 3.3 naveden je pregled prethodnih istraživanja u kojima su korišćene *web* metrike, *web log* podaci i *clickstream* podaci za analizu ponašanja potrošača.

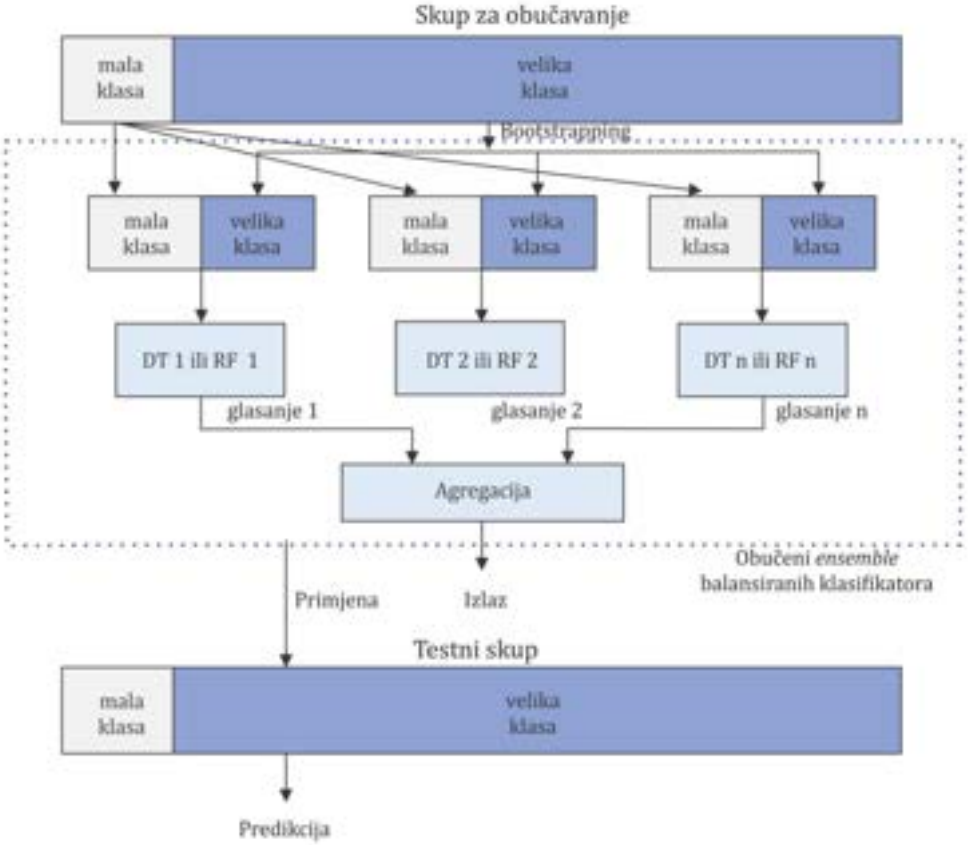
U literaturi je predloženo nekoliko metoda koje se bave problemom neravnoteže klasa, dok je najčešće korišćena metoda poduzorkovanje. Kao što je već istaknuto, jedan od glavnih nedostataka ove metode je da ignoriše veliki broj primjera glavne klase, koja sadrži značajne informacije za razdvajanje klasa. Dakle, minimiziranje gubitka ove važne informacije za diferencijaciju klasa i smanjenje pristrasnosti uzorka, može se postići korišćenjem *ensemble* tehnika u kombinaciji s balansiranim klasifikatorima (Kang et al., 2012; Miguéis et al., 2017), što je i razlog za izbor ovog pristupa u navedenoj studiji. Ova procedura uzima nasumične podskupove glavne klase u više iteracija, koji su po veličini jednaki manjoj klasi i generiše različite modele klasifikacije za izabrane podskupove. Na kraju, svi rezultati se objedinjuju da bi se dobio konačni rezultat. Dokazano je da ovaj pristup ima nekoliko prednosti, kao što je balansiranje klasa, poboljšanje tačnosti predviđanja i izbjegavanje problema prekomjernog prilagođavanja modela (Dietterich, 2002; Zhang & Ma, 2012).

Prethodna istraživanja su utvrdila važnost *web* metrika za različite vrste analitike u digitalnom marketingu. Kao što je ranije istaknuto, neka istraživanja su potvrdila važnost *web* metrika za rano predviđanje *online* naručivanja, pri čemu se problem neravnoteže klasa rješava uglavnom nasumičnim poduzorkovanjem ili

preuzorkovanjem. Međutim, treba imati na umu da je stopa odgovora za skupove podataka koje su koristili uglavnom veća od 5%. Uzimajući u obzir mogućnost veoma male stope odgovora, manje od 0,5%, kao i prethodno navedene prednosti pristupa, u ovom dijelu rada se predlaže kombinovanje slučajnog poduzorkovanja sa *ensemble* meta-algoritmom, kako bi se povećala tačnost predviđanja takvih izuzetno neuravnoteženih klasa. Takođe, predlaže se kvantitativno mjerenje uticaja *web* metrika na tačnost predviđanja i poređenje sa uticajem podataka o proizvodu i kupovnog ponašanja na performanse modela.

Kao i u prethodnim modelima, predloženi klasifikatori će biti ocijenjeni korišćenjem sljedećih mjera performansi: površina ispod krive (AUC), broj tačno pozitivnih instanci (TP), broj lažno pozitivnih instanci (FP), preciznost, senzitivnost (eng. *sensitivity/recall*) i *fallout*.

Slično kao u prethodno predloženim prediktivnim procedurama, obučavanjem modela se pronalazi optimalna kombinacija parametara modela, čime se postiže maksimalni prediktivni učinak. Ovo se postiže kombinovanjem *Grid-Search* tehnologije sa *k-fold* kros-validacijom. Zatim, obučeni model se primjenjuje na testnom skupu podataka (Slika 35).



Slika 35. Ilustracija prediktivne procedure *ensemble* modela odgovora na kampanju

Ensemble metode primijenjene u ovoj studiji, kao što su *Bagging* i *Random Forest*, prepoznate su u literaturi kao efikasne u povećanju prediktivnih performansi slabijih klasifikatora (Miguéis et al., 2017; Zhang & Ma, 2012). Stoga, u ovom radu, klasifikatori su kombinovani sa iterativnim slučajnim balansiranim poduzorkovanjem, formirajući *ensemble* balansiranih klasifikatora. Krajnji cilj je da se uporedi koliko je *ensemble* balansiranih klasifikatora (*Balanced Bagging* DT i *Balanced* RF) bolji od samostalnog balansiranog DT klasifikatora (*Balanced* DT), kao i da li je i koliko bolji od pristupa baziranog na SVM pretprocesiranju iz prethodnog odjeljka.

4.5.5 Koncept modela za predikciju profitabilnosti kupaca baziran na SVR metodi

U tradicionalnim regresionim modelima (Liu & Shih, 2005; Monalisa et al., 2019; Stone, 1995; Yao & Xiong, 2011), fokus je na prosječnom kupcu, dok se heterogenost kupaca obično ne uzima u obzir. Na ovaj način se ne prave dovoljno precizne razlike između vrijednih i manje vrijednih kupaca.

U skladu s navedenim nedostatkom standardnih regresionih metoda, neka od prethodnih istraživanja (Heldt et al., 2019; Rogic & Kascelan, 2019) pokazala su da se ove metode teško mogu uspješno nositi sa asimetričnom distribucijom i dati tačna predviđanja, posebno za grupu vrlo visokoprofitabilnih kupaca. Ovaj problem se još više ističe ako segmentacija ili grupisanje kupaca nije prethodno izvršeno.

U ovom radu, za procjenu profitabilnosti kupaca biće korišćena *Support Vector Regression* metoda (Vapnik et al., 1997). Već je naglašeno da kao i SVM, ova metoda smatra podatke vektorima u *n*-dimenzionalnom prostoru (ulazni prostor) i u slučaju da je odnos između regresora i zavisne promjenljive nelinearan, vektori se mapiraju u prostor veće dimenzije. U tom novom *n+1* dimenzionalnom prostoru, moguće je pronaći hiperravan koja linearno može modelirati vezu između regresora i zavisne varijable. SVR metoda ima dva cilja: prvo, minimizirati grešku u procjeni zavisne varijable i drugo, učiniti model u većoj dimenziji što je moguće ravnijim, kako bi se povećala njegova tačnost predviđanja na nepoznatom skupu podataka. Zahvaljujući tome, kao što je već istaknuto, metoda je robustna za ekstremne vrijednosti i ima izvanredne mogućnosti generalizacije.

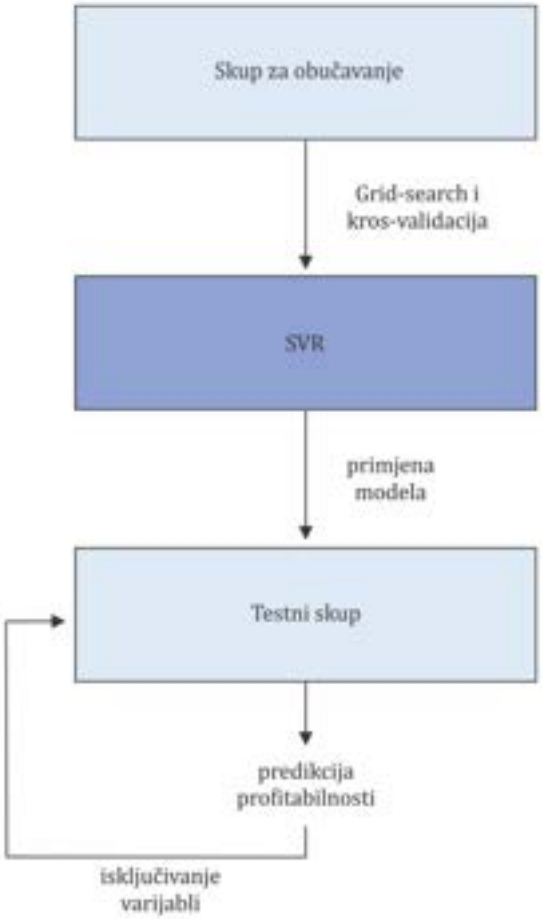
Pošto je distribucija profitabilnosti kupaca obično iskrivljena, tačnost predviđanja se smanjuje kako se vrijednost profitabilnosti kupaca povećava. Kada je heterogenost uzorka prisutna zbog postojanja ekstremnih vrijednosti, *data mining* metode obično precjenjuju predviđenu vrijednost profitabilnosti. U nekim prethodnim istraživanjima, takvi kupci sa ekstremnim vrijednostima profitabilnosti su uklonjeni iz baze podataka (Kim et al., 2008; Lei et al., 2018; Wang & Lan, 2020), kako bi se poboljšala tačnost predviđanja. Međutim, podaci o kupcima sa

ekstremnim vrijednostima profitabilnosti sadrže značajne informacije za cjelokupan segment visokoprofitabilnih kupaca, što je od posebnog značaja za donosioce odluka. S tim u vezi, s obzirom na njene gore istaknute prednosti, SVR metoda će se koristiti u cilju prevazilaženja problema asimetrične raspodjele profitabilnosti kupaca. Osim toga, SVR dobro funkcioniše i s manjim veličinama uzorka (Tange et al., 2017) i u ovom radu će biti testiran na malom skupu podataka (jer su iz baze kupaca za predikciju profitabilnosti izdvojeni samo kupci koji su imali transakcije u razmatranim kampanjama koje su trajale nekoliko mjeseci; takođe, treba napomenuti da je riječ o malom crnogorskom tržištu).

Metodološki postupak za predviđanje profitabilnosti kupaca se sastoji od sljedećih koraka:

1. Priprema baze podataka – pretprocesiranje podataka, izračunavanje RFM atributa i normalizacija (opseg 0-1);
2. Obučavanje SVR modela za predviđanje profitabilnosti korisnika. Ovaj korak uključuje pronalaženje optimalnih parametara za SVR model (C, γ i ε), korišćenjem *5-fold* kros-validacije i *Grid-Search* tehnike;
3. Nakon izbora optimalnih parametara, model se generiše pomoću dobijenih vrijednosti parametara, a zatim se primjenjuje na neviđene podatke (testni skup);
4. Prethodni korak se ponavlja isključivanjem grupa promjenljivih pojedinačno, kako bi se procijenio njihov uticaj na tačnost predviđanja, odnosno značaj isključenih atributa za predviđanje profitabilnosti.

Šematski prikaz prediktivne procedure predstavljen je na Slici 36.



Slika 36. Ilustracija prediktivne procedure *ensemble* modela odgovora na kampanju

Što se tiče koraka 4, u ovom radu će se, osim osnovnog modela za predikciju profitabilnosti kupaca, u predviđanju profitabilnosti korisnika pomoću SVR metode, procijeniti i uticaj RFM atributa, i kao grupe i pojedinačno. Osim RFM atributa, u zavisnosti od dostupnosti različitih kategorija podataka, varijable koje opisuju proizvode, *web* metrike i podaci o klijentima takođe će se analizirati u smislu uticaja na tačnost predviđanja profitabilnosti kupaca. Ovaj pristup se ne oslanja na potencijalno subjektivne procjene različitih varijabli, već, procjenjujući performanse modela, objektivno pokazuje koje su varijable dominantne u predviđanju profitabilnosti kupaca.

Za ocjenu prediktivnih performansi modela, osim standardnih regresionih pokazatelja koji su ranije istaknuti (sekcija 4.1.2), biće korišćeni i procenti pojedinačnih odstupanja stvarnih vrijednosti od predviđenih koji su manji od 10% i 20%. Ova mjera je posebno važna radi procjene modela, tj. da se procijeni da li model precjenjuje ili potcjenjuje najprofitabilnije kupce, tj. da li je robustan za ekstremne vrijednosti.

Predložena procedura biće implementirana korišćenjem *Rapid Miner* softvera u vidu gotovog procesa spremnog za upotrebu.

5. EMPIRIJSKO TESTIRANJE PREDLOŽENIH PREDIKTIVNIH MODELA

U ovom poglavlju najprije će biti predstavljeni skupovi podataka koji se koriste za testiranje, a zatim i rezultati testiranja predloženih koncepata. Na osnovu rezultata biće potvrđene hipoteze i dati odgovori na istraživačka pitanja.

5.1 Opis korišćenih skupova podataka

U ovom dijelu rada biće detaljno opisani skupovi podataka korišćeni za obučavanje i testiranje predloženih modela.

5.1.1 Opis podataka za modele prediktivne RFM segmentacije

Primarni cilj prediktivne segmentacije potrošača u direktnom marketingu je dobijanje uvida u vrijednost potrošača, koja će odrediti da li će kupac biti targetiran ili ne. U ovoj sekciji rada biće predstavljen opis podataka korišćen za predloženu prediktivnu proceduru iz sekcije 4.5.1 i sekcije 4.5.2.

Prvi korak je prikupljanje podataka o transakcijama, tj. obavljenim kupovinama iz prethodnih direktnih kampanja, koji mogu obuhvatati podatke o kupcima kao što su: pol, starost, region, nivo dohotka, itd., podatke o proizvodima kao što su vrsta, kategorija, namjena i podatke o ponašanju u kupovini, tj. *Recency*, *Frequency* i *Monetary* vrijednosti kupovine. Ovi podaci će biti korišćeni kao skup za obučavanje prediktivnog modela.

Za empirijsko testiranje u ovom dijelu rada korišćen je skup podataka o *online* kupovnim transakcijama iz prethodnih direktnih kampanja kompanije *Sport Vision Montenegro* (kompanija je dio sistema *Sport Vision* i vodeći je prodavac sportske opreme na Balkanu), za period od početka septembra 2018. do kraja januara 2019. godine (sezona jesen/zima). Baza podataka sastoji se od 1.605 zapisa (transakcija) i ima sljedeće atribute: ID narudžbe, popust, cijena, datum narudžbe, pol kupca, tip

proizvoda, namjena proizvoda po polu, kategorija proizvoda, namjena proizvoda po starosti i brend proizvoda. Tip proizvoda predstavlja maloprodajnu klasifikaciju proizvoda na: obuću (patike, cipele, čizme itd.), odjeću (majice, duksevi, helanke, itd.) i opremu (torbe, tegovi, rukavice itd.). S druge strane, kategorija proizvoda je drugi oblik klasifikacije, zasnovan na svrsi aktivnosti (na primjer, kategorija "trčanje" - za patike za trčanje, "*lifestyle*" – za pamučne majice i slično). Namjena proizvoda po polu sastoji se od pet vrijednosti: proizvodi za žene, muškarce, dječake, djevojčice i uniseks proizvodi. Pored toga, namjena po starosti opisuje starosnu grupu kojoj su proizvodi namenjeni (za bebe, djecu, odrasle itd.). Konačno, brend proizvoda dijeli proizvode u dvije glavne grupe - A brendovi (proizvodi brendova za koje ova kompanija ima pravo distribucije) i brendovi licence (brendovi za koje ova kompanija ima pravo proizvodnje i distribucije) i malu grupu brendova „ostalo". Generalno, A brendovi su dobro poznati i afirmisani sportski brendovi, koji su obično skuplji (*Nike, Adidas, Converse* i slično), dok su brendovi s licencom pristupačniji, ali nemaju tako jak imidž i prepoznatljivost brenda (*Kronos, Lonsdale, Champion* i sl).

Nakon prikupljanja i pretprocesiranja podataka, baza se kompletira izračunavanjem RFM atributa na sljedeći način: *Recency* se definiše kao datum posljednje porudžbine, *Frequency* - kao ukupan broj porudžbina u razmatranom periodu i *Monetary* - kao novčani iznos koji je kupac potrošio u razmatranom periodu, izražen u eurima. Atribut *Recency* je kodiran tako što se za 20% najnovijih datuma dodjeljuje ocjena 5, za sljedećih 20% datuma daje se ocjena 4 i tako dalje do ocjene 1 - za najstarije transakcije u bazi. Atributi *Frequency* i *Monetary* zadržavaju se u izvornom obliku. Na kraju, svi atributi se normalizuju transformacijom opsega 0-1. Tabela 5 prikazuje distribuciju atributa u početnom skupu podataka.

Tabela 5. Distribucija atributa u početnom skupu podataka

Atribut	Statistika	Opseg
Cust_gend	most = M (891), least = F (714)	F (714), M (891)
Discount	avg = 0.371 +/- 0.107	[0.000 ; 0.500]
Prod_type	most = Footwear (784), least = Equipment (181)	Footwear (784), Equipment (181), Apparel (640)

Prod_gender	most = For men (786), least = For girls (67)	For women (399), For boys (210), For men (786), Unisex (143), For girls (67)
Prod_category	most = Lifestyle (869), least = Handball (1)	Lifestyle (869), Fitness (231), Running (119), Football (70), Skiing (103), Outdoor (85), Basketball (103), Other (5), Boxing (3), Tennis (12), Accessories (2), Handball (1), Volleyball (1), Skateboarding (1)
Prod_brand	most = A brands (853), least = Other (74)	A brands (853), Licence (678), Other (74)
Prod_age	most = For adults (1272), least = For all (23)	For adults (1272), For babies (0-4) (62), For teens (8-14) (127), For younger kids (4-10) (121), For all (23)
R	avg = 3,143 +/- 1.353	[1 ; 5]
F	avg = 3,616 +/- 3.401	[1 ; 17]
M	avg = 100,081 +/- 78,211	[9,60 ; 352,00]

U svrhu testiranja prediktivnih performansi modela, korišćeni su isti tipovi podataka iz godine kasnije, ali iz iste sezone (jesen/zima), kada je slična ponuda dostupna potrošačima. Odluka o izboru skupa podataka za testiranje modela motivisana je činjenicom da je sezonalnost u ovoj industriji veoma izražena, te direktno utiče i definiše aktuelnu ponudu. Na primjer, u sezoni jesen/zima marketing fokus je na „vraćanju u školu" i na skijaškim kampanjama, dok je tokom sezone proljeće/ljeto fokus na ljetnjim aktivnostima. Stoga, ima smisla upoređivati performanse istih sezona i različitih godina, sa istim dostupnim atributima.

Podaci za testni skup podataka su pripremljeni na isti način kao i za skup za obučavanje (izračunati su RFM atributi i normalizovani su svi atributi istom transformacijom od 0 do 1). Tabela 6 prikazuje distribuciju atributa u testnom skupu podataka.

Tabela 6. Distribucija atributa u početnom skupu podataka

Atribut	Statistika	Opseg
Cust_gend	most = M (2622),least = F (2596)	F (2622), M (2596)
Discount	avg = 0.376 +/- 0.131	[0.000 ; 0.700]
Prod_type	most = Footwear (2367), least = Equipment (487)	Footwear (2367), Equipment (487), Apparel (2365)
Prod_gender	most = For men (2615), least = For girls (367)	For women (1169), For boys (688), For men (2615), Unisex (380), For girls (367)
Prod_category	most = Lifestyle (3100), least = Skateboarding (1)	Lifestyle (3100), Fitness (556), Outdoor (358), Football (315), Skiing (311), Running (288), Basketball (185), Other (51), Boxing (20), Volleyball (13), Accessories (10), Tennis (7), Handball (4), Skateboarding (1).
Prod_brand	most = A brands (3134), least = Other (72)	A brands (3134), Licence (2013), Other (72)
Prod_age	most = For adults (3982), least = For all (85)	For adults (3982), For teens (8-14) (502), For younger kids (4-10) (484), For babies (0-4) (166), For all (85)
R	avg = 3,003 +/- 1.413	[1 ; 5]
F	avg = 4,004 +/- 4.608	[1 ; 29]
M	avg = 106,09 +/- 104,64	[4,50 ; 667,10]

Da bi se potvrdila primjenljivost predloženog modela na podacima iz drugih privrednih djelatnosti i s drugim atributima, on će biti validiran na javno dostupnom skupu podataka „*Customer transaction dataset*". Ovaj skup podataka je dostupan u *Kaggle online* bazi (Kaggle, n.d.) i sadrži podatke o prodaji opreme za biciklizam. Ovaj skup podataka se sastoji od 20.000 prodajnih transakcija za 3.500 kupaca, a odnosi se na period od januara do decembra 2017. godine. Skup podataka je prečišćen uklanjanjem nedostajućih vrijednosti, nakon čega konačna verzija ove baze podataka sadrži 19.765 instanci, koje su podijeljene na skup za obučavanje (70%) i testni skup (30%).

Atribut *Recency* je izračunat na osnovu datuma transakcije, *Frequency* kao ukupan broj transakcija u posmatranom periodu, dok je *Monetary* obuhvatio ukupan zbir vrijednosti ostvarenih transakcija, što je identična procedura kao i u originalnom skupu podataka na kojem je obučen model. Distribucija atributa iz ovog skupa podataka predstavljena je u Tabeli 7.

Tabela 7. Distribucija atributa u javno dostupnom skupu podataka *Customer transaction dataset*

Naziv atributa	*Min*	*Max*	*Avg*	*Dev*	Vrijednosti
transaction_id	1	20000			
order_status					Approved (19588), Cancelled (177)
brand					Solex (4245), Giant Bicycles (3307), WeareA2B (3287), OHM Cycles (3039), Trek Bicycles (2983), Norco Bicycles (2904)
product_line					Standard (14151), Road (3961), Touring (1233), Mountain (420)
product_class					medium (13797), high (3011), low (2957)
product_size					medium (12965), large (3969), small (2831)
customer_id	1	3500			
gender					Female (9908), Male (9419), U (438)
age	18,00	120,00	44,41	16,88	
job_industry_category					Manufacturing (3976), Financial Services (3836), n/a (3192), Health (3073), Retail (1746),

					Property (1278), IT (1052), Entertainment (694), Agriculture (566), Telecommunications (352)
wealth_segment					Mass Customer (9923), High Net Worth (5040), Affluent Customer (4802)
owns_car					Yes (9950), No (9815)
region					NSW (10083), VIC (4509), QLD (4214), New South Wales (480), Victoria (479)
Recency	1,00	5,00	4,67	0,63	
Frequency	1,00	14,00	6,67	2,34	
Monetary	60,34	19071,32	7380,20	3018,18	

Iz Tabele 7 se može uočiti da su kupci opisani atributima: pol, godine starosti, zanimanje, kategorija dohotka, region i eventualno vlasništvo automobila, što može biti interesantan podatak pri opisivanju kupaca bicikla. Dodatno, prosječna vrijednost *Recency* atributa je relativno visoka i iznosi 4,67 (od maksimalnih 5), dok je prosječni broj kupovina 6,67, a prosječna vrijednost trasakcija 7380,20.

U narednom dijelu rada biće opisan skup podataka korišćen za modele predikcije odgovora kupca.

5.1.2 Opis podataka za modele predikcije odgovora kupca

Jedna od bazičnih karakteristika *online* kampanja direktnog marketinga je podsticanje korisnika da preduzme određenu radnju koja se može mjeriti, kao što je klik na link ka *web* sajtu, kupovina proizvoda, korišćenje koda za popust, itd. Ova karakteristika direktnog *online* marketinga čini odgovore kupaca mjerljivim, a dodatno, omogućeno je njihovo praćenje, što doprinosi izgradnji velikih baza podataka o kupcima (Chun, 2012). Da bi ovi podaci bili korisni, kompanije mogu da

izgrade modele za predikciju odgovora kupaca, koji mogu pomoći u identifikaciji kupaca koji će, s velikom vjerovatnoćom, odgovoriti na sljedeću planiranu kampanju. Pored toga, takve analize mogu pružiti informacije o profitabilnosti kampanje, kao i pomoći u donošenju relevantnih odluka u marketingu. U ovom modelu odgovora kupca, kao odgovor se uzima samo završena kupovina.

Za empirijsko testiranje predloženih modela odgovora kupaca, dobijen je skup podataka od vodećeg distributera sportske opreme iz Crne Gore, kao i u prethodnom slučaju. Skup podataka sadrži posjete *web* sajtu e-trgovine upućene sa sponzorisanih objava na društvenim mrežama za četiri mjeseca, od oktobra 2018. do januara 2019. godine. Dakle, u skupu podataka nema indikatora izvora saobraćaja, jer su za analizu odabrane samo sesije koje se odnose na sponzorisane objave na društvenim mrežama. Dakle, kao odgovor se računa transakcija izvršena direktno preko linkova društvenih mreža (pozitivna klasa). U posmatranom periodu bilo je 12.990 jedinstvenih korisnika *web* sajta, koji su pratili link s targetirane objave na društvenim mrežama *Instagram* ili *Facebook*, koji ih čini potencijalnim kupcima, jer su, samim klikom na link, iskazali interesovanje za predstavljenu ponudu. Ukupan broj sesija je bio 33.662 tokom šest *online* kampanja direktnog marketinga plasiranih putem pomenutih društvenih mreža. Konačni skup podataka (baza kupaca) je rezultat spajanja nekoliko baza podataka: sopstvene baze podataka o proizvodima i kupcima kompanije, *Google Analytics* i *Facebook Business Manager*, nakon čega je uslijedilo pretprocesiranje i priprema skupa podataka za analizu modela odgovora kupaca, agregiranjem podataka po kupcu, tj. posjetiocu sajta, za odgovarajući vremenski period. Skup podataka sadrži sljedeće grupe atributa: *web* metrike, podatke o opisu proizvoda, kao i podatke o prethodnoj istoriji kupovine u smislu RFM atributa.

Opis atributa u ovom skupu podataka, kao i odgovarajuće statistike za ukupni razmatrani period date su u Tabeli 8.

Tabela 8. Opis podataka za model odgovora na kampanju

Naziv atributa	Opis atributa	*Min*	*Max*	*Avg*

Customer_ID	šifra kupca	1	9660	/
Camp_Sessions_avg	prosječan broj sesija u svim kampanjama	1	37,00	2,267
Camp_Avg Sess duration	prosječno trajanje sesija u svim kampanjama	1,88	4755,00	190,785
Camp_Avg bounce rate	prosječni *bounce rate* za sve posjete sajtu tokom kampanja	0	0,90	0,154
Cons_Reg_Central	broj sesija ostvarenih iz centralnog regiona	0	26	1,005
Cons_Reg_South	broj sesija ostvarenih iz južnog regiona	0	6	0,179
Cons_Reg_North	broj sesija ostvarenih iz sjevernog regiona	0	6	0,092
Cons_Dev_Desktop	broj ostvarenih sesija korišćenjem desktop uređaja	0	29	0,030
Cons_Dev_Mobile	broj ostvarenih sesija korišćenjem mobilnog uređaja	0	79	2,534
Cons_Dev_Tablet	broj ostvarenih sesija korišćenjem tablet uređaja	0	25	0,023
Cons_OS_Android	broj ostvarenih sesija korišćenjem Android operativnog sistema	0	79	2,160
Cons_OS_Ios	broj ostvarenih sesija korišćenjem iOS operativnog sistema	0	64	0,397
Cons_OS_Windows	broj ostvarenih sesija korišćenjem Windows operativnog sistema	0	29	0,030
Prod_Apparel	broj kupljenih proizvoda iz kategorije odjeća	0	19	0,008
Prod_Footwear	broj kupljenih proizvoda iz kategorije obuća	0	6	0,014
Prod_Equipment	broj kupljenih proizvoda iz kategorije oprema	0	2	0,000
Prod_Gen_For boys	broj kupljenih proizvoda za dječake	0	3	0,003
Prod_Gen_For girls	broj kupljenih proizvoda za djevojčice	0	3	0,001
Prod_Gen_For men	broj kupljenih proizvoda za muškarce	0	24	0,012
Prod_Gen_For women	broj kupljenih proizvoda za žene	0	4	0,005
Prod_Gen_unisex	broj *unisex* kupljenih proizvoda	0	2	0,001

Prod_Type_Performance	broj kupljenih proizvoda iz *performance* kategorije	0	4	0,005
Prod_Type_Lifestyle	broj kupljenih proizvoda iz *lifestyle* kategorije	0	25	0,015
Prod_Type_Outdoor	broj kupljenih proizvoda za *outdoor* aktivnosti	0	5	0,002
Prod_Br_A brand	broj kupljenih proizvoda A brendova (viši nivo cijene)	0	4	0,012
Prod_Br_Licence	broj kupljenih proizvoda brendova s licencom (niži nivo cijene)	0	25	0,010
Prod_Age_For adults	broj kupljenih proizvoda za odrasle	0	24	0,017
Prod_Age_For kids	broj kupljenih proizvoda za djecu	0	3	0,003
Prod_Age_For teens	broj kupljenih proizvoda za tinejdžere	0	3	0,002
Prod_Age_For all	broj kupljenih proizvoda za sve uzraste	0	2	0,000
Prod_Disc_<30%	broj kupljenih proizvoda sa popustom manjim od 30%	0	14	0,010
Prod_Disc_30-50%	broj kupljenih proizvoda sa popustom između 30% i 50%	0	11	0,013
R1	*Recency* dobijen podjelom skupa podataka na pet jednakih djelova od najstarijih do najnovijih transakcija	0	5	0,042
R2	*Recency* dobijen dodjeljivanjem brojeva od 2 do 5 na osnovu posljednje kampanje iz koje je kupac naručio	0	5	0,042
F1	broj kampanja sa ostvarenim transakcijama	0	3	0,014
F2	ukupan broj porudžbina u svim kampanjama	0	6	0,016
F3	broj porudžbina u posljednjoj plasiranoj kampanji	0	1	0,002
M1	prosječan iznos transakcija u svim kampanjama	0	119,75	0,607
M2	prosječan iznos transakcija u posljednjoj plasiranoj kampanji	0	103,20	0,093
M3	ukupan iznos realizovanih transakcija	0	430,00	0,742

Response	Ciljna varijabla – odgovor na kampanju	0 (9620)	1 (40)	/

Napomena: Opis mjere *Google* analitike – *bounce rate* je sesija na jednoj stranici podijeljena sa svim sesijama, ili procenat svih sesija na *web* sajtu u kojima su korisnici pregledali samo jednu stranicu i pokrenuli samo jedan zahtjev na serveru Analitike (Google, 2021).

Iz Tabele 8 može se vidjeti da je najveća grupa kupaca posjetila *web* sajt e-trgovine sa svojih mobilnih telefona, uglavnom korišćenjem *Android* uređaja, a zatim *iOS* uređaja. Najviše sesija je realizovano iz centralnog regiona (koji je najveći region u Crnoj Gori), a zatim iz južnog i sjevernog regiona. Ovo je u skladu s trenutnim stanjem ekonomskog razvoja tri crnogorska regiona.

Cilj modela odgovora na kampanju je da se, na osnovu podataka o prethodnoj istoriji kupovine (RFM), kao i podataka o proizvodima i *web* metrikama, predvidi da li će potencijalni kupac odgovoriti na sljedeću kampanju. Da bi se sprovela procedura predviđanja, kompletan skup podataka je podijeljen na skupove podataka za obuku i testiranje.

Skup podataka za obučavanje modela sadrži istoriju ponašanja na *web* sajtu kompanije i kupovine 9.660 posjetilaca *web* sajta od kampanje 1 do kampanje 4, kao i indikator njihovog odgovora na sljedeću kampanju 5 (samo 40 kupaca je odgovorilo direktno na ponudu, tj. kupilo u ovoj kampanji, što daje stopu odgovora od samo 0,41%). Skup za testiranje modela sadrži iste kategorije podataka kao i skup za obuku, za 7.929 posjetilaca od kampanje 1 do kampanje 5 i odgovor koji pokazuje da li je klijent odgovorio na kampanju 6 (i u ovoj kampanji je bilo 40 odgovora), ne uključujući nove posjetioce, koji su se prvi put pojavili prateći ponudu iz kampanje 5 ili kampanje 6. Pored toga, iz oba skupa podataka, isključeni su podaci za posjetioce koji su u prosjeku proveli manje od 30 sekundi u sesiji.

U skladu sa istraživačkom praksom kod prediktivnog modeliranja, za validaciju predloženog modela korišćen je i javno dostupni skup podataka - *Direct Marketing Educational Foundation 3 (DMEF3)*, koji obuhvata podatke iz kataloške prodaje za

period od 12 godina. Prateći proceduru prethodne obrade podataka, koju su u svom radu predstavili autori *Malthouse* i *Blattberg* (2005), sadašnji trenutak je definisan kao 1. avgust 1990. godine, a takođe je primijenjena jednoprocentna vinzorizacija (eng. 1% *Winzorisation)* i transformacija kvadratnim korijenom za sve kvantitativne promjenljive. Uzorak kupaca prije ovog datuma bio je 41.669, što je odredilo bazne i ciljne periode od šest godina. U posljednjem koraku pripreme podataka, uzorak je podijeljen na skupove podataka za obuku (odgovori kupaca za period od šest godina do 1. avgusta 1990. godine) i testiranje približno iste veličine (odgovori kupaca za period od šest godina poslije 1. avgusta 1990. godine).

Ovaj skup podataka sadrži sljedeće atribute koji opisuju korisnika: ID korisnika, dan i godina unosa u bazu podataka, vrijeme u datoteci (eng. *time on file*); RFM atributi: broj mjeseci od posljednje porudžbine, iznosi prodaje i broj porudžbina po klasama proizvoda u osnovnom periodu, ukupan broj porudžbina i iznos prodaje u baznom periodu, da li je bilo narudžbina tokom svake posmatrane godine, da li je kupac poručio u dvije ili tri godine zaredom, broj godina s narudžbama, vještačke *Recency* varijable formirane na osnovu broja mjeseci od posljednje transakcije i *Recency* kvantili 1-20; cjeloživotni atributi: cjeloživotna prodaja (zavisna promjenljiva). Horizont predviđanja u ovom skupu podataka je šest godina, što je značajno duže od prvog skupa podataka.

Tabela 9. Opis podataka za validaciju modela odgovora na kampanju

Naziv atributa	Opis atributa	*Min*	*Max*	*Avg*
conv_year	godina unosa u bazu podataka	8,426	9,487	9,098
tof	vrijeme u datoteci	8,414	17,377	12,256
RECMON	broj mjeseci od posljednje porudžbine	0	7,810	4,648
SALCLS1	kupovina proizvoda klase 1 za pet godina	0	10,987	0,312
SALCLS2	kupovina proizvoda klase 2 za pet godina	0	17,124	1,513
SALCLS3	kupovina proizvoda klase 3 za pet godina	0	16,309	1,216
SALCLS4	kupovina proizvoda klase 4 za pet godina	0	6,468	0,054

SALCLS5	kupovina proizvoda klase 5 za pet godina	0	12,020	0,086
SALCLS6	kupovina proizvoda klase 6 za pet godina	0	16,037	1,481
SALCLS7	kupovina proizvoda klase 7 za pet godina	0	13,143	1,121
monetar5	ukupna vrijednost porudžbina za pet godina	0	24,072	5,383
ORDCLS1	broj poružbina proizvoda iz klase 1	0	1	0,043
ORDCLS2	broj poružbina proizvoda iz klase 2	0	1,732	0,198
ORDCLS3	broj poružbina proizvoda iz klase 3	0	1,732	0,164
ORDCLS4	broj poružbina proizvoda iz klase 4	0	1	0,009
ORDCLS5	broj poružbina proizvoda iz klase 5	0	1,414	0,161
ORDCLS6	broj poružbina proizvoda iz klase 6	0	1,732	0,223
ORDCLS7	broj poružbina proizvoda iz klase 7	0	1,732	0,174
orders5	ukupan broj porudžbina za pet godina	0	2,646	0,655
PROMy1_resp	odgovor na promotivne kampanje u prvoj godini	0	1	0,080
PROMy2_resp	odgovor na promotivne kampanje u drugoj godini	0	1	0,080
PROMy3_resp	odgovor na promotivne kampanje u trećoj godini	0	1	0,073
PROMy4_resp	odgovor na promotivne kampanje u četvrtoj godini	0	1	0,124
resp12	odgovor na promotivne kampanje u prvoj i drugoj godini	0	1	0,026
resp13	odgovor na promotivne kampanje u prvoj i trećoj godini	0	1	0,022
resp23	odgovor na promotivne kampanje u drugoj i trećoj godini	0	1	0,026
resp123	odgovor na promotivne kampanje u prvoj, drugoj i trećoj godini	0	1	0,015
nyPROM_resp	broj godina sa odgovorima na kampanju	0	1	0,290
r_0to26	*Recency* - 0 do 26 mjeseci od posljednje transakcije	0	1	0,202
r_21to23	*Recency* - 21 do 23 mjeseca od posljednje transakcije	0	1	0,048

r_33to35_45to47	*Recency* - 33 do 47 mjeseca od posljednje transakcije	0	1	0,051
r_lastfall	posljednja transakcija u pretposljednjoj jesenjoj kampanji	0	1	0,082
r_fall	posljednja transakcija u posljednjoj jesenjoj kampanjii	0	1	0,208
rec_20coding	kodirana *Recency* varijabla	0	4,472	2,600
TOTORD	ukupan broj porudžbina	1,414	7,416	2,954
TOTSALE	ukupna vrijednost porudžbina	5,831	45,127	17,006

Validacija predloženog modela na DMEF3 skupu podataka potvrdiće da se on može primjenjivati ne samo za *online* kampanje, već i za kataloške i druge vrste kampanja, te u drugim industrijama s različitim dostupnim atributima.

5.1.3 Opis podataka za model predikcije profitabilnosti kupca baziran na SVR metodi

Za empirijsko testiranje predloženog pristupa korišćen je isti skup podataka kao u prethodnoj sekciji.

Kao što je već navedeno, ovaj skup podataka dobijen je od kompanije koja se bavi prodajom sportske opreme. Konkretno, podaci koji su korišćeni za ovaj dio rada odnose se na ostvarenu prodaju proizvoda preko *web* sajta, odnosno e-trgovine u toku pomenutih šest kampanja. Preciznije, izdvojeni su samo oni korisnici koji su odgovorili na kampanju kupovinom proizvoda.

Konačni skup podataka je prethodno obrađen i pretvoren u odgovarajući oblik za predviđanje profitabilnosti korisnika, kao u Tabeli 8 iz prethodne sekcije, te je podijeljen na skupove podataka za obuku i testiranje približno iste veličine. Skup podataka za obuku sadrži istoriju ponašanja na internetu i kupovine za 130 kupaca iz kampanja od K1 do K4, a pokazatelj njihove profitabilnosti bila je vrijednost svih transakcija ostvarenih zaključno s kampanjom 5 (K5). Testni skup sadrži istoriju ponašanja na mreži i kupovine za 158 ispitanika iz kampanja K1 do K5, a pokazatelj

profitabilnosti bila je vrijednost svih transakcija od početka posmatranog perioda, zaključno s kampanjom 6 (K6). Dakle, predviđa se cjeloživotna profitabilnost (eng. *lifetime profitability*), pri čemu je horizont predviđanja sljedeća kampanja. U skladu s tim, zavisna (ciljna) varijabla je atribut M3 iz Tabele 8.

Za validaciju predloženog modela korišćen je DMEF3 skup podataka, koji je detaljno opisan u prethodnoj sekciji. Za predikciju profitabilnosti, kao ciljna varijabla uzeta je TOTSALE, koja predstavlja ukupnu vrijednost prodaje za kupca koja je zabilježena u bazi za razmatrani vremenski period (za period od šest godina do 1. avgusta 1990. godine - kod skupa za obuku i za period od šest godina poslije 1. avgusta 1990. godine - kod skupa za testiranje).

5.2 Testiranje modela prediktivne RFM segmentacije

Za empirijsko testiranje predložene prediktivne procedure, za ovaj model korišćen je skup podataka o *online* kupovnim transakcijama iz direktnih kampanja kompanije za prodaju sportske opreme, *Sport Vision*, u periodu od početka septembra 2018. godine do kraja januara 2019. godine. Podaci su pripremljeni izračunavanjem RFM atributa po proceduri predloženoj u sekciji 4.5.1 (korak 1).

Klasterizacijom polaznog skupa podataka *k-means* metodom pomoću normalizovanih RFM atributa, dobijeni su rezultati prikazani u Tabeli 10. Kao što se može vidjeti, najbolji *Davis-Bouldin* indeks ima model sa tri klastera. Ovaj klaster model prikazan je u Tabeli 11.

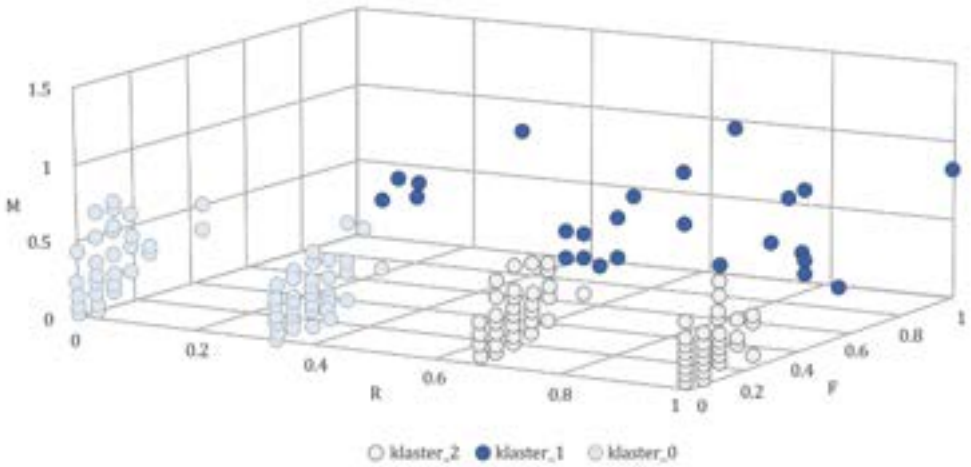
Tabela 10. Izbor broja klastera (parametra *k*) kod *k-means* klasterizacije

K	2	3	4	5	6	7	8	9	10
DB	-1,025	**-0,811**	-0,983	-0,958	-0,909	-1,02	-0,98	-0,976	-0,96

Tabela 11. Centroid klaster model za RFM segmentaciju kupaca

	R	F	M	Broj primjera
cluster_0	0,766807	0,565126	0,677733	238
cluster_1	0,735348	0,088065	0,179499	819
cluster_2	0,137318	0,101734	0,211345	548

Iz Tabele 11 se može vidjeti da *cluster_0* čine najfrekventniji i najprofitabilniji kupci, koji su najskorije obavili posljednje transakcije (CV-nivo=1), *cluster_1* čine kupci koji su skoro kupovali, ali su manje frekventni i manje profitabilni (CV-nivo=2), dok *cluster_2* čine kupci koji su relativno davno kupovali, koji su manje frekventni i manje profitabilni (CV-nivo=3). Na Slici 37 prikazan je dijagram rasipanja, koji ilustruje efikasnost klasterizacije, odnosno segmentacije kupaca. Na dijagramu se jasno izdvajaju različiti klasteri, te nema preklapanja između grupa, što potvrđuje preciznost predloženog modela i odabira broja klastera na osnovu DB indeksa.



Slika 37. Ilustracija RFM segmentacije za tri klastera

Koristeći optimizaciju parametara pomoću *Grid-Search* metode i *10-fold* kros-validacije, za predviđanje CV-nivoa generisan je odgovarajući DT model. Zatim je, u istu svrhu, obučen SVM model, tj. izvršen izbor parametara C i γ pomoću *Grid-Search* metode i *10-fold* kros-validacije i utvrđene klasifikacione performanse ovog modela. Predviđene vrijednosti za *Cluster* (tj- CV-nivo) su dodate polaznom skupu podataka kao nova oznaka klase – *SVM Cluster* (*class label*). Zatim je generisan DT model s

takvom oznakom klase uz *Grid-Search* izbor parametara i *10-fold* kros-validaciju (SVM-RE model). Optimalni parametri i klasifikacione performanse ovih modela prikazani su u Tabeli 12.

Tabela 12. Rezultati testiranja procedure prediktivne klasifikcije

Parametri	Rezultati *10-fold* kros-validacije
DT	
Decision Tree.criterion = gini_index	accuracy: **61,27% +/- 2,58% (mikro: 61,28%)**
Decision Tree.minimal_size_for_split = 5	mean_recall: 47,34% +/- 2,49% (mikro: 47,35%)
Decision Tree.minimal_leaf_size = 6	mean_precision: 46,63% +/- 7,42% (mikro: 52,03%)
Decision Tree.maximal_depth = 11	class recall: **4,27%**,80,00%, 57,77%
Decision Tree.minimal_gain = 0.1	class precision: **33,33%**,62,69%,60,08%
SVM	
SVM.gamma = 400.0	accuracy: **61,31% +/- 3,28% (mikro: 61,31%)**
SVM.C = 400.0	mean_recall: 50,14% +/- 2,93% (mikro: 50,15%)
	mean_precision: 57,76% +/- 7,22% (mikro: 58,72%)
	class recall: **19,75%**, 81,44%, 49,27%
	class precision: **51,65%**, 61,42%, 63,08%
SVM-RE – Pretprocesirani DT	
Decision Tree.criterion = accuracy	accuracy: **85,73% +/- 2,52% (mikro: 85,75%)**
Decision Tree.minimal_size_for_split = 7	mean_recall: 77,84% +/- 6,32% (mikro: 77,77%)
Decision Tree.minimal_leaf_size = 2	mean_precis: 83,25% +/- 5,85% (mikro: 83,39%)
Decision Tree.maximal_depth = 15	class recall: **62,70%**, 93,07%, 77,53%
Decision Tree.minimal_gain = 0.3	class precision: **77,45%**,86,27%, 86,46%

Iz gornje tabele se može uočiti da SVM-RE metoda ima značajno bolje klasifikacione performanse od DT metode. DT metoda je korektno targetirala samo 4% najvrednijih kupaca, dok je SVM-RE uspješno targetirala čak 63% najvrednijih kupaca. To praktično znači da će, među postojećim kupcima, DT modelom biti identifikovano samo 9 od 238 najvrednijih kupaca. Metoda SVM-RE će targetirati za kampanju 150 od 238 najvrednijih kupaca i povećati mogući prihod od kampanje za 16,7 puta u odnosu na targetiranje metodom DT.

Uz to, sve razmatrane klasifikacione performanse su bolje kod SVM-RE nego kod DT metode. *Class precision* najvrednijih kupaca za DT je samo 33%, što znači da će kompanija imati nepotrebne troškove kampanje za 67% pogrešno klasifikovanih kupaca. *Class precision* SVM-RE modela za najmanju, a najvredniju klasu je 77%, što znači da će biti upućeno samo 23% ponuda na koje vrlo vjerovatno neće biti odgovora. Dakle, SVM-RE će, u odnosu na DT, smanjiti troškove kampanje oko 3 puta.

Može se zaključiti da je, uz visoku tačnost predviđanja CV-nivoa (86%), predložena SVM-RE metoda uspjela da riješi problem nebalansiranosti klasa. Rezultati pokazuju da je SVM, kao pretrocesor podataka o kupovnim transakcijama, riješio problem preklapanja klasa i obezbijedio precizniju klasifikaciju. Naime, tačnost DT klasifikacije je povećana za 25% - tačnost za polazni DT iznosi 61%, a za pret-procesirani DT 86%; *mean class recall* za DT iznosi 47%, a nakon SVM pret-procesiranja podataka 78%; *mean class precision* je nakon SVM pretprocesiranja porastao sa 52% na 77%. S obzirom na visok stepen tačnosti na nivou od 86% (za kros-validacionu tačnost ekstrakcije pravila iz SVM predikcije), pravila izvedena pretprocesiranim DT modelom validno i efikasno interpretiraju SVM model.

U Tabeli 13 su prikazana neka od 39 izvedenih primarnih pravila (prikazana su pravila koja pokrivaju veći broj primjera).

Tabela 13. Izdvojena prediktivna pravila izvedena pomoću pretprocesiranog DT-a

Rule	Prod_gend	Prod_cat	Prod_age	Prod_brand	Prod_type	Gender	Discount	Cluster[1]	Conf.
R2	For boys	Lifestyle	For teens (8-14)	Licence	Apparel	M		0 (0 / 9 / 1)	90%
R3	For boys	Basket	For teens (8-14)	Licence	Footwear			0 (0 / 6 / 0)	100%
R4	For men		For adults	A brands	Apparel	M	35% - 45%	0 (2 / 7 / 0)	77,78 %
R5	For men	Lifestyle	For adults	Licence	Apparel	M	25% - 45%	0 (0 / 32 / 0)	100%
R6	For women	Lifestyle	For adults	Licence	Equipment	M		0 (0 / 6 / 0)	100%
R9	For men		For adults	A brands	Apparel	M	25% - 35%	1 (80 / 3 / 1)	96,39 %
R12	For men		For adults	A brands	Footwear	M	<35%	1 (66 / 0 / 3)	95,65 %
R14	For men	Lifestyle	For adults	Licence	Apparel		35% - 45%	1 (63 / 0 / 0)	100%
R15	For men	Lifestyle	For adults	Licence	Footwear		25% - 45%	1 (51 / 0 / 0)	100%
R31	For men	Lifestyle	For adults	A brands	Footwear	M	>45%	2 (0 / 0 / 43)	100%
R36	For women	Lifestyle	For adults	A brands	Footwear		>45%	2 (0 / 0 / 32)	100%

[1]U okviru ove kolone, u zagradama je dat broj primjera koji je svrstan u odgovarajući klaster, redosljedom klastera 1 / 0 / 2.

Na osnovu izvedenih pravila može se uočiti da su kupci sa CV-nivoom=1 (*cluster_0*), kupci muškog pola koji uglavnom kupuju: *lifestyle* odjeću za odrasle muškarce, licenciranih brendova i sa popustom od 25% do 45%; odjeću i patike za košarku, za dječake tinejdžerskog uzrasta, licenciranih brendova; kao i sportsku opremu za žene

licenciranih brendova. Kupci CV-nivoa=2 (*cluster_1*) uglavnom kupuju odjeću i patike A brendova, za odrasle muškarce sa popustom manjim od 35%, ili *lifestyle* odjeću i patike licenciranih brendova za odrasle muškarce sa popustom od 35%-45%. Najmanje vrijedni kupci iz trećeg klastera su uglavnom kupci patika za odrasle A brendova sa visokim popustom (>45%). U direktnoj kampanji, kupcima iz odgovarajućeg klastera mogu se ponuditi proizvodi koji su identifikovani pravilima. Takođe, na osnovu karakteristika kupca i proizvoda koji mu se nude može se predvidjeti CV-nivo za tog kupca sa vjerovatnoćom od oko 86% a samim tim i da li će se ovakav kupac targetirati u budućim kampanjama.

Diskusija rezultata i potvrda hipoteza H1 i H2

Sprovođenjem procesa segmentacije i odabirom broja definisanih potrošačkih segmenata na osnovu objektivnog parametra, izbjegava se subjektivna procjena u analizi i segmentaciji kupaca. U ovoj studiji je korišćen *Davies-Bouldin* indeks, koji precizno i objektivno, na osnovu raspoloživih podataka dijeli bazu kupaca na odgovarajući broj podgrupa. U tom slučaju, segmentacija se vrši na osnovu podataka o prethodnom kupovnom ponašanju – RFM atributa, te se, na ovaj način, obezbjeđuje maksimalna homogenost kupaca u okviru pojedinačnih segmenata, maksimalna heterogenost između različitih segmenata (ilustrovano na Slici 37) i definiše se optimalan broj segmenata, čime je potvrđena hipoteza H1.

Pored unapređenja procesa segmentacije kupaca, cilj ovog rada bio i kreiranje kombinacije SVM i DT metoda, odnosno hibridne SVM-RE metode, koja će biti testirana u pogledu tačnosti klasifikacije minorne klase i koja će moći da generiše pravila klasifikacije, uzimajući u obzir da je sama SVM tzv. „*black box*" metoda, koja se ne može interpretirati. Rezultati testiranja su pokazali da ova metoda omogućava uspješnu selekciju i targetiranje najvrednijih kupaca, što potvrđuje ostvarena vrijednost *class recall* metrike za minornu klasu od 62,79%, dok je samostalni DT model identifikovao svega 4,27% najvrednijih kupaca, a zanemario gotovo 96% kupaca, što može drastično smanjiti mogući prihod od kampanje. Dakle, SVM-RE je targetirao 16 puta više vrlo vjerovatnih respondenata nego DT i značajno povećao

šanse za visoki prihod od kampanje. Osim toga, za isti segment, *class precision* kod SVM-RE modela iznosi 77,45%, što znači da bi manje od 23% kupaca koji bi bili odabrani za targetiranje u narednim kampanjama vrlo vjerovatno ostali nerespondenti. S druge strane, samostalni DT model za istu grupu kupaca ostvario je svega 33,33% preciznosti, čime bi gotovo 67% kupaca bilo neefikasno targetirano u narednim aktivnostima direktnog marketinga. U tom smislu, SVM-RE metoda, u odnosu na DT, trostruko smanjuje troškove kampanja, što je posebno značajno u situacijama ograničenih marketing budžeta.

Predloženi model je prevazišao problem minorne klase, odnosno riješio problem preklapanja klasa i obezbijedio precizniju klasifikaciju kupaca. U ovom modelu, ostvarena je ukupna tačnost od 85,73%, dok je samostalni DT ostvario tačnost od svega 61,25%. Uz to, generisana pravila pomoću hibridnog modela detaljno opisuju svaki pojedinačni segment (uključujući i onaj najmanji s najvrednijim kupcima, koji bi većina klasifikatora potpuno zanemarila ili veoma slabo opisala), što olakšava planiranje kampanja direktnog marketinga i buduću selekciju kupaca. Izvedena pravila imaju *confidence* veći od 80% (Tabela 13), čime se potvrđuje da precizno interpretiraju SVM model.

Dakle, SVM-RE je za minornu klasu ostvario *class recall* od 62,70% i *class precision* od 77,45%, čime je potvrđena hipoteza H2.1. Takođe, kod minorne klase, *class recall* DT modela unaprijeđen je za 58,43 procentna poena, a *class presission* za 44,12 procentna poena, čime je potvrđena hipoteza H2.2. Na kraju, DT je, pri izvođenju pravila iz SVM predikcije, ostvario *confidence* veći od 80%, čime je potvrđena i hipoteza H2.3.

S tim u vezi, u ovoj studiji je potvrđena efikasnost targetiranja i predikcije najvrednijih kupaca kod RFM segmentacionog modela baziranog na SVM metodi, čime se smanjuju nepotrebni troškovi kampanje, povećavaju ukupni prihodi i na osnovu generisanih pravila formira profil segmenta najvrednijih kupaca, koji omogućava efikasniju interakciju s njima, čime je potvrđena hipoteza H2.

Sumarna razmatranja u vezi sa SVM-RE baziranim modelom prediktivne RFM segmentacije

U ovom dijelu disertacije predložena je efikasna metoda za predikciju vrijednosti kupaca u direktnom marketingu i njihovu selekciju, tj. targetiranje na osnovu te vrijednosti. Predložena prediktivna procedura podrazumijeva klasfikaciju klasterizacije kupaca. S tim u vezi, najprije se kupci klasterizuju (pomoću *k-means* algoritma) na osnovu njihovog kupovnog ponašanja (tačnije, pomoću RFM atributa). Kupci koji pripadaju različitim klasterima imaju veći ili manji CV-nivo, odnosno stepen vrijednosti za kompaniju, a samim tim i veću ili manju vjerovatnoću odgovora na kampanju direktnog marketinga u kojoj su targetirani. Zatim se pomoću SVM-RE metode, na osnovu karakteristika kupca i podataka o proizvodima koje je kupovao, predviđa pripadnost kupca jednom od klastera, tj. njemu odgovarajući CV-nivo, na osnovu čega se može donijeti odluka da li će taj kupac biti targetiran za naredne kampanje ili ne. Uz to, SVM-RE metoda vrši generisanje klasifikacionih pravila na osnovu kojih se mogu targetirati novi kupci u kampanji i ponuditi odgovarajući proizvodi kroz prilagođene ponude.

Empirijsko testiranje je pokazalo da je predložena metoda uspješno riješila problem nebalansiranosti klasa, koji često dovodi do pogrešne klasifikacije najmanje klase (najvrednijih kupaca). Povećanjem vrijednosti za *class recall* i *class precision* za klasu najvrednijih kupaca, SVM-RE metoda može značajno povećati prihode i smanjiti troškove direktne marketing kampanje. Takođe je dokazano da se SVM metoda može koristiti kao pretprocesor podataka, koji uspješno rješava problem nebalansiranosti i preklapanja klasa i unapređuje klasifikacione performanse.

U budućim istraživanjima, ova metoda se može testirati na drugim skupovima podataka kako bi se potvrdila ili unaprijedila njena efikasnost (uključivanjem više karateristika kupaca mogu se dobiti jasnija pravila za targetiranje novih kupaca).

5.3 Testiranje ensemble baziranog modela prediktivne RFM segmentacije

Poboljšanje performansi modela iz prethodne sekcije pomoću *ensemble* metoda podrazumijeva implementaciju prediktivne procedure prikazane na Slici 33 iz sekcije 4.5.2. U prvom koraku predložene procedure, u fazi obučavanja generisan je klaster model na isti način kao u prethodnoj sekciji.

Tabela 14. Izbor broja klastera (parametar *k*) za *k-means* klasterizaciju

K	2	3	4	5	6	7	8	9	10
DB	-1,025	**-0,811**	-0,983	-0,958	-0,909	-1,02	-0,98	-0,976	-0,96

Tabela 14 pokazuje distribuciju učestalosti kupovine, recentnosti kupovine i profitabilnosti kupaca u ovim segmentima na nivou vrijednosti kupca.

Tabela 15. Centroid klaster model za RFM segmentaciju kupaca

	Recency	*Frequency*	*Monentary*	Broj primjera
cluster_0	0,766807	0,565126	0,677733	238
cluster_1	0,735348	0,088065	0,179499	819
cluster_2	0,137318	0,101734	0,211345	548

Napomena: U tabeli su predstavljene normalizovane vrijednosti centroida (transformacija ranga 0-1)

Kao i u prethodnom slučaju, može se primijetiti da *cluster_0* čine kupci koji su najskorije obavili transakcije, koji ih obavljaju često, kao i kupci s transakcijama najveće vrijednosti (*Customer Value* - CV-nivo=1), *cluster_1* čine kupci koji su skoro obavljali transakcije, ali s manjom frekvencijom i vrijednošću transakcije (CV-nivo=2), dok *cluster_2* čine kupci koji nisu skoro obavljali transakcije, kao i oni koji ne kupuju često i čije su vrijednosti traksakcije male (CV-level=3). Najvredniji klaster kupaca sadrži znatno manje kupaca od druga dva klastera (238 naspram 819 i 548), tako da je problem neravnoteže klasa evidentan.

Tabela 16 pokazuje distribuciju učestalosti kupovine, recentnosti kupovine i profitabilnosti kupaca u ovim segmentima na nivou vrijednosti kupca. Može se primijetiti da su kupci iz najvrednijeg segmenta kupci koji su posljednju transakciju obavili skoro, da su kupovali u prosjeku 10 puta u posmatranom periodu i da im je prosječan iznos trgovine oko 242 eura. Nasuprot tome, kupci s nivoom vrijednosti 3 kupuju u prosjeku najviše 3 puta, s prosječnom vrijednošću transakcije od oko 82 eura.

Tabela 16. CV-nivo definisanih segmenata kupaca

CV-nivo	*Recency*	*Frequency*	*Monetary*
1	Avg: 4 Min: 3 Max: 5	Avg: 10 Min: 3 Max: 17	Avg: 241,65 € Min: 113,4 € Max: 352 €
2	Avg: 4 Min: 3 Max: 5	Avg: 2,4 Min: 1 Max: 7	Avg: 71,06 € Min: 9,6 € Max: 199,5 €
3	Avg: 1,54 Min: 1 Max: 2	Avg: 2,6 Min: 1 Max: 9	Avg: 81,96 € Min: 12,5 € Max: 239,5 €

Faza obučavanja prediktivnog modela

U nastavku faze obučavanja, u cilju pretprocesiranja koje će smanjiti nebalansiranost klasa, najprije je generisan *Bagging SVM* model za predviđanje CV-nivoa, koji je dobijen početnom klasterizacijom kupaca. Korišćenjem *Grid-Search* tehnike i *10-fold* kros-validacije utvrđena je optimalna kombinacija parametara za SVM i *Bagging*: *SVM.C = 400,6, SVM.gamma = 200,006, Bagging.sample_ratio = 0,9 i Bagging.iterations = 10*. Model je zatim generisao predviđanje na nivou CV-nivoa,

koje se uzima kao oznaka klase (eng. *class label*) skupa podataka za obučavanje modela.

U sljedećem koraku, kreiran je poduzorkovan skup za obučavanje modela, koji je isključio sva predviđanja čiji je *Confidence <= 0,9*, odnosno one rezultate za koje je samo 90% modela i manje glasalo u proceduri *Bagging SVM*. U Tabeli 17 prikazan je tako dobijen novi skup podataka za obuku.

Tabela 17. Izmjene u skupu za obučavanje modela u prediktivnoj proceduri

Skup podataka za obučavanje modela	Oznaka klase (eng. *class label*)	Distribucija oznake klase	Broj primjera
Početni skup	*CV-Level*	CV-Level 1 (238) CV-Level 2 (819) CV-Level 3 (548)	1605
Skup dobijen kao *Bagging* SVM izlaz	*Bagging SVM predicted CV-Level*	CV-Level 1 (132) CV-Level 2 (981) CV-Level 3 (492)	1605
Poduzorkovan skup (Conf. > 0.9)	*Bagging SVM predicted CV-Level*	CV-Level 1 (82) CV-Level 2 (834) CV-Level 3 (360)	1276

Nakon poduzorkovanja, obučavan je balansirani *Bagging* DT klasifikator na novom skupu podataka za obučavanje modela. *Grid-Search* tehnikom i *10-fold* kros-validacijom utvrđena je optimalna kombinacija parametara: *Bagging.sample_ratio = 0,9, Bagging.iteration = 304, balancing_proportion: 82:500:100, split_criterion = gain_ratio, min_size_for_split = 4, min_leaf_size = 2, max_depth = 15, confidence = 0,2, min_gain = 0,01*.

Za potrebe poređenja, samostalni DT klasifikator je obučen sa sljedećom optimalnom kombinacijom parametara: *split_criterion = gini_index, min_size_for_split = 4, min_leaf_size = 16, max_depth = 15, confidence = 0,1, min_gain = 0,01*.

Performanse klasifikacije svih obučenih modela date su u Tabeli 18.

Tabela 18. Rezultati faze obučavanja modela (kros-validacione performanse)

Model	Tačnost	*Class Recall*	*Class Precision*
Bagging SVM[4]	61,00%	**21,01%**[1], 81,07%[2], 48,36%[3]	**50,51%**[1], 61,37%[2], 62,50%[3]
Balansirani *Bagging* DT na *Bagging* SVM izlazu[5]	88,71%	**69,51%**, 96,76%, 74,44%	**80,28%**, 87,91%, 93,38%
Samostalni DT[6]	60,69%	**4,20%**, 80,34%, 55,84%	**27,78%**, 61,90%, 60,47%

[1] Performanse klase za CV-nivo 1 (najvredniji kupci – minorna klasa); [2] Performanse klase za CV-nivo 2; [3] Performanse klase za CV-nivo 3; [4] Ovaj model je pretprocesor podataka; [5] Performanse ovog modela su zapravo performanse konačnog modela za klasifikaciju segmenata kupaca pod nazivom *ensemble* SVM-RE; [6] Samostalni DT je generisan u svrhu poređenja rezultata.

Iz prethodne tabele može se uočiti da *Balansirani Bagging DT na Bagging SVM izlazu* model (u daljem tekstu: *ensemble SVM-RE* model) ima značajno bolje klasifikacione performanse od samostalnog DT modela. Naime, samostalni DT model tačno klasifikuje svega 4% najvrednijih kupaca, dok *ensemble SVM-RE* tačno targetira 69% kupaca iz ovog segmenta. Nadalje, sve posmatrane klasifikacione performanse sa *ensemble SVM-RE* modelom su bolje u poređenju sa samostalnim DT modelom. Za klasu najvrednijih kupaca, *class precision* DT modela je samo 28%, što implicira da bi kompanija imala nepotrebnih troškova kampanje za 72% pogrešno klasifikovanih kupaca. Preciznost *ensemble SVM-RE* modela za ovaj segment kupaca je 80%, što značajno smanjuje troškove za pogrešno targetirane kupce, kojih je u ovom slučaju svega 20%. Stoga, *ensemble SVM-RE* može, u poređenju sa samostalnim DT modelom, značajno smanjiti troškove kampanje za one kupce od kojih se odgovor na kampanju ne očekuje, odnosno, za one koji ne pripadaju klasi najvrednijih kupaca. Osim toga, može se zaključiti da predložena *ensemble* SVM-RE metoda s velikom stopom tačnosti predviđanja CV-nivoa (89%) uspješno rješava problem nebalansiranosti klasa.

Rezultati pokazuju da je *Bagging SVM*, kao pretprocesor podataka o kupovnim transakcijama, eliminisao šum u podacima, tako da je moguća preciznija klasifikacija. Tačnost DT klasifikacije je povećana za 28 procentnih poena - tačnost za samostalni DT je 61%, dok je *ensemble SVM-RE* ostvario tačnost od 89%. Prosječni *class recall* za samostalni DT je 47%, a nakon pretprocesiranja podataka i korišćenja *ensemble* metode, ova metrika ima vrijednost 80%. Srednja preciznost klase (eng. *class precision*) je povećana sa 50% na 87% nakon pretprocesiranja.

S obzirom na visoku tačnost kros-validacije ekstrakcije pravila iz *Bagging SVM* izlaza (89%), pravila valjano tumače *Bagging SVM* klasifikaciju. Tabela 19 prikazuje neka od 81 izvedenih pravila, koja su prepoznata kao najvažnija, odnosno pravila koja pokrivaju veliki broj primjera (*Support* ~ 1% i više, osim za minornu klasu, gdje je minimalni iznos za *Support* 0,3%), imaju visoku tačnost (*Confidence* > 80%) i dobro povjerenje u odnosu na ukupan skup podataka (*Lift* > 1).

Tabela 19. Najznačajnija klasifikaciona pravila generisana iz *ensemble* SVM-RE modela

CV-nivo	Pravilo	*Confidence* (>80%)	*Support* (>1%*)	*Lift* (>1)
CV-nivo 1	if Discount > 45% and Prod_category = Football and Prod_age = For younger kids (4-10)	100%	0,3%	8,33
CV-nivo 1	if Discount > 45% and Prod_category = Outdoor and Prod_type = Footwear and Prod_brand = Licence	100%	0,4%	8,33
CV-nivo 1	if Discount ≤ 45% and Discount > >35% and Cons_gender = F and Prod_type = Apparel and Prod_brand = A brands and Prod_gender = For women	100%	0,3%	8,33
CV-nivo 1	if Discount ≤ 45% and Discount > 25% and Cons_gender = M and Prod_type = Apparel and Prod_age = For adults and Prod_gender = For men and	100%	0,4%	8,33

	Prod_category = Basketball and Prod_brand = Licence			
CV-nivo 1	if Discount ≤ 35% and Discount > 25% and Cons_gender = M and Prod_type = Apparel and Prod_age = For adults and Prod_gender = For men and Prod_category = Football	100%	0,4%	8,33
CV-nivo 1	if Discount ≤ 45% and Discount > 35% and Cons_gender = M and Prod_type = Apparel and Prod_age = For adults and Prod_gender = For men and Prod_category = Lifestyle and Prod_brand = A brands	100%	1%	8,33
CV-nivo 1	if Discount ≤ 35% and Discount > 25% and Cons_gender = M and Prod_type = Apparel and Prod_age = For adults and Prod_gender = For men and Prod_category = Lifestyle and Prod_brand = Licence	100%	4%	8,33
CV-nivo 1	if Discount ≤ 45% and Discount > 25% and Cons_gender = M and Prod_type = Apparel and Prod_age = For teens (8-14)	80%	1%	6,67
CV-nivo 1	if Discount ≤ 45% and Discount > 10% and Cons_gender = M and Prod_type = Equipment and Prod_brand = Licence and Prod_gender = For women	100%	1%	8,33
CV-nivo 2	if Discount > 45% and Prod_category = Skiing	86%	2%	1,17
CV-nivo 2	if Discount ≤ 35% and Discount > 25% and Cons_gender = F and Prod_type = Apparel and Prod_brand = A brands	100%	6%	1,37

CV-nivo 2	if Discount ≤ 45% and Discount > 25% and Cons_gender = F and Prod_type = Apparel and Prod_brand = Licence	100%	11%	1,37
CV-nivo 2	if Discount ≤ 45% and Discount > 25% and Cons_gender = F and Prod_type = Equipment	94%	3%	1,29
CV-nivo 2	if Discount ≤ 45% and Discount > 35% and Cons_gender = F and Prod_type = Footwear	98%	7%	1,34
CV-nivo 2	if Discount ≤ 35% and Discount > 25% and Cons_gender = F and Prod_type = Footwear and Prod_gender = For men	100%	2%	1,37
CV-nivo 2	if Discount ≤ 45% and Discount > 35% and Cons_gender = M and Prod_type = Apparel and Prod_age = For adults and Prod_gender = For men and Prod_category = Lifestyle and Prod_brand = Licence	100%	4%	1,37
CV-nivo 2	if Discount ≤ 35% and Discount > 25% and Cons_gender = M and Prod_type = Apparel and Prod_age = For adults and Prod_gender = For men and Prod_category = Lifestyle and Prod_brand = A brands	100%	3%	1,37
CV-nivo 2	if Discount ≤ 45% and Discount > 10% and Cons_gender = M and Prod_type = Footwear and Prod_age = For adults and Prod_category = Lifestyle	100%	8%	1,37
CV-nivo 3	if Discount > 45% and Prod_category = Basketball	88%	1%	5,83
CV-nivo 3	if Discount > 45% and Prod_category = Fitness and Cons_gender = M and Prod_brand = A brands	100%	1%	6,67

CV-nivo 3	if Discount > 45% and Prod_category = Lifestyle and Prod_brand = A brands and Cons_gender = M	95%	3%	6,33
CV-nivo 3	if Discount > 45% and Prod_category = Lifestyle and Prod_brand = Licence and Prod_age = For adults and Prod_type = Apparel and Prod_gender = For men	100%	1%	6,67
CV-nivo 3	if Discount > 45% and Prod_category = Lifestyle and Prod_brand = Licence and Prod_age = For adults and Prod_type = Footwear and Cons_gender = F	100%	1%	6,67
CV-nivo 3	if Discount > 45%and Prod_category = Running and Prod_brand = A brands	91%	2%	6,06
CV-nivo 3	if Discount ≤ 25% and Discount > 10% and Cons_gender = M and Prod_type = Apparel and Prod_gender = For men	100%	1%	6,67

Napomena: kriterijum za „*Support*" za izabrana pravila je ~1% i više, osim za minornu klasu, gdje je minimalni *Support*=0,3%.

Na osnovu izvedenih pravila, može se konstatovati da su kupci sa CV-nivoom=1 uglavnom kupci muškog pola, koji najviše kupuju košarkašku odjeću za muškarce licenciranih brendova (brendovi za koje *Sport Vision* ima licencu za proizvodnju i distribuciju, kao što su: *Champion, Umbro, Lonsdale, Ellesse, Slazenger, Sergio Tacchini* itd.) i s popustom od 25% do 45%; odjeću za odrasle muškarce, bilo za fudbal ili *lifestyle* kategoriju licenciranih brendova, s popustom između 25% i 35%; pored toga, u ovaj segment spadaju i muškarci koji kupuju odjeću za tinejdžere s popustom od 25-45%, ili žensku opremu s popustom od 10-45%.

Kupci koji imaju CV-nivo=2 su uglavnom žene, koje ili kupuju odjeću A brendova (brendovi višeg ranga, čiji je kompanija distributer, kao što su: *Adidas, Nike, Under Armour, Reebok, Converse* itd.), uz popust od 25% do 35%, ili odjeću licenciranih brendova i opreme, na sniženju od 25% do 45%. Pored toga, u ovaj segment kupaca spadaju i žene, koje kupuju mušku obuću s popustom od 25% do 35%. Kupci muškog

pola u ovoj kategoriji uglavnom kupuju *lifestyle* odjeću za odrasle muškarce ili licenciranih brendova sa 35% do 45% popusta ili A brendova od 25% do 35% popusta.

CV-nivo=3 predstavlja grupu najmanje vrijednih kupaca. Kupci koji pripadaju ovoj grupi uglavnom kupuju proizvode s popustom većim od 45% (kupci koji dominantno kupuju na "sniženjima"). Muški kupci u ovoj kategoriji uglavnom kupuju *lifestyle* proizvode ili opremu A brendova. Kupci u ovom segmentu kupuju *lifestyle* obuću za odrasle od licenciranih brendova.

Dakle, sa *ensemble* SVM-RE modelom mogu se generisati pravila koja objašnjavaju segmente kupaca. Također, rješavanje problema male klase omogućava efikasniji opis segmenta najvrednijih kupaca s većim brojem važnih pravila.

Faza testiranja modela

Da bi se odredile prediktivne performanse modela na testnom skupu, na ovom skupu podataka je sprovedena klasterizacija korišćenjem modela klastera generisanog u fazi obuke (vidjeti Sliku 33). Na ovaj način, svakom korisniku se dodjeljuje odgovarajući CV-nivo, koji će biti korišćen za određivanje prediktivnih performansi u fazi testiranja. Sljedeći korak je pretprocesiranje testnog skupa korišćenjem *Bagging SVM* modela, koji je obučen u prvoj fazi (fazi obuke), a zatim i poduzorkovanje testnog skupa, uzimajući primjere sa oznakom klase za koju je glasalo više od 90% SVM modela tokom *Bagging* procedure. Karakteristike testnog skupa prije i poslije pretprocesiranja prikazane su u Tabeli 20.

Tabela 20. Izmjene testnog skupa podataka u fazi testiranja

Testni skup	Oznaka klase	Distribucija oznake klase	Broj primjera
Početni skup	*CV-Level*	CV-Level 1 (286) CV-Level 2 (2906) CV-Level 3 (2027)	5219
Skup dobijen kao *Bagging* SVM izlaz	*Bagging SVM predicted CV-Level*	CV-Level 1 (491) CV-Level 2 (3399) CV-Level 3 (1329)	5219
Poduzorkovan skup (*Undersampled* - Conf. > 0.9)	*Bagging SVM predicted CV-Level*	CV-Level 1 (406) CV-Level 2 (3155) CV-Level 3 (1043)	4604

Važno je napomenuti da je *Bagging SVM* dopunio najmanji segment kupaca, tako da on sadrži 491 instancu nakon pretprocesiranja podataka. Pored toga, ovaj pretprocesor je uklonio preklapanja između segmenata, kako bi buduća klasifikacija bila što preciznija. Nakon poduzorkovanja, na ovaj testni skup podataka primijenjen je obučeni ensemble SVM-RE model. Dobijeni rezultati prikazani su u Tabeli 21.

Tabela 21. Performanse *ensemble* SVM-RE modela na nepoznatim podacima

Model	Tačnost	*Class Recall*	*Class Precision*
***Ensemble* SVM-RE**	85,79%	**93,84%**[1], 84,44%[2], 86,77%[3]	**79,38%**[1], 94,57%[2], 69,24%[3]

[1] Performanse klase za CV-nivo 1 (najvredniji kupci – minorna klasa); [2] Performanse klase za CV-nivo 2; [3] Performanse klase za CV-nivo 3;

Iz gornje tabele se može vidjeti da je ukupna tačnost *ensemble SVM-RE* modela na prethodno nepoznatim podacima 85,79%, što je dobro u poređenju sa ostvarenom tačnošću u kros-validaciji, koja je iznosila 88,71%. Osim održavanja slične ukupne tačnosti kao kod kros-validacije modela, nepoznati podaci pokazuju i dobre performanse za najmanju - najvredniju klasu kupaca (*class precision* od oko 80% i *class recall* od oko 94%), što je od izuzetne važnosti, jer su postojeći i predviđeni potencijalno najvredniji kupci precizno identifikovani.

Ensemble SVM-RE modelom bi se targetiralo ukupno 480 najvrednijih kupaca za kampanju. Od toga, 381 najvredniji kupac bi bio tačno targetiran (94% svih takvih kupaca u testnom skupu), a 99 kupaca bilo bi targetirano bez vjerovatnog odgovora. Dakle, oko 80% poslatih ponuda bilo bi potencijalno profitabilno, dok bi 20% moglo biti bezuspješno. Konfuziona matrica prikazana je u Tabeli 22.

Tabela 22. Konfuziona matrica za *ensemble* SVM-RE klasifikaciju na testnom skupu podataka

	True CV-Level 1	***True CV-Level 2***	***True CV-Level 3***	***Total***	***Class Precision***
Pred. CV-Level 1	381	91	8	480	79,38%
Pred. CV-Level 2	23	2664	130	2817	94,57%
Pred. CV-Level 3	2	400	905	1307	69,24%
Total	406	3155	1043	4604	
Class Recall	93,84%	84,44%	86,77%		**Tačnost:** 85,79%

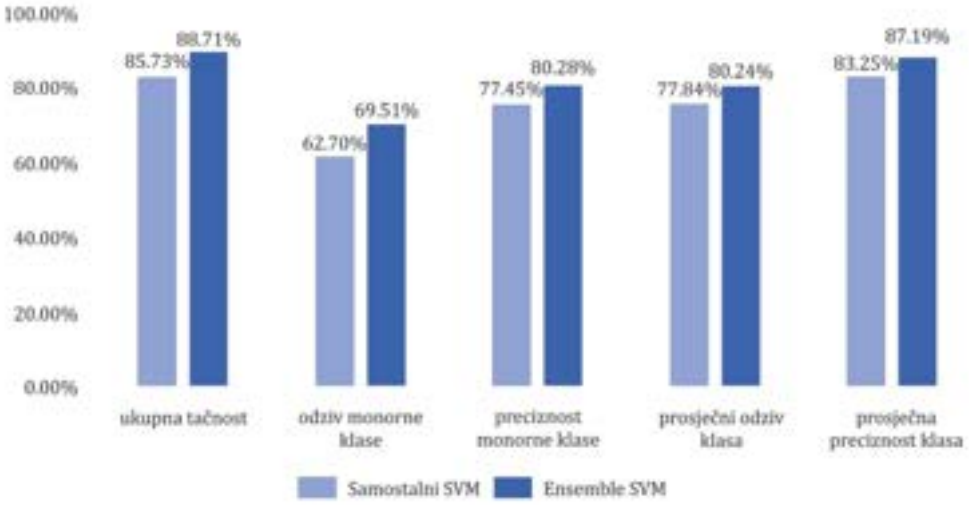
Dakle, kao rezultat faze testiranja, može se zaključiti da je predloženi *ensemble* SVM-RE model efikasan prediktor za CV-nivo, odnosno adekvatan model za klasifikaciju segmenta kupaca.

Poređenje sa SVM-RE baziranim metodom segmentacije i odgovor na IP1

U poređenju rezultata pretprocesiranja podataka samostalnom SVM-RE metodom (Tabela 12 iz prethodne sekcije) u odnosu na *ensemble* SVM-RE, može se zaključiti da su poboljšane sve prediktivne performanse:

- Ukupna tačnost je povećana sa 85,73% na 88,71%;
- Odziv minorne klase je povećan sa 62,70% % na 69,51%;
- Preciznost minorne klase je povećana sa 77,45% na 80,28%;
- Prosječni odziv klasa je povećan sa 77,84% na 80,24%;
- Prosječna preciznost klasa je povećana sa 83,25% na 87,19%.

Na Slici 38 prikazana je komparacija rezultata samostalne SVM-RE metode u odnosu na *ensemble* SVM metodu.



Slika 38. Komparacija rezultata samostalne i *ensemble* SVM metode

Kao što se na osnovu rezultata vidi, na ovaj način je, kada je u pitanju model prediktivne RFM segmentacije kupaca, na IP1 odgovoreno pozitivno, jer su *ensemble* tehnike popravile performanse modela. Takođe se može primijetiti da je riječ o svega nekoliko procenata poboljšanja. Međutim, i nekoliko procenata najvrednijih kupaca može biti značajno kada govorimo o profitabilnosti kampanje. Koliko je zaista došlo do poboljšanja rezultata pokazaće poređenje u terminima profitabilnosti u nastavku. Stoga se *ensemble* SVM-RE model može smatrati konačnim predlogom za prediktivnu RFM segmentaciju, pa će, u nastavku ovog rada, u cilju odgovora na IP2, ovaj model biti upoređen sa samostalnim *ensemble* metodama. Osim poboljšanja prediktivnih performansi, može se uočiti i unapređenje opisa segmenta najvrednijih kupaca. Naime, u Tabeli 19 broj pravila koji opisuju klasu CV-nivo 1 je devet, dok ih je u Tabeli 13 ukupno pet, što znači da je *ensemble* SVM-RE dao semantički bogatiji opis.

Poređenje s drugim metodama za balansiranje klasa i odgovor na IP2

U cilju poređenja predloženog *ensemble SVM-RE* modela sa samostalnim *ensemble* modelima po pitaju efikasnosti u rješavanju problema minorne klase, testirana je kombinacija DT klasifikatora s različitim balansiranim *ensemble* tehnikama (*ensemble* u kombinaciji s balansiranim poduzorkovanjem, koje izjednačava broj primjera u svakoj klasi s brojem primjera u minornoj). Prvo je na početnom skupu podataka za obučavanje generisan Balansirani *Bagging DT* model sa sljedećim parametrima: *Bagging.sample_ratio = 0,7, Bagging.iterations = 108, balancing_proportion: 238: 238: 238, criterion = gain_ratio, min_size_for_split = 4, min_leaf_size = 2, max_depth = 15, confidence = 0,2, min_gain = 0,01*. Zatim je generisan Balansirani *AdaBoost DT* model sa sljedećim parametrima: *AdaBoost.iterations = 3001, balancing_proportion: 238: 238: 238, criterion = gain_ratio, min_size_for_split = 4, min_leaf_size = 2, max_depth = 15, confidence = 0,2, min_gain = 0,01*. Konačno, generisan je i Balansirani *Random Forest* model s parametrima: *RandomForest.sample_ratio = 1,0, Random.Forest.iterations = 75, balancing_proportion: 238: 238: 238, criterion = gain_ratio, max_depth = 10*. Optimalna kombinacija parametara dobijena je korišćenjem *Grid-Search* tehnike i *10-fold* kros-validacije.

Performanse balansiranja klasa prethodno pomenutih modela date su u narednoj tabeli.

Tabela 23. Klasifikacione performanse samostalnih *ensemble* modela

Model	Tačnost	*Class Recall*	*Class Precision*
Kros-validacione performanse			
Balansirani *Bagging* DT	56,69%	**44,12%**[1], 51,28%[2], 70,26%[3]	**34,20%**[1], 68,74%[2], 56,04%[3]
Balansirani *AdaBoost* DT	55,32%	**41,60%**, 64,22%, 57,30%	**30,43%**, 69,44%, 57,46%
Balansirani *Random Forest*	51,03%	**53,36%**, 41,76%, 63,87%	**28,93%**, 69,94%, 51,70%
Performanse na testnom skupu (nepoznati podaci)			
Balansirani *Bagging* DT	49,55%	**26,22%**, 43,53%, 61,47%	**8,35%** , 66,20%, 51,70%
Balansirani *AdaBoost* DT	46,14%	**35,66%**, 46,73, 46,77%	**7,53%**, 65,48%, 52,93%
Balansirani *Random Forest*	44,51%	**34,62%**, 37,89%, 55,40%	**7,12%**, 67,01%, 51,37%

[1] Performanse klase za CV-nivo 1 (najvredniji kupci – minorna klasa); [2] Performanse klase za CV-nivo 2; [3] Performanse klase za CV-nivo 3.

Upoređujući rezultate *ensemble* SVM-RE metode iz Tabele 21 sa samostalnim balansiranim *ensemble* metodama *Bagging DT, AdaBoost DT* i *Random Forest* u Tabeli 23, može se zaključiti da ova metoda nadmašuje njihove mogućnosti u pogledu balansiranja klasa, odnosno pružanja rješenja za problem minorne klase. Naime, dok su za *ensemble* SVM-RE *class recall* i *class precision* iznosili 69,51% odnosno 80,28%, najbolji *class recall* za minornu klasu ostvaren je u *Random Forest* modelu (53,36%), a najbolji *class precision* za minornu klasu ostvaren je u balansiranom *Bagging DT* modelu (34,20%). Takođe, maksimalna ukupna tačnost samostalnih balansiranih *ensemble* modela (55,69%) je značajno manja od modela *ensemble* SVM-DT (88,71%). Kada se uporede rezultati na testnom skupu (Tabela 23), superiornost *ensemble* SVM-RE modela je još izraženija.

Konačno, može se zaključiti da SVM u kombinaciji sa *ensemble* meta-algoritmom *Bagging* efikasnije rješava probleme neravnoteže klasa od samostalnih balansiranih

ensemble metoda koje se koriste za ovu svrhu, čime je, kada je u pitanju model prediktivne RFM segmentacije, dat pozitivan odgovor na IP2.

Poređenje na osnovu kriterijuma profitabilnosti

Za poređenje modela u smislu potencijalno ostvarivog maksimalnog profita od kampanje na osnovu predviđanja minorne klase (tj. segmenta najvrednijih kupaca), u Tabeli 24 prikazan je obračun ovog indikatora pojedinačno, po modelima. Indikator profita se izračunava po formuli:

$$Profit = True\ Predicted\ \times R - (True\ Predicted + False\ Predicted) \times C$$

gdje je *True Predicted* – broj tačno predviđenih kupaca iz najvrednijeg segmenta; R - potencijalni prihod jednog klijenta u kampanji; *False Predicted* – broj netačno predviđenih kupaca najvrednijeg segmenta iz modela i C – procijenjeni trošak kampanje po jednom kupcu. Pretpostavlja se da je potencijalni prosječan prihod ostvaren u prethodnim kampanjama na nivou najvrednijeg segmenta 242 eura (vidjeti Tabelu 24), dok je procijenjeni trošak po kampanji po kupcu 1 euro. Broj tačno predviđenih i netačno predviđenih članova najvrednijeg segmenta kupaca se daje proporcionalno učešću ove klase u početnom skupu podataka za obučavanje modela (s obzirom na to da je skup za obučavanje dobijen na izlazu *Bagging SVM* modela poduzorkovan).

Tabela 24. Poređenje modela prema potencijalnoj profitabilnosti kampanje

Model	*True Predicted (TP)*[1]	*False Predicted (FP)*[2]	Prihodi[3]	Troškovi[4]	Profit[5]
SVM-RE[6]	150	44	36.300	194	36.106
Ensemble SVM-RE	165	41	39.930	206	39.724
Balansirani Bagging DT	105	202	25.410	307	25.103
Balansirani AdaBoost DT	126	288	30.492	414	30.078
Balanced Random Forest	127	312	30.734	439	30.295

[1] Broj tačno predviđenih kupaca najvrednijeg segmenta
[2] Broj netačno predviđenih kupaca najvrednijeg segmenta

[3] TP * Potencijalni prihod pojedinačnog klijenta (€ 242)
[4] (TP+FP) * Procijenjeni trošak kampanje za jednog klijenta (€ 1)
[5] Prihodi - Troškovi
[6] Rezultati za samostalni SVM-RE preuzeti su iz Tabele 12 iz prethodne sekcije

Na osnovu proračuna u gornjoj tabeli, primjećuje se da se maksimalni profit može očekivati na osnovu predikcije u *ensemble* SVM-RE modelu. Poboljšanje samostalne SVM-RE metode pomoću *ensemble* meta-algoritma može dovesti do povećanja profita u kampanji za 3.618 eura. Od samostalnih balansiranih *ensemble* metoda, najveći očekivani profit od 30.295 eura se postiže predviđanjem pomoću *Random Forest* modela, što je 9.429 eura manje od očekivane dobiti s predviđanjem u okviru *ensemble* SVM-RE.

Uzimajući u obzir poređenje prema kriterijumu prediktivnog učinka iz prethodnog odjeljka, kao i na osnovu kriterijuma profitabilnosti, može se zaključiti da je *ensemble* SVM-RE značajno bolja metoda balansiranja klasa od ostalih razmatranih metoda.

Validacija metoda testiranjem na javno dostupnom skupu podataka

Predloženi pristup validiran je na javno dostupnom skupu podataka *„Customer transaction dataset"*, koji je opisan u sekciji 5.1.1.

Kao i u slučaju osnovnog skupa podataka, u prvom koraku predložene procedure, izvršena je klasterizacija na osnovu RFM atributa. Na osnovu minimalne ostvarene apsolutne vrijednosti DB indeksa (-0,879) za optimalnu vrijednost k=3, formirana su tri klastera. Model centroida klastera za ovaj skup podataka predstavljen je u Tabeli 25.

Tabela 25. Model centroida klastera za javno dostupni skup podataka

	Recency	*Frequency*	*Monetary*	Broj instanci
Cluster 0	0,668	0,302	0,282	4.264
Cluster 1	0,969	0,623	0,549	7.034
Cluster 2	1,000	0,347	0,301	8.467

Napomena: Prikazane su normalizovane vrijednosti centroida (izvršena je transformacija za rang 0-1)

Klaster najvrednijih kupaca (*Cluster 1*) ima CV-nivo=1 i sadrži 7.034 primjera, klaster srednje vrijednih kupaca (*Cluster 2*) ima CV-nivo=2 i sadrži 8.467 primjera, dok je klaster najmanje vrijednih kupaca (Cluster 0) u ovom slučaju najmanji i ima 4.264 kupaca sa CV-nivoom koji iznosi 3. Uzimajući u obzir razlike koje postoje među brojem primjera u svakom od klastera, očigledno je i ovdje prisutan problem neuravnoteženih klasa, s tim što je u ovom slučaju minorna klasa klaster najmanje vrijednih kupaca.

Ponavljanjem iste prediktivne procedure definisane na Slici 33, u fazi obuke kros-validacijom i u fazi testiranja - testiranjem na nepoznatom skupu podataka, dobijeni su rezultati prikazani u Tabeli 26.

Tabela 26. Prediktivne performanse modela na javno dostupnom skupu podataka

Model	Parametri	Performanse kros-validacije	Performanse u testiranju modela
Bagging SVM	SVM.gamma = 0,0325 SVM. C = 1000,0 Bagging.iterations = 10 Bagging.sample_ratio = 0,8	accuracy: 46,15% class recall: 21,94%[1], 44,25%[2], 59,93%[3] class precision: 33,87%, 46,65%, 49,12%	accuracy: 47,29% class recall: 23,69%[1], 46,16%[2], 60,12%[3] class precision: 37,04%, 47,56%, 49,85%
Ensemble SVM-RE	DT.criterion = gain_ratio DT.min_size_for_split = 4 DT.minimal_leaf_size = 2 DT.maximal_depth = 15 DT.confidence = 0,1 DT.minimal_gain = 0,01 Bagging.sample_ratio = 0,9 Bagging.iterations = 100	accuracy: **88,69%** class recall: **61,61%**, 85,50%, 93,74% class precision: 74,24%, 88,71%, 90,03%	**accuracy: 90,32%** class recall: **70,07%**, 86,67%, 94,55% class precision: 72,03%, 91,57%, 91,41%

	Bagging.Balancig_propor tion: 435:1000:2000		
Samostalni DT	DT.criterion = gain_ratio DT.min_size_for_split = 4 DT.minimal_leaf_size = 2 DT.maximal_depth = 15 DT.confidence = 0,1 DT.minimal_gain = 0,01	accuracy: **43,01%** class recall: **0,50%**, 0,35%, 99,87% class precision: 78,95%, 80,95%, 42,90%	**accuracy: 43,08%** class recall: **0,55%**, 0,33%, 100% class precision: 100%, 100%, 42,94%

[1] Performanse klase za CV-nivo 1 (najvredniji kupci – minorna klasa); [2] Performanse klase za CV-nivo 2; [3] Performanse klase za CV-nivo 3.

Podaci u prethodnoj tabeli ukazuju da je *ensemble* SVM-RE i na ovom skupu podataka uspješno riješio problem netačne klasifikacije minorne klase, koji se javlja usljed neravnoteže i preklapanja klasa. Na skupu za obuku i testiranje, samostalni DT je potpuno pogrešno klasifikovao najbolje kupce (*class recall* je samo oko 0,3%) i najgore kupce (*class recall* oko 0,5%), a ukupna tačnost modela je oko 43%. Tačnost modela *ensemble* SVM-RE na nepoznatim podacima je oko 90%, *class recall* za najbolje kupce je oko 87%, a za najmanje vrijedan klaster oko 70%. *Bagging SVM* model je ostvario *class recall* ispod 50%, ne samo za najmanji klaster (najmanje vrijednih kupaca), već i za klaster najvrednijih kupaca, koji je sadržao značajno više primjera. Dakle, u ovom skupu podataka, pored problema minorne klase, postoji i problem preklapanja klasa (šuma), koji je *Bagging SVM* pretprocesor uspješno riješio.

Diskusija ostvarenih rezultata

Cilj predloženog modela je bio da testira nekoliko poboljšanja postojećih metoda za prediktivnu klasifikaciju kupaca u direktnom marketingu, kao što su objektivna segmentacija kupaca sa indikatorom za optimalan broj klastera, opis segmenata u pogledu karakteristika kupaca i proizvoda, predviđanje vrijednosti kupaca za nove kupce s nepoznatim kupovnim ponašanjem i, konačno - najvažnije, smanjenje

pogrešne klasifikacije za segment najvrednijih kupaca, odnosno rješenje problema neravnoteže klasa.

Za razliku od nekih prethodnih studija, u kojima je sprovedeno kodiranje RFM atributa, kao i sortiranje na bazi kodiranih vrijednosti i subjektivna selekcija segmenata za kampanju (Drozdenko & Drake, 2002; Hughes, 1994; McCarty & Hastak, 2007), ova disertacija predlaže sofisticiranu i objektivnu proceduru, zasnovanu na *data mining* tehnikama. Primjena *k-means* klasterizacije omogućava postizanje algoritamske segmentacije korišćenjem mjere Euklidske udaljenosti, kako bi se obezbijedila maksimalna sličnost unutar segmenata i razlika između segmenata. Umjesto jednoobraznog kodiranja RFM atributa, koje kupca ne tretira pojedinačno, već ga identifikuje s grupom kojoj pripada, što je karakteristika mnogih dosadašnjih studija (Cheng & Chen, 2009; Hosseini et al., 2010; Sarvari et al., 2016), u ovoj disertaciji je predložena klasterizacija s nekodiranim atributima, uzimajući u obzir da postoji veliki broj numeričkih varijabli koje omogućavaju nesmetano funkcionisanje algoritma klasterizacije. Ovo sprečava gubitak važnih informacija na nivou svakog pojedinačnog kupca, što ga može razlikovati od drugih. Za razliku od metode predložene u radu autora *Cheng* i *Chen* (2009), koja uključuje subjektivnu procjenu i testiranje optimalnog broja klastera, metoda predložena u ovom radu definiše optimalan broj klastera na objektivan način, upotrebom DB indeksa, što značajno pojednostavljuje proceduru i osigurava tačnost modela.

Klasična RFM segmentacija uključuje predviđanje budućeg ponašanja kupaca, koje je zasnovano samo na ova tri atributa, te se ne može primijeniti kod traženja novih kupaca, jer informacije o transakcijama novih kupaca nisu dostupne u bazi podataka (McCarty & Hastak, 2007). U svom istraživanju, *Cheng* i *Chen* (2009) su tokom segmentacije kupaca koristili sofisticirane tehnike rudarenja podataka, ali, pored karakteristika kupaca, i sami RFM atributi su uključeni kao prediktivni atributi, tako da se predloženi model ne može koristiti za nove kupce za koje su ovi atributi nepoznati. U ovom istraživanju, podaci o proizvodu i karakteristike kupaca se isključivo koriste kao prediktivni atributi, jer se očekuje da će se dobiti pravila

predviđanja s prikladnijim informacijama za targetiranje novih i nepoznatih potencijalnih kupaca (Tsai & Chiu, 2004).

Za prediktivnu klasifikaciju kupaca, ovo istraživanje predlaže SVM-RE metodu u kombinaciji sa *ensemble* tehnikama, koje unapređuju prediktivne performanse modela. Rezultati su pokazali da *ensemble* SVM efikasno vrši pretprocesiranje podataka, tj. rješava problem preklapanja i nebalansiranosti klasa. Prvo, pomjeranjem margine ka najbližim (a samim tim i najsličnijim) primjerima veće klase i njihovim klasifikovanjem u manju klasu, SVM prevazilazi problem šuma u podacima, odnosno preklapanja klasa i dopunjuje manju klasu najrelevantnijim primjerima. Zatim, objedinjavanjem rezultata više SVM modela u *ensemble* proceduri, instance koje se pridružuju najmanjoj klasi preciznije se identifikuju (primjer se pridružuje klasi za koju je SVM model najviše glasao). I na kraju, uzimajući samo rezultate za koje je glasalo više od 90% SVM modela, biraju se predstavnici klasa koji joj najvjerovatnije pripadaju, odnosno oni koji su najudaljeniji jedan od drugog i između kojih je margina najšira, što dovodi do maksimalnog razdvajanja klasa. Primjenom balansiranog *ensemble* DT modela za ekstrakciju pravila iz takvog prethodno obrađenog skupa podataka (*ensemble* SVM-RE), stopa pogrešne klasifikacije najvrednijeg segmenta kupaca je smanjena za 66%, što je značajno bolji rezultat od rezultata dobijenog u radu *Đurišić et al.* (2020), gdje je korišćenjem samostalnog SVM pretprocesora i samostalnog DT ekstraktora, ova stopa pogrešne klasifikacije smanjena za 37%.

Za testni skup, *ensemble* SVM-RE metoda je postigla *Balanced Correction Rate* (BCR) – korijen proizvoda odziva klasa (eng. *rooted product of class recall of all classes*) od 83%, što je 15% bolje od najboljeg rezultata, koji su u svom radu postigli *Kang et al.* (2012), a koji je dobijen primjenom poduzorkovanja zasnovanog na klasterizaciji i *ensemble* tehnikama. Upoređujući najbolji ostvaren *class recall* minorne klase (88%), koji je dobijen u radu autora *Marinakos* i *Daskalaki* (2017) korišćenjem poduzorkovanja zasnovanog na klasterima i k-NN klasifikatora, s rezultatom postignutim metodom predloženom u ovom radu (94%), zaključuje se da je superiornost ovog modela očigledna. Pored toga, *class recall* za većinsku klasu u

istom istraživanju je prilično nizak (63%), dok u ovom radu *class recall* za druge dvije (veće) klase iznosi iznad 84%. U svom radu, *Kim et al.* (2013) su kao najbolju vrijednost za odziv minorne klase dobili 73%, za skup podataka sa umjerenim stepenom neravnoteže klasa, koristeći slučajno poduzorkovanje sa odnosom klasa od 2:1 SVM klasifikatorom. I u ovom slučaju, rezultat dobijen u ovom istraživanju je značajno bolji, sa ostvarnih 94% za odziv minorne klase na testnom skupu podataka.

Osim dokazane efikasnosti, automatsko balansiranje klasa korišćenjem *ensemble* SVM-RE metoda je manje složeno za praktičnu primjenu (nema nepoznanica u vezi sa izborom primjera koje treba ukloniti, izborom optimalnog odnosa klasa, itd.) u poređenju s tehnikama reuzorkovanja u sličnim studijama iz direktnog marketinga (Kang et al., 2012; Kim et al., 2013; Lawi et al., 2018; Marinakos & Daskalaki, 2017; Miguéis et al., 2017). Balansiranje podataka na ovaj način nudi stabilno rješenje koje prevazilazi problem pristrasnosti uzorkovanja i pretjerane zavisnosti od podataka za obuku (Sun et al., 2009).

Metoda *ensemble* SVM-RE imala je pogrešnu klasifikaciju najvrednijeg segmenta kupaca od oko 6% na testnom skupu, što je odličan rezultat. Sličan rezultat, odnosno stopa pogrešne klasifikacije od 4% postignuta je u radu autora *Zou et al.* (Zou et al., 2010), u kome je korišćena metoda zasnovana na prilagođavanju SVM algoritma uvođenjem troškovno-senzitivnog učenja i slučajnog poduzorkovanja. Međutim, prednost *ensemble* SVM-RE metode je u tome što njena primjena u praksi ne zahtijeva opsežno poznavanje SVM algoritma, koje bi bilo potrebno za njegovo prilagođavanje.

Upoređujući rezultate samostalne SVM-RE metode iz sekcije 5.2 i *ensemble* SVM-RE, može se vidjeti da je *ensemble* pristup poboljšao ukupnu tačnost za 2,98%, *class recall* za minornu klasu za 6,81%, kao i preciznost za 2,83%. Iako ova poboljšanja izgledaju mala, s obzirom na to da je identifikacija najvrednijih kupaca poboljšana za oko 7%, a njihovo predviđanje za oko 3%, to može dovesti do velikog povećanja profita generisanog kampanjom (vidjeti Tabelu 24). Treba imati na umu da jedan tačno odabran i targetiran visokoprofitabilan kupac može da generiše više prihoda u kampanji od svih ostalih kupaca zajedno.

Predložena metoda je dodatno testirana na javno dostupnom skupu podataka, gdje je potvrđena njena superiornost. Ukupna tačnost klasifikacije na nepoznatim podacima je poboljšana sa 43% (koja je ostvarena samostalnim DT modelom) na 90%. Metoda *ensemble* SVM-RE na testnom skupu imala je pogrešnu klasifikaciju najvrednijeg segmenta kupaca od oko 13%, za razliku od samostalnog DT modela, koji je imao pogrešnu klasifikaciju od čak 97%, zbog preklapanja ovog segmenta s kupcima iz segmenta srednje vrijednosti. Što se tiče problema minorne klase, odnosno najmanje vrijednih kupaca u ovom slučaju, *ensemble* SVM je smanjio grešku pogrešne klasifikacije sa 94,5% na 29% na nepoznatim podacima. Stoga, može se zaključiti da je metoda uspješno riješila problem neravnoteže i preklapanja klasa i na skupu podataka za validaciju predloženog koncepta.

Doprinosi postojećoj teoriji i literaturi

Imajući u vidu navedeno poređenje i istaknute prednosti predložene *ensemble* SVM-RE metode u odnosu na ranije primijenjene metode, može se zaključiti da ova disertacija, kada je u pitanju model prediktivne RFM segmentacije, doprinosi postojećoj teoriji i znanju u oblasti direktnog marketinga na više načina:

1. Umjesto RFM segmentacije zasnovane na subjektivnoj procjeni, predlaže se objektivna RFM segmentacija zasnovana na *k-means* klasterizaciji i procjena optimalnog broja klastera na osnovu DB indeksa, što pojednostavljuje primjenu i garantuje veću tačnost modela;
2. Umjesto razvrstavanja kupaca u uniformno kodirane grupe, klasterizacija se vrši na nivou pojedinačnog kupca, čime se sprečava gubitak značajnih informacija za segmentaciju;
3. Umjesto RFM atributa, kao prediktori se koriste karakteristike kupaca i proizvoda, pa se metoda može koristiti i za klasifikaciju nepoznatih kupaca;
4. Umjesto uzorkovanja ili prilagođavanja algoritma učenja, predlaže se automatsko balansiranje i uklanjanje preklapanja klasa korišćenjem *ensemble* SVM metode, što dovodi do stabilnog rješenja bez pristrasnosti uzorkovanja, pretjeranog

prilagođavanja podacima i opsežnog poznavanja metoda učenja od strane marketing stručnjaka;

5. Umjesto pretprocesiranja podataka korišćenjem samostalne SVM metode, predložen je *ensemble* SVM, koji povećava efikasnost balansiranja i razdvajanja klasa;
6. Umjesto ekstrakcije pravila korišćenjem samostalnog klasifikatora, ova studija predlaže ekstrakciju pravila korišćenjem DT klasifikatora u kombinaciji s meta-algoritmom balansirane *ensemble* metode, koji daje bolje prediktivne performanse u poređenju s korišćenjem samostalnog DT modela kao ekstraktora pravila;
7. Predložena metoda ima manju pogrešnu klasifikaciju manjinske klase (segmenta najvrednijih kupaca) od samostalnih balansiranih *ensemble* metoda, kao i metode uzorkovanja ili prilagođavanja algoritama, korišćenih u prethodnim studijama, uz održavanje visoke ukupne tačnosti;
8. Predložena metoda izdvaja pravila koja efektivno opisuju segmente kupaca (uključujući i onaj najmanji s najvrednijim kupcima, za koji pravila mogu biti izostavljena ako se ne adresira pogrešna klasifikacija manjinskih klasa). Ova pravila su semantički bogatija, jer sadrže karakteristike kupaca i proizvoda i pogodna su za targetiranje postojećih i novih kupaca u narednoj kampanji;
9. Za razliku od većine prethodnih studija, metoda je u ovoj studiji testirana na realnom skupu podataka, koji nije prečišćen i posebno pripremljen za analizu. Zatim je metoda potvrđena testiranjem na javno dostupnom skupu podataka.

Praktične implikacije

Pored teorijskog doprinosa, predložena metoda je značajna i za praktičnu primjenu i može pomoći donosiocima odluka u marketingu prilikom planiranja direktnih kampanja. Veoma kreativna i inovativna ponuda može rezultirati niskom stopom odgovora ako se targetiranje ne izvrši precizno, dok, s druge strane, loše formulisana i srednje kreativna ponuda pravoj ciljnoj grupi može smanjiti, ali ne i eliminisati, odgovor željenih potrošača (Stone & Jacobs, 2008). S tim u vezi, razumijevanje preferencija i potreba potrošača je važniji faktor u kreiranju kampanje nego

kreativni proces i način komuniciranja ponude. Pored toga, poslovna inteligencija i rudarenje podataka mogu poboljšati konkurentsku prednost kompanija na savremenim tržištima (Bach et al., 2018). Ovo je u skladu sa aktuelnim i tekućim trendom digitalne transformacije u kompanijama, koja se sprovodi u cilju držanja koraka s konkurencijom (Furjan et al., 2020) i poboljšanja korisničkog iskustva (Rekettye & Rekettye, 2019).

Koristeći predloženi model, moguće je prevazići bezličnu prirodu tradicionalnog marketinga, jer omogućava kompanijama da tretiraju slične grupe kupaca na jedinstven način. Prednosti koje ovaj model pruža praktičarima ogledaju se kroz precizno targetiranje, minimiziranje rasipanja poruka i profitabilnije kampanje. Na taj način im je omogućeno da objektivno segmentiraju tržište, prilagode sadržaj pojedinim segmentima i izgrade pouzdan i lojalan odnos s kupcima. U ovom dijelu rada je pokazano da korišćenje predviđanja pomoću *ensemble* SVM-RE modela za najvredniji segment rezultira najvećim brojem pravih predviđenih kupaca, kao i najmanjim brojem lažno predviđenih kupaca u okviru tog segmenta. U tom smislu, predloženim modelom u praksi direktnog marketinga može se ostvariti najveći profit u poređenju sa ostalim razmatranim modelima i može se smanjiti rasipanje marketing resursa.

Na osnovu rezultata iz ovog prediktivnog modela, mogu se kreirati razrađenije strategije segmentacije i efikasnije targetiranje u budućim kampanjama direktnog marketinga. Generisana pravila iz predloženog modela omogućavaju praktičarima da saznaju korisne informacije o svojim najvrednijim potrošačima, što je od velike važnosti, imajući u vidu da je zadržavanje postojećih kupaca često šest do deset puta isplativije od sticanja novih (Colgate & Danaher, 2000; Verbeke et al., 2011), posebno u uslovima fragmentacije tržišta i stepena tržišne konkurencije. Pored toga, eksplicitna pravila koja opisuju najvrednije potrošače omogućavaju akviziciju upravo onih kupaca koji su najsličniji ovoj grupi, kroz različite strategije targetiranja. Dakle, kupci iz različitih klastera i različite vrijednosti za kompaniju mogu biti targetirani u prilagođenim i personalizovanim promotivnim aktivnostima. Drugim riječima, targetiranje se može sprovesti na objektivan i precizan način, čime

se poboljšava profitabilnost svake kampanje, kao i ukupna efektivnost aktivnosti direktnog marketinga.

Još jedna prednost za marketing menadžere je jednostavnost upotrebe predložene metode. S obzirom na to da se koristi automatsko balansiranje podataka, nema potrebe da se sprovode složene procedure uzorkovanja. Takođe, praktičari ne moraju da znaju detalje algoritma učenja niti da angažuju eksterne eksperte u te svrhe.

Konačno, treba istaći da je model realizovan u *Rapid Miner* alatu u vidu gotovog korisničkog procesa spremnog za primjenu na podacima, tj. za predikciju i targetiranje kupaca koji će odgovoriti na kampanju (Prilozi 5 i 6). S obzirom na to da njegova primjena zavisi od podataka dostupnih u bazi kupaca, važno je napomenuti da se procesi moraju u skladu s tim modifikovati.

Sumarna razmatranja u vezi s predloženim *ensemble* baziranim modelom prediktivne RFM segmentacije

U ovom dijelu disertacije je osmišljena efikasna metoda za klasifikaciju kupaca u direktnom marketingu s ciljem njihovog targetiranja u kampanji. Prikazani prediktivni postupak podrazumijeva klasifikaciju klasterizovanih kupaca. Koristeći *k-means* klasterizaciju, kupci su podijeljeni u segmente na osnovu vrijednosti njihovih RFM atributa (prethodno kupovno ponašanje). Različiti klasteri imaju različite nivoe vrijednosti kupaca, kao i različitu vjerovatnoću odgovora na kampanju direktnog marketinga. Prateći ovu proceduru, pripadnost kupaca jednom od klastera (kao i odgovarajući CV-nivo potrošača) predviđa se metodom *ensemble* SVM-RE, korišćenjem podataka o karakteristikama kupaca i proizvodima koje preferiraju.

Rezultati empirijskog testiranja pokazuju da se problem neravnoteže klasa može prevazići, čime se poboljšava klasifikacija manjinske i najvrednije klase. Kombinovanje više SVM modela sa *ensemble* meta-algoritmom može poboljšati pretprocesiranje podataka i odvojiti segmente kupaca efikasnije od samostalnog SVM modela. Primjena balansiranih *ensemble* klasifikatora na takav prethodno

obrađeni skup podataka za obučavanje modela poboljšava prediktivne indikatore za manjinsku klasu i, posljedično, povećava efikasnost prediktivne segmentacije (posebno za segment najvrednijih kupaca), kao i šanse za ostvarivanje većeg profita u kampanji. Kombinovanje *ensemble* metode, zasnovano na slučajnom poduzorkovanju (s vraćanjem) većih klasa, odnosno na *bootstrapping* metodi, sa SVM pretprocesiranjem podataka i ekstrakcijom pravila, balansira klase bolje od samostalnih *ensemble* balansiranih metoda u klasifikaciji segmenta kupaca.

Glavni doprinos predložene metode je da ona bolje tretira problem neravnoteže klasa koja se javlja pri klasifikovanju kupaca u direktnom marketingu, nego metode uzorkovanja i adaptacije algoritama, koje su primijenjene u prethodnim radovima iz ove oblasti. Naime, u poređenju s prethodnim rezultatima, postignuta je manja pogrešna klasifikacija manjinske klase (segment kupaca visoke vrijednosti) s visokom ukupnom tačnošću. Procedura balansiranja klasa automatizovana je pretprocesiranjem podataka, čime se prevazilaze nedostaci prethodno primijenjenih metoda (pristrasnost uzorkovanja, prekomjerno prilagođavanje, potreba za obimnim poznavanjem metoda učenja). Konačno, sama primjena modela je pojednostavljena.

Pored naučnog doprinosa, predloženi pristup je od praktičnog značaja jer može značajno pomoći marketing menadžerima da povećaju efikasnost i profitabilnost kampanja direktnog marketinga i da održe dobre odnose s kupcima. Rezultati metode mogu pomoći u odluci da li postojeće (i nove) kupce treba targetirati u sljedećim kampanjama direktnog marketinga, s obzirom na to da su dva ključna elementa upravljanja odnosima s kupcima upravo privlačenje i zadržavanje kupaca (Lazar, 2018). Ova metoda izvlači i generiše pravila klasifikacije, koja se mogu koristiti za poboljšanje odnosa s postojećim kupcima i ciljanje novih potencijalnih kupaca, na osnovu njihovih karakteristika i ponuđenih proizvoda. Dodatno, ova metoda može značajno povećati prihode kampanje, kao i smanjiti njene troškove.

Međutim, ovaj pristup takođe ima nekoliko ograničenja i nedostataka. Prvo, korišćeni su skupovi podataka za obučavanje s relativno malim brojem instanci. Za veliki skup za obučavanje modela, sam proces obučavanja SVM modela, tj.

definisanje odgovarajućih parametara zahtijeva puno kompjuterskog vremena (Cui & Curry, 2005). Drugo, kao pretprocesor, SVM je kombinovan samo s *Bagging* metodom, pa ostaje otvoreno pitanje da li bi SVM bolje balansirao klase u kombinaciji s nekom drugom *ensemble* tehnikom. Treće, za ekstrakciju pravila iz *ensemble* SVM prethodno obrađenog skupa podataka, testirana je samo kombinacija *Bagging* i DT klasifikatora. *Bagging* tehnika koristi nasumično poduzorkovanje s vraćanjem, što može potencijalno biti manje efikasno od nekih drugih tehnika, kao što je poduzorkovanje zasnovano na klasterima. Stoga, ostaje pitanje da li bi ekstrakcija pravila dala bolje rezultate sa, recimo, *ensemble* metodom, koja koristi poduzorkovanje na bazi klastera (Kang et al., 2012; Wong et al., 2020) ili kombinovanjem standarnog poduzorkovanja na osnovu klastera s različitim klasifikatorima (Kang et al., 2012; Marinakos & Daskalaki, 2017), kao i kombinovanjem neke druge *ensemble* tehnike (npr. *Adaptive Boosting*) s različitim klasifikatorima (Lawi et al., 2018). Konačno, uspjeh *data mining* metode u velikoj mjeri zavisi od kvaliteta podataka. Skup podataka koji se ovdje koristi uključuje samo neke karakteristike kupaca. Uključivanjem više atributa kupaca, poput njihovog nivoa dohotka, broja članova domaćinstva, zanimanja i godina starosti, mogu se dobiti jasnija pravila za targetiranje novih kupaca u budućim marketing aktivnostima.

U budućim istraživanjima ova metoda se može testirati na drugim skupovima podataka, kako bi se provjerila ili poboljšala njena efikasnost. U ovoj studiji metoda je testirana u klasifikaciji segmenata kupaca u kojima je udio manjinske klase oko 15%. Bilo bi interesantno testirati njene performanse u slučaju da je udio minorne klase značajno niži, čak i ispod 5%.

Da bi se generisala pravila iz *ensemble* SVM modela, mogle bi se testirati neke druge *ensemble* tehnike, kao što je *Random Forest*, kao i tehnike na nivou algoritma. S obzirom na to da je poduzorkovanje zasnovano na klasterima ranije u literaturi potvrđeno kao uspješna tehnika balansiranja klasa (Kang et al., 2012; Marinakos & Daskalaki, 2017), kombinacija ovih tehnika mogla bi se testirati za generisanje pravila (nezavisno ili unutar *ensemble* procedure) s različitim klasifikatorima.

Također, iako je u ovoj studiji *Bagging SVM* potvrđen kao pretprocesor koji uspješno balansira podatke iz domena direktnog marketinga, u budućim istraživanjima njegovi rezultati bi mogli da se uporede s poduzorkovanjem zasnovanim na klasterima na istom skupu podataka. Bilo bi korisno testirati da li bi neka druga *ensemble* tehnika, kao što je *Adaptive Boosting*, u kombinaciji sa SVM metodom, bolje obrađivala, tj. efikasnije izbalansirala podatke.

Dâ se zaključiti da postoji veliki broj mogućnosti za testiranje različitih kombinacija *data mining* metoda, koji prevazilazi okvire ove disertacije. U svakom slučaju, otvorena je problematika i predloženo jedno od efikasnih rješenja koje je i potvrđeno objavljivanjem u referentnim međunarodnim publikacijama (Rogić & Kašćelan, 2020; Rogic & Kascelan, 2019).

5.4 Testiranje SVM-RE baziranog modela odgovora na kampanju

U skladu s prediktivnom procedurom sa Slike 34 u sekciji 4.5.3, najprije je obučen SVM pretprocesor na balansiranom skupu podataka za obučavanje (gdje je iz klase nerespondenata slučajno izabrano onoliko primjera koliko ima respondenata, tj. klase su balansirane) i dobijeni su optimalni parametri: C = 10000; *gamma* = 0.01; *epsilon* = 0.001. Zatim je SVM, sa ovim parametrima, primijenjen na cjelokupnom skupu podataka za obučavanje, kako bi se dobilo B-SVM oznaka klase, tj. da bi se podaci pretprocesirali (balansiranje i razdvajanje klasa).

Prateći dalje proceduru predloženu u okviru sekcije 4.5.3, testirane su performanse sljedećih klasifikatora: LR, GBT, RF, k-NN i DT, kako prije B-SVM pretprocesiranja i balansiranja podataka, tako i nakon ovog postupka. Rezultati prediktivnih performansi svih testiranih klasifikatora dati su u Tabeli 27. Svaki obučavani model je kros-validiran na početnom i pretprocesiranom skupu za obučavanje i potom primijenjen na testnim podacima. U tabeli su prikazani rezultati dobijeni na testnim podacima.

Tabela 27. Prediktivne performanse klasifikacionih algoritama bez i sa SVM pretprocesiranjem

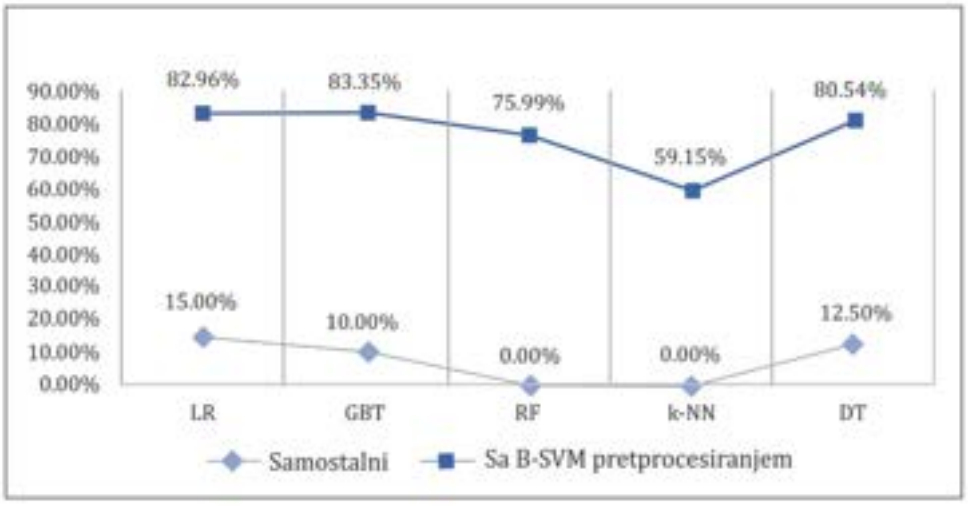
Klasifikator	Tačnost	Senzitivnost	AUC	*Fallout*
LR	99,23%	15,00%	0,680	0,34%
GBT	99,18%	10,00%	0,727	0,37%
RF	99,48%	0,00%	0,827	0,01%
k-NN	99,50%	0,00%	0,593	0,00%
DT	99,43%	12,50%	0,608	0,13%
B-SVM	87,15%	67,50%	0,832	12,75%
B-SVM+LR	88,21%	82,96%	0,954	11,01%
B-SVM+GBT	89,97%	83,35%	0,950	9,03%
B-SVM+RF	89,27%	75,99%	0,921	8,74%
B-SVM+k-NN	85,53%	59,15%	0,831	10,51%
B-SVM+DT	90,96%	80,54%	0,898	8,16%

Kada se nalazi predloženog pristupa balansiranja podataka iz Tabele 27 uporede s rezultatima nezavisnih klasifikatora, evidentno je da ih ova metoda nadmašuje u smislu balansiranja klasa, odnosno rješenja problema najmanje klase.

Iz Tabele 27, može se uočiti da su nakon prethodne obrade podataka, korišćenjem B-SVM pristupa, senzitivnost i AUC poboljšani u svim modelima. Na primjer, prije balansiranja podataka, RF je ostvario 0% senzitivnost, dok je B-SVM+RF ostvario 75,99%. Pored toga, AUC metrika je za neke modele, poput k-NN i DT, bila relativno niska: 0,539 i 0,608, respektivno, što je previše blizu vrijednosti 0,5, koja sugeriše da model ne može efikasno da napravi razliku između pozitivne i negativne klase. Najniža AUC vrijednost kod modela sa B-SVM pretprocesiranjem je 0,831 (B-SVM+k-NN) i kreće se do maksimalne vrijednosti od 0,954 (B-SVM+LR), što ukazuje na odlične performanse modela.

Visoki nivoi tačnosti u svim samostalnim modelima su rezultat pristrasnosti modela prema većoj klasi. Stoga, uzimajući u obzir važnost pravilnog identifikovanja upravo onih kupaca koji će odgovoriti na kampanju direktnog marketinga, odnosno

stvarnih pozitivnih instanci iz baze podataka, u ovoj studiji fokus je stavljen na metriku senzitivnosti, a ne na ukupnu tačnost. Iz Tabele 27 se može vidjeti da je *Balanced* SVM+GBT postigao najbolje performanse u pogledu senzitivnosti: 83,35%. Ovaj rezultat ukazuje na potencijalno poboljšanje profitabilnosti budućih kampanja, jer kompanija može precizno targetirati grupu kupaca s velikom vjerovatnoćom odgovora. Na primjer, u ovom skupu podataka, samostalni GBT bi tačno identifikovao samo 10% takvih potencijalnih kupaca. Pored toga, B-SVM+LR, model s drugim najboljim performansama senzitivnosti, targetirao bi 82,96% potencijalnih kupaca s visokom vjerovatnoćom stope odgovora, a samostalni LR samo 15%. Poboljšanja nivoa senzitivnosti nakon B-SVM procedure prikazana su na Slici 39.



Slika 39. Vrijednost metrike senzitivnosti prije i nakon pretprocesiranja podataka

Još jedna važna metrika za planiranje kampanje direktnog marketinga i njenog budžeta je *fallout*. Uzimajući u obzir da rezultat ove metrike pokazuje procenat kupaca koji bi bili targetirani i koji ne bi odgovorili na ponudu, ključno je da ova metrika bude što niža. Stoga, *Balanced* SVM+DT, s metrikom *fallout* od 8,16% sugeriše da bi ovaj procenat kupaca bio pogrešno klasifikovan kao vjerovatni respondenti. Ovaj procenat je od praktičnog značaja, imajući u vidu posebno one kompanije koje se suočavaju sa ograničenim budžetima za marketing. Rezultati ove studije pokazuju da bi alokacija budžeta za marketing bila efikasna.

Validacija modela na javno dostupnom skupu podataka

Kao i kod modela prediktivne RFM segmentacije, da bi se potvrdila njegova robustnost, predloženi pristup modeliranju odgovora kupaca je validiran na javno dostupnom skupu podataka. Skup podataka za validaciju je *"Direct Marketing Education Foundation 3"* (DMEF3). Kako je i detaljno opisano u sekciji 5.1.2, ovaj skup podataka sastoji se od 106.284 podataka o transakcijama kupaca iz kompanije za katalošku prodaju.

Zavisna varijabla je broj porudžbina, pri čemu je iz originalnog oblika transformisana u binominalni, gdje je broj naloga veći ili jednak 1 (respondenti)

je kodiran sa 1, a nerespondenti sa 0. Prateći proceduru koju su u svom radu opisali autori *Malthouse* i *Blattberg* (2005), sadašnji trenutak je postavljen na 1. avgust 1990. godine, što je rezultiralo skupovima podataka za obuku i testiranje približno iste veličine, a odgovor na ponudu u ciljnom periodu je korišćen kao zavisna promjenljiva. Stopa odgovora u ovom skupu podataka je 5,4%. Rezultati su prikazani u Tabeli 28.

Tabela 28. Prediktivni učinak klasifikacionih algoritama bez i sa B-SVM pretprocesiranjem za DMEF3 skup podataka

Klasifikator	Tačnost	Senzitivnost	AUC	*Fallout*
LR	88,72%	65,39%	0,856	3,67%
GBT	89,53%	63,56%	0,861	2,00%
RF	89,22%	60,59%	0,851	1,44%
k-NN	87,47%	62,01%	0,827	4,22%
DT	89,42%	63,66%	0,819	2,17%
B-SVM	87,30%	69,21%	0,815	6,79%
B-SVM+LR	99,88%	99,94%	1,000	0,14%

B-SVM+GBT	99,89%	100,00%	0,999	0,15%
B-SVM+RF	99,89%	100,00%	1,000	0,14%
B-SVM+k-NN	98,90%	95,84%	0,995	0,23%
B-SVM+DT	99,97%	99,87%	0,999	0,00%

Iz Tabele 28 može se primijetiti da, slično prvom skupu podataka, B-SVM dovodi do značajnog poboljšanja performansi modela. U svim modelima nakon pretprocesiranja i balansiranja podataka, metrika senzitivnosti je iznosila 95% ili više. Pored toga, u slučaju B-SVM+GBT i B-SVM+RF, modeli su dobili 100% senzitivnosti. S druge strane, *fallout* metrika je smanjena u svim B-SVM modelima na manje od 0,3%.

Nakon prethodne obrade, AUC je u svim testiranim modelima bio blizu 1, što znači da su modeli imali savršenu sposobnost da razlikuju klase respondenata i nerespondenata. Takođe, ukupna tačnost u svim B-SVM modelima je bila oko 99%.

U pogledu dobijene metrike senzitivnosti, B-SVM+GBT model je postigao najbolje performanse na ovom skupu podataka, kao i na prvom skupu podataka, predstavljenom u Tabeli 27. Na prvom skupu podataka dobijena senzitivnost je bila 83,35%, dok je u DMEF3 skupu podataka model postigao 100% senzitivnosti. S druge strane, najslabiji rezultati su dobijeni korišćenjem B-SVM+k-NN za oba skupa podataka. Naime, na prvom skupu podataka, rezultat senzitivnosti ovog modela iznosio je 59,15%, dok je za DMEF3 skup podataka iznosio 95,84%.

Dakle, B-SVM pretprocesiranje, sa značajnim poboljšanjima u svim modelima, pokazalo se kao moćna tehnika za balansiranje podataka i model je uspješno validiran na dodatnom skupu podataka. Takođe je potvrđeno da napredni klasifikatori koji uključuju *ensemble* meta-algoritam (GBT i RF) daju bolje rezultate od klasičnih, na prethodno obrađenim podacima. Generalno, bolje performanse modela na ovom skupu podataka su ostvarene zbog veće stope odgovora, kao i zbog toga što baza podataka kupaca sadrži istoriju ponašanja prilikom kupovine tokom

12 godina, dok je prvi korišćeni skup podataka uključivao podatke za svega nekoliko kampanja i šest meseci. Ova validacija je pokazala da se ovaj pristup može koristiti ne samo u upravljanju *online* direktnom marketing kampanjom, već i *offline,* kao i u različitim sektorima s različitim bazama kupaca i podataka s kojim raspolažu.

Potvrda hipoteze H3 i odgovor na IP1

Iz rezultata za oba skupa podataka se može vidjeti da je pristup koji kombinuje SVM pretprocesiranje i različite klasifikatore superiorniji od samostalnih klasifikatora.

Kada je u pitanju kombinacija SVM+DT, tj. SVM-RE metoda na osnovnom skupu podataka za testiranje, ostvarena senzitivnost je 80,45%, a specifičnost 91,84%, dok je AUC 0,898. SVM-RE metoda je na skupu podataka za validaciju imala senzitivnost, specifičnost i AUC od 99,97%, 99,87% i 0,999, respektivno. Dakle, može se konstatovati je da je hipoteza H3 potvrđena.

Takođe, kod osnovnog skupa podataka najbolji rezultat ima SVM-GBT klasifikator, a kod drugog skupa SVM-GBT i SVM-RF, što znači da kada se DT unaprijedi sa *ensemble* metodama, dolazi do unapređenja performansi predloženog modela. Tako, na primjer, kod osnovnog testnog skupa podataka, SVM-GBT je u odnosu na SVM-DT povećao AUC za 5% a senzitivnost za 2,81% (što, kada su u pitanju najvredniji kupci, može biti značajno, kako je već objašnjeno kod modela za prediktivnu RFM segmentaciju). Dakle, kada je u pitanju model odgovora na kampanju zasnovan na SVM-RE, *ensemble* metode poboljšavaju njegove prediktivne performanse, što znači da je odgovor na IP1 pozitivan.

Diskusija ostvarenih rezultata

Predloženi pristup modeliranju odgovora kupaca je osmišljen da testira mogućnosti za poboljšanje postojećih metoda za predikciju odgovora kupaca na kampanju direktnog marketinga. Kao jedan od osnovnih ciljeva predložene procedure, ističe se smanjenje pogrešne klasifikacije za segment respondenata, tj. predlog za rješenje problema neravnoteže klasa na skupu podataka sa izuzetno niskom stopom

odgovora. U korišćenom skupu podataka stopa odgovora iznosila je 0,41%, što je značajno niže u poređenju sa skupovima podataka korišćenim u prethodno objavljenoj literaturi.

Upoređujući rezultate iz ovog rada s rezultatima iz prethodnih studija, može se konstatovati da predloženi pristup nadmašuje prediktivne performanse prethodnih studija u modeliranju odgovora kupaca, pri čemu efikasno funkcioniše na skupu podataka s najmanjom stopom odgovora. U prethodnim radovima s najmanjim uočenim stopama odgovora, *Lee et al.* (2021) i *Kim et al.* (2013) su pokazali nivo senzitivnosti od 73,92% i 23,8%, respektivno, što je prikazano u Tabeli 2. Najbolji analizirani rezultat je postignut u studiji, koju su sproveli autori *Asare-Frempong* i *Jayabalan* (2017), koji su ostvarili 90,2% senzitivnosti i 0,927 AUC, sa stopom odgovora od 11,63%, koristeći balansirani RF. Rezultati naše studije na skupu podataka distributera sportske opreme pokazali su slabiji učinak u nivoima senzitivnosti 83,35%, ali su postigli viši AUC od 0,950, koristeći B-SVM+GBT. Međutim, stopa odgovora u ovoj studiji bila je znatno niža. Na skupu podataka DMEF3, model je postigao senzitivnost i AUC od 100% i 0,999, respektivno, nadmašujući sve prethodne studije. I *Chaudhuri et al.* (2021) su dobili visok stepen senzitivnosti od 96% i AUC od 0,89, ali u njihovom radu nema indikacija o stopi odgovora (ili konverzije) u korišćenom skupu podataka.

Ova studija otkriva da korišćenje B-SVM pristupa u kombinaciji s tehnikama klasifikacije poboljšava prediktivnu sposobnost modela odgovora na kampanju. Rezultati su ukazali na to da B-SVM efikasno pretprocesira podatke, rješavajući probleme poput preklapanja i neravnoteže klasa. Dakle, B-SVM smanjuje preklapanje klasa i dopunjuje manju klasu s najrelevantnijim instancama, pomjeranjem margine ka najbližim, a samim tim i najsličnijim primjerima veće klase i njihovim svrstavanjem u manju klasu respondenata. Na taj način se manja klasa dopunjava instancama, tj. kupcima iz grupe visoko vjerovatnih respondenata. Zahvaljujući tome, kompanije mogu da targetiraju širu grupu potencijalnih respondenata, bez trošenja marketing budžeta na nasumičan ili subjektivan izbor.

U skladu s tim, pristup predložen u ovoj disertaciji na više načina doprinosi postojećoj literaturi iz oblasti modeliranja odgovora kupaca:

1. Predložen je model targetiranja kupaca koji ne samo da identifikuje ispitanike iz baze korisnika, već i one za koje je vrlo vjerovatno da će biti budući respondenti, prepoznajući njihovu sličnost s postojećim;
2. Predloženi model ima bolje prediktivne performanse u poređenju s modelima iz prethodnih studija;
3. Model je validiran na osnovu podataka o *online* i *offline* kupcima, što znači da se može primijeniti u oba slučaja. Takođe, model je validiran na podacima sa atributima koji se razlikuju od osnovnog skupa za testiranje, što potvrđuje da je robustan, bez obzira na to kako baza kupaca izgleda, tj, koje podatke sadrži;
4. Potvrđene su mogućnosti balansirane SVM metode za prečišćavanje podataka i balansiranje u modeliranju odgovora kupaca sa izuzetno niskim stopama odgovora;
5. Upoređeni su rezultati naprednih i klasičnih klasifikatora na pretprocesiranom skupu podataka i empirijski su potvrđene prednosti naprednih klasifikatora (koji uključuju *ensemble* metode) u ovom kontekstu;
6. Zbog poduzorkovanja, vremenska i tehnološka složenost u implementaciji pretprocesiranja podataka je smanjena, čime je i primjena predložene metode pojednostavljena.

Praktične implikacije

Svi modeli, nakon B-SVM pretprocesiranja, pokazali su značajna poboljšanja performansi, a rezultati se mogu koristiti za podršku odlučivanju u planiranju aktivnosti direktnog marketinga, kao i u upravljanju kampanjama za precizno i tačno targetiranje potencijalnih kupaca. Tako je, na primjer, najbolji klasifikator targetirao samo 10% respondenata bez pretprocesiranja, a nakon predložene procedure čak 83,35% vrlo vjerovatnih respondenata i potencijalnih kupaca. To znači da je predloženom metodom identifikovano 7,3 puta više mogućih respondenata, što može dovesti do značajnog povećanja profitabilnosti kampanje.

Istovremeno, ostvareno je manje od 10% pogrešno targetiranih kupaca, što ukazuje na to da će biti relativno malo nepotrebnih troškova kampanje. Na taj način, kompanije mogu prilagoditi ponudu i selektovati upravo one kupce s velikom vjerovatnoćom odgovora, što će biti troškovno efikasno i profitabilno. Zasićena tržišta i stanje hiperkonkurekcije dovode do toga da su kupci targetirani brojnim ponudama za koje nisu zainteresovani, pa ovaj pristup može pomoći kompanijama da mudro selektuju samo one kupce za koje će ponuda biti relevantna.

U skladu s tim, *Stone* i *Jacobs* (2008) navode da veoma kreativna i originalna ponuda može rezultirati niskom stopom odgovora ako targetiranje nije urađeno kako treba, dok loše strukturirana i umjereno kreativna ponuda usmjerena ka odgovarajućoj ciljnoj grupi može smanjiti, ali ne i eliminisati očekivani odgovor kupaca. Stoga, donosioci odluka u direktnom marketingu mogu imati koristi od ovog pristupa, jer omogućava preciznije targetiranje, manje rasipanja poruka i profitabilnije kampanje.

Još jedan pozitivan aspekt predložene metode za praktičare direktnog marketinga je njena jednostavnost. S obzirom na to da se koristi automatizovano balansiranje podataka, nema potrebe da se rade komplikovane operacije reuzorkovanja. Štaviše, od marketing menadžera se ne traži da razumiju specifičnosti algoritma učenja ili da angažuju dodatne specijaliste ili eksterne eksperte da sprovode postupak.

Konačno, važno je istači da je model realizovan u *Rapid Miner* alatu, u vidu dva gotova korisnička procesa - za obuku i za predikciju (Prilozi 7-9), uz napomenu da je potrebna modifikacija ulaza, jer procesi zavise od formata podataka u bazi kupaca konkretne kompanije.

Sumarna razmatranja u vezi s predloženim SVM-RE baziranim modelom odgovora na kampanju

Potreba odabira relevantnih kupaca za efikasan direktni marketing značajno je porasla. Zasićena tržišta i konkurentski pritisci smanjuju odziv kupaca i povećavaju troškove marketinga (Mandapaka et al., 2014). Kao rezultat ovog izazova, nameće

se potreba za poboljšanim modelima odgovora s prilagođenim pristupom, koji omogućavaju kompanijama da ulažu u direktan marketing, uz pravilan i efikasan odabir kupaca. Kako je profitabilnost kampanje direktnog marketinga u velikoj mjeri određena brojem respondenata, odnosno koliko potrošača se odazove na plasiranu ponudu, identifikacija ciljnih kupaca je jedan od najznačajnijih koraka u procesu planiranja kampanje. Da bi se ispunili različiti ciljevi kompanije i maksimizirala profitabilnost kampanje, izbor potencijalnih kupaca mora biti optimizovan.

Svrha ovog dijela istraživanja bila je da se tretira problem neravnoteže klasa u modeliranju odgovora kupaca, što je jedan od najčešćih izazova kada se koriste algoritmi mašinskog učenja u direktnom marketingu i upravljanju kampanjama. Ovaj problem je posebno prisutan kada su u pitanju *online* kupci, kod kojih stopa odgovora može biti veoma niska zbog velikog broja posjeta *web* sajtu koje ne rezultiraju obavljenom transakcijom. Balansirana SVM metoda, kao pretprocesor, korišćena je za otkrivanje rješenja za veoma neuravnotežene podatke. U poređenju s rezultatima opisanim u prethodnim odjeljcima, predloženi pristup pokazuje odlične prediktivne performanse. U kombinaciji sa *ensemble* klasifikatorima, ovaj pristup najbolje predviđa potencijalne kupce u *online* kupovini. Jedan od ključnih doprinosa ovog istraživanja je da predloženi pristup bolje tretira i rješava problem klasne neravnoteže, koji se javlja pri klasifikovanju kupaca u direktnom marketingu, od metoda predstavljenih u ranijim studijama. Konkretno, u poređenju s rezultatima iz prethodne literature, rezultati su pokazali manji stepen pogrešne klasifikacije najmanje klase. Takođe, pretprocesiranje podataka automatizuje tehniku balansiranja klasa i, u konačnom, kompletna aplikacija je pojednostavljena.

S tim u vezi, kako savremeni kupci postaju sve više korisnici e-trgovine, donosioci odluka u marketingu mogu da se fokusiraju na kreiranje prilagođenog sadržaja, kao i da unaprijede svoje sisteme targetiranja na osnovu predloženog pristupa, koji odražava praktični značaj predložene metode. Imajući ovo na umu, prilagođene i adaptirane objave na društvenim mrežama mogu biti moćno sredstvo za

povezivanje s kupcima na mreži, a uz ispravnu selekciju i targetiranje, proces može rezultirati sticanjem novih i zadržavanjem starih kupaca.

Međutim, predloženi pristup također ima nekoliko nedostataka i ograničenja. Prvo, zbog slučajnog poduzorkovanja, pretprocesiranje podataka može dovesti do gubitka informacija koje su važne za model i bolju identifikaciju kupaca najsličnijih respondentima. Drugo, prvi korišćeni skup podataka odnosi se na kratak period od nekoliko mjeseci, tako da sezonalnost transakcija nije uzeta u obzir. Također, model predviđa ponašanje kupaca nakon izvršene transakcije, a ne tokom same trgovine, što bi moglo biti korisnije u smislu razvijanja sistema za preporuke ili stimulisanja kupca tokom *web* sesije.

U skladu sa ovim ograničenjima, buduća istraživanja bi mogla testirati tehnike pretprocesiranja, koje kombinuju klasterizaciju veće klase, *ensemble* i poduzorkovanje, slično proceduri koju su u svom radu predstavili *Kang et al.* (2012), kako bi se obezbijedio što reprezentativniji uzorak za ovu klasu i smanjila mogućnost da se zbog slučajnog poduzorkovanja neki nerespondenti slični respondentima zanemare i izgube. Također, metodu treba testirati na drugim skupovima podataka koji pokrivaju duži vremenski period i više kampanja i analizirati uticaj sezonalnosti u bazama podataka. Bilo bi interesantno ispitati mogućnosti predložene metode kao dijela sistema preporuka, koji bi predvidio odgovor kupaca tokom sesije *online* kupovine.

Rezultati predloženog modela potvrđeni su publikovanjem u referentnom međunarodnom časopisu (Rogić, Kašćelan, & Pejić Bach, 2022).

5.5 Testiranje modela odgovora na kampanju baziranog na ensemble metodama

U cilju testiranja mogućnosti *ensemble* tehnika u odnosu na SVM-RE metodu po pitanju balansiranja klasa, u ovom dijelu rada biće predstavljeni rezultati i diskusija

predloženog pristupa za predikciju odgovora na kampanju predstavljenog u sekciji 4.5.4.

Jedan od ciljeva ovog pristupa je rješavanje problema klasne neravnoteže. Ovaj problem je prevaziđen korišćenjem metode balansiranog poduzorkovanja, kao i *ensemble* tehnika za klasifikaciju. Zatim su rezultati dobijeni na taj način poređeni s rezultatima iz prethodne sekcije, da bi se utvrdilo koji pristup ima bolje prediktivne performanse.

U Tabeli 29 prikazane su prediktivne perfomanse modela za tri klasifikatora, *Balanced* DT, *Balanced Bagging* DT i *Balanced* RF. Svi klasifikacioni algoritmi su primijenjeni na cijeli testni skup (svi atributi), a, osim toga, svaki od tri skupa atributa (*web* metrike, RFM i podaci o proizvodu) su isključeni posebno, kako bi se pokazao njihov značaj za prediktivne performanse modela. Modeli testirani na svim atributima pokazali su dobre performanse, dok su najveći rezultati tačnosti (78,72%) i AUC (0,839) dobijeni za *Balanced* RF model, kao i najmanji *fallout* (21,26%). Na osnovu rezultata AUC indikatora, može se konstatovati da ovaj model može razlikovati pozitivnu i negativnu klasu u 83,9% slučajeva. Ovaj rezultat ukazuje na dobre performanse modela, jer je postignuti rezultat od 83,9% značajno veći od AUC od 50%, što bi značilo da model nema diferencijacijski kapacitet između klasa.

Tabela 29. Prediktivne performanse modela primijenjenih klasifikacionih algoritama

Algoritam	Indikatori performansi	Svi atributi	*Web* atributi isključeni	RFM atributi isključeni	Atributi o proizvodima isključeni
Balanced DT	*Accuracy*	73,04%	97,81%	73,04%	73,04%
	AUC	0,780	0,641	0,780	0,780
	TP	30	12	30	30
	FP	2.128	146	2.128	2.128
	Sensitivity	75,00%	30,00%	75,00%	75,00%
	Fallout	26,97%	1,85%	26,97%	26,97%
Balanced Bagging DT	*Accuracy*	76,86%	97,81%	76,87%	76,86%
	AUC	0,827	0,641	0,827	0,827
	TP	32	12	32	32
	FP	1.827	146	1.826	1.827

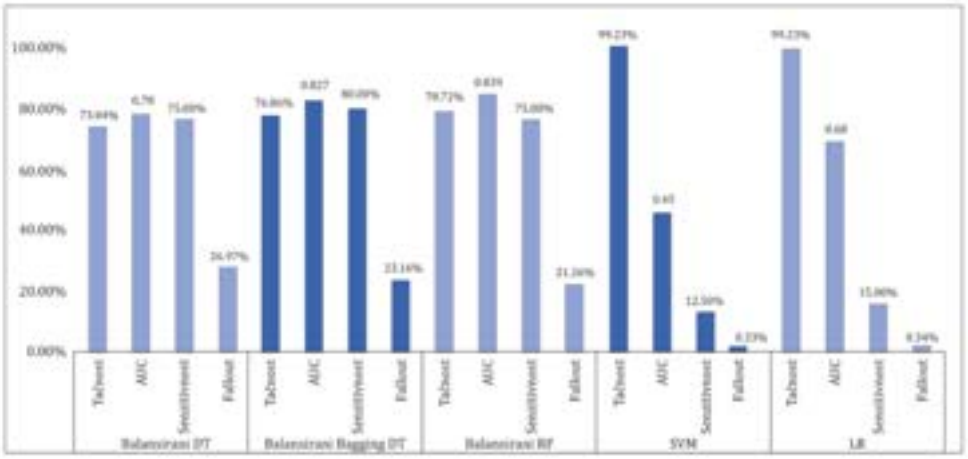
	Sensitivity	80,00%	30,00%	80,00%	80,00%
	Fallout	23,16%	1,85%	23,15%	23,16%
Balanced Random Forest	*Accuracy*	78,72%	97,81%	78,37%	77,87%
	AUC	0,839	0,642	0,840	0,843
	TP	30	12	30	30
	FP	1.677	146	1.705	1.745
	Sensitivity	75,00%	30,00%	75,00%	75,00%
	Fallout	21,26%	1,85%	21,61%	22,12%
SVM	*Accuracy*	99,23%			
	AUC	0,450			
	TP	5			
	FP	35			
	Sensitivity	12,50%			
	Fallout	0,33%			
LR	*Accuracy*	99,23%			
	AUC	0,680			
	TP	6			
	FP	27			
	Sensitivity	15,00%			
	Fallout	0,34%			

Napomena: U tabeli je prikazan učinak dobijen na testnom skupu (rezultati predviđanja odgovora na šestu kampanju na osnovu istorije *web* ponašanja, kao i prethodnog ponašanja pri kupovini za posmatranih pet kampanja).

Iako je tačnost visoka za sva tri modela, važnija metrika učinka, kako bi se dobile vrijedne informacije o potencijalnim respondentima s većom vjerovatnoćom namjere kupovine, jeste senzitivnost. Analizirajući sve atribute, vrijednost senzitivnosti je 75% (*Balanced* DT i *Balanced* RF) i 80% (*Balanced Bagging* DT), što bi potencijalno moglo da poboljša buduće targetiranje u kampanjama direktnog marketinga, s ciljem smanjenja grešaka u selekciji kupaca koje nastaju tako što se ne biraju kupci s visokom vjerovatnoćom kupovine. Pored toga, rezultati dobijeni u *Balanced* RF modelu pokazuju da je algoritam pogrešno klasifikovao 21,25% nerespondenata kao respondente. Uzimajući u obzir budući izbor i targetiranje kupaca, ovih 21,25% potencijalno predstavlja kupce koji će biti selektovani i targetirani, a koji vrlo vjerovatno neće odgovoriti na kampanju. Manji procenat znači niže troškove kampanje, pa je poželjno da *fallout* metrika bude što manja. Primjeri FP instanci, prikazani u Tabeli 29 – 2.128, 1.827, 1.677 za *Balanced* DT, *Balanced Bagging* DT i *Balanced* RF, respektivno, predstavljaju tačno one kupce za koje će

ulaganje u kampanju vjerovatno biti neisplativo. Dakle, u tom smislu, *Balanced* RF model ima najbolje performanse.

Upoređujući rezultate sva tri modela (Slika 40), može se zaključiti da primjena *ensemble* meta-algoritma (*Bagging* i *Random Forest*) poboljšava performanse dobijene samo balansiranim poduzorkovanjem, odnosno korišćenjem *Balanced* DT modela. Ovaj pristup ne samo da povećava ukupnu preciznost i AUC, već i povećava senzitivnost i smanjuje *fallout*, što je podjednako važno iz perspektive plasiranja direktne marketing kampanje. Sličan problem je tretirao i *Dou* (2020) - autor je takođe uzeo neuravnoteženi skup podataka s *web* stranice za e-trgovinu i koristio *ensemble* metode da predvidi buduće kupovine određenih proizvoda. Kao rezultat je dobio senzitivnost od 84,84% i tačnost od 88,51%, na skupu podataka sa 15,49% primjera male klase. Kada se ovaj rezultat uporedi sa onim dobijenim u ovoj studiji sa značajno većim klasnim disbalansom, možemo konstatovati da su modeli pokazali dobre performanse.



Slika 40. Poređenje performansi modela

Da bi se demonstrirale prednosti balansiranih *ensemble* metoda, izvršeno je poređenje rezultata s drugim metodama predviđanja odgovora kupaca, kao što su SVM i logistička regresija (Slika 40). Ukupna tačnost ovih metoda je oko 99%, *fallout* 0,33% i 0,34%, što pokazuje da je segment nerespondenata klasifikovan s velikom preciznošću. Međutim, veoma je malo tačno klasifikovanih respondenata

(senzitivnost je 12,5% i 15%), za razliku od balansiranih *ensemble* algoritama, gdje je senzitivnost iznad 75%. Može se zaključiti da standardne metode za modele odgovora potpuno pogrešno klasifikuju malu klasu (respondente), te da se ne mogu primijeniti kada je riječ o izuzetno niskoj stopi odgovora, kao što je slučaj u ovom skupu podataka.

Tabela 29 takođe pokazuje razlike u performansama modela nakon isključivanja svake grupe atributa iz analize. Na osnovu prikazanih rezultata, može se konstatovati da, izuzimajući *web* atribute, sva tri klasifikatora pokazuju najslabije performanse, što ukazuje na značaj *web* metrika za razvoj modela odgovora kupaca. Sa senzitivnošću od samo 30% za sva tri modela i sa smanjenjem AUC na 0,64, može se uočiti da algoritmi ne mogu da razlikuju pozitivnu i negativnu klasu na efikasan način, kao kada su ovi atributi uključeni u analizu. Svega 12 kupaca je tačno identifikovano kao respondenti. S druge strane, ukupna tačnost je porasla na 98%, a *fallout* je pao na čak 1,85%, što ukazuje na veoma visoku tačnost klasifikacije nerespondenata, odnosno pristrasnost algoritama prema većinskoj klasi. Ovako značajan uticaj *web* atributa je u skladu s prethodnim istraživanjem, jer je u prethodnoj literaturi potvrđeno da *clickstream* podaci *web* metrike imaju važnu ulogu u predviđanju ponašanja kupaca (Esmeli, Mohasseb, et al., 2020; Noviantoro & Huang, 2021; Van den Poel & Buckinx, 2005). Pored ovoga, autori *Liao et al.* (2011) istakli su značaj *web* metrika za analizu spremnosti kupaca da odgovore na *online* oglašavanje.

Za razliku od ponašanja na mreži, izuzimanje RFM atributa, kao i atributa o proizvodima ne utiče značajno na prediktivne performanse sva tri modela. Dakle, iako je kupovno ponašanje potvrđeno u prethodnim istraživanjima kao značajan prediktor odgovora na kampanju (Guido et al., 2013; Hauser et al., 2011), kada su u pitanju kupci u elektronskoj trgovini, *web* ponašanje ima značajniji uticaj, što je potvrđeno u drugim nedavnim istraživanjima (Esmeli, Bader-El-Den, et al., 2020; Lee et al., 2002; Rho et al., 2011). U uslovima porasta korišćenja *online* kampanja direktnog marketinga, čini se da nekadašnji značaj kupovnog ponašanja za predviđanje odgovora sve više preuzima *web* ponašanje, što rezultati ove empirijske studije eksperimentalno i potvrđuju.

Komparacija s modelom odgovora baziranim na SVM-RE i odgovori na IP2 i IP3

Poređenje rezultata predstavljenih u sekcijama 5.4 i 5.5 ukazuje na superiornost SVM-RE metode za balansiranje klasa, odnosno pretprocesiranje podataka, u odnosu na *ensemble* metode. U Tabeli 30 prikazani su isječci iz empirijskog testiranja performansi SVM-RE i *ensemble* balansiranih metoda za podatke iz direktnog marketinga.

Tabela 30. Komparacija ostvarenih rezultata primjenom SVM i *ensemble* metoda za pretprocesiranje podataka

	B-SVM+GBT	B-SVM+RF	*Balanced* DT	*Balanced Bagging* DT	*Balanced* RF
Tačnost	89,97%	89,27%	73,04%	76,86%	78,72%
AUC	0,95	0,921	0,78	0,827	0,839
Senzitivnost	83,35%	75,99%	75,00%	80,00%	75,00%
Fallout	9,03%	8,74%	26,97%	23,16%	21,26%

Nakon SVM-RE pretprocesiranja, klasifikatori koji su na testnim podacima pokazali najbolje prediktivne performanse su GBT i RF. U oba slučaja, tačnost je iznosila preko 89%, AUC 0,95 i 0,921, a *fallout* 9,03% i 8,74%, respektivno. Metrika senzitivnosti za GBT klasifikator iznosila je 83,35%, dok je za RF ostvarena vrijednost 75,88%.

S druge strane, tačnost za sva tri klasifikatora u slučaju *ensemble* balansiranja iznosi ispod 80%, najveća vrijednost za AUC ostvarena je za *Balanced* RF i iznosila je 0,839, dok je *fallout* za sva tri klasifikatora značajno veći u poređenju sa SVM-RE pretprocesiranjem (preko 20%). Slične vrijednosti za oba tipa balansiranja ostvarene su za metriku senzitivnosti (između 75% i 80%).

Ukoliko se za potrebe komparacije uporede performanse RF klasifikatora nakon SVM-RE i *ensemble* balansiranja klasa, uočava se da je tačnost veća za preko 10 procentnih poena za SVM-RE balansiranje, AUC je veći za 0,082, dok je *fallout* vrijednost manja za 12,52 procentna poena. Slične performanse nakon dva tipa

balansiranja ostvarene su za metriku senzitivnosti, gdje je predložena metoda ostvarila veću senzitivnost za 0,99 procentnih poena.

Dakle, u domenu podataka u direktnom marketingu, SVM-RE metoda balansiranja je pokazala značajno bolje performanse, te je na ovaj način dat pozitivan odgovor na IP2.

Osim toga, u ovoj empirijskoj studiji istražene su prediktivne performanse modela nakon isključivanja određenih grupa atributa iz analize, kako bi se utvrdio njihov individualni uticaj na metrike performansi. S tim u vezi, na osnovu pokazatelja iz Tabele 29 može se uočiti da su predloženi klasifikatori najslabije performanse pokazali nakon isključivanja *web* atributa. U tom slučaju, uočena je velika pristrasnost algoritama prema većinskoj klasi – tačnost je porasla sa 73% na 98%, a *fallout* sa 26,97% pao je na 1,85%. Dodatno, senzitivnost je iznosila svega 30%, u poređenju sa ostvarenih 75% u slučaju analize sa svim atributima, čime se potvrđuje značaj *web* atributa za predikciju odgovora na kampanju direktnog marketinga.

S druge strane, isključivanje RFM atributa, odnosno podatka o prethodnom kupovnom ponašanju nije značajno promijenilo prediktivne performanse. Tako je, na primjer, za *Balanced* RF tačnost manja za svega 0,35 procentnih poena, AUC povećan za 0,001, dok je senzitivnost ostala na istom nivou. Ovim je potvrđeno da za analizu, odnosno predikciju odgovora na buduće kampanje direktnog marketinga, koje su plasirane putem digitalnih kanala, prethodno kupovno ponašanje nema velikog značaja. Međutim, kao što je prethodno navedeno, *web* atributi su u ovom slučaju od presudne važnosti za buduće targetiranje, što predstavlja i odgovor na IP3.

Sumarna razmatranja u vezi s predloženim *ensemble* baziranim modelom odgovora na kampanju

Nivo konkurencije u globalizovanoj i digitalnoj ekonomiji definitivno pritiska kompanije da inoviraju i unaprijede svoje aktivnosti ka privlačenju i zadržavanju kupaca. Kompanije, s druge strane, mogu ostvariti značajne benefite proučavanjem podataka o potrošačima, kako bi razumjele njihove preferencije i na taj način poboljšale sistem podrške donošenju odluka u marketingu. Učinak kompanija za e-

trgovinu može se poboljšati plasiranjem odgovarajućih ponuda i prilagođenog sadržaja, kako bi se ispunili zahtjevi potrošača, jer uspjeh *web* prodavnice u značajnoj mjeri zavisi od poboljšanja kvaliteta informacija i usluga, kako bi se klijentima pristupilo na pravi način, a što je omogućeno bihejvioralnim targetiranjem.

Na osnovu uvida iz podataka o ponašanju pri kupovini, kompanije mogu kreirati strategije za precizno prilagođavanje sadržaja za definisanu grupu ciljnih kupaca, što će im omogućiti da izgrade pouzdane i smislene odnose s kupcima. Imajući to na umu, targetirane objave na društvenim mrežama mogu se koristiti kao glavno sredstvo za povezivanje s kupcima na mreži, a razumijevanjem namjera *online* kupaca i uz precizno targetiranje, ove aktivnosti mogu privući nove i zadržati stare kupce.

Cilj predloženog *ensemble* baziranog pristupa bio je da se riješi problem neravnoteže klasa, koji je jedan od najčešćih problema za primjenu algoritama mašinskog učenja u direktnom marketingu i upravljanju kampanjama (posebno kada su u pitanju *online* kupci, kod kojih stopa odgovora može biti izuzetno niska zbog velikog broja posjeta *web* sajtu). Da bi se pronašlo rješenje za značajno neuravnotežene podatke, primijenjena je tehnika poduzorkovanja, kao i *ensemble* metode. Rezultati prediktivnih modela pokazali su da predloženi *ensemble* bazirani klasifikatori imaju dobre prediktivne performanse.

Pored toga, rezultati pokazuju da bi kompanije za elektronsku trgovinu mogle da se u velikoj mjeri oslone na *web* metriku radi analize i predviđanja budućeg ponašanja svojih kupaca. Imajući to u vidu, segmenti kupaca se ne mogu u potpunosti opisati samo korišćenjem *web* podataka, ali je ovaj tip podataka pokazao svoju vrijednost za klasifikaciju kupaca na respondente i nerespondente, što je potvrđeno i u prethodnim istraživanjima.

Međutim treba istaći i neka ograničenja i nedostatke u vezi sa ovim pristupom i njegovim testiranjem. Naime, skup podataka korišćen za testiranje modela dobijen je od kompanije koja se bavi elektronskom trgovinom, i to za prvih sedam mjeseci od pokretanja *web* prodavnice. To je dovelo do male stope konverzije i nedostatka

dodatnih podataka o klijentima. U skladu s tim, u budućim istraživanjima bi se mogao koristiti skup podataka za duži vremenski period (jedna godina), čime bi se uzeo u obzir i problem sezonalnosti u maloprodaji odjeće i obuće. Pored toga, s obzirom na izuzetno nisku stopu odgovora, možda bi neke metode, kao što je *stacking ensemble*, mogle dati bolje rezultate, koji bi se mogli istražiti i uporediti sa ovim u budućnosti.

Iako je ovaj *ensemble* bazirani model izgrađen zbog poređenja, kako bi se ukazalo na superiornost SVM-RE baziranog pristupa, on se kod jako izražene neravnoteže klasa pokazao značajno boljim od klasičnih klasifikatora (kao što su SVM ili LR). Takođe, bio je od koristi da se utvrdi značaj *web* metrika za modele odgovora na kampanju. Rezultati ovog modela prezentovani su na renomiranoj međunarodnoj konferenciji i objavljeni u referentnoj publikaciji (Rogić & Kaščelan, 2022).

5.6 Testiranje SVR baziranog modela za targetiranje najprofitabilnijih kupaca

U ovom dijelu rada biće predstavljen pregled rezultata testiranja pristupa predloženog u sekciji 4.5.5, kao i diskusija o nalazima primijenjenog modela.

U Tabeli 32 predstavljeni su rezultati dobijeni na testnim skupovima (neviđenim podacima) za oba skupa empirijskih podataka opisanih u sekciji 5.1.3. Na oba skupa podataka primijenjena je *Support Vector Regression* metoda: prvo, na svim atributima, a zatim i na selekciji atributa, isključivanjem jedne po jedne grupe ili pojedinačnog atributa iz analize. Na ovaj način može se procijeniti uticaj pojedinačnih atributa na performanse modela, i to: *web* atributa, RFM atributa i kao grupe i pojedinačno, atributa o proizvodima i atributa o kupcima. Praćenjem rezultata u svakoj iteraciji, mogu se utvrditi grupe atributa koje su najznačajnije za predikciju profitabilnosti kupaca i odgovoriti na IP3 kada je u pitanju ovaj model.

Tabela 31. Uporedni prikaz rezultata LR i SVR modela za predikciju profitabilnosti kupaca

	LR	SVR
RE	28,64%	9,19%
R^2	0,840	0,899
RE TOP20%	18,17%	13,45%
DEV <10% TOP20%	41,94%	48,39%
FP	9,45% (12)	2,36% (3)
FN	6,45% (2)	12,9% (4)

Komparacija s linearnom regresijom i potvrda hipoteze H4

Na osnovu rezultata dobijenih pomoću SVR, korišćenjem prvog skupa podataka distributera sportske opreme, može se uočiti da je relativna greška (RE) na cijelom skupu podataka na nivou 9,19%, dok je koeficijent determinacije (eng. *squared correlation* - R^2) u vrijednosti 0,899. Ovi rezultati ukazuju na generalno dobre performanse modela i potvrđuju hipotezu H4 na nivou svih kupaca. Međutim, kao što hipoteza H4 ističe, fokus je na kategoriji najvrednijih kupaca.

Kada je u pitanju 20% najvrednijih kupaca, RE iznosi 13,45%. Procenat odstupanja predikcije manjih od 10% u ovoj kategoriji kupaca je 48,39%. Dodatno, stopa pogrešne klasifikacije 20% najvrednijih kupaca je svega 2,36%.

U poređenju sa običnim linearno regresionim modelom, SVR je dao bolje rezultate. Tabela 31 prikazuje uporedne rezultate za RE, R^2, RE za grupu od 20% najprofitabilnijih kupaca, procenat odstupanja manjih od 10% za ovu grupu, stopu pogrešne klasifikacije 20% najvrednijih kupaca (FP) i stopu pogrešne klasifikacije ostalih 80% kupaca, koji su predikcijom prepoznati kao najvredniji (FN), za LR i SVR modele, respektivno.

Za gotovo sve analizirane parametre, osim FN, model LR je pokazao slabije performanse, s relativnom greškom od 18,17% za najboljih 20% kupaca u odnosu

na 13,45% koje je ostvario SVR. Iz tabele se može zaključiti da SVR metoda s visokom tačnošću predviđa profitabilnost najvrednijih kupaca, s niskom stopom RE za ovaj segment (13,45%) i niskom stopom njihove pogrešne klasifikacije (2,36%), pa je hipoteza H4 potvrđena i za ovaj segment.

Testiranje uticaja atributa i odgovor na IP3

Ukoliko se isključene grupe atributa posmatraju pojedinačno, može se primijetiti da je SVR model imao najgore performanse isključivanjem RFM atributa kao grupe, što je značajno povećalo RE na 29,46%, a smanjilo R^2 na 0,709. Još jedna značajna promjena u performansama modela uočava se prilikom isključivanja *Monetary* varijable (pojedinačno) – RE raste do nivoa 31,72%, dok se R^2 smanjuje na 0,737.

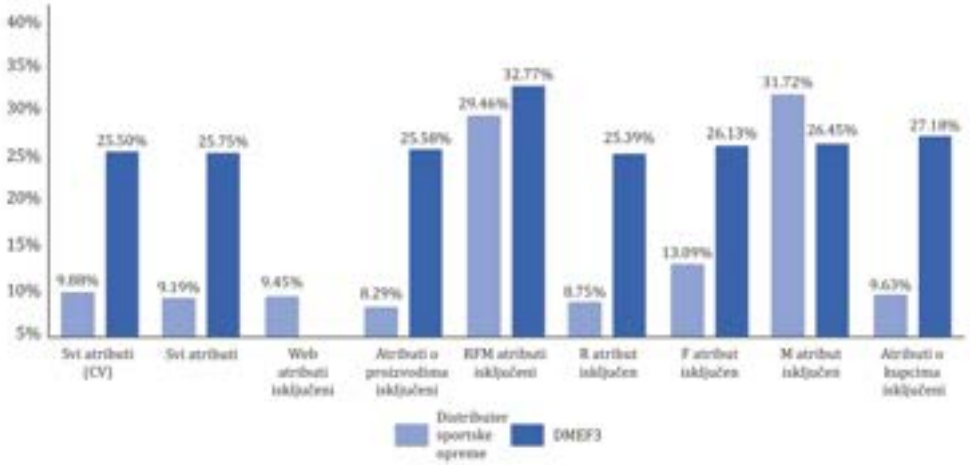
S druge strane, isključivanje *web* atributa i atributa o kupcima samo neznatno utiče na performanse modela, sa RE od 9,45% i 9,63%, kao i R^2 od 0,851 i 0,859, respektivno. Konačno, prilikom isključivanja atributa o proizvodima, postiže se unaprijeđen rezultat relativne greške u odnosu na model sa svim podacima: 8,29%. S druge strane, u ovom slučaju ostvarena vrijednost R^2 je 0,791.

Na drugom skupu podataka – DMEF3, testiranjem modela na svim podacima, za najznačajnije mjere performansi modela ostvareni su sljedeći rezultati: RE=25,25% i R^2=0,473. Veći nivo greške je u slučaju ovog skupa podataka očekivan, uzimajući u obzir značajno duži horizont predikcije (šest godina). *Malthouse* i *Blattberg* (2005) su u svom radu koristili ovaj skup podataka za predikciju profitabilnosti korišćenjem linearne regresije i ostvarili stopu pogrešne klasifikacije od 55% za 20% najvrednijih kupaca. U istom radu, autori su postavili pitanje mogućnosti tačnog predviđanja profitabilnosti za tako dug horizont predikcije.

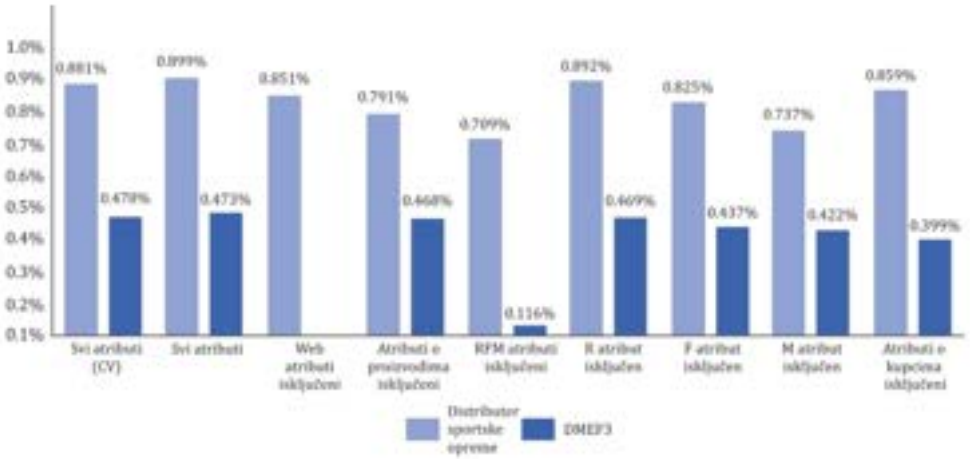
Analizom rezultata korišćenih atributa za kraći i duži vremenski period mogu se generisati različiti zaključci, što opravdava korišćenje ovog validacionog skupa podataka u te svrhe.

Slično kao i u prvom skupu podataka, isključivanje RFM atributa iz modela dovodi do značajnog pogoršanja njegovih performansi, sa RE i R^2 u iznosu od 32,77% i

0,116, respektivno. Međutim, u ovom slučaju atributi o kupcima imaju mnogo veći uticaj na performanse modela. Kao što se iz Tabele 31 vidi, nakon rezultata isključivanjem RFM atributa, uklanjanje podataka o kupcima iz analize dovodi do drugog najgoreg rezultata (RE=27,18% i R^2=0,399). Neznatne promjene u rezultatima modela uočavaju se prilikom isključivanja podataka o proizvodima, kao pojedinačnim isključivanjem RFM atributa. Kada se posmatraju ovi atributi pojedinačno, najveći uticaj na model ima isključivanje *Monetary* varijable, zatim *Frequency*, dok je najmanji uticaj *Recency* varijable na rezultate modela. S tim u vezi, važno je istaći razliku u performansama prilikom isključivanja ovih pojedinačnih atributa u poređenju s prvim skupom podataka. Ostvarene razlike u metrikama performansi modela u ovom slučaju su manje, a direktna su posljedica dužeg horizonta predviđanja. Ostvarene vrijednosti relativne greške i koeficijenta determinacije na svim testiranim modelima predstavljeni su na slikama 41 i 42.



Slika 41. Vrijednosti relativne greške (RE) za oba testirana modela za oba skupa podataka



Slika 42. Vrijednosti koeficijenta determinacije (R^2) u svim testiranim modelima za oba skupa podataka

Rezultati pokazuju da se adekvatno predviđanje može napraviti korišćenjem samih RFM atributa, ako je to sve što je kompaniji dostupno. RFM atributi sadrže značajne informacije o istoriji kupovine kupaca, a u oba testirana skupa podataka ove varijable su se pokazale kao najznačajniji prediktor profitabilnosti kupaca. Ovo je u skladu s rezultatima prethodnih istraživanja, koja ističu značaj RFM atributa za procjenu vrijednosti kupaca i predviđanje profitabilnosti (Ait Daoud et al., 2015; Wei-jiang et al., 2011).

Analizirajući RFM atribute pojedinačno, oni bi bili rangirani po sljedećem redosljedu: M, F, R. U slučaju oba skupa podataka, atribut *Monetary* promjenljiva najviše određuje profitabilnost korisnika. Ovo nije u skladu s većinom prethodnih istraživanja, koja su predstavljena u sekciji 3.1.1 (Tabela 1). U nekim prethodnim istraživanjima (Khajvand et al., 2011; Monalisa et al., 2019; Safari et al., 2016) *Monetary* varijabla bila je druga po važnosti, dok je u radovima druge grupe autora (Chen et al., 2017; Liu & Shih, 2005) bila najmanje važna od tri RFM varijable. Međutim, autori *Aggarwal* i *Delhi* (2020) su u svojoj studiji tehnikom *Fuzzy-AHP* dostigli isti redosljed važnosti RFM varijabli, kao u ovoj disertaciji.

Dakle, važnost i značaj RFM varijabli kao grupe se ne dovodi u pitanje, međutim, njihov pojedinačni uticaj na profitabilnost kupaca mora se detaljno istražiti za svaki pojedinačni slučaj, privrednu djelatnost ili sektor (Dursun & Caber, 2016). Iako

rezultati ovog empirijskog istraživanja ukazuju na to da *Monetary* varijabla ima najznačajniji uticaj na predviđanje profitabilnosti, sa dužim horizontom predviđanja, njena vodeća pozicija se smanjuje, pa sva tri RFM atributa imaju sličnu važnost u tom slučaju. Dakle, veći uticaj *Monetary* varijable u prvom skupu podataka može se objasniti kraćim horizontom predviđanja – ako je kupac imao transakcije većeg obima i veće vrijednosti u prethodnim kampanjama, ovi iznosi će se prenijeti na zbir transakcija u sljedećem periodu. S druge strane, ako je cilj predviđanje na duži rok (na primjer nekoliko godina), iznosi transakcija iz prvog dijela posmatranog perioda ne bi dugoročno diktirali krajnju profitabilnost.

Kada su u pitanju *web* metrike, ovo istraživanje nije dokazalo njihov značaj za predviđanje profitabilnosti kupaca. U prethodnim istraživanjima, autori su istakli da *web log* podaci predstavljaju dominantne faktore za segmentaciju kupaca na osnovu njihove istorije kupovina (Van den Poel & Buckinx, 2005). Osim toga, *web* metrike su se pokazale kao značajne u modeliranju odgovora kupaca i predviđanju budućih kupovina (Esmeli et al., 2020; Martínez et al., 2020), što je i potvrđeno u prethodnom poglavlju ove distertacije.

Stoga se podaci i *web* metrike mogu koristiti u nekoliko drugih tipova analiza kupaca, ali, u slučaju predviđanja profitabilnosti, ova grupa atributa se nije pokazala značajnom. U tom smislu, ako se, na primjer, u predviđanju odgovora korisnika na buduću kampanju koristi prosječno trajanje sesije, moglo bi se očekivati da će taj atribut uticati na ovo predviđanje - može odrediti da li će se transakcija dogoditi ili ne, ali je manja vjerovatnoća da će ona odrediti samu vrijednost transakcije. Obično, kada procjenjuju značaj atributa, menadžeri imaju u vidu odgovor kupca i na osnovu toga donose odluke. Međutim, kada je u pitanju predviđanje profitabilnosti na osnovu vrijednosti transakcije, onda je važnije da li je kupac potrošio mnogo ili malo u prošlosti. Stoga, rezultati ovog istraživanja sugerišu da bi za predviđanje profitabilnosti, u poređenju sa predikcijom odgovora na kampanju, trebalo usmjeriti fokus na RFM, a posebno *Monetary* atribut u kratkom roku, čime je dat odgovor na IP3 kada je u pitanju ovaj model.

Korišćenje različitih kategorija podataka može pomoći u profilisanju i segmentiranju kupaca, predviđanju njihove profitabilnosti i kreiranju personalizovanih poruka i ponuda u cilju maksimizacije njihove vrijednosti.

Sumarna razmatranja u vezi sa SVR baziranim modelom za targetiranje najprofitabilnijih kupaca

U ovom dijelu rada testiran je model za targetiranje kupaca baziran na SVR predikciji njihove profitabilnosti. Takođe, u ovom dijelu istraživanja testiran je i značaj RFM atributa za predviđanje profitabilnosti kupaca u odnosu na *web* metrike, podatke o kupcima i kupljenim proizvodima. Za predviđanje profitabilnosti, kao kontinuirane varijable, implementiran je SVR pristup. Ova studija jedinstveno koristi nekoliko različitih kategorija podataka i izvora i uključuje potencijalno nove faktore predviđanja performansi, što u prethodnim istraživanjima nije često sprovođeno.

Rezultati potvrđuju da je SVR robustan u slučaju modeliranja podataka s distribucijom koja nema karakteristike normalne raspodjele. S tim u vezi, SVR je efikasno odgovorio na izazov predikcije profitabilnosti za najvrednije kupce. Dakle, uprkos izraženoj asimetričnosti distribucije profitabilnosti, SVR je, sa ostvarenim vrijednostima relativne greške od 13,45% za 20% najvrednijih kupaca, potvrdio svoju efikasnost za predikciju profitabilnosti najvrednije grupe kupaca, za razliku od klasičnog regresionog modela, čija je RE za šest procentnih poena veća kod ove grupe.

Takođe, rezultati ove studije ukazuju na to da bi se kompanije mogle osloniti isključivo na RFM atribute za analizu i predviđanje profitabilnosti svojih kupaca. Međutim, podaci o ponašanju pri kupovini mogu se koristiti u kombinaciji s drugim kategorijama podataka, kao što su *web* metrike ili podaci o kupcima, iako je ovaj individualni tip podataka dokazao svoju vrijednost za predviđanje profitabilnosti kupaca, što je takođe potvrđeno u prethodnoj literaturi. Ova studija je istakla *Monetary* atribut za kratkoročno predviđanje profitabilnosti kupaca. U slučaju prvog skupa podataka (distributer sportske opreme), isključenje *Monetary* atributa dovelo je do povećanja relativne greške na vrijednost od 31,72%, što je bio najveći nivo

greške u poređenju sa svim ostalim testiranim modelima. Osim toga, rezultati ukazuju na smanjenje značaja ovog atributa s povećanjem horizonta predviđanja. U dugoročnim predviđanjima, sva tri RFM atributa pokazala su sličan efekat na performanse predviđanja. Kada je u pitanju *Monetary* atribut i njegov uticaj u skupu podataka DMEF3, nivoi RE po svim RFM atributima su slični (26,13%, 26,45% i 27,18%, respektivno), što naglašava dobijeni zaključak.

Sa aspekta marketinga, prezentovani rezultati mogu biti osnov za segmentaciju kupaca i odabir ciljne grupe za targetiranje u budućim kampanjama na osnovu profitabilnosti, kao i za druge buduće aktivnosti direktnog marketinga. I ovaj prediktivni model je, kao i u slučaju prethodna dva, realizovan u *Rapid Miner* alatu, u vidu dva korisnička procesa spremna za obučavanje i predikciju (Prilozi 10 i 11).

Kao preporuka za buduća istraživanja, može se sprovesti kombinacija predviđanja profitabilnosti kupaca i predikcije odgovora kupaca na kampanju, kako bi se dobila potpuna slika o najprofitabilnijim kupcima, kao i onima s najvećom vjerovatnoćom odgovora. Podaci iz takvog istraživanja mogu biti posebno korisni u kompanijama koje posluju sa ograničenim marketing budžetom.

Rezultati ovog modela prezentovani su na međunarodnoj konferenciji i objavljeni u monografiji izdavača *Springer* (Rogic & Kascelan, 2021).

Tabela 32. Prediktivne performanse modela na dva skupa podataka

Skup podataka	Performanse	Skup za obučavanje	Testni skup							
		Svi atributi (Cross-Validation)	Svi atributi	*Web* atributi isključeni	Atributi o proizvodima isključeni	RFM atributi isključeni	*Recency* isključen	*Frequency* isključen	*Monetary* isključen	Podaci o kupcima isključeni
Distributer sportske opreme (e-trgovina)	*RMSE*	13,759 +/- 7,634	18,246	22,238	25,968	30,541	18,843	23,826	28,580	21,629
	Absolute error	7,315 +/- 4,129	7,528 +/- 16,620	8,412 +/- 20,586	8,625 +/- 24,494	16,427 +/- 25,748	7,456 +/- 17,306	10,437 +/- 21,419	16,628 +/- 23,245	8,482 +/- 19,896
	Relative error	9,88% +/- 2,49%	9,19% +/- 14,40%	9,45% +/- 16,02%	8,29% +/- 20,02%	29,46% +/- 28,87%	8,75% +/- 14,51%	13,09% +/- 16,77%	31,72% +/- 32,44%	9,63% +/- 15,48%
	Correlation	0,938 +/- 0,040	0,948	0,922	0,889	0,842	0,944	0,908	0,858	0,927
	Squared Correlation	0,881 +/- 0,075	0,899	0,851	0,791	0,709	0,892	0,825	0,737	0,859
	Spearman Rho	0,938 +/- 0,047	0,937	0,925	0,899	0,737	0,94	0,916	0,715	0,929
DMEF3	*RMSE*	5,407 +/- 0,081	5,334	-	5,360	6,952	5,359	5,517	5,581	5,721
	Absolute error	4,024 +/- 0,043	3,995 +/- 3,535	-	4,029 +/- 3,536	5,183 +/- 4,633	4,015 +/- 3,550	4,131 +/- 3,656	4,181 +/- 3,697	4,308 +/- 3,765
	Relative error	25,50% +/- 0,52%	25,25% +/- 21,84%	-	25,58% +/- 22,16%	32,77% +/- 28,95%	25,39% +/- 21,90%	26,13% +/- 22,68%	26,45% +/- 23,07%	27,18% +/- 22,46%
	Correlation	0,691 +/- 0,014	0,688	-	0,684	0,340	0,685	0,661	0,650	0,632
	Squared Correlation	0,478 +/- 0,019	0,473	-	0,468	0,116	0,469	0,437	0,422	0,399
	Spearman Rho	0,478 +/- 0,019	0,647	-	0,638	0,360	0,642	0,622	0,621	0,555

Zaključak

Digitalni trag, koji korisnici interneta ostavljaju za sobom može se posmatrati kao nepresušni izvor podataka o njima, njihovom *online* ponašanju, preferencijama i željama. Istovremeno, zbog raznorodnosti, kompleksnosti, različitih izvora iz kojih se podaci prikupljaju, složenog manipulisanja i zahtjevne obrade, organizacije se suočavaju sa izazovom da te podatke pretvore u informacije i znanje. Da bi takvi podaci bili obrađeni i pretočeni u relevantne informacije, neophodno je korišćenje *data mining* rješenja. S tim u vezi, pronalaženje skrivenih obrazaca u podacima omogućava kompanijama da ih valorizuju kroz svoje poslovne odluke, sprovođenjem objektivne, precizne i na podacima zasnovane segmentacije, selekcije i targetiranja kupaca, što predstavlja osnovne aktivnosti direktnog marketinga.

Kako je jedan od osnovnih ciljeva direktnog marketinga predviđanje ponašanja potrošača, a *data mining* metode su uglavnom prediktivne, u ovom radu su definisani i empirijski testirani prediktivni modeli odlučivanja u direktnom marketingu, zasnovani na *data mining* metodama. Rezultati su pokazali da su predloženi modeli bazirani na SVM pretprocesiranju podataka, u kombinaciji sa odgovarajućim prediktivnim klasifikatorima, uspješno riješili probleme nebalansiranosti klasa i asimetričnosti distribucije profitabilnosti, koji nastaju zbog manjeg broja respondenata u odnosu na nerespondente i koji kod najvećeg broja prediktivnih modela dovode do loše predikcije respondenata.

Svi ciljevi postavljeni u uvodnom poglavlju ove distertacije ostvareni su kroz potvrdu hipoteza na testnim podacima i podacima za validaciju koji su javno dostupni, što je detaljno obrazloženo u nastavku. Međutim, treba istaći da su postojala i određena ograničenja i nedostaci predloženih modela, o čemu će takođe detaljnije biti riječi u nastavku.

Kod modela prediktivne RFM segmentacije, prateći proceduru konceptualnog modela, korišćeni su objektivni parametri za definisanje broja potrošačkih segmenata, koji se

zasnivaju na *Davies-Bouldin* (DB) indeksu. S tim u vezi, segmentacija na osnovu prethodnog kupovnog ponašanja (RFM atributa, pomoću *k-means* metode) i izbor broja segmenata na osnovu DB indeksa, obezbijedili su maksimalnu homogenost kupaca u okviru pojedinačnih segmenata, maksimalnu heterogenost između različitih segmenata, kao i optimalan broj segmenata, čime je potvrđena hipoteza H1.

Takođe, kod ovog modela, empirijski je, korišćenjem SVM-RE metode, potvrđena efikasnost predikcije najvrednijih kupaca, što doprinosi kvalitetnijem targetiranju ove grupe kupaca u budućim kampanjama direktnog marketinga. Na ovaj način, smanjuje se efekat rasipanja u kampanji, što povećava ukupne prihode generisane kroz ovu marketing aktivnost. Osim toga, ekstrakcijom pravila pomoću DT modela, omogućen je precizan opis segmenta najvrednijih kupaca i kreiranje njihovog profila, što olakšava buduću komunikaciju sa ovom važnom grupom. Ostvarene prediktivne performanse potvrdile su efikasnost modela za najmanji i najvredniji potrošački segment, rješavajući efikasno problem minorne klase najvrednijih kupaca. Na taj način, predloženi model povećava efikasnost targetiranja najvrednijih potrošača, povećavajući moguće prihode, a smanjujući troškove kampanje, kao i povećavajući efikasnost interakcije s najvrednijim segmentom kupaca, zahvaljujući otkrivanju njihovog profila, čime je potvrđena hipoteza H2.

U disertaciji je empirijski testiran i prediktivni model odgovora kupca na kampanju direktnog marketinga. Predloženi model ostvario je bolje prediktivne performanse u poređenju s modelima iz prethodnih studija, a validiran je i na podacima iz *online* i *offline* transakcija, tako da se može primijeniti u različitim industrijama. S tim u vezi, predloženi model povećava efikasnost targetiranja i predikcije kupaca koji će najvjerovatnije odgovoriti na kampanju direktnog marketinga, te omogućava kreiranje profila respondenata i olakšava buduću interakciju s njima, čime je potvrđena hipoteza H3.

Empirijskim testiranjem modela za predikciju profitabilnosti, potvrđena je robustnost SVR metoda kada je u pitanju asimetrična distribucija varijable koja se predviđa. Dakle, SVR je efikasno odgovorio na izazov predikcije profitabilnosti za najmanju i najvredniju grupu kupaca, ne smanjujući značajno tačnost predikcije ekstremno visoke profitabilnosti kod najmanje i najvrednije grupe kupaca, za razliku od metoda za predikciju profitabilnosti koje su ranije primjenjivane. Na ovaj način, predložena metoda, identifikujući najprofitabilnije kupce s visokom preciznošću, povećava profitabilnost kampanje, čime se potvrđuje hipoteza H4.

Upoređivanjem performansi modela prediktivne RFM segmentacije baziranog na SVM metodi s modelom koji uključuje i *ensemble* tehnike, utvrđeno je poboljšanje performansi korišćenjem *ensemble* tehnika. Pored toga, prediktivne performanse modela su unaprijeđene korišćenjem *ensemble* metoda i pri empirijskom testiranju SVM-RE modela odgovora na kampanju, što je potvrdilo da SVM-RE u kombinaciji sa *ensemble* pristupom značajno povećava tačnost predikcije, pa je odgovor na IP1 pozitivan.

Za rješavanje nebalansiranosti klasa u ranijim radovima primjenjivane su uglavnom *ensemble* metode u kombinaciji s balansiranim uzorkovanjem. Upoređivanjem prediktivnih performansi SVM-RE baziranih modela RFM segmentacije i odgovora na kampanju sa *ensemble* modelima baziranim na balansiranom uzorkovanju, evaluirani su modeli po pitanju rješavanja problema nebalansiranosti klasa. Prilikom testiranja oba modela, SVM-RE metod dao je bolje rezultate za balansiranje klasa od *ensemble* metoda, čime je dat pozitivan odgovor na IP2.

Značaj *web* i kupovnog ponašanja za predikciju u direktnom marketingu istraživan je u dva pravca – u modelu odgovora na kampanju, kao i u modelu za targetiranje najprofitabilnijih kupaca. Pri testiranju modela odgovora na kampanju, pokazalo se da su *web* atributi najrelevantniji za predikciju i buduće targetiranje kupaca. S druge strane, testiranje SVR baziranog modela za targetiranje najprofitabilnijih kupaca,

pokazalo je da najznačajniji uticaj na prediktivne performanse modela imaju atributi koji opisuju kupovno ponašanje, odnosno RFM atributi. Osim toga, rezultati ukazuju da *Monetary* atribut u kratkom roku predviđanja ima najveći uticaj na performanse, čime je odgovoreno na IP3 kada su u pitanju ova dva modela.

Kada je u pitanju preispitivanje validnosti istraživanja i ograničenja u ovom radu, u prvom redu treba istaći da je SVM, kao pretprocesor podataka, kombinovan isključivo sa *Bagging ensemble* metodom, dok se za opisivanje SVM izlaza koristio samo DT metod. Korišćenjem drugih *emsemble* metoda i tehnika za generisanje pravila iz SVM izlaza, moglo bi dovesti do drugačijih rezultata. Zbog složene i vremenski zahtjevne ekstrakcije realnog skupa podataka za testiranje modela, u ovoj disertaciji modeli su testirani na podacima iz jedne kompanije, dok su, dodatno, validirani na javno dostupnim skupovima podataka. Testiranjem na drugim skupovima podataka s različitim distribucijama klasa mogli bi se potencijalno dobiti drugačiji nalazi. Na kraju, uspjeh *data mining* metoda u velikoj mjeri zavisi od kvaliteta podataka. Skupovi podataka koji su se koristili u ovoj disertaciji uključuju samo neke karakteristike kupaca, uzimajući u obzir da su baze kupaca za direktni marketing kod naših kompanija još uvijek nedovoljno razvijene. Uključivanjem više atributa kupaca mogu se dobiti jasnija pravila za targetiranje novih kupaca.

Osnovni naučni doprinos ove disertacije ostvaren je prevazilaženjem nedostataka postojećih i poznatih modela za segmentaciju, selekciju i targetiranje kupaca u direktnom marketingu (koji se odnose na probleme zanemarivanja minorne klase i nepreciznog predviđanja asimetrične distribucije), kroz predlog koncepta i realizacije pet novih prediktivnih modela u ovom domenu.

Takođe, s obzirom na to da se problemi nebalansiranosti klasa i asimetrične distribucije kod *data mining* zasnovanih prediktivnih procesa u direktnom marketingu do sada nisu rješavali pomoću SVM metode, ova disertacija je pokrila ovaj teorijski jaz. Dodatno, iako su društvene mreže dominantan medij za plasiranje targetiranih kampanja direktnog marketinga u praksi, u najvećem dijelu literature se direktni maketing i dalje

poistovjećuje s direktnom poštom (tradicionalnom poštom ili *e-mail* kampanjama). Takođe, *online* kampanje plasirane putem društvenih mreža još više potenciraju gore pomenute probleme, zbog male stope odziva i broja najvrednijih kupaca u ogromnim bazama *online* potrošača. Stoga, ova disertacija dopunjava postojeću teoriju novim konceptom prediktivnih modela, koji efikasno podržavaju odlučivanje prilikom plasiranja kampanja putem društvenih mreža i uopšte u direktnom marketingu.

Uporedo s naučnim doprinosom, predloženi prediktivni modeli imaju praktični značaj, jer povećavaju moguću profitabilnost kampanje i omogućavaju individualni pristup kupcu, podržavajući tako lakše i efikasnije odlučivanje marketing menadžerima. Naime, povećanje tačnosti predikcije vrlo vjerovatnih respondenata dovodi do potencijalno većih prihoda, dok smanjenje prediktivne greške smanjuje troškove koji se odnose na neefikasno targetiranje, tj. na nerespondente. Osim toga, definisanje pravila koja opisuju profil najvrednijih potrošača i respondente u kampanjama direktnog marketinga, omogućava kreiranje individualnih ponuda, što je u skladu s razvojem "jedan na jedan" pristupa u marketingu.

Suočeni sa smanjenjem budžeta, donosioci odluka u marketingu su pod sve većim pritiskom da maksimiziraju rezultate i optimizuju investicije u marketing aktivnostima. Stoga, da bi ostali konkurentni na dinamičnom tržištu, oni moraju svoje odluke zasnivati na objektivnim podacima i koristiti prediktivnu analitiku u domenu poslovnog odlučivanja. Kao rezultat toga, donosioci odluka u marketingu mogu imati povjerenja da će odluke koje donesu na ovaj način dovesti do krajnjih rezultata kojim teže. Iako postoji mnogo mogućnosti za dalja usavršavanja, modeli koji su ovoj disertaciji predloženi i empirijski testirani, potvrđuju da je odlučivanje bazirano na podacima i prediktivnoj analitici efikasno. Modeli nisu samo testirani, već i praktično realizovani u *Rapid Miner* alatu, u vidu gotovih procesa koje treba prilagoditi bazi kupaca konkretne kompanije (Prilozi 5-11). To znači da sofisticirani analitičar, koristeći ovaj alat, podešava parametre DM operatora na osnovu podataka iz ove baze i u okviru definisanih procesa. Za obučavanje se koriste podaci iz prethodnih kampanja (istorija

kupovnog i *web* ponašanja kupaca, kao i njihov o dgovor na kampanju). Pronađeni parametri se prenose u procese za pretprocesiranje i predikciju, koji se zatim primjenjuju na trenutne podatke o kupcima. Rezultat je predikcija - da li će konkretni kupac odgovoriti na buduću kampanju ili ne, kao i kolika je očekivana profitabilnost za pojedine kupce iz baze.

Dakle, s praktičnog aspekta, ova disertacija doprinosi promociji i primjeni prediktivne analitike u procesima odlučivanja u direktnom marketingu u kompanijama. Ona je od najveće koristi menadžerima kampanja, s obzirom na to da potvrđuje da oni mogu imati povjerenja u donošenje odluka na osnovu podataka. Ovo posebno važi za upravljanje *online* kampanjama, koje karakterišu ogromne baze potencijalnih kupaca, koje s jedne strane predstavljaju izvor vrijednih informacija, ali s druge strane i težak zadatak kada odlučivanje nije podržano *data mining* rješenjima.

Uzimajući u obzir efikasnost SVM metode u balansiranju podataka, potvrđenu u ovom istraživanju, u budućim istraživanjima bi se mogle testirati mogućnosti ovog pristupa za nadogradnju CRM sistema i razvoja individualnog pristupa kupcu. U novije vrijeme, u prediktivnoj analitici sve više je aktuelan *text mining*, koji omogućava izvlačenje važnih informacija iz teksta (dokumenata, komentara korisnika i slično), kao i analiza sentimenta (eng. *sentiment analysis*), koja omogućava otkrivanje osjećaja (stavova) izraženih u tekstu. S obzirom na to da je kroz CRM omogućena i pisana komunikacija s kupcima, ove tehnike prediktivne analitike bi mogle pomoći da se ona iskoristi za izvlačenje vrijednih informacija o stavu svakog pojedinačnog kupca.

Analiza sentimenta, odnosno osjećanja i stavova kupaca na osnovu tekstualnih podataka, tj. korišćenjem *text-mining* i NLP tehnika (tehnike obrade prirodnog jezika), u kombinaciji s prediktivnim klasifikatorima, omogućava prediktivnu klasifikaciju poruka u pozitivne, negativne i neutralne. S obzirom na to da je broj negativnih poruka obično manji u odnosu na pozitivne i neutralne, prilikom predviđanja sentimenta takođe dolazi do problema nebalansiranosti klasa. Stoga bi se kod predikcije sentimenta mogao testirati pristup kombinovanja SVM pretprocesiranja s različitim

klasifikatorima, da bi se povećala tačnost predikcije. Pri tome, prednost treba dati onim klasifikatorima koji daju opise klasa tj. generišu pravila (kao što je DT), jer, pored predikcije setimenta, za CRM sisteme značajno je i otkrivanje termina koji se vežu za pozitivne ili negativne stavove kupaca. Na taj način, kompanije koje već imaju uspostavljene CRM sisteme i direktan kontakt s kupcima, te koje bilježe transkripte interakcije s kupcima prilikom prodaje (ili njihov *feedback*), mogu primijeniti analizu sentimenta, kako bi tačnije predvidjeli stavove kupaca prema proizvodu koji se nudi, kao i koji su značajni termini koji se povezuju s pozitivnim i negativnim stavovima kupaca prema ponudi. S tim u vezi, može se utvrditi da li kupci bolje reaguju na jednokratni popust, produženje garancije, poklon uz kupovinu, *cash-back* ili druge podsticaje i elemente ponude. Dakle, integracijom analize sentimenta u CRM sisteme, mogu se pratiti povratne informacije kupaca i iskoristiti za predikciju njihovog stava prema proizvodu, kao i za prilagođavanje ponude ciljnoj grupi, što može uticati na zadovoljstvo kupaca i izgradnju dugoročne lojalnosti. Na ovaj način, mogli bi se dobiti efikasniji CRM sistemi, koji kompaniji donose više prihoda i omogućavaju kreiranje baze lojalnih kupaca.

PRILOZI

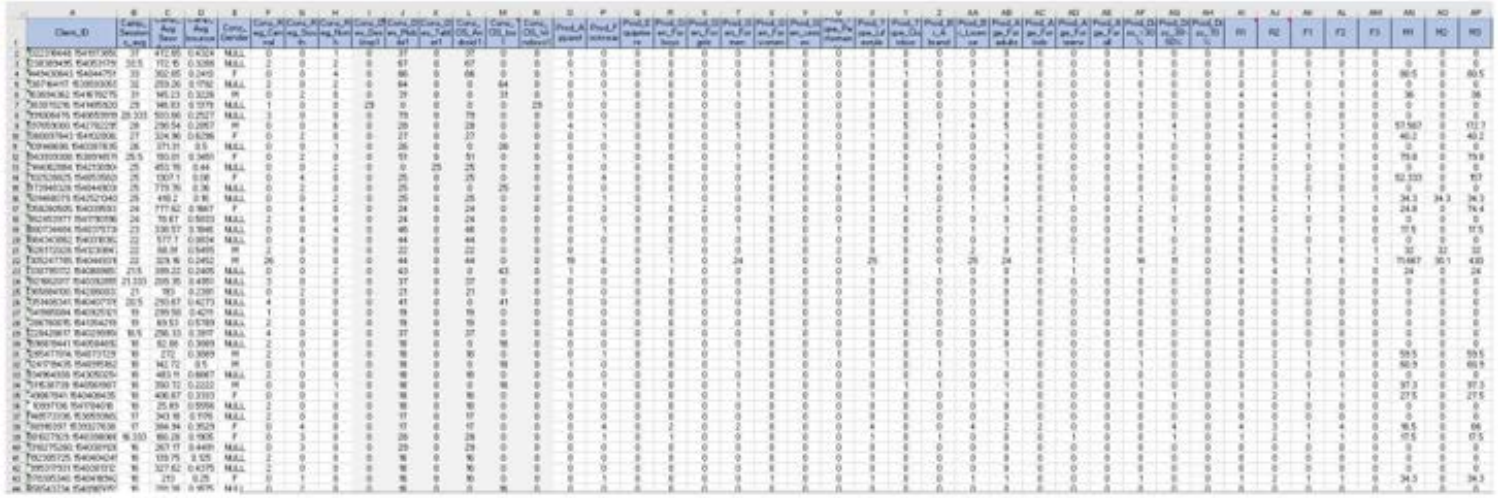
Prilog 1 – isječak iz baze podataka obavljenim kupovinama iz prethodnih direktnih kampanja kompanije *Sport Vision*

	A	B	C	D	E	F	G	H	I	J	K
1	Order_ID	Cons_gender	Discount	Prod_type	Prod_gender	Prod_category	Prod_brand	Prod_age	R	F	M
2	28045	M	0.40	Footwear	For men	Running	A brands	For adults	1	2	117
3	28046	F	0.40	Equipment	For men	Accessories	Licence	For adults	1	1	27
4	28050	F	0.40	Footwear	For women	Lifestyle	A brands	For adults	1	1	18
5	28053	M	0.50	Footwear	For boys	Football	A brands	For teens (8-14)	1	1	16
6	28054	M	0.00	Footwear	For boys	Lifestyle	A brands	For teens (8-14)	1	1	44
7	28055	M	0.00	Footwear	For men	Outdoor	A brands	For adults	1	1	79
8	28057	M	0.00	Equipment	Unisex	Fitness	A brands	For teens (8-14)	1	2	60
9	28057	M	0.00	Equipment	Unisex	Lifestyle	A brands	For all	1	2	60
10	28059	F	0.50	Footwear	For women	Lifestyle	A brands	For adults	1	1	29.5
11	28060	F	0.50	Apparel	For boys	Lifestyle	Licence	For younger kids (4-10)	1	2	42
12	28065	F	0.00	Footwear	For boys	Basketball	A brands	For babies (0-4)	1	3	94.6
13	28067	F	0.00	Footwear	For girls	Lifestyle	A brands	For babies (0-4)	1	1	29
14	28068	F	0.50	Footwear	For women	Lifestyle	Licence	For adults	1	1	19.5
15	28070	F	0.50	Footwear	For girls	Outdoor	A brands	For teens (8-14)	1	1	29.5
16	28073	M	0.40	Footwear	For men	Running	A brands	For adults	1	2	96.8
17	28073	M	0.00	Apparel	For men	Lifestyle	A brands	For adults	1	2	96.8
18	28077	F	0.50	Apparel	For girls	Lifestyle	A brands	For teens (8-14)	1	1	34.5
19	28079	F	0.40	Footwear	For girls	Lifestyle	A brands	For teens (8-14)	1	2	48.3
20	28084	F	0.30	Footwear	For women	Lifestyle	A brands	For adults	1	8	306.2

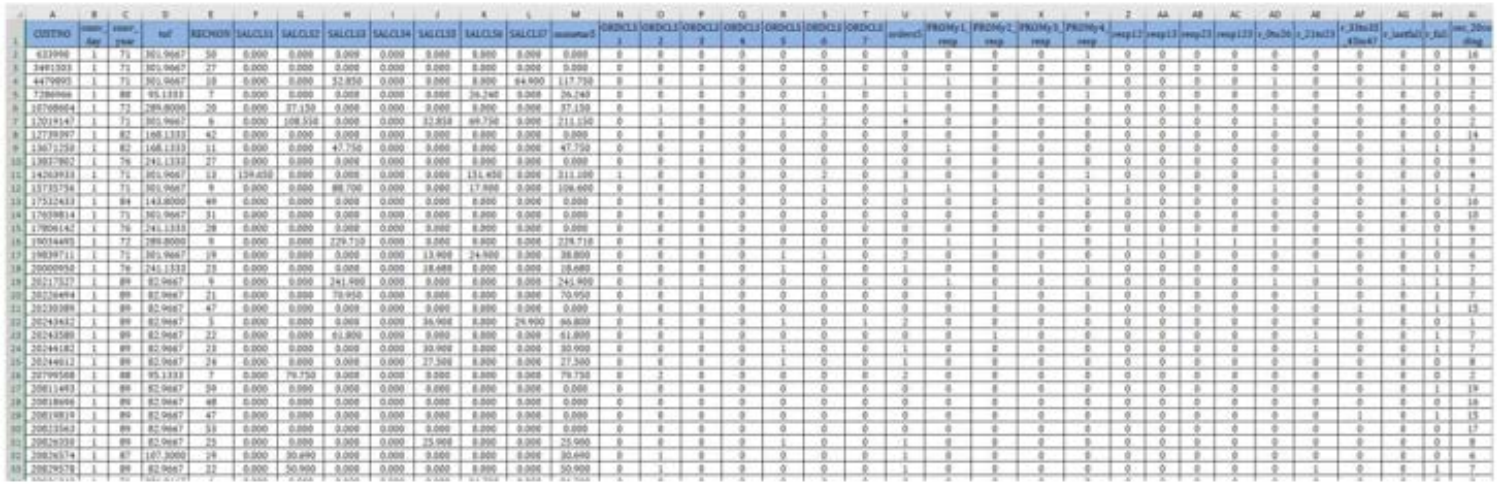
Prilog 2 – isječak iz baze podataka „*Customer transaction dataset*" o prodaji opreme za biciklizam

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	transaction_id	product_id	customer_id	transaction_date	online_order	order_status	brand	product_line	product_class	product_size	list_price	gender	age	job_industry_category	wealth_segment	owns_car	region	R	F	M
2	1	2	2950	25-02-2017	FALSE	Approved	Solex	Standard	medium	medium	71.49	Male	65	Financial Services	Mass Customer	Yes	VIC	4	3	1963.15
3	2	3	3120	21-05-2017	TRUE	Approved	Trek Bicycles	Standard	medium	large	2091.47	Female	41	Health	Mass Customer	Yes	NSW	5	7	7351.7
4	3	37	402	16-10-2017	FALSE	Approved	OHM Cycles	Standard	low	medium	1793.43	Male	43	Retail	Affluent Customer	No	NSW	5	8	8148.22
5	4	88	3135	31-08-2017	FALSE	Approved	Norco Bicycles	Standard	medium	medium	1198.46	Male	58	Financial Services	Mass Customer	No	NSW	4	7	7539.81
6	5	78	787	01-10-2017	TRUE	Approved	Giant Bicycles	Standard	medium	large	1765.3	Female	23	Retail	Affluent Customer	Yes	NSW	4	10	11292.54
7	6	25	2339	08-03-2017	TRUE	Approved	Giant Bicycles	Road	medium	medium	1538.99	Male	61	Property	Affluent Customer	Yes	NSW	5	9	10497.39
8	7	22	1542	21-04-2017	TRUE	Approved	WeareA2B	Standard	medium	medium	66.34	Male	42	Health	Mass Customer	Yes	NSW	4	5	5826.49
9	8	15	2459	15-07-2017	FALSE	Approved	WeareA2B	Standard	medium	medium	1292.84	Female	34	Manufacturing	High Net Worth	No	QLD	5	11	11877.19
10	9	67	1305	10-08-2017	FALSE	Approved	Solex	Standard	medium	large	1971.23	Male	58	Financial Services	High Net Worth	Yes	QLD	4	5	4962.79
11	10	12	3262	30-08-2017	TRUE	Approved	WeareA2B	Standard	medium	medium	1231.15	Male	61	Entertainment	Mass Customer	No	QLD	5	10	9899.64
12	11	1	1986	17-01-2017	FALSE	Approved	Trek Bicycles	Mountain	low	medium	574.84	Female	38	Retail	Mass Customer	No	NSW	5	6	6472.25
13	12	61	2783	05-01-2017	TRUE	Approved	OHM Cycles	Standard	low	medium	71.16	Male	42	Retail	Affluent Customer	Yes	NSW	4	11	9019.44
14	13	35	1243	26-02-2017	TRUE	Approved	Trek Bicycles	Standard	low	medium	1957.51	Male	67	Retail	Mass Customer	No	NSW	5	10	11620.08
15	14	16	2717	10-09-2017	FALSE	Approved	Norco Bicycles	Standard	high	small	1661.92	Male	39	Health	Mass Customer	No	NSW	5	10	10140.5
16	15	12	247	11-06-2017	FALSE	Approved	Giant Bicycles	Standard	medium	large	1765.3	Female	37	Financial Services	High Net Worth	Yes	NSW	4	7	9779.37
17	16	3	2961	10-10-2017	FALSE	Approved	Trek Bicycles	Standard	medium	large	2091.47	Male	45	Health	Mass Customer	Yes	VIC	5	8	8983.52
18	17	79	2426	03-04-2017	FALSE	Approved	Norco Bicycles	Standard	medium	medium	1555.58	U	120	IT	Affluent Customer	Yes	NSW	5	5	6399.89
19	18	33	1642	02-06-2017	FALSE	Approved	Giant Bicycles	Standard	medium	small	1311.44	Male	43	Financial Services	Mass Customer	No	NSW	3	5	6542.54
20	19	54	2268	06-04-2017	TRUE	Approved	WeareA2B	Standard	medium	medium	1292.84	Female	55	Health	Mass Customer	No	VIC	5	8	9258.43
21	20	25	3092	28-01-2017	TRUE	Approved	Giant Bicycles	Road	medium	medium	1538.99	Male	63	Financial Services	Mass Customer	Yes	NSW	5	7	7261.75
22	21	27	1582	05-10-2017	FALSE	Approved	Trek Bicycles	Standard	medium	medium	499.53	Male	42	Property	Mass Customer	No	NSW	5	5	5262.89
23	22	37	696	29-06-2017	TRUE	Approved	OHM Cycles	Standard	low	medium	1793.43	Male	47	Financial Services	Mass Customer	No	QLD	3	3	3268.42
24	23	37	2661	08-04-2017	TRUE	Approved	OHM Cycles	Standard	low	medium	1793.43	Female	43	Manufacturing	Mass Customer	Yes	NSW	4	8	10016.79
25	24	82	515	16-10-2017	FALSE	Approved	Giant Bicycles	Road	medium	medium	1538.99	Female	31	IT	High Net Worth	No	NSW	5	9	12264.8
26	25	89	2822	11-06-2017	FALSE	Approved	WeareA2B	Touring	medium	large	1262.99	Female	39	n/a	Mass Customer	No	VIC	5	9	9764.17
27	26	64	2696	10-01-2017	FALSE	Approved	Trek Bicycles	Standard	medium	large	1469.44	Female	44	n/a	Affluent Customer	Yes	NSW	4	5	7674.39
28	27	64	2666	11-04-2017	TRUE	Approved	Trek Bicycles	Standard	medium	large	1469.44	Male	49	Health	Mass Customer	No	NSW	2	4	4534.9
29	28	19	76	23-12-2017	FALSE	Approved	Trek Bicycles	Mountain	low	medium	574.84	Female	42	Retail	High Net Worth	Yes	QLD	5	8	7425.84

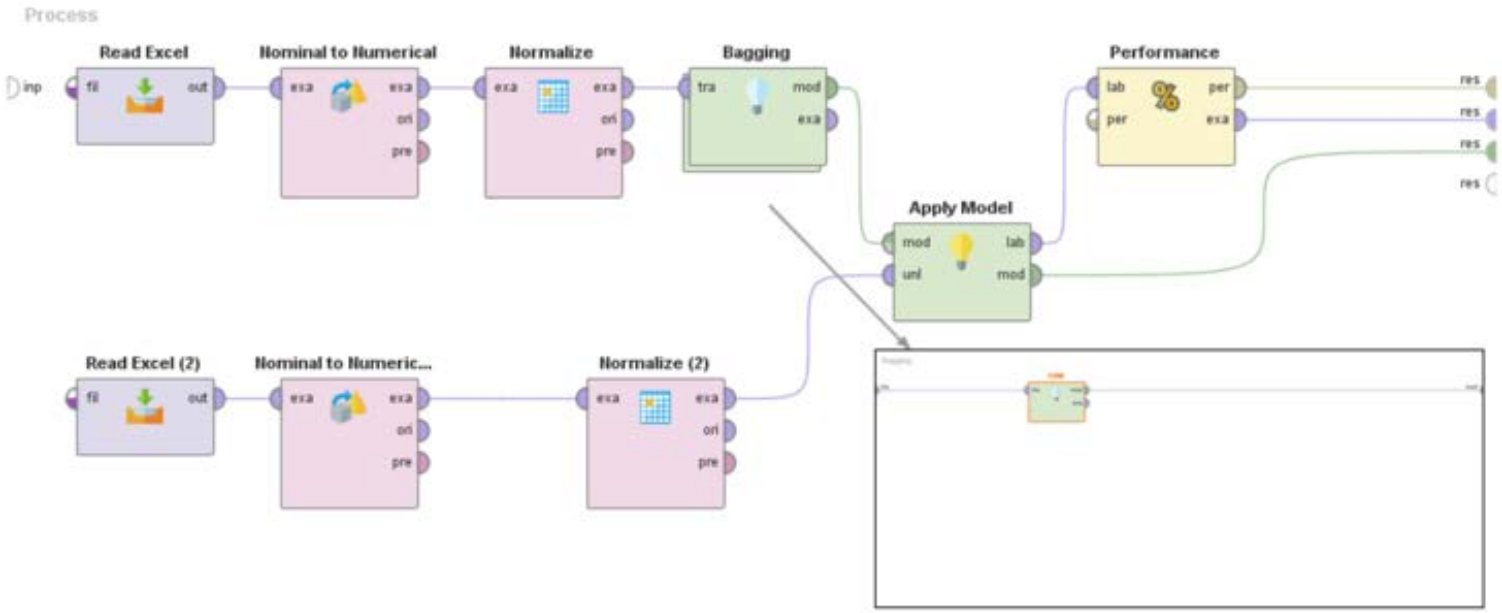
Prilog 3 – isječak iz baze podataka o posetama *web* sajtu e-trgovine upućene sa sponzorisanih objava na društvenim mrežama kompanije *Sport Vision* (o ostvarenoj prodaji proizvoda preko web sajta kompanije Sport Vision)



Prilog 4 – isječak iz baze podataka „*Direct Marketing Educational Foundation 3* (DMEF3)" o kataloškoj prodaji

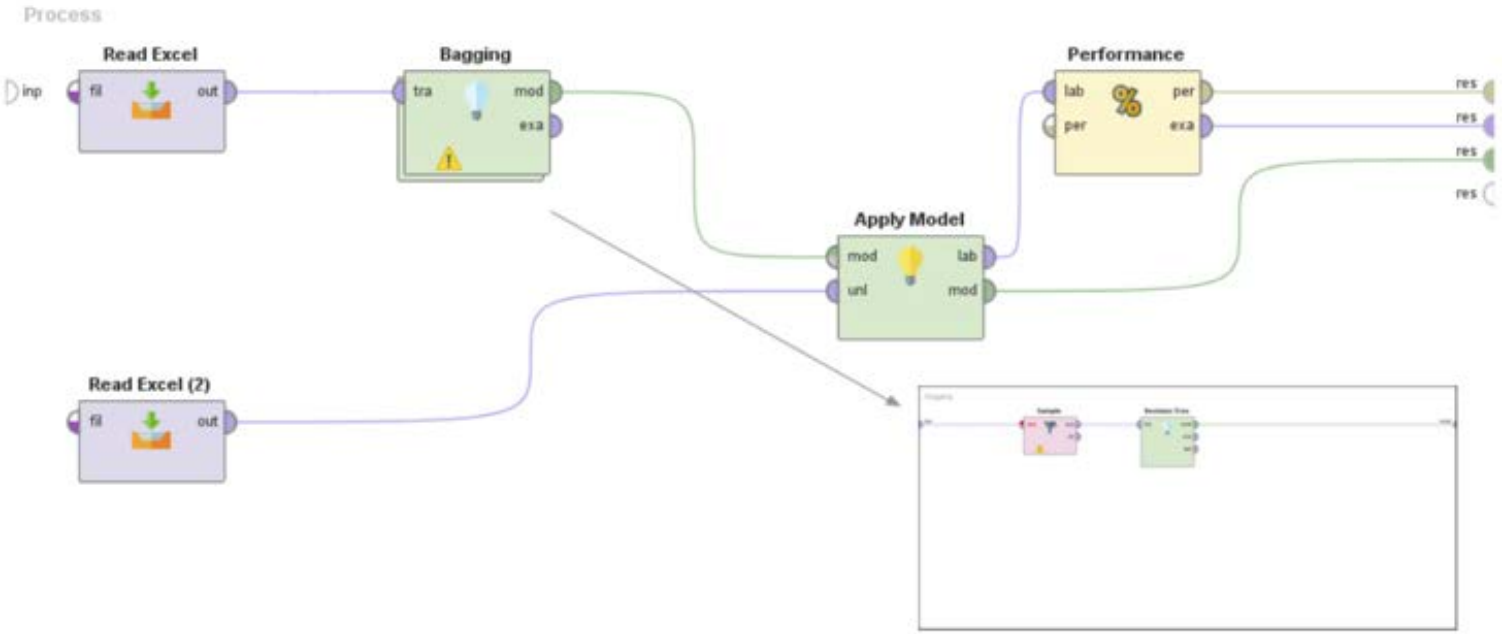


Prilog 5 – *Rapid Miner* proces za pretprocesiranje podataka za predikciju kod predloženog modela RFM segmentacije



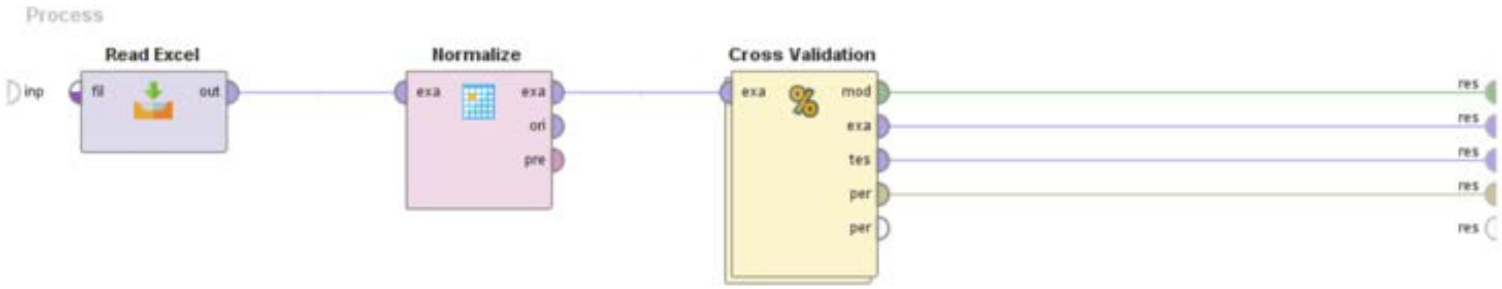
Napomena: U gornji *Read Excel* operator učitavaju se podaci za obuku, a u donji podaci za predikciju.

Prilog 6 – *Rapid Miner* proces predikcije na pretprocesiranim podacima kod predloženog modela RFM segmentacije



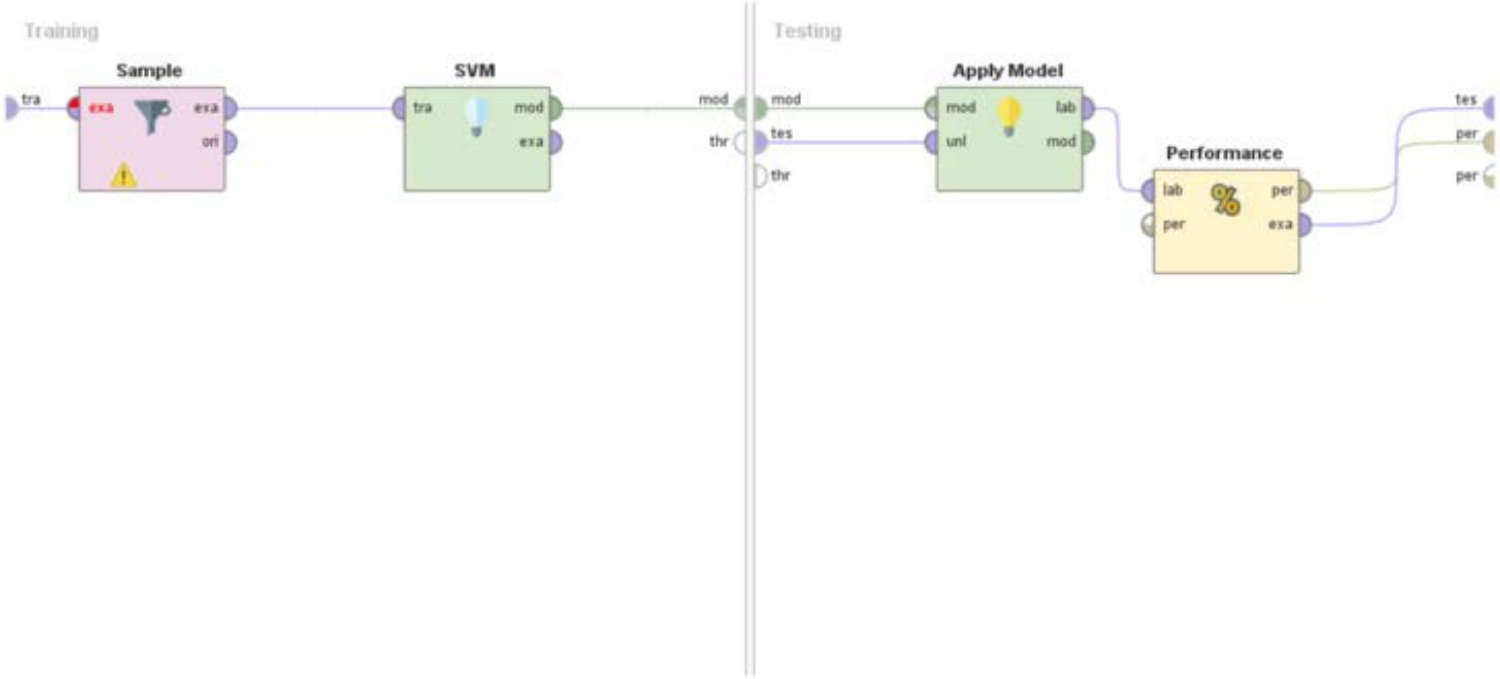
Napomena: U gornji *Read Excel* operator učitavaju se pretprocesirani podaci za obuku, a u donji pretprocesirani podaci za predikciju.

Prilog 7 – *Rapid Miner* proces za obučavanje balansiranog SVM pretprocesora kod predloženog modela odgovora kupca (1/2)

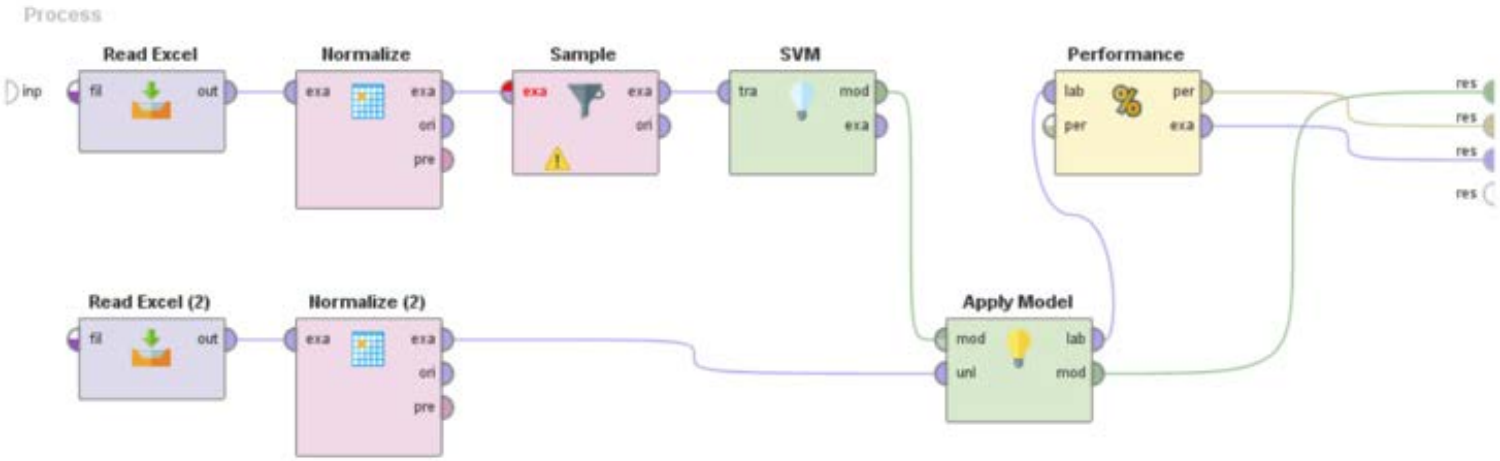


Napomena: U *Read Excel* operator učitavaju se podaci za obuku.

Prilog 7 – *Rapid Miner* proces za obučavanje balansiranog SVM pretprocesora kod predloženog modela odgovora kupca (2/2)

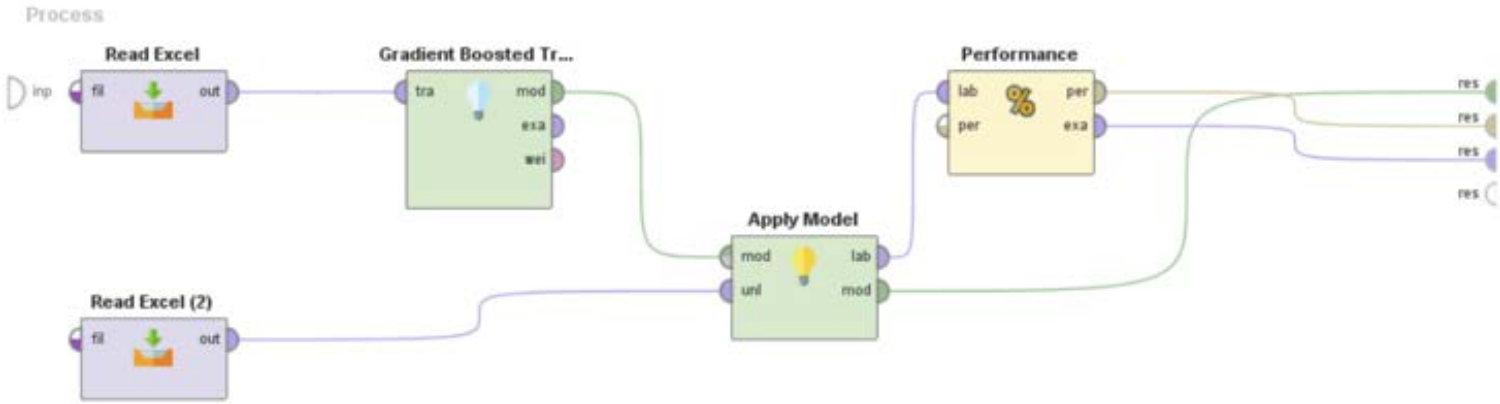


Prilog 8 - *Rapid Miner* proces za pretprocesiranje podataka za predikciju kod modela odgovora kupca



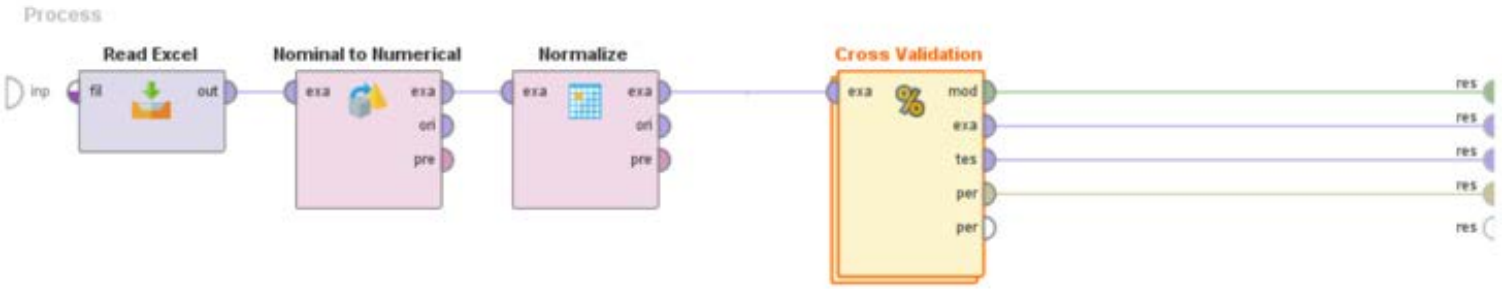
Napomena: U gornji *Read Excel* operator učitavaju se podaci za obuku, a u donji podaci za predikciju.

Prilog 9 - *Rapid Miner* proces za predikciju odgovora kupca na kampanju



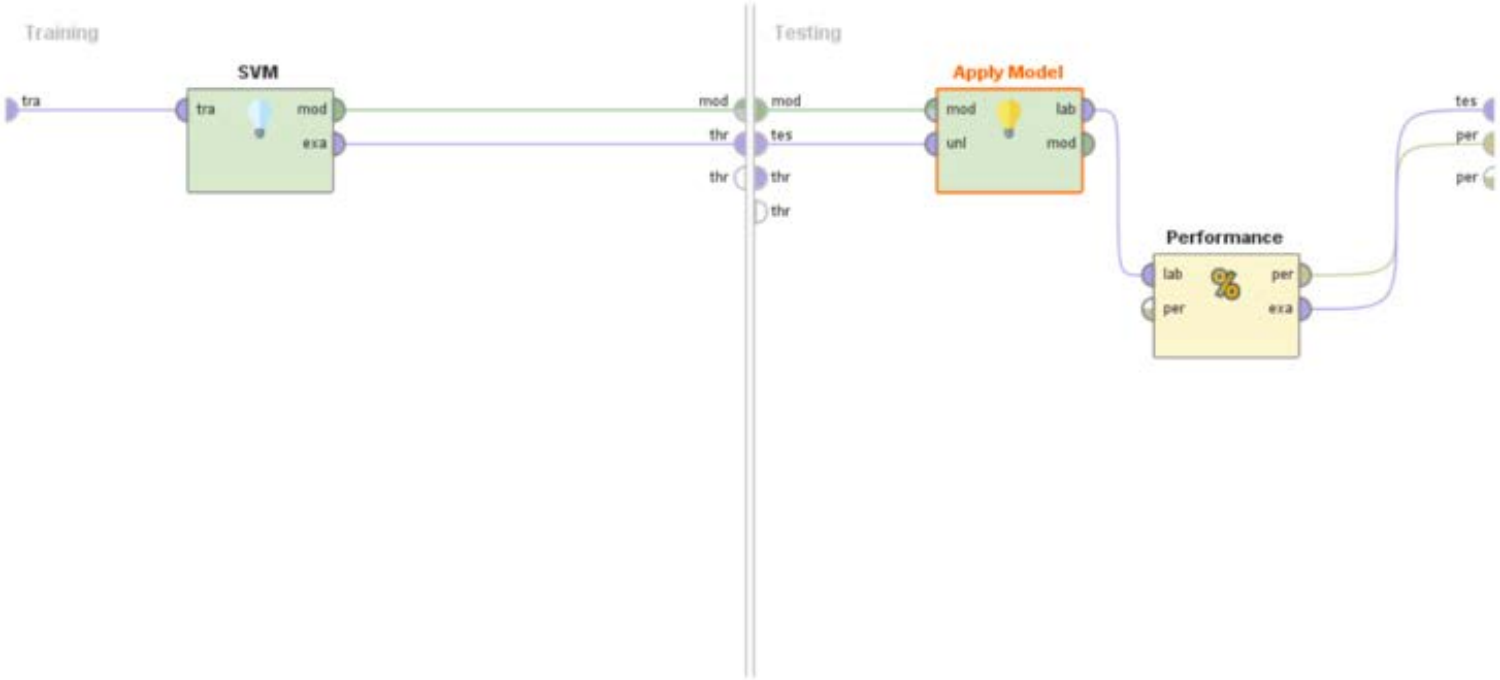
Napomena: U gornji *Read Excel* operator učitavaju se pretprocesirani podaci za obuku, a u donji pretprocesirani podaci za predikciju.

Prilog 10 - *Rapid Miner* proces za obučavanje modela za targetiranje na osnovu profitabilnosti (1/2)



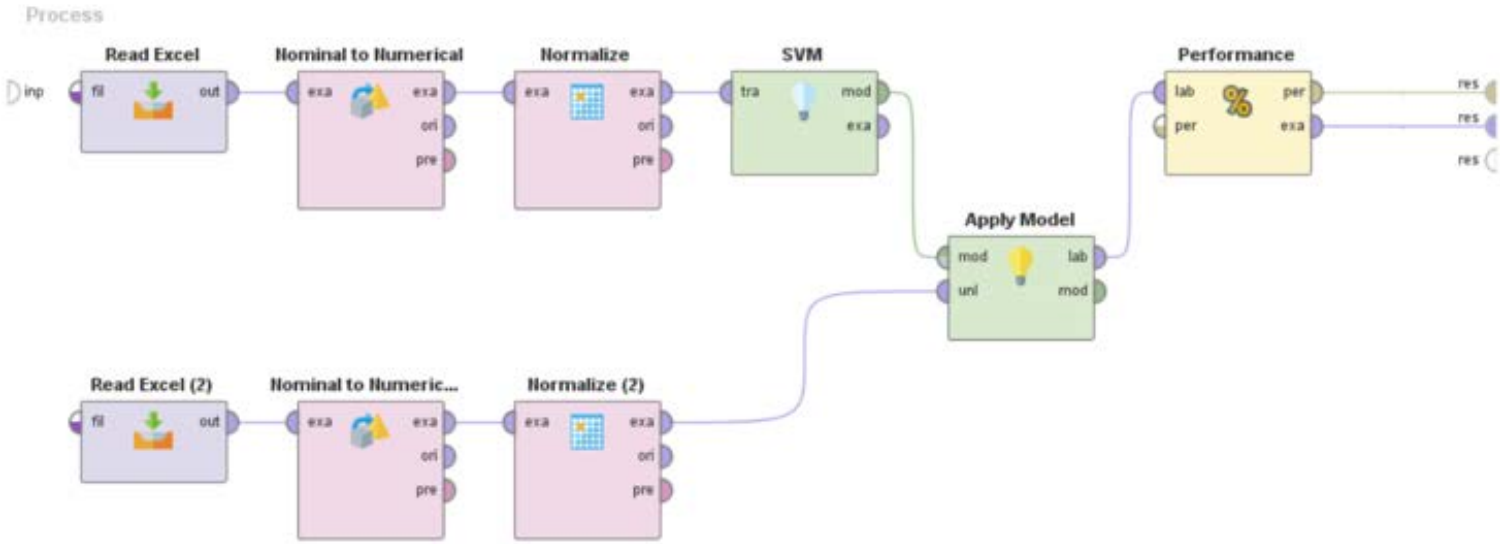
Napomena: U *Read Excel* operator učitavaju se podaci za obuku.

Prilog 10 – *Rapid Miner* proces za obučavanje modela za targetiranje na osnovu profitabilnosti (2/2)



Napomena: Korišćen je *„epsilon SVR"* tip SVM operatora.

Prilog 11 – *Rapid Miner* proces za predikciju profitabilnosti kupca



Napomena: U gornji *Read Excel* operator učitavaju se podaci za obuku, a u donji podaci za predikciju.

LITERATURA

Abdi, F., & Abolmakarem, S. (2019). Customer Behavior Mining Framework (CBMF) using clustering and classification techniques. *Journal of Industrial Engineering International, 15*. https://doi.org/10.1007/s40092-018-0285-3

AdAge. (2003). *History: 1970s*. https://adage.com/article/adage-encyclopedia/history-1970s/98703

AdAge. (2013). *Telemarketing: Overview*. https://adage.com/article/adage-encyclopedia/telemarketing-overview/98900

Aggarwal, A. G., & Delhi, N. (2020). Customer Segmentation Using Fuzzy-AHP and RFM Model. *8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 77–80.

Aghdaie, M. H. (2016). *A New Perspective on RFM Analysis*. *October*, 1–4. https://doi.org/10.4018/978-1-5225-0997-4.ch001

Air Marketing Group. (2018). *It wasn't always a piece of cake – The history of telemarketing*. https://www.air-marketing.co.uk/2018/11/23/it-wasnt-always-a-piece-of-cake-the-history-of-telemarketing/

Ait Daoud, R., Bouikhalene, B., Amine, A., & Lbibb, R. (2015). Combining RFM Model and Clustering Techniques for Customer Value Analysis of a Company selling online. *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*.

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. *European Conference on Machine Learning*, 39–50.

Aliabadi, A. N., & Berenji, H. (2013). Hybrid model of customer response modeling through combination of neural networks and data pre-processing. *IEEE International Conference on Fuzzy Systems*. https://doi.org/10.1109/FUZZ-IEEE.2013.6622378

AlizadehZoeram, Ali, & KarimiMazidi, Ahmadreza. (2018). A New Approach for Customer Clustering by Integrating the LRFM Model and Fuzzy Inference System. *Iranian Journal of Management Studies*, *11*(2), 351–378. https://doi.org/10.22059/ijms.2018.242528.672839

American Association of Advertising Agencies. (2022). *The Golden Age of Advertising*. https://www.aaaa.org/timeline-event/golden-age-advertising/?cn-reloaded=1

American Express. (n.d.). *Our History*. https://about.americanexpress.com/our-history/

Ammerman, W. (2019). *The Invisible Brand - Marketing in the Age of Automation, Big Data and Machine Learning*. McGraw Hill Professional.

Ansari, A., & Riasi, A. (2016). Taxonomy of Marketing Strategies Using Bank Customers' Clustering. *International Journal of Business and Management*, *11*(7), 106. https://doi.org/10.5539/ijbm.v11n7p106

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Asare-Frempong, J., & Jayabalan, M. (2017a). Predicting customer response to bank direct telemarketing campaign. *2017 International Conference on Engineering Technology and Technopreneurship, ICE2T 2017*, *2017-Janua*(September), 1–4. https://doi.org/10.1109/ICE2T.2017.8215961

Asare-Frempong, J., & Jayabalan, M. (2017b). Predicting customer response to bank direct telemarketing campaign. *2017 International Conference on Engineering Technology and Technopreneurship, ICE2T 2017*, *2017*(September), 1–4. https://doi.org/10.1109/ICE2T.2017.8215961

Athanassopoulos, A. D. (2000). Customer Satisfaction Cues To Support Market Segmentation and Explain Switching Behavior. *Journal of Business Research*, *2963*(47), 191–207.

Au, T., Chin, M. L. I., & Ma, G. (2010). Mining rare events data by sampling and boosting:

A case study. In S. K. Prasad, H. M. Sahni, S. Jaiswal, & M. P. B. Thipakorn (Eds.), *Communications in Computer and Information Science* (ICISTM 201, Vol. 54, pp. 373–379). Springer. https://doi.org/10.1007/978-3-642-12035-0_38

Bach, M. P., Jaklič, J., & Vugec, D. S. (2018). Understanding impact of business intelligence to organizational performance using cluster analysis: Does culture matter? *International Journal of Information Systems and Project Management*, *6*(3), 63–86. https://doi.org/10.12821/ijispm060304

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Baesens, Bart, Viaene, S., Poel, D. Van Den, Vanthienen, J., & Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, *138*, 191–211.

Ballestar, M. T., Grau-Carles, P., & Sainz, J. (2019). Predicting customer quality in e-commerce social networks: a machine learning approach. *Review of Managerial Science*, *13*(3), 589–603. https://doi.org/10.1007/s11846-018-0316-x

Barakat, N., & Bradley, A. P. (2010). Rule extraction from support vector machines: A review. *Neurocomputing*, *74*(1–3), 178–190. https://doi.org/10.1016/j.neucom.2010.02.016

Barwise, P., & Farley, J. U. (2005). The state of interactive marketing in seven countries: Interactive marketing comes of age. *Journal of Interactive Marketing*, *19*(3), 67–80. https://doi.org/10.1002/dir.20044

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support Vector Regression. *Neural Information Processing - Letters and Reviews*, *11*(10), 203–224.

Basye, A. (2008). *Opportunities in Direct Marketing Careers* (Revised ed). McGraw-Hill.

Baumann, A., Haupt, J., Gebert, F., & Lessmann, S. (2019). The Price of Privacy: An

Evaluation of the Economic Value of Collecting Clickstream Data. *Business and Information Systems Engineering*, *61*(4), 413–431. https://doi.org/10.1007/s12599-018-0528-2

Behera, R. K., Gunasekaran, A., Gupta, S., Kamboj, S., & Bala, P. K. (2020). Personalized digital marketing recommender engine. *Journal of Retailing and Consumer Services*, *53*(101799), 1–24. https://doi.org/10.1016/j.jretconser.2019.03.026

Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*, *12*(1), 17–30. https://doi.org/10.1002/(SICI)1520-6653(199824)12:1<17::AID-DIR3>3.0.CO;2-K

Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, *39*(227), 357–365.

Bermejo, P., Gámez, J. A., & Puerta, J. M. (2011). Improving the performance of Naive Bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets. *Expert Systems with Applications*, *38*(3), 2072–2080. https://doi.org/10.1016/j.eswa.2010.07.146

Berrar, D. (2018). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1–3*(January 2018), 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Berry, M. J., & Linoff, G. S. (2004). Data mining techniques: for marketing, sales, and customer relationship management. In *Enero* (Second Edi). John Wiley & Sons. https://www.bancomundial.org/es/news/press-release/2018/01/09/global-economy-to-edge-up-to-3-1-percent-in-2018-but-future-potential-growth-a-concern

Bhattacharya, C. . B. (1994). When Customers Are Members : Customer Retention in Paid Membership Contexts. *Journal of the Academy of Marketing Science*, *26*(1), 31–44.

Binu, D., & Rajakumar, B. R. (2021). *Artificial Intelligence in Data Mining: Theories and Applications*. Elsevier.

Birant, D. (2011). Data mining using RFM analysis. In *Knowledge-oriented applications in data mining* (pp. 91–108). InTechOpen.

Bird, D. (1989). *Commonsense Direct Marketing* (2nd editio). Kogan Page.

Blattberg, R. C. (1987). Research opportunities in direct marketing. *Journal of Direct Marketing*, *1*(1), 7–14.

Blattberg, R. C., Kim, B.-D., & Neslin, S. A. (2008). Database Marketing. In *Direct Marketing in Practice*. Springer. https://doi.org/10.1016/b978-0-7506-2428-2.50008-9

Boone, D. S., & Roehm, M. (2002). Evaluating the Appropriateness of Market Segmentation Solutions Using Artificial Neural Networks and the Membership Clustering Criterion. *Marketing Letters*, *13*(4), 317–333. https://doi.org/10.1023/A:1020321132568

Bose, I., & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, *195*(1), 1–16. https://doi.org/10.1016/j.ejor.2008.04.006

Bramer, M. (2020). *Principles of Data Mining* (I. Mackie (ed.); fourth edi, Vol. 30, Issue 7). Springer-Verlag London Ltd. https://doi.org/10.1007/978-1-4471-7493-6

Breiman, L. (1984). *Classification and regression trees*. Wadsworth International Group.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Breiman, Leo. (1996). Bagging Predictors. *Machine Learning*, *24*(421), 123–140. https://doi.org/10.1007/BF00058655

Breur, T. (2011). Data analysis across various media: Data fusion, direct marketing,

clickstream data and social media. *Journal of Direct, Data and Digital Marketing Practice*, *13*(2), 95–105. https://doi.org/10.1057/dddmp.2011.32

Brewer, J., & Hunter, A. (1989). *Multimethod research: A synthesis of styles*. Sage Publications, Inc.

Brewer, J., & Hunter, A. (2006). *Foundations of Multimethod Research: Synthesizing Styles*. SAGE Publications.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis : partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, *164*(1), 252–268. https://doi.org/10.1016/j.ejor.2003.12.010

Bucklin, R. E., Lehmann, D. R., & Little, J. D. C. (1998). From decision support to decision automation: A 2020 vision. *Marketing Letters*, *9*(3), 235–246. https://doi.org/10.1023/A:1008047504898

Bult, J. R., & Wansbeek, T. (1995). Optimal Selection for Direct Mail. *Marketing Science*, *14*(4), 378–394. https://doi.org/10.1287/mksc.14.4.378

Bult, R. J., Van Der Scheer, H., & Wansbeek, T. (1997). Interaction between target and mailing characteristics in direct marketing, with an application to health care fund raising. *International Journal of Research in Marketing*, *14*(4), 301–308. https://doi.org/10.1016/s0167-8116(97)00012-8

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3 PART 1), 4626–4636. https://doi.org/10.1016/j.eswa.2008.05.027

Carosella, G., Findling, S., Fitzgerald, S., Jimenez, D.-Z., Little, G., O'Brien, A., Parker, R., Pucciarelli, J. C., & Yorifuji, Y. (2021). *IDC FutureScape: Worldwide Digital Transformation 2022 Predictions*. International Data Corporation. https://www.idc.com/getdoc.jsp?containerId=US47115521

Carr, S. (2021). *How Many Ads Do We See A Day In 2021?* PPC Protect.

https://ppcprotect.com/blog/strategy/how-many-ads-do-we-see-a-day/

Carter, R. (2021). *The Ultimate List of Big Data Statistics for 2022*. FindStack. https://findstack.com/big-data-statistics/

Case, A. N. (2015). 'The solid gold mailbox'': direct mail and the changing nature of buying and selling in the postwar United States.' *History of Retailing and Consumption*, *1*(1), 28–46. https://doi.org/10.1080/2373518x.2015.1012863

Centro de Documentación Publicitaria. (2020). *Lester Wunderman*. 23.08.2020.

Chaffey, D. (2012). *Marketing campaign response rates*. Smart Insights. https://www.smartinsights.com/managing-digital-marketing/planning-budgeting/marketing-campaign-response-rates/

Chagas, B. N. R., Viana, J., Reinhold, O., Lobato, F. M. F., Jr, A. F. L. J., & Alt, R. (2020). *A literature review of the current applications of machine learning and their practical implications. 1*, 1–15. https://doi.org/10.3233/WEB-200429

Chan, C.-C. H. (2005). Online auction customer segmentation using a neural network model. *International Journal of Applied Science and Engineering*, *3*(2), 101–109.

Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert Systems with Applications*, *34*(4), 2754–2762. https://doi.org/10.1016/j.eswa.2007.05.043

Chang, H. C., & Tsai, H. P. (2011). Group RFM analysis as a novel framework to discover better customer consumption behavior. *Expert Systems with Applications*, *38*(12), 14499–14513. https://doi.org/10.1016/j.eswa.2011.05.034

Chaudhuri, N., Gupta, G., Vamsi, V., & Bose, I. (2021). On the platform but will they buy? Predicting customers' purchase behavior using deep learning. *Decision Support Systems*, *149*(December 2020), 113622. https://doi.org/10.1016/j.dss.2021.113622

Chen, L. da, Gillenson, M. L., & Sherrell, D. L. (2002). Enticing online consumers: An extended technology acceptance perspective. *Information and Management*, *39*(8),

705–719. https://doi.org/10.1016/S0378-7206(01)00127-6

Chen, Q., Zhang, M., & Zhao, X. (2017). Analysing customer behaviour in mobile app usage. *Industrial Management and Data Systems*, *117*(2), 425–438. https://doi.org/10.1108/IMDS-04-2016-0141

Chen, W. C., Hsu, C. C., & Hsu, J. N. (2011). Optimal selection of potential customer range through the union sequential pattern by using a response model. *Expert Systems with Applications*, *38*(6), 7451–7461. https://doi.org/10.1016/j.eswa.2010.12.078

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, *36*(3 PART 1), 4176–4184. https://doi.org/10.1016/j.eswa.2008.04.003

Cheung, K. W., Kwok, J. T., Law, M. H., & Tsui, K. C. (2003). Mining customer product ratings for personalized marketing. *Decision Support Systems*, *35*(2), 231–243. https://doi.org/10.1016/S0167-9236(02)00108-2

Christmann, A. (2004). An approach to model complex high?dimensional insurance data. *Allgemeines Statistisches Archiv*, *88*(4), 375–396. https://doi.org/10.1007/s101820400178

Chun, Y. H. (2012). Monte Carlo analysis of estimation methods for the prediction of customer response patterns in direct marketing. *European Journal of Operational Research*, *217*(3), 673–678. https://doi.org/10.1016/j.ejor.2011.10.008

Colgate, M. R., & Danaher, P. J. (2000). Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution. *Journal of the Academy of Marketing Science*, *28*(3), 375–387. https://doi.org/10.1177/0092070300283006

Constantinides, E., & Fountain, S. J. (2008). Web 2.0: Conceptual foundations and marketing issues. *Journal of Direct, Data and Digital Marketing Practice*, *9*(3), 231–244. https://doi.org/10.1057/palgrave.dddmp.4350098

Coussement, K., Bossche, F. A. M. Van Den, & Bock, K. W. De. (2014). Data accuracy ' s impact on segmentation performance : Benchmarking RFM analysis , logistic regression , and decision trees. *Journal of Business Research*, *67*(1), 2751–2758. https://doi.org/10.1016/j.jbusres.2012.09.024

Coussement, K., Harrigan, P., & Benoit, D. F. (2015). Improving direct mail targeting through customer response modeling. *Expert Systems with Applications*, *42*(22), 8403–8412. https://doi.org/10.1016/j.eswa.2015.06.054

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, *34*(1), 313–327. https://doi.org/10.1016/j.eswa.2006.09.038

Craven, M. W., & Shavlik, J. W. (1995). Extracting Thee-Structured Representations of Thained Networks. *Advances in Neural Information Processing Systems*, *8*, 24–30.

Cresswell, J. W. (1999). Mixed-Method Research: Introduction and Application. In *Handbook of Educational Policy* (pp. 455–472). Academic press.

Csikósová, A., Antošová, M., & Čulková, K. (2014). Strategy in Direct and Interactive Marketing and Integrated Marketing Communications. *Procedia - Social and Behavioral Sciences*, *116*, 1615–1619. https://doi.org/10.1016/j.sbspro.2014.01.444

Cuadros, A. J., & Domínguez, V. E. (2014). Customer segmentation model based on value generation for marketing strategies formulation. *Estudios Gerenciales*, *30*(130), 25–30. https://doi.org/10.1016/j.estger.2014.02.005

Cui, D., & Curry, D. (2005). Prediction in marketing using the support vector machine. *Marketing Science*, *24*(4), 595–615. https://doi.org/10.1287/mksc.1050.0123

Cui, G., Wong, M. L., & Lui, H. K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, *52*(4), 597–612. https://doi.org/10.1287/mnsc.1060.0514

Cui, G., Wong, M. L., & Wan, X. (2015). Targeting High Value Customers While Under Resource Constraint: Partial Order Constrained Optimization with Genetic Algorithm. *Journal of Interactive Marketing*, *29*, 27–37. https://doi.org/10.1016/j.intmar.2014.09.001

D'Haen, J., Van Den Poel, D., & Thorleuchter, D. (2013). Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. *Expert Systems with Applications*, *40*(6), 2007–2012. https://doi.org/10.1016/j.eswa.2012.10.023

Danaher, P. J., & Rossiter, J. R. (2011). Comparing perceptions of marketing communication channels. *European Journal of Marketing*, *45*(1), 6–42. https://doi.org/10.1108/03090561111095586

Daneshmandi, M., & Ahmadzadeh, M. (2013). A Hybrid Data Mining Model to Improve Customer Response Modeling in Direct Marketing. *Indian Journal of Computer Science and Engineering (IJCSE)*, *3*(6), 844–855. http://www.doaj.org/doaj?func=fulltext&aId=1216062

Dave, V. S., & Dutta, K. (2014). Neural network based models for software effort estimation: A review. *Artificial Intelligence Review*, *42*(2), 295–307. https://doi.org/10.1007/s10462-012-9339-x

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Deichmann, J., Eshghi, A., Haughton, D., Sayek, S., & Teebagy, N. (2002). Application of multiple adaptive regression splines (mars) in direct response modeling. *Journal of Interactive Marketing*, *16*(4), 15–27. https://doi.org/10.1002/dir.10040

Deighton, J., & Kornfeld, L. (2009). Interactivity's Unanticipated Consequences for Marketers and Marketing. *Journal of Interactive Marketing*, *23*(1), 4–10. https://doi.org/10.1016/j.intmar.2008.10.001

Deloitte Consumer Review. (2016). *CX Marks the Spot: Rethinking Customer Experience to Win.*

Delua, J. (2021). *Supervised vs. Unsupervised Learning: What's the Difference?* IBM. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

DeSarbo, W. S., & Ramaswamy, V. (1994). CRISP: Customer response based iterative segmentation procedures for response modeling in direct marketing. *Journal of Direct Marketing*, *8*(3), 7–20. https://doi.org/10.1002/dir.4000080304

Dibb, S., & Simkin, L. (1996). *The market segmentation workbook: target marketing for marketing managers.* Routledge.

Diederich, J. (2008). Rule Extraction from Support Vector Machines: An Introduction. In J. Diederich (Ed.), *Rule Extraction from Support Vector Machines. Studies in Computational Intelligence, vol 80.* (pp. 3–31). Springer.

Dietterich, T. G. (2002). Ensemble Learning. In *The handbook of brain theory and neural networks*.

Dimensional Research. (2013). *Customer Service and Business Results: A Survey of Customer Service From Mid-Size Companies.* https://d16cvnquvjw7pr.cloudfront.net/resources/whitepapers/Zendesk_WP_Customer_Service_and_Business_Results.pdf

Direct Marketing Association. (2009). *Future of Direct Marketing.*

Djurisic, V., Kascelan, L., Rogic, S., & Melovic, B. (2020). Bank CRM Optimization Using Predictive Classification Based on the Support Vector Machine Method. *Applied Artificial Intelligence*, *00*(00), 1–15. https://doi.org/10.1080/08839514.2020.1790248

Dobilas, S. (2020). *Support Vector Regression (SVR) — One of the Most Flexible Yet Robust Prediction Algorithms.* Towards Data Science. https://towardsdatascience.com/support-vector-regression-svr-one-of-the-most-flexible-yet-robust-prediction-algorithms-4d25fbdaca60

Doğan, O., Ayçin, E., & Bulut, Z. A. (2018). Customer Segmentation by Using RFM Model and Clustering Methods: A Case Study in Retail Industry. *International Journal of Contemporary Economics and Administrative Sciences*, *8*(1), 1–19. www.ijceas.com

Domenico, D., & Nunan, D. (2013). Market research & the ethics of big data. *International Journal of Market Research*, *55*(4), 505–520.

Donio, J., Massari, P., & Passiante, G. (2006). Customer satisfaction and loyalty in a digital environment: An empirical test. *Journal of Consumer Marketing*, *23*(7), 445–457. https://doi.org/10.1108/07363760610712993

Donkers, B., Verhoef, P. C., & de Jong, M. G. (2007). Modeling CLV: A test of competing models in the insurance industry. *Quantitative Marketing and Economics*, *5*(2), 163–190. https://doi.org/10.1007/s11129-006-9016-y

Dou, X. (2020). Online Purchase Behavior Prediction and Analysis Using Ensemble Learning. *IEEE 5th International Conference on Cloud Computing and Big Data Analytics*, 532–536.

Drozdenko, R., & Drake, P. (2002). *Optimal database marketing: Strategy, development, and data mining*. Sage Publications.

Dubinets, L. (n.d.). *Top Down Decision Tree Inducers*. https://sites.math.washington.edu/~morrow/336_15/papers/lev.pdf

Duffett, R. G. (2015). Facebook advertising's influence on intention-to-purchase and purchase amongst millennials. *Internet Research*, *25*(4), 498–526. https://doi.org/10.1108/IntR-01-2014-0020

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*(3), 32–57. https://doi.org/10.1080/01969727308546046

Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling prof table hotel customers: An application of RFM analysis. *Tourism Management Perspectives*, *18*, 153–160. https://doi.org/10.1016/j.tmp.2016.03.001

Dutta, S., Bhattacharya, S., & Guin, K. K. (2015). Data mining in market segmentation: A literature review and suggestions. *Advances in Intelligent Systems and Computing*, *335*, 87–98. https://doi.org/10.1007/978-81-322-2217-0_8

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Elsalamony, H. A. (2014). Bank Direct Marketing Analysis of Data Mining Techniques. *International Journal of Computer Applications*, *85*(7), 12–22.

EMC Education Services. (2015). Data Science and Big Data Analytics: discovering, analyzing, visualizing and presenting data. In *Book*. John Wiley & Sons.

Ertugan, A. (2017). Using statistical reasoning techniques to describe the relationship between Facebook advertising effectiveness and benefits gained. *Procedia Computer Science*, *120*, 132–139. https://doi.org/10.1016/j.procs.2017.11.220

Esmeli, R., Bader-El-Den, M., & Abdullahi, H. (2020). Towards early purchase intention prediction in online session based retailing systems. *Electronic Markets*. https://doi.org/10.1007/s12525-020-00448-x

Esmeli, R., Mohasseb, A., & Bader-El-Den, M. (2020). Analysing the Effect of Platform and Operating System Features on Predicting Consumers' Purchase Intent using Machine Learning Algorithms. *12th International Joint Conference on Knowledge Discovery. SciTePress.*, 333–340. https://doi.org/10.5220/0010176803330340

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). Cluster analysis. In *Quality and Quantity* (5th editio, Vol. 14, Issue 1). John Wiley & Sons. https://doi.org/10.1007/BF00154794

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, *42*(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

Fahy, J., & Jobber, D. (2022). *Foundations of Marketing* (7th editio). McGraw Hill.

Fang, K., Jiang, Y., & Song, M. (2016). Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. *Computers and Industrial*

Engineering, 101, 554–564. https://doi.org/10.1016/j.cie.2016.09.011

Farquad, M. A. H., & Bose, I. (2012). Preprocessing unbalanced data using support vector machine. *Decision Support Systems, 53*(1), 226–233. https://doi.org/10.1016/j.dss.2012.01.016

Fill, C., & Turnbull, S. (2016). *Marketing Communications: Discovery, Creation and Conversations* (7th editio). Pearson Education Ltd.

Finances Online. (2019). *75 Essential Ecommerce Statistics: 2020 Data and Market Share Analysis.* https://financesonline.com/40-essential-ecommerce-statistics-2019-analysis-of-trends-data-and-market-share/

Fix, E., & Hodges Jr., J. L. (1989). Discriminatory analysis-nonparametric discrimination: consistency properties. *International Statistical Review, 57*(3), 238–247.

Fletcher, K. P., & Peters, L. D. (1997). Trust and direct marketing environments: A consumer perspective. *Journal of Marketing Management, 13*(6), 523–539. https://doi.org/10.1080/0267257X.1997.9964491

Forbes. (2017). *Finding Brand Success In The Digital World.* https://www.forbes.com/sites/forbesagencycouncil/2017/08/25/finding-brand-success-in-the-digital-world/#40e4617a626e

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics,* 1189–1232.

Furjan, M. T., Tomičić-Pupek, K., & Pihir, I. (2020). Understanding Digital Transformation Initiatives: Case Studies Analysis. *Business Systems Research, 11*(1), 125–141. https://doi.org/10.2478/bsrj-2020-0009

G. Weiss. (2004). Mining with rarity: A unifying framework. *SIGKDD Explorations, 6*(1), 7–19. http://storm.cis.fordham.edu/gweiss/papers/sigkdd04.pdf

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man and Cybernetics Part C:*

Applications and Reviews, *42*(4), 463–484. https://doi.org/10.1109/TSMCC.2011.2161285

Gandhi, R. (2018). *K-Means Clustering — Introduction to Machine Learning Algorithms*. Towards Data Science. https://towardsdatascience.com/k-means-clustering-introduction-to-machine-learning-algorithms-c96bf0d5d57a

Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the customer base of service providers: An examination of the differences between switchers and stayers. *Journal of Marketing*, *64*(3), 65–87. https://doi.org/10.1509/jmkg.64.3.65.18028

Garcia-Dias, R., Vieira, S., Lopez Pinaya, W. H., & Mechelli, A. (2020). Clustering analysis. In A. Mechelli & S. Vieira (Eds.), *Machine Learning: Methods and Applications to Brain Disorders* (pp. 227–247). Elsevier Inc. https://doi.org/10.1016/B978-0-12-815739-8.00013-4

Gázquez-Abad, J. C., Canniére, M. H. De, & Martínez-López, F. J. (2011). Dynamics of Customer Response to Promotional and Relational Direct Mailings from an Apparel Retailer: The Moderating Role of Relationship Strength. *Journal of Retailing*, *87*(2), 166–181. https://doi.org/10.1016/j.jretai.2011.03.001

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, *70*, 320–328.

Genkin, A., Lewis, D., & Madigan, D. (2007). Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, *49*, 291–304.

Glady, N., Baesens, B., & Croux, C. (2008). Modeling Churn Using Customer Lifetime Value. *Expert Systems with Applications*, *197*(1), 402–411. http://www.ghbook.ir/index.php?name=مجموعه مقالات دومین هم اندیشی سراسری رسانه و تلویزیون سکولاریسم&option=com_dbook&task=readonline&book_id=13629&page=108&chkhashk=03C706812F&Itemid=218&lang=fa&tmpl=component

Godin, S. (1999). *Permission marketing: Turning strangers into friends and friends into customers*. Simon and Schuster.

Goel, A. (2020). *Are you solving ML Clustering problems using K-Means?* Towards Data Science. https://towardsdatascience.com/are-you-solving-ml-clustering-problems-using-k-means-68fb4efa5469

Goli, S., Mahjub, H., Faradmal, J., Mashayekhi, H., & Soltanian, A. R. (2016). Survival Prediction and Feature Selection in Patients with Breast Cancer Using Support Vector Regression. *Computational and Mathematical Methods in Medicine, 2016*. https://doi.org/10.1155/2016/2157984

Gončarovs, P. (2018). Using Data Analytics for Customers Segmentation: Experimental Study at a Financial Institution. *59th International Scientific Conference on Information Technology and Management Science of Riga Technical University, ITMS 2018 - Proceedings*, 1–5. https://doi.org/10.1109/ITMS.2018.8552951

González Abril, L., Velasco Morente, F., Gavilán Ruiz, J. M., & Sánchez-Reyes Fernández, L. M. (2010). The similarity between the square of the coefficient of variation and the Gini index of a general random variable. *Revista de Métodos Cuantitativos Para La Economía y La Empresa*, *10*, 5–18.

Google. (2021). *Bounce Rate*. https://support.google.com/analytics/answer/1009409?hl=en

Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management*, *62*(April 2017), 100–107. https://doi.org/10.1016/j.indmarman.2016.08.003

Gordon, A. D. (1999). *Classification* (2nd editio). Chapman and Hall/CRC.

Gordon, J., & Perrey, J. (2015). The dawn of marketing's new golden age. In *McKinsey Quarterly: Vol. February*.

Govidarajan, M. (2013). A hybrid framework using RBF and SVM for direct marketing. *International Journal of Advanced Computer Science and Applications, 4*(4).

Govindarajan, M. (2016). Ensemble Strategies for Improving Response Model in Direct Marketing. *International Journal of Computer Science and Information Security, 14*(9), 108–114.

Graeber, C. (2013). *Trends 2013: Five trends shaping the next generation of North American digital banking*. Forrester Research.

Guido, G., Prete, I. M., Miraglia, S., & De, I. (2013). Targeting direct marketing campaigns by neural networks. *Journal of Marketing Management, 29*(9–10), 992–1006. https://doi.org/10.1080/0267257X.2010.543018

Gupta, S., & Lehmann, D. R. (2003). Customers As Assets. *Journal of Interactive Marketing, 17*(1), 9–24.

Han, S. H., Lu, S. X., & Leung, S. C. H. (2012). Segmentation of telecom customers based on customer value by decision tree model. *Expert Systems with Applications, 39*(4), 3964–3973. https://doi.org/10.1016/j.eswa.2011.09.034

Hasouneh, A. B. I., & Ayed Alqeed, M. (2010). Measuring the Effectiveness of E-mail Direct Marketing in Building Customer Relationship. *International Journal of Marketing Studies, 2*(1), 48–64. http://web.ebscohost.com.esc-web.lib.cbs.dk/ehost/pdfviewer/pdfviewer?hid=13&sid=fcac4489-2941-4836-8ea5-8de613e94547@sessionmgr10&vid=11

Hauser, W. J., Orr, L., & Daugherty, T. (2011). Customer response models: What data predicts best, hard or soft? *Marketing Management Journal, 21*(1), 1–15.

He, H., & Garcia, E. A. (2009). *Learning from Imbalanced Data. 21*(9), 1263–1284.

He, H., & Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms and Applications* (1st editio). Wiley-IEEE Press.

He, J., Hu, H., Harrison, R., Tai, P. C., Pan, Y., & Member, S. (2006). Rule Generation for Protein Secondary Structure Prediction With Support Vector Machines and

Decision Tree. *IEEE TRANSACTIONS ON NANOBIOSCIENCE*, *5*(1), 46–53.

Heldt, R., Silveira, C. S., & Luce, F. B. (2019). Predicting customer value per product: From RFM to RFM/P. *Journal of Business Research*, *May*. https://doi.org/10.1016/j.jbusres.2019.05.001

Henley Centre. (1995). *Dataculture 2000*.

Hofgesang, P. I., & Kowalczyk, W. (2006). Analysing clickstream data: From anomaly detection to visitor profiling. *Belgian/Netherlands Artificial Intelligence Conference*.

Hong, S. H., & Park, M. J. (2020). Dynamics of marketing automation adoption for organisational marketing process transformation: the case of Microsoft. *International Journal of Electronic Customer Relationship Management*, *12*(3), 205–224.

Horita, Y., & Yamashita, H. (2019). Bayesian network considering the clustering of the customers in a hair salon. *Cogent Business and Management*, *6*(1), 1–15. https://doi.org/10.1080/23311975.2019.1641897

Hosseini, M., & Shabani, M. (2015). New approach to customer segmentation based on changes in customer value. *Journal of Marketing Analytics*, *3*(3), 110–121. https://doi.org/10.1057/jma.2015.10

Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, *37*(7), 5259–5264. https://doi.org/10.1016/j.eswa.2009.12.070

Howard, J., & Bowles, M. (2012). *The two most important algorithms in predictive modeling today*. Strata Conference presentation, February (Vol. 28).

Hsu, F. M., Lu, L. P., & Lin, C. M. (2012). Segmenting customers by transaction data with concept hierarchy. *Expert Systems with Applications*, *39*(6), 6221–6228. https://doi.org/10.1016/j.eswa.2011.12.005

Huang, C., Li, Y., Loy, C. C., & Tang, X. (2020). Deep imbalanced learning for face

recognition and attribute prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(11), 2781–2794. https://doi.org/10.1109/TPAMI.2019.2914680

Huang, J., & Tzeng, G. (2007). Marketing segmentation using support vector clustering. *Expert Systems with Applications, 32*, 313–317. https://doi.org/10.1016/j.eswa.2005.11.028

Huang, Y. M., Hung, C. M., & Jiau, H. C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications, 7*(4), 720–747. https://doi.org/10.1016/j.nonrwa.2005.04.006

Hughes, A. M. (1994). *Strategic database marketing: the masterplan for starting and managing a profitable, customer-based marketing program*. Irwin.

Hughes, A. M. (1996). Boosting Response With RFM. *Marketing Tools*, 4–8.

Huysmans, J., Baesens, B., & Vanthienen, J. (2006). *Using rule extraction to improve the comprehensibility of predictive models* (No. 0612). http://dx.doi.org/10.2139/ssrn.961358

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications, 26*(2), 181–188. https://doi.org/10.1016/S0957-4174(03)00133-7

Iacobucci, D., Petrescu, M., Krishen, A., & Bendixen, M. (2019). The state of marketing analytics in research and practice. In *Journal of Marketing Analytics* (Vol. 7, Issue 3). Palgrave Macmillan UK. https://doi.org/10.1057/s41270-019-00059-2

IBM. (2021). *CRISP-DM Help Overview*. https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview

Jackson, G., & Ahuja, V. (2016). Dawn of the digital age and the evolution of the marketing mix. *Journal of Direct, Data and Digital Marketing Practice, 17*(3), 170–

186. https://doi.org/10.1057/dddmp.2016.3

Jamsa, K. (2021). *Introduction to Data Mining and Analytics*. Jones & Bartlett Learning.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*, 429–449.

Järvinen, J. (2016). The Use of Digital Analytics for Measuring and Optimizing Digital Marketing Performance. In *Dissertation*.

Jayachandran, S., Sharma, S., Kaufman, P., & Raman, P. (2004). The Role of Relational Information Processes and Technology Use in Customer Relationship Management. *Journal of Marketing*, *69*(4), 177–192.

Jeffery, M. (2010). *Data-Driven Marketing*. John Wiley & Sons, Inc.

Jiang, T., Yang, J., Yu, C., & Sang, Y. (2018). A Clickstream Data Analysis of the Differences between Visiting Behaviors of Desktop and Mobile Users. *Data and Information Management*, *2*(3), 130–140. https://doi.org/10.2478/dim-2018-0012

Jonker, J. J., Piersma, N., & Van Den Poel, D. (2004). Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. *Expert Systems with Applications*, *27*(2), 159–168. https://doi.org/10.1016/j.eswa.2004.01.010

Jonker, J., Piersma, N., & Potharst, R. (2006). A decision support system for direct mailing decisions. *Decision Support Systems*, *42*, 915–925. https://doi.org/10.1016/j.dss.2005.08.006

Joseph, R. (2018). *Grid Search for Model Tuning*. Towards Data Science. https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e

Jüttner, U., & Wehrli, H. P. (1994). Relationship Marketing from a Value System Perspective. *International Journal of Service Industry Management*, *5*(5), 54–73. https://doi.org/10.1007/s11576-007-0007-8

Kaatz, C., Brock, C., & Figura, L. (2019). Journal of Retailing and Consumer Services Are you still online or are you already mobile ? – Predicting the path to successful

conversions across di ff erent devices. *Journal of Retailing and Consumer Services*, *50*(October 2018), 10–21. https://doi.org/10.1016/j.jretconser.2019.04.005

Kaggle. (n.d.). *Customer Transaction Dataset*. https://www.kaggle.com/archit9406/customer-transaction-dataset

Kang, P., Cho, S., & MacLachlan, D. L. (2012). Improved response modeling based on clustering, under-sampling, and ensemble. *Expert Systems with Applications*, *39*(8), 6738–6753. https://doi.org/10.1016/j.eswa.2011.12.028

Kaniewska-Seba, A., & Pilarczyk, B. (2014). Negative Effects of Personalization in Direct Marketing. *International Journal of Arts & Sciences*, *07*(02), 89–98.

Kantardzic, M. (2020). *Data Mining: Concepts, Models, Methods, and Algorithms* (third edit). Wiley-IEEE Press. https://doi.org/10.1002/9781118029145

Kašćelan, L., Kašćelan, V., & Jovanović, M. (2015). Hybrid support vector machine rule extraction method for discovering the preferences of stock market investors: Evidence from Montenegro. *Intelligent Automation and Soft Computing*, *21*(4), 503–522. https://doi.org/10.1080/10798587.2014.971500

Kašćelan, L., & Rogić, S. (2022). Data Analytics for Marketing Knowledge Advancement: A Market Segmentation Example Using Support Vector Machine. In G. Schiuma & A. Bassi (Eds.), *IFKAD: Knowledge Drivers for Resilience and Transformation*.

Kašćelan, V., Kašćelan, L., & Burić, M. N. (2016). A nonparametric data mining approach for risk prediction in car insurance: A case study from the Montenegrin market. *Economic Research-Ekonomska Istrazivanja* , *29*(1), 545–558. https://doi.org/10.1080/1331677X.2016.1175729

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *29*(2), 119. https://doi.org/10.2307/2986296

Kastratović, V., & Kašćelan, L. (2009). Sistem poslovne inteligencije u uslužnoj djelatnosti radio-televizije. *InfoM*, *30*, 39–44.

Kaufman, L., & Rousseuw, P. J. (1991). Finding Groups in Data: An Introduction to Cluster Analysis. In *Biometrics* (Vol. 47, Issue 2). John Wiley & Sons. https://doi.org/10.2307/2532178

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, *3*, 57–63. https://doi.org/10.1016/j.procs.2010.12.011

Khalili-Damghani, K., Abdi, F., & Abolmakarem, S. (2018). Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. *Applied Soft Computing Journal*, *73*, 816–828. https://doi.org/10.1016/j.asoc.2018.09.001

Khreich, W., Granger, E., Miri, A., & Sabourin, R. (2010). Iterative Boolean combination of classifiers in the ROC space: An application to anomaly detection with HMMs. *Pattern Recognition*, *43*(8), 2732–2752. https://doi.org/10.1016/j.patcog.2010.03.006

Kim, D., Lee, H. joo, & Cho, S. (2008). Response modeling with support vector regression. *Expert Systems with Applications*, *34*(2), 1102–1108. https://doi.org/10.1016/j.eswa.2006.12.019

Kim, G., Chae, B. K., & Olson, D. L. (2013). A support vector machine (SVM) approach to imbalanced datasets of customer responses: Comparison with other customer response models. *Service Business*, *7*(1), 167–182. https://doi.org/10.1007/s11628-012-0147-9

Kim, S., Shin, K. S., & Park, K. (2005). An application of support vector machines for customer churn analysis: Credit card case. *Lecture Notes in Computer Science*, *3611*(PART II), 636–647. https://doi.org/10.1007/11539117_91

Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications*, *31*(1), 101–107.

https://doi.org/10.1016/j.eswa.2005.09.004

Kim, Y. S., & Street, W. N. (2004). An intelligent system for customer targeting: A data mining approach. *Decision Support Systems*, *37*(2), 215–228. https://doi.org/10.1016/S0167-9236(03)00008-3

Kim, Y. S., Street, W. N., Russell, G. J., & Menczer, F. (2005). Customer targeting: A neural network approach guided by genetic algorithms. *Management Science*, *51*(2), 264–276. https://doi.org/10.1287/mnsc.1040.0296

Kirsch, K. (2021). *40 Ad Blocker Stats Brands Need to Know in 2021*. Hubspot. https://blog.hubspot.com/marketing/ad-blocking-stats#:~:text=39%25 of people believe ads,companies from collecting personal data.

Koetsier, J. (2020). *Google Is Tracking You On 86% Of The Top 50,000 Websites On The Planet*. Forbes. https://www.forbes.com/sites/johnkoetsier/2020/03/11/google-is-tracking-you-on-86-of-the-top-50000-websites-on-the-planet/?sh=7a23c77c750f

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Appears in the International Joint Conference on Artificial Intelligence*, *14*(2), 1137–1145. https://doi.org/10.1067/mod.2000.109032

Korstanje, J. (2020). *GridSearch: the ultimate Machine Learning Tool*. Towards Data Science. https://towardsdatascience.com/gridsearch-the-ultimate-machine-learning-tool-6cd5fb93d07

Kotler, P. (1972). A Generic Concept of Marketing. *Journal of Marketing*, *36*(2), 46–54. https://doi.org/10.1177/002224297203600209

Kotler, P., & Armstrong, G. (1996). *Principles of Marketing, 7th edition*. Prentice Hall.

Kotler, P., & Armstrong, G. (2008). *Principles of Marketing* (12th editi). Person Education Inc.

Kumar, S. (2020). *Understanding K-Means, K-Means++ and, K-Medoids Clustering Algorithms*. Towards Data Science.

https://towardsdatascience.com/understanding-k-means-k-means-and-k-medoids-clustering-algorithms-ad9c9fbf47ca

Kumar, V., Venkatesan, R., & Reinartz, W. (2008). Performance Implications of Adopting a Customer-Focused Sales Campaign. *Journal of Marketing, 72*(5), 50–68.

Kurnia, Y., & Kusuma, K. (2018). Comparison of C4.5 Algorithm, Naive Bayes and Support Vector Machine (SVM) in Predicting Customers that Potentially Open Deposits. *Bit-Tech, 1*(2), 40–47. https://doi.org/10.32877/bt.v1i2.46

Kurniawan, I., Abdussomad, Akbar, M. F., Saepudin, D. F., Azis, M. S., & Tabrani, M. (2020). Improving the Effectiveness of Classification Using the Data Level Approach and Feature Selection Techniques in Online Shoppers Purchasing Intention Prediction. *Journal of Physics: Conference Series, 1641*(1). https://doi.org/10.1088/1742-6596/1641/1/012083

Kytö, E., Virtanen, M., & Mustonen, S. (2019). From intention to action: Predicting purchase behavior with consumers' product expectations and perceptions, and their individual properties. *Food Quality and Preference, 75*(February), 1–9. https://doi.org/10.1016/j.foodqual.2019.02.002

Lador, S. M. (2017). *What metrics should be used for evaluating a model on an imbalanced data set? (precision + recall or ROC=TPR+FPR).* Towards Data Science. https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba

Ładyżyński, P., Żbikowski, K., & Gawrysiak, P. (2019). Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications, 134*, 28–35. https://doi.org/10.1016/j.eswa.2019.05.020

Lam, S. (2018). The ensemble of neural network and gradient boosting for the prediction of customer profitability: A two-stage modeling approach. *Model Assisted Statistics and Applications, 13*(4), 329–340. https://doi.org/10.3233/MAS-180443

Lameski, P., Zdravevski, E., Mingov, R., & Kulakov, A. (2015). SVM parameter tuning with grid search and its impact on reduction of model over-fitting. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9437*, 464–474. https://doi.org/10.1007/978-3-319-25783-9_41

Landau, S., & Barthel, S. (2010). Recursive Partitioning. *International Encyclopedia of Education*, *3rd editio*, 383–389. https://doi.org/10.1016/B978-0-08-044894-7.01314-2

Larivière, B., & Van Den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, *29*(2), 472–484. https://doi.org/10.1016/j.eswa.2005.04.043

Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics* (2nd editio). John Wiley & Sons, Inc.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, *22*(1), 45–55.

Laudon, K. C., & Traver Guercio, C. (2017). *E-commerce 2017: business, technology, society*. Pearson.

Lawi, A., Velayaty, A. A., & Zainuddin, Z. (2018). On identifying potential direct marketing consumers using adaptive boosted support vector machine. *Proceedings of the 2017 4th International Conference on Computer Applications and Information Processing Technology, CAIPT 2017*, *2018-Janua*, 1–4. https://doi.org/10.1109/CAIPT.2017.8320691

Lazar, E. (2018). Customer Churn Prediction Embedded in an Analytical CRM Model. *SSRN Electronic Journal*, 24–30. https://doi.org/10.2139/ssrn.3281605

Lazović, V., & Đuričković, T. (2018). *Digitalna ekonomija*. Autorsko izdanje.

Lee, D., Hosanagar, K., & Nair, H. S. (2018). Advertising content and consumer engagement on social media: Evidence from Facebook. *Management Science*,

64(11), 5105–5131. https://doi.org/10.1287/mnsc.2017.2902

Lee, J. H., & Park, S. C. (2005). Intelligent profitable customers segmentation system based on business intelligence tools. *Expert Systems with Applications, 29*(1), 145–152. https://doi.org/10.1016/j.eswa.2005.01.013

Lee, J., Jung, O., Lee, Y., Kim, O., & Park, C. (2021). A comparison and interpretation of machine learning algorithm for the prediction of online purchase conversion. *Journal of Theoretical and Applied Electronic Commerce Research, 16*(5), 1472–1491. https://doi.org/10.3390/jtaer16050083

Lee, K. C., Kim, J. S., Chung, N. H., & Kwon, S. J. (2002). Fuzzy cognitive map approach to web-mining inference amplification. *Expert Systems with Applications, 22*(3), 197–211. https://doi.org/10.1016/S0957-4174(01)00054-9

Lei, M., Jiang, G., Yang, J., Mei, X., Xia, P., & Shi, H. (2018). Improvement of the regression model for spindle thermal elongation by a Boosting-based outliers detection approach. *International Journal of Advanced Manufacturing Technology, 99*(5–8), 1389–1403. https://doi.org/10.1007/s00170-018-2559-8

Leick, R. (2007). *Building Airline Passenger loyalty through an understanding of customer value*. Cranfield University.

Lejeune, M. A. P. M. (2001). Measuring the impact of data mining on churn management. *Internet Research, 11*(5), 375–387. https://doi.org/10.1108/10662240110410183

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research, 43*(2), 276–286. https://doi.org/10.1509/jmkr.43.2.276

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing, 80*(6), 69–96. https://doi.org/10.1509/jm.15.0420

LeSueur, J. (2007). *Marketing Automation: Practical Steps to More Effective Direct*

Marketing. John Wiley & Sons, Inc.

Levin, N., & Zahavi, J. (1998). Continuous predictive modeling - A comparative analysis. *Journal of Interactive Marketing*, *12*(2), 5–22. https://doi.org/10.1002/(SICI)1520-6653(199821)12:2<5::AID-DIR2>3.0.CO;2-D

Liao, S. H., Chen, Y. J., & Hsieh, H. H. (2011). Mining customer knowledge for direct selling and marketing. *Expert Systems with Applications*, *38*(5), 6059–6069. https://doi.org/10.1016/j.eswa.2010.11.007

Liashchynskyi, P., & Liashchynskyi, P. (2019). *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS*. *2017*, 1–11. http://arxiv.org/abs/1912.06059

Lin, J. B., Liang, T. H., & Lee, Y. G. (2012). Mining important association rules on different customer potential value segments for life insurance database. *Proceedings - 2012 IEEE International Conference on Granular Computing, GrC 2012*, *00*, 283–288. https://doi.org/10.1109/GrC.2012.6468569

Lipyanina, H., Sachenko, A., Lendyuk, T., Nadvynychny, S., & Grodskyi, S. (2020). Decision tree based targeting model of customer interaction with business page. *CEUR Workshop Proceedings*, *2608*, 1001–1012.

Liu, D. R., & Shih, Y. Y. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information and Management*, *42*(3), 387–400. https://doi.org/10.1016/j.im.2004.01.008

Liu, J., & Zio, E. (2019). Integration of feature vector selection and support vector machine for classification of imbalanced data. *Applied Soft Computing Journal*, *75*, 702–711. https://doi.org/10.1016/j.asoc.2018.11.045

Liu, X., Lee, D., & Srinivasan, K. (2019). Large Scale Cross Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning. *Journal of Marketing Research*, *56*(6), 918–943.

Loh, W. Y., & Shin, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*(4), 815–840.

Lopez-Garcia, P., Masegosa, A. D., Osaba, E., Onieva, E., & Perallos, A. (2019). Ensemble classification for imbalanced data based on feature space partitioning and hybrid metaheuristics. *Applied Intelligence*, *49*(8), 2807–2822. https://doi.org/10.1007/s10489-019-01423-6

Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature Communications*, *10*(1), 1–9. https://doi.org/10.1038/s41467-019-09311-w

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*(14), 281–297. http://books.google.de/books?hl=de&lr=&id=IC4Ku_7dBFUC&oi=fnd&pg=PA281&dq=MacQueen+some+methods+for+classification&ots=nNTcK1IdoQ&sig=fHzdVcbvmYJ-lTNHu1HncmOFOkM#v=onepage&q=MacQueen some methods for classification&f=false

Mahdiloo, M., Noorizadeh, A., & FarzipoorSaen, R. (2014). Optimal direct mailing modelling based on data envelopment analysis. *Expert Systems*, *31*(2), 101–109. https://doi.org/10.1111/exsy.12011

Majovski, I., Janevski, Z., & Petkoviski, V. (2018). Digital Skills Readiness of Selected Western Balkan Countries. *Economic Development*, *3*, 41–54.

Maldonado, S., & López, J. (2014). Imbalanced data classi fi cation using second-order cone programming support vector machines. *Pattern Recognition*, *47*(5), 2070–2079. https://doi.org/10.1016/j.patcog.2013.11.021

Malina, M. A., Nrreklit, H. S. O., & Selto, F. H. (2011). Lessons learned: Advantages and disadvantages of mixed method research. *Qualitative Research in Accounting and Management*, *8*(1), 59–71. https://doi.org/10.1108/11766091111124702

Mallin, M. L., & Finkle, T. A. (2009). Social entrepreneurship and direct marketing. In *Direct Marketing: An International Journal*. https://doi.org/10.1108/17505930710756833

Malthouse, E. (1999). Ridge Regression and Direct Marketing Scoring Models. *Journal of Interactive Marketing*, *13*, 19–23.

Malthouse, E. C. (1999). Ridge regression and direct marketing scoring models. *Journal of Interactive Marketing*, *13*(4), 10–23. https://doi.org/10.1002/(SICI)1520-6653(199923)13:4<10::AID-DIR2>3.0.CO;2-3

Malthouse, E. C., & Blattberg, R. C. (2005). Can we predict customer lifetime value? *Journal of Interactive Marketing*, *19*(1), 2–16. https://doi.org/10.1002/dir.20027

Mandapaka, A. K., Singh Kushwah, A., & Chakraborty, D. (2014). *Role of customer response models in customer solicitation center's direct marketing campaign*. *July*, 1–12.

Marinakos, G., & Daskalaki, S. (2017). Imbalanced customer classification for bank direct marketing. *Journal of Marketing Analytics*, *5*(1), 14–30. https://doi.org/10.1057/s41270-017-0013-7

Markethink. (n.d.). *Lester Wunderman*. http://www.markethink.guru/en/markethinkers/736-lester-wunderman

Martens, D., Baesens, B., Gestel, T. Van, & Vanthienen, J. (2007). Comprehensible Credit Scoring Models Using Rule Extraction From Support Vector Machines. *Decision Sciences*, *183*(3), 1466–1476.

Martens, D., Huysmans, J., Setiono, R., Vanthienen, J., & Baesens, B. (2008). Rule extraction from support vector machines: An overview of issues and application in credit scoring. *Studies in Computational Intelligence*, *80*(2008), 33–63. https://doi.org/10.1007/978-3-540-75390-2_2

Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual

setting. *European Journal of Operational Research, 281*(3), 588–596. https://doi.org/10.1016/j.ejor.2018.04.034

Martínez, R. G., Carrasco, R. A., García-Madariaga, J., Gallego, C. P., & Herrera-Viedma, E. (2019). A comparison between Fuzzy Linguistic RFM Model and traditional RFM model applied to Campaign Management. Case study of retail business. *Procedia Computer Science, 162*(Itqm), 281–289. https://doi.org/10.1016/j.procs.2019.11.286

Maxwell, J. (2013). *Demystifying the online shopper: 10 myths of multichannel retailing*. https://www.pwc.com/us/en/retail-consumer/publications/assets/pwc-multi-channel-shopper-survey.pdf

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks, 21*(2–3), 427–436. https://doi.org/10.1016/j.neunet.2007.12.031

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research, 60*(6), 656–662. https://doi.org/10.1016/j.jbusres.2006.06.015

McCorkell, G. (1992). Direct Marketing – a new industry or a new idea? In B. Halsey (Ed.), *The Practitioner's Guide to Direct Marketing*. Institute of Direct Marketing.

McPartlin, S., & Dugal, L. F. (2012). *Understanding how US online shoppers are reshaping the retail experience*. PricewaterhouseCoopers. https://www.pwc.com/us/en/retail-consumer/publications/assets/pwc-us-multichannel-shopping-survey

Melović, B., Pavičić, J., Gnjidić, V., & Drašković, N. (2019). *Strategijski marketing*. Ekonomski fakultet. Podgorica i Ekonomski fakultet Zagreb.

Mero, J., Tarkiainen, A., & Tobon, J. (2020). Effectual and causal reasoning in the adoption of marketing automation. *Industrial Marketing Management,*

86(December), 212–222. https://doi.org/10.1016/j.indmarman.2019.12.008

Microsoft. (2015). *Attention Spans: Consumer Insights*. http://dl.motamem.org/microsoft-attention-spans-research-report.pdf

Microsoft. (2017). *State of Global Customer Service Report*. http://info.microsoft.com/rs/157-GQE-382/images/EN-CNTNT-Report-DynService-2017-global-state-customer-service-en-au.pdf

Miglautsch, J. (2002). Application of RFM principles : What to do with 1 – 1 – 1 customers ? *Journal of Database Marketing & Customer Strategy Management*, *9*(4), 319–324.

Miguéis, V. L., Camanho, A. S., & Borges, J. (2017). Predicting direct marketing response in banking: comparison of class imbalance methods. *Service Business*, *11*(4), 831–849. https://doi.org/10.1007/s11628-016-0332-3

Migueis, V., & Teixeira, R. (2020). Predicting Market Basket Additions as a Way to Enhance Customer Service Levels. In H. Nóvoa, M. Dragoicea, & N. Kühl (Eds.), *Exploring Service Science* (Issue January, pp. 121–134). Springer Nature Switzerland AG 2020. https://doi.org/10.1007/978-3-030-38724-2

Minaee, S. (2019). *20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics*. Towards Data Science. https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce

Monalisa, S., Nadya, P., & Novita, R. (2019). Analysis for customer lifetime value categorization with RFM model. *Procedia Computer Science*, *161*, 834–840. https://doi.org/10.1016/j.procs.2019.11.190

Morgan, J., & Sonquist, J. (1963). *Problems in the Analysis of Survey Data, and a Proposal*. *58*(302), 415–434.

Moro, S., Laureano, R. M. S., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the CRISP-DM methodology. *ESM 2011 - 2011*

European Simulation and Modelling Conference: Modelling and Simulation 2011, Figure 1, 117–121.

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *Handbook of Social Psychology* (Vol 2). Addison-Wesley.

Mrzic, E., & Zaimovic, T. (2020). Data science methods and machine learning algorithm implementations for customized pratical usage. *TEM Journal*, *9*(3), 1179–1185. https://doi.org/10.18421/TEM93-44

Mulhern, Francis. (1999). Customer Profitability Analysis: Measurement, Concentration and Research Directions. *Journal of Interactive Marketing*, *13*(1), 25–40. http://dx.doi.org/10.1002/(SICI)1520-6653(199924)13:1%3C25::AID-DIR3%3E3.0.CO;2-L

Mulhern, Frank. (2009). Integrated marketing communications: From media channels to digital connectivity. *Journal of Marketing Communications*, *15*(2–3), 85–101. https://doi.org/10.1080/13527260902757506

Mulhern, Frank. (2010). Direct and Interactive Marketing. In J. N. Sheth & N. K. Malhotra (Eds.), *Wiley International Encyclopedia of Marketing*. John Wiley & Sons Ltd.

Mulvenna, M., Norwood, M., & Buchner, A. G. (1998). Data Driven Marketing. *Electronic Markets*, *8*(3), 32–37. https://doi.org/10.1007/978-3-658-31959-5_5

Mutula, S. M. (2009). Digital economies: SMEs and e-readiness. In *Digital Economies: SMEs and E-Readiness*. https://doi.org/10.4018/978-1-60566-420-0

Nagel, T. (2007). International dialog for successful acquisition of new customers. In R. D. Krafft M., Hesse J., Höfling J., Peters K. (Ed.), *International Direct Marketing*. Springer.

Naik, P. A., & Tsai, C. L. (2004). Isotonic single-index model for high-dimensional database marketing. *Computational Statistics and Data Analysis*, *47*(4), 775–790. https://doi.org/10.1016/j.csda.2003.11.023

Nash, E. L. (1984). *The Direct Marketing Handbook* (2nd editio). McGraw-Hill.

Nemhauser, M. (2014). *The Real Mad Men: The 1960s—A Golden Age of Advertising*. (Advanced History Seminar in Historical Research and Writing, Issue April). http://s3.amazonaws.com/academia.edu.documents/33896914/Max_Nemhauser_-_Golden_Age_of_Advertising_-_FINAL_DRAFT.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1496390117&Signature=bFU%252F84UPoEejFRd7FKy%252BQA8%252FOhw%253D&response-content-disposition=

Newman, D. (2021). *The Move Away From Third-Party Data Is Imminent: Are You Ready?* Forbes. https://www.forbes.com/sites/danielnewman/2021/12/08/the-move-away-from-third-party-data-is-imminent-are-you-ready/?sh=6da4f9c62093

Nextiva. (2021). *100 Essential Customer Service Statistics and Trends for 2022*. https://www.nextiva.com/blog/customer-service-statistics.html

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2 PART 2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

Ngo, Q. (2019). *Customer Response Analysis for Facebook Marketing – Case Study : Cheapsleep Hostel Helsinki*.

Noviantoro, T., & Huang, J.-P. (2021). Applying Data Mining Techniques to Investigate Online Shopper Purchase Intention Based on Clickstream Data. *Review of Business, Accounting & Finance*, *01*(02), 130–159.

Olson, D. L., & Chae, B. (2012). Direct marketing decision support through predictive customer response modeling. *Decision Support Systems*, *54*(1), 443–451. https://doi.org/10.1016/j.dss.2012.06.005

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques* (1st ed.). Springer-Verlag Berlin Heidelberg.

Otter, P. W., van der Scheer, H., & Wansbeek, T. (2006). *Optimal selection of households*

for direct marketing by joint modeling of the probability and quantity of response (No. 200606.).

Palmer, A., & Koenig-Lewis, N. (2009). An experiential, social network-based approach to direct marketing. *Direct Marketing: An International Journal*, *3*(3), 162–176. https://doi.org/10.1108/17505930910985116

Panigrahi, A., & Patnaik, M. C. (2020). Customer Deposit Prediction Using Neural Network Techniques. *International Journal of Applied Engineering Research*, *15*(3), 253–258.

Parise, S., Guinan, P. J., & Kafka, R. (2016). Solving the crisis of immediacy: How digital technology can transform the customer experience. *Business Horizons*, *59*(4), 411–420. https://doi.org/10.1016/j.bushor.2016.03.004

Park, C. H., & Park, Y. H. (2016). Investigating purchase conversion by uncovering online visit patterns. *Marketing Science*, *35*(6), 894–914. https://doi.org/10.1287/mksc.2016.0990

Parr Rud, O. (2001). *Data mining cookbook: modeling data for marketing, risk, and customer relationship management*. John Wiley & Sons.

Peppers, D., & Rogers, M. (1997). *The One to One Future: Building Relationships One Customer at a Time*. Bantam Doubleday Dell Publishing.

Peters, L. (1998). The new interactive media: One-to-one, but who to whom? *Marketing Intelligence & Planning*, *16*(1), 22–30. https://doi.org/10.1108/02634509810199472

Peterson, R. A., Balasubramanian, S., & Bronnenberg, B. J. J. A. . (1997). Exploring the implications of the internet for consumer marketing. *Journal of the Academy of Marketing Science*, *25*(4), 329–346.

Pisner, D. A., & Schnyer, D. M. (2019). Support vector machine. In *Machine Learning: Methods and Applications to Brain Disorders* (pp. 101–121). Elsevier Inc. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Pitta, D., Franzak, F., & Fowler, D. (2006). A strategic approach to building online customer loyalty: Integrating customer profitability tiers. *Journal of Consumer Marketing*, *23*(7), 421–429. https://doi.org/10.1108/07363760610712966

Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. (2011). A Survey on Graphical Methods for Classification Predictive Performance Evaluation. *IEEE Transactions on Knowledge and Data Engineering*, *23*(11), 1601–1618.

Protect, P. (2020). *How Many Ads Do We See A Day In 2020?* https://ppcprotect.com/how-many-ads-do-we-see-a-day/

Provost, F., & Fawcett, T. (2013). *Data science for business - what you need to know about data mining and data-analytic thinking* (1st editio). O'Reilly Media, Inc.,.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Quinlan, J. R. (1992). *C4.5 - programs for machine learning*. Kaufmann.

Raj, A. (2020). *Unlocking the True Power of Support Vector Regression*. Towards Data Science. https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0

Raphaeli, O., Goldstein, A., & Fink, L. (2017). Analyzing Online Consumer Behavior in Mobile and PC Devices: A Novel Web Usage Mining Approach. *Electronic Commerce Research and Applications*, *26*, 1–12. https://doi.org/10.1016/j.elerap.2017.09.003

Raskutti, B., & Kowalczyk, A. (2004). Extreme Re-balancing for SVMs: a case study. *ACM Sigkdd Explorations Newsletter*, *6*(1), 60–69.

Ravald, A., & Grönroos, C. (1996). The value concept and relationship marketing. *European Journal of Marketing*, *30*(2), 19–30. https://doi.org/10.1108/03090560210430782

Rekettye, G., & Rekettye, G. (2019). The Effects of Digitalization on Customer Experience. *SSRN Electronic Journal*, *September*, 340–346. https://doi.org/10.2139/ssrn.3491767

Reutterer, T., Mild, A., Natter, M., & Taudes, A. (2006). A Dynamic Segmentation Approach for Targeting and Customizing Direct Marketing Campaigns. *Journal of Interactive Marketing*, *20*(3–4), 43–57. https://doi.org/10.1002/dir

Rhee, E., & Russell, G. J. (2009). Forecasting household response in database marketing: A latent trait approach. *Advances in Business and Management Forecasting*, *6*(508), 109–131. https://doi.org/10.1108/S1477-4070(2009)0000006008

Rho, J. J., Moon, B.-J., Kim, Y.-J., & Yang, D.-H. (2011). Internet Customer Segmentation Using Web Log Data. *Journal of Business & Economics Research (JBER)*, *2*(11), 59–74. https://doi.org/10.19030/jber.v2i11.2940

Rhys, H. I. (2020). *Machine Learning with R, the tidyverse, and mlr*. Manning Publications Co.

Ripley, B. D. (1994). Neural Networks and Related Methods for Classification. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(3), 409–437. https://doi.org/10.1111/j.2517-6161.1994.tb01990.x

Roach, G. (2009). Consumer perceptions of mobile phone marketing: A direct marketing innovation. *Direct Marketing*, *3*(2), 124–138. https://doi.org/10.1108/17505930910964786

Robert M., M., & Shelby D., H. (1994). The Commitment-Trust theory of Relationship Marketing.pdf. In *Journal of Marketing* (Vol. 58, Issue July, pp. 20–38).

Rogic, S., & Kascelan, L. (2019). Customer Value Prediction in Direct Marketing Using Hybrid Support Vector Machine Rule Extraction Method. *Communications in Computer and Information Science*, *1064*, 283–294. https://doi.org/10.1007/978-3-030-30278-8_30

Rogic, S., & Kascelan, L. (2020). Class balancing in customer segments classification using support vector machine rule extraction and ensemble learning. *Computer Science and Information Systems*, *18*(3), 893–925. https://doi.org/10.2298/csis200530052r

Rogic, S., & Kascelan, L. (2021). Estimating Customers' Profitability - Influence of RFM Attributes,Web Metrics and Product Data. In *Marketing and Smart Technologies* (Vol. 1, pp. 293–304). Springer. https://doi.org/10.1007/978-981-15-1564-4

Rogić, S., & Kašćelan, L. (2021). Segmentation Approach for Athleisure and Performance Sport Retailers Based on Data Mining Techniques. *International Journal of E-Services and Mobile Applications*, *13*(3), 71–85. 10.4018/IJESMA.2021070104

Rogić, S., & Kašćelan, L. (2022). Customer Response Modeling Using Ensemble of Balanced Classifiers: Significance of Web Metrics. In K. Arai (Ed.), *Intelligent Computing. SAI 2022. Lecture Notes in Networks and Systems, vol 506* (pp. 433–448). Springer, Cham. https://doi.org/10.1007/978-3-031-10461-9_30

Rogić, S., Kašćelan, L., Kašćelan, V., & Đurišić, V. (2022). Automatic customer targeting: a data mining solution to the problem of asymmetric profitability distribution. *Information Technology and Management*. https://doi.org/10.1007/s10799-021-00353-5

Rogić, S., Kašćelan, L., & Pejić Bach, M. (2022). Customer Response Model in Direct Marketing: Solving the Problem of Unbalanced Dataset with a Balanced Support Vector Machine. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(3), 1003–1018. https://doi.org/10.3390/jtaer17030051

Rokach, L., & Maimon, O. (2015). *Data Mining With Decision Trees Theory and Applications* (2nd editio). World Scientific Publishing Co. Pte. Ltd.

Roque, C. (2014). *99 Years of Content Marketing: How American Express Became a Major American Publisher*. Contently. https://contently.com/2014/03/20/99-years-of-content-marketing-how-american-express-became-a-major-american-publisher/

Rosenbloom, B. (2003). *Marketing Channels: A Management View* (Revised Ed). Cengage South-Western.

Rowe, C. W. (1989). A Review of Direct Marketing and How It Can be Applied to the Wine Industry. *European Journal of Marketing*, *23*(9), 5–14.

Rrustemi, V., Podvorica, G., & Jusufi, G. (2020). Digital Marketing Communication in Developing Countries. *LeXonomica*, *12*(2), 243–260. https://doi.org/10.18690/lexonomica.12.2.243-260.2020

Rust, R. T., Kumar, V., & Venkatesan, R. (2011). Will the frog change into a prince? Predicting future customer profitability. In *International Journal of Research in Marketing* (Vol. 28, Issue 4). https://doi.org/10.1016/j.ijresmar.2011.05.003

Rust, R. T., Zahorik, A. J., & Keiningham, T. L. (1995). Return on Quality (ROQ): Making Service Quality Financially Accountable. *Journal of Marketing*, *59*(2), 58–70. https://doi.org/10.1177/002224299505900205

Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. *International Journal of Advanced Computer Science and Applications*, *9*(2), 273–281. https://doi.org/10.14569/IJACSA.2018.090238

Safari, F., Safari, N., & Montazer, G. A. (2016). Customer lifetime value determination based on RFM model. *Marketing Intelligence and Planning*, *34*(4), 446–461. https://doi.org/10.1108/MIP-03-2015-0060

Sagala, N. T. M., & Permai, S. D. (2021). Predictive Model using SVM to Improve the Effectiveness of Direct Marketing. *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, 1–6.

Salehinejad, H., & Rahnamayan, S. (2016). Customer Shopping Pattern Prediction: A Recurrent Neural Network Approach. *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–6.

Salesforce Research. (2016). *State of the Connected Customer*.

Sanderson, M. (2010). Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. *Natural Language Engineering*, *16*(1), 100–103.

Sanne, P. N. C., & Wiese, M. (2018). The theory of planned behaviour and user engagement applied to Facebook advertising. *SA Journal of Information*

Management, *20*(1), 1–10. https://doi.org/10.4102/sajim.v20i1.915

Sarvari, P., Ustundag, A., & Takci, H. (2016). Performance Evaluation of Different Customer Segmentation Approaches Based on RFM and Demographics Analysis. *Kybernetes*, *45*(7), 1129–1157. https://doi.org/http://dx.doi.org/10.1108/K-07-2015-0180

Schafer, J. Ben, Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, *5*(1–2), 115–153. https://doi.org/10.1007/978-1-4615-1627-9_6

Schnack, H. G., Van Haren, N. E. M., Nieuwenhuis, M., Pol, H. E. H., Cahn, W., & Kahn, R. S. (2016). Accelerated brain aging in schizophrenia: A longitudinal pattern recognition study. *American Journal of Psychiatry*, *173*(6), 607–616. https://doi.org/10.1176/appi.ajp.2015.15070922

Seller, M., & Gray, P. (1999). *A Survey of Database Marketing*.

Semerádová, T., & Weinlich, P. (2019). Computer Estimation of Customer Similarity With Facebook Lookalikes : Advantages and Disadvantages of Hyper-Targeting. *IEEE Access*, *7*, 153365–153377. https://doi.org/10.1109/ACCESS.2019.2948401

Shanahan, T., Tran, T. P., & Taylor, E. C. (2019). Journal of Retailing and Consumer Services Getting to know you : Social media personalization as a means of enhancing brand loyalty and perceived quality. *Journal of Retailing and Consumer Services*, *47*(October 2018), 57–65. https://doi.org/10.1016/j.jretconser.2018.10.007

Shankar, V., & Hollinger, M. (2007). *Online and Mobile Advertising: Current Scenario , Emerging Trends, and Future Directions Online and Mobile Advertising: Current Scenario, Emerging Trends, and Future Directions*.

Shankar, V., & Malthouse, E. C. (2006). Moving interactive marketing forward. *Journal of Interactive Marketing*, *20*(1), 2–4. https://doi.org/10.1002/dir.20057

Shankar, V., & Malthouse, E. C. (2007). The Growth of Interactions and Dialogs in

Interactive Marketing. *Journal of Interactive Marketing*, *21*(2), 2–4. https://doi.org/10.1002/dir.20080

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems*, *31*(1), 127–137. https://doi.org/10.1016/S0167-9236(00)00123-8

Shchutskaya, V. (2021). *What is Predictive Performance Models and Why Their Performance Evaluation is Important*. InData Labs. https://indatalabs.com/blog/predictive-models-performance-evaluation-important

Sheikh, A., Ghanbarpour, T., & Gholamiangonabadi, D. (2019). A Preliminary Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting. *Journal of Business-to-Business Marketing*, *26*(2), 197–207. https://doi.org/10.1080/1051712X.2019.1603420

Shen, C. C., & Chuang, H. M. (2009). A study on the applications of data mining techniques to enhance customer lifetime value. *WSEAS Transactions on Information Science and Applications*, *6*(2), 319–328.

Sheshasaayee, A., & Logeshwari, L. (2018). IMPLEMENTATION OF CLUSTERING TECHNIQUE BASED RFM ANALYSIS. *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, *Icoei*, 1166–1170.

Shin, H., & Cho, S. (2006). *Response modeling with support vector machines*. *30*(4), 746–760. https://doi.org/10.1016/j.eswa.2005.07.037

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. J. (2018). *Data Mining for Business Analytics - Concepts, Techniques, and Applications in R*. John Wiley & Sons, Inc.

Sidyakov, M. (n.d.). *Davies-Bouldin Index for K-Means Clustering Evaluation in Python*. PyShark. https://pyshark.com/davies-bouldin-index-for-k-means-clustering-evaluation-in-python/

Sigma. (2021). *Breaking Down The Data Language Barrier.* https://www.sigmacomputing.com/resources/data-language-barrier/

Sin, L. Y. M., Tse, A. C. B., & Yim, F. H. K. (2005). CRM: Conceptualization and scale development. *European Journal of Marketing, 39*(11–12), 1264–1290. https://doi.org/10.1108/03090560510623253

Singh, D. (2019). *What is Predictive Model Performance Evaluation*. Medium. https://medium.com/@divyacyclitics15/what-is-predictive-model-performance-evaluation-8ef117ae0e40

Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing, 21*(1), 3–8. https://doi.org/10.4135/9781483329864.n5

Soni, D. (2018). *Supervised vs. Unsupervised Learning*. Towards Data Science. https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

Sprague, L. (2022). *Direct Mail Response Rates.* The Mail Shark. https://www.themailshark.com/resources/articles/direct-mail-response-rates-and-how-to-improve-them/#:~:text=Normal Rates of Return on Direct Mail Campaigns&text=5 to 2%25 return rate,mailers with a standard campaign.

Statista. (2020). *Number of digital buyers worldwide from 2014 to 2021.* https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/

Statista. (2021a). *Average daily time spent using the internet by online users worldwide as of 3rd quarter 2020, by region.* https://www.statista.com/statistics/1258232/daily-time-spent-online-worldwide/

Statista. (2021b). *Number of monthly active Facebook users worldwide as of 3rd quarter 2021.* https://www.statista.com/statistics/264810/number-of-monthly-active-

facebook-users-worldwide/

Steinholtz, O. S. (2018). *A Comparative Study of Black-box Optimization Algorithms for Tuning of Hyper-parameters in Deep Neural Networks* [Luleå University of Technology]. http://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-69865

Stewart-Knox, B. J., Markovina, J., Rankin, A., Bunting, B. P., Kuznesof, S., Fischer, A. R. H., ..., & Frewer, L. J. (2016). Making personalised nutrition the easy choice: creating policies to break down the barriers and reap the benefits. *Food Policy*, *63*, 134–144.

Stojanovic, B., & Kostic, Z. (2018). Convergence Challenges in Digital Business Environment of Western Balkan Countries. In I. Domazet, M. Radović-Marković, & A. Bradić-Martinović (Eds.), *Digital Transformation New Challenges and Opportunities* (Issue April 2020, pp. 31–52). Silver and Smith Publishers.

Stone, B. (1995). *Successful Direct Marketing Methods*. NTC Business Books.

Stone, B., & Jacobs, R. (2008). *Successful Direct Marketing Methods* (8th editio). McGraw Hill.

Stone, M. D., & Woodcock, N. D. (2014). Interactive, direct and digital marketing: A future that depends on better use of business intelligence. *Journal of Research in Interactive Marketing*, *8*(1), 4–17. https://doi.org/10.1108/JRIM-07-2013-0046

Stone, R. N., & Mason, J. B. (1997). Relationship Management: Strategic Marketing's Next Source of Competitive Advantage. *Journal of Marketing Theory and Practice*, *5*(2), 8–19. https://doi.org/10.1080/10696679.1997.11501761

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning. *Psychol Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973.An

Strömgren, B. (1956). Two-dimensional spectral classification of F stars through photoelectric photometry with interference filters. *Vistas in Astronomy*, *2*(C), 1336–1346. https://doi.org/10.1016/0083-6656(56)90060-5

Sujah, A. M. A., & Rathnayaka, & R. M. K. T. (2019). Mining Profitability of Telecommunication Customers and Customer Segmentation With Novel Data

Mining Approach. *Proceedings of 9th International Symposium (Full Paper), South Eastern University of Sri Lanka, Oluvil. 27th – 28th November 2019*, *November*, 978–955.

Suki, N. M., & Suki, N. M. (2013). Innovation in the High-Tech Economy. In P. Chuan, V. Khachidze, I. Lai, Y. Liu, S. Siddiqui, & T. Wang (Eds.), *Innovation in the High-Tech Economy: Contributions to Economics*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41585-2_12

Sun, M., Chen, Z. Y., & Fan, Z. P. (2014). A multi-task multi-kernel transfer learning method for customer response modeling in social media. *Procedia Computer Science*, *31*, 221–230. https://doi.org/10.1016/j.procs.2014.05.263

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(4), 687–719. https://doi.org/10.1142/S0218001409007326

Svozil, D., Kvasnieka, V., & Pospichal, J. (1997). *Introduction to multi-layer feed-forward neural networks*. *39*, 43–62.

Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *14*(4), 1502. https://doi.org/10.12928/telkomnika.v14i4.3956

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (2nd editio). Pearson.

Tange, R. I., Rasmussen, M. A., Taira, E., & Bro, R. (2017). Benchmarking support vector regression against partial least squares regression and artificial neural network: Effect of sample size on model performance. *Journal of Near Infrared Spectroscopy*, *25*(6), 381–390. https://doi.org/10.1177/0967033517734945

Tapp, A., Whitten, I., & Housden, M. (2014). *Principles of Direct , Database and Digital Marketing* (Fifth edit). Pearson.

Tharwat, A. (2019). Parameter investigation of support vector machine classifier with kernel functions. *Knowledge and Information Systems*, *61*(3), 1269–1302. https://doi.org/10.1007/s10115-019-01335-4

Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables Published by: Econometric Society OF RELATIONSHIPS FOR LIMITED DEPENDENT VARIABLES '. *Econometrica*, *26*(1), 24–36.

Trappey, R. J., & Woodside, A. G. (2005). Consumer responses to interactive advertising campaigns coupling short-message-service direct marketing and TV commercials. *Journal of Advertising Research*, *45*(4), 382–401. https://doi.org/10.1017/S0021849905050476

Tsai, C. Y., & Chiu, C. C. (2004). A purchase-based market segmentation methodology. *Expert Systems with Applications*, *27*(2), 265–276. https://doi.org/10.1016/j.eswa.2004.02.005

Tsiptsis, K., & Chorianopoulos, A. (2010). Data Mining Techniques in CRM: Inside Customer Segmentation. In *Data Mining Techniques in CRM: Inside Customer Segmentation*. https://doi.org/10.1002/9780470685815

University of Cincinatti. (n.d.). *K-means Cluster Analysis*. UC Business Analytics R Programming Guide. https://uc-r.github.io/kmeans_clustering#optimal

Vadrevu, P. K., Suggala, R. K., & Varma, G. T. (2016). Big Data Expansion and Challenges. *International Journal of Engineering Research & Technology*, *4*(34), 1–4.

Valero-Fernandez, R., Collins, D. J., Lam, K. P., Rigby, C., & Bailey, J. (2017). Towards Accurate Predictions of Customer Purchasing Patterns. *IEEE CIT 2017 - 17th IEEE International Conference on Computer and Information Technology*, 157–161. https://doi.org/10.1109/CIT.2017.58

Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research*, *166*(2), 557–575. https://doi.org/10.1016/j.ejor.2004.04.022

Van der Sheer, H. R. (1998). *Quantitative approaches for profit maximization in direct marketing*. University of Groningen, Netherlands.

van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, *6*, 129–140.

Vapnik, V., Golowich, S. E., & Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, 281–287.

Vapnik, V. N. (2010). *The nature of statistical learning theory*. Springer.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Vassiljeva, K., Tepljakov, A., Petlenkov, E., & Netsajev, E. (2017). Computational intelligence approach for estimation of vehicle insurance risk level. *Proceedings of the International Joint Conference on Neural Networks*, *2017-May*, 4073–4078. https://doi.org/10.1109/IJCNN.2017.7966370

Vecchia, M. D., & Peter, M. K. (2018). Marketing automation. In R. Dornberger (Ed.), *Business Information Systems and Technology 4.0* (Vol. 141, pp. 117–130). https://doi.org/10.29235/1818-9857-2021-6-45-48

Vellido, A., Lisboa, P. J. G., & Meehan, K. (1999). Segmentation of the on-line shopping market using neural networks. *Expert Systems with Applications*, *17*(4), 303–314. https://doi.org/10.1016/S0957-4174(99)00042-1

Venkatesan, R., & Kumar, V. (2004). A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. *Journal of Marketing*, *68*, 106–125.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, *38*(3), 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.023

Verhoef, P. C. (2003). Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development. *Journal of Marketing*, *67*(4), 30–45. https://doi.org/10.1509/jmkg.67.4.30.18685

Verhoef, P. C., & Donkers, B. (2001). Predicting Customer Potential Value: an Application in the Insurance Industry. *Decision Support Systems*, *32*(2), 189–199.

Verhoef, P. C., Spring, P. N., Hoekstra, J. C., & Leeflang, P. S. H. (2002). The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. *Decision Support Systems*, *34*(4), 471–481. https://doi.org/10.1016/S0167-9236(02)00069-6

Verhoef, P., Kooge, E., & Walk, N. (2016). Creating Value with Big Data Analytics Introduction : Edwin Kooge. *The Customer Connection*. http://thecustomerconnection.nl/docs/member94427/14092016 Presentatie Edwin Kooge Big Data.pdf

Vernon, J. (2019). *Direct Marketing vs Brand Marketing – Which one?* Marketinghy. https://marketinghy.com/2019/01/direct-marketing-vs-brand-marketing-which-one/

Vest, C. (2013). *Facebook: the new Direct Marketing frontier*. https://3qdigital.com/blog/facebook-the-new-direct-marketing-frontier/

Villanueva, J., Yoo, S., & Hanssens, D. M. (2007). The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth. *Journal of Marketing Research*, *45*(1), 48–59.

Vriens, M., van der Scheer, H. R., Hoekstra, J. C., & Roelf Bult, J. (1998). Conjoint experiments for direct mail response optimization. *European Journal of Marketing*, *32*(3/4), 323–339. https://doi.org/10.1108/03090569810204625

Wang, B., & Pineau, J. (2016). Online Bagging and Boosting for Imbalanced Data Streams. *IEEE Transactions on Knowledge and Data Engineering*, *28*(12), 3353–3366. https://doi.org/10.1109/TKDE.2016.2609424

Wang, C. L. (2021). New frontiers and future directions in interactive marketing: Inaugural Editorial. *Journal of Research in Interactive Marketing*, *15*(1), 1–9. https://doi.org/10.1108/JRIM-03-2021-270

Wang, K. E., Zhou, S., Yang, Q., & Yeung, J. S. M. (2005). Mining Customer Value: From Association Rules to Direct Marketing. *Data Mining and Knowledge Discovery*, *11*(1), 57–79. https://link.springer.com/article/10.1007/s10618-005-1355-x

Wang, K., & Lan, H. (2020). Robust support vector data description for novelty detection with contaminated data. *Engineering Applications of Artificial Intelligence*, *91*(April 2019), 103554. https://doi.org/10.1016/j.engappai.2020.103554

Wang, Q. (2013). *Customer selection for direct marketing: bi- objective optimization using support vector machine*. Lingnan University.

Wasson, C. S. (2005). *System analysis, design, and development: concepts, principles, and practices* (Volume 22). John Wiley & Sons.

Watjatrakul, B., & Drennan, J. (2005). Factors Affecting E-Mail Marketing Sourcing Decisions: A Transaction Cost Perspective. *Journal of Marketing Management*, *21*(7–8), 701–723. https://doi.org/10.1362/026725705774538444

Webber, R. (2013). The evolution of direct, data and digital marketing. *Journal of Direct, Data and Digital Marketing Practice*, *14*(4), 291–309. https://doi.org/10.1057/dddmp.2013.20

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations*. Dordrecht: Kluwer Academic.

Wei-jiang, L., Shu-yong, D., Xue, Y., & Xiao-feng, W. (2011). Determination of customer value measurement model RFM index weights. *African Journal of Business Management*, *5*(14), 5567–5572. https://doi.org/10.5897/AJBM11.290

Wei, J.-T., Lin, S.-Y., & Wu, H.-H. (2010). A review of the application of RFM model. *African Journal of Business Management December Special Review*, *4*(19), 4199–4206. http://www.academicjournals.org/AJBM

Winer, R. (2001). A framework for customer relationship management. *California Management Review*, *43*(4), 89–105. https://doi.org/10.2307/41166102

Wolford, B. (n.d.). *What is GDPR, the EU's new data protection law?* GDPR. https://gdpr.eu/what-is-gdpr/

Wong, M. L., Seng, K., & Wong, P. K. (2020). Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain. *Expert Systems with Applications*, *141*(112918).

Wu, C. H., Ho, J. M., & Lee, D. T. (2004). Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems*, *5*(4), 276–281. https://doi.org/10.1109/TITS.2004.837813

Xiahou, J., Xu, Y., Zhang, S., & Liao, W. (2016). Customer profitability analysis of automobile insurance market based on data mining. *ICCSE 2016 - 11th International Conference on Computer Science and Education*, *Iccse*, 603–609. https://doi.org/10.1109/ICCSE.2016.7581649

Xu, D. J. J. (2005). The importance of personalization in affecting consumer attitude toward mobile advertising in China. *The Journal of Computer Information Systems*, *47*(2), 9–19.

Yao, L., & Xiong, J. (2011). Customers segmentation using RFM and two-step clustering. *Advanced Materials Research*, *268–270*, 631–635. https://doi.org/10.4028/www.scientific.net/AMR.268-270.631

Yao, Z., Sarlin, P., Eklund, T., & Back, B. (2014). Combining visual customer segmentation and response modeling. *Neural Computing and Applications*, *25*(1), 123–134. https://doi.org/10.1007/s00521-013-1454-3

Yeh, I. C., Yang, K. J., & Ting, T. M. (2009). Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, *36*(3), 5866–5871. https://doi.org/10.1016/j.eswa.2008.07.018

Yu, C., Zhang, Z., Lin, C., & Jim, Y. (2020). Can data-driven precision marketing promote

user ad clicks? Evidence from advertising in WeChat moments. *Industrial Marketing Management*, *90*(October 2020), 481–492. https://doi.org/10.1016/j.indmarman.2019.05.001

Zahay, D., Mason, C. ., & Schibrowsky, J. . (2009). The Present and Future of IMC and Database Marketing. *International Journal of Integrated Marketing Communications*, *1*(2), 13–30.

Zahedi, L., Mohammadi, F. G., Rezapour, S., Ohland, M. W., & Amini, M. H. (2021). Search Algorithms for Automated Hyper-Parameter Tuning. *ArXiv Preprint ArXiv:2104.14677.*, 1–10. http://arxiv.org/abs/2104.14677

Zeithaml, V. A., Rust, R. T., & Lemon, K. N. (2001). The Customer Pyramid: Creating and Serving Profitable Customers. *California Management Review*, *43*(4), 118–142.

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: methods and applications*. Springer Science & Business Media.

Zhang, F., Deb, C., Lee, S. E., Yang, J., & Shah, K. W. (2016). Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique. *Energy and Buildings*, *126*, 94–103. https://doi.org/10.1016/j.enbuild.2016.05.028

Zhang, F., & O'Donnell, L. J. (2019). Support vector regression. In *Machine Learning: Methods and Applications to Brain Disorders* (pp. 123–140). Elsevier Inc. https://doi.org/10.1016/B978-0-12-815739-8.00007-9

Zhang, X. (2009). *Improving the profitability of direct marketing : a quantile regression approach*. Lingnan University.

Zhang, Y., Kimberg, D. Y., Coslett, H. B., Schwartz, M. F., & Wang, Z. (2014). Multivariate lesion-symptom mapping using support vector regression. *Human Brain Mapping*, *35*(12), 5861–5876. https://doi.org/10.1002/hbm.22590

Zhou, Z.-H. (2012). Ensemble methods: foundations and algorithms. In R. Herbrich & T. Graepel (Eds.), *SEAISI Quarterly (South East Asia Iron and Steel Institute)* (Vol. 13,

Issue 2). Chapman & Hall / CRC Press.

Zhu, G., & Gao, X. (2019). The Digital Sales Transformation Featured by Precise Retail Marketing Strategy. *Expert Journal of Marketing*, *7*(1), 72–76.

Zhu, Z. B., & Song, Z. H. (2010). Fault diagnosis based on imbalance modified kernel Fisher discriminant analysis. *Chemical Engineering Research and Design*, *88*(8), 936–951. https://doi.org/10.1016/j.cherd.2010.01.005

Zou, P., Hao, Y., & Li, Y. (2010). Customer value segmentation based on cost-sensitive learning support vector machine. *International Journal of Services, Technology and Management*, *14*(1), 126–137. https://doi.org/10.1504/IJSTM.2010.032888

Zumstein, D., Oswald, C., Gasser, M., Lutz, R., & Schoepf, A. (2021). *Lead Generation and Lead Qualification Through Data-Driven Marketing in B2B*. https://doi.org/10.21256/zhaw-2402

Biografija

Sunčica Rogić je rođena 23.10.1992. godine u Podgorici, gdje je završila osnovnu i srednju školu. Ekonomski fakultet u Podgorici na Univerzitetu Crne Gore upisala je 2011. godine, a diplomirala je 2015. godine kao jedan od najboljih studenata generacije. Peti semestar je, kao stipendista CEEPUS programa, provela na Univerzitetu za ekonomiju i biznis u Beču (*Wirtschaftsuniversität Wien*). Magistarske studije je, na Ekonomskom fakultetu, završila sa prosječnom ocjenom 10, odbranivši magistarski rad na temu "*Uticaj sponzorstva na percepciju brenda*" u martu 2018. godine. Na jesen iste godine upisala je doktorske studije na Ekonomskom fakultetu u Podgorici.

Nakon završetka osnovnih studija, angažovana je na Ekonomskom fakultetu Podgorici kao saradnik u nastavi, na predmetima iz oblasti poslovne informatike, marketinga, digitalne i međunarodne ekonomije. Autor je preko dvadeset naučnih radova, a učestvovala je u većem broju konferencija, kao i naučnih i profesionalnih projekata, kao dio projektnog tima sa Ekonomskog fakulteta.

Boravila je na brojnim značajnim fakultetima u cilju usavršavanja kroz Erasmus+ program, među kojima se ističu *University of Beira Interior* u Portugalu, kao i *University of Cordoba* i *University of Vigo* u Španiji. Kao gostujući predavač, posredstvom istog programa mobilnosti, održala je predavanja na univerzitetima u Poljskoj, Rumuniji i Letoniji (*University of Applied Sciences Nysa, Rzeszow University of Technology, Alexandru Ioan Cuza University of Iași, Riga Technical University*).

EF

ORIGINALITY REPORT

2%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.comsis.org Internet	1218 words — 1%
2	www.mdpi.com Internet	228 words — < 1%
3	fedora.ucg.ac.me Internet	176 words — < 1%
4	Sunčica Rogić, Ljiljana Kašćelan. "Chapter 30 Customer Response Modeling Using Ensemble of Balanced Classifiers: Significance of Web Metrics", Springer Science and Business Media LLC, 2022 Crossref	158 words — < 1%
5	"New Trends in Databases and Information Systems", Springer Science and Business Media LLC, 2019 Crossref	58 words — < 1%
6	dokumen.pub Internet	53 words — < 1%
7	zir.nsk.hr Internet	35 words — < 1%
8	Stjepan Lakusic. "Influence of structural system on the construction time and cost of residential	26 words — < 1%

projects", Journal of the Croatian Association of Civil Engineers, 2019
Crossref

9	union.edu.rs Internet	23 words — < 1%
10	nardus.mpn.gov.rs Internet	22 words — < 1%
11	repositorio.uchile.cl Internet	22 words — < 1%
12	www.etran.rs Internet	21 words — < 1%
13	infom.fon.bg.ac.rs Internet	19 words — < 1%
14	linkup.rs Internet	19 words — < 1%
15	"A simple formulation for early-stage cost estimation of building construction projects", Journal of the Croatian Association of Civil Engineers, 2021 Crossref	17 words — < 1%
16	Andrés García-Medina, Toan Luu Duc Huynh. "What Drives Bitcoin? An Approach from Continuous Local Transfer Entropy and Deep Learning Classification Models", Entropy, 2021 Crossref	17 words — < 1%
17	biblioteka.tfbor.bg.ac.rs Internet	17 words — < 1%
18	www.docsity.com Internet	13 words — < 1%

19	Caporali de Andrade, Rodrigo. "Data-Driven Operations Management for Multichannel Customer Support Services.", Stevens Institute of Technology, 2020 ProQuest	12 words — < 1%
20	www.ucg.ac.me Internet	12 words — < 1%
21	Antonios Chorianopoulos. "Effective CRM Using Predictive Analytics", Wiley, 2015 Crossref	11 words — < 1%
22	Savanovic, Marija. "PRILOG RAZVOJU METODOLOGIJA IZRADE OPTIMALNIH PROJEKATA LOKALNIH GEODETSKIH MREZA METROA.", University of Novi Sad (Serbia), 2020 ProQuest	11 words — < 1%
23	Vladimir Djurisic, Ljiljana Kascelan, Suncica Rogic, Boban Melovic. "Bank CRM Optimization Using Predictive Classification Based on the Support Vector Machine Method", Applied Artificial Intelligence, 2020 Crossref	11 words — < 1%
24	open.uct.ac.za Internet	11 words — < 1%
25	vbs.rs Internet	11 words — < 1%
26	Melita Jovanović Tončev, Marija Kostić, Vladimir Džamić. "Od tradicionalnog ka elektronskom Word-of-mouth marketingu", Proceedings of the International Scientific Conference - Synthesis 2015, 2015 Crossref	10 words — < 1%

27	Mojicevic, Marija. "Antifungalni potencijal streptomiceta izolovanih iz rizosfera medicinski znaaajnih biljaka: karakterizacija i optimizacija biosinteze staurosporina, produkta metabolizma Streptomyces sp. BV410.", University of Novi Sad (Serbia), 2020 ProQuest	10 words — < 1%
28	commons.und.edu Internet	10 words — < 1%
29	parlamentfbih.gov.ba Internet	10 words — < 1%
30	repozitorij.fsb.unizg.hr Internet	10 words — < 1%
31	severnobacki.okrug.gov.rs Internet	10 words — < 1%
32	vasic.info Internet	10 words — < 1%
33	www.cek.ef.uni-lj.si Internet	10 words — < 1%
34	www.researchgate.net Internet	10 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 10 WORDS