UNIVERZA V LJUBLJANI

EKONOMSKA FAKULTETA

MAGISTRSKO DELO

NAPOVEDOVANJE OSIPA BANČNIH STRANK S POMOČJO SODOBNIH ORODIJ NAPOVEDNE ANALITIKE

ma

Ljubljana, september 2016 MAJA PROSEN

IZJAVA O AVTORSTVU

Podpisana Maja Prosen, študentka Ekonomske fakultete Univerze v Ljubljani, avtorica predloženega dela z naslovom Napovedovanje osipa bančnih strank s pomočjo sodobnih orodij napovedne analitike, pripravljenega v sodelovanju s svetovalcem prof. dr. Markom Pahorjem

IZJAVLJAM

1. da sem predloženo delo pripravila samostojno;

2. da je tiskana oblika predloženega dela istovetna njegovi elektronski obliki;

3. da je besedilo predloženega dela jezikovno korektno in tehnično pripravljeno v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani, kar pomeni, da sem poskrbela, da so dela in mnenja drugih avtorjev oziroma avtoric, ki jih uporabljam oziroma navajam v besedilu, citirana oziroma povzeta v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani;

4. da se zavedam, da je plagiatorstvo – predstavljanje tujih del (v pisni ali grafični obliki) kot mojih lastnih – kaznivo po Kazenskem zakoniku Republike Slovenije;

5. da se zavedam posledic, ki bi jih na osnovi predloženega dela dokazano plagiatorstvo lahko predstavljalo za moj status na Ekonomski fakulteti Univerze v Ljubljani v skladu z relevantnim pravilnikom;

6. da sem pridobila vsa potrebna dovoljenja za uporabo podatkov in avtorskih del v predloženem delu in jih v njem jasno označila;

7. da sem pri pripravi predloženega dela ravnala v skladu z etičnimi načeli in, kjer je to potrebno, za raziskavo pridobil/–a soglasje etične komisije;

8. da soglašam, da se elektronska oblika predloženega dela uporabi za preverjanje podobnosti vsebine z drugimi deli s programsko opremo za preverjanje podobnosti vsebine, ki je povezana s študijskim informacijskim sistemom članice;

9. da na Univerzo v Ljubljani neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravico shranitve predloženega dela v elektronski obliki, pravico reproduciranja ter pravico dajanja predloženega dela na voljo javnosti na svetovnem spletu preko Repozitorija Univerze v Ljubljani;

10. da hkrati z objavo predloženega dela dovoljujem objavo svojih osebnih podatkov, ki so navedeni v njem in v tej izjavi.

V Ljubljani, dne 1.9.2016 Podpis študentke:_____

KAZALO

KAZALO TABEL

KAZALO SLIK

UVOD

V zadnjih letih je v podjetjih prišlo do spremembe v načinu razmišljanja, saj si želijo vpeljati več predvidljivosti v svoje poslovanje in načrtovati prihodnje poslovanje. V ospredje prihaja pogled tako na skupine strank (segmente) kot tudi na posamezne stranke. Zato se vidik, koliko strank smo izgubili v preteklem letu, spreminja v vprašanje, za koliko strank lahko pričakujemo, da jih bomo izgubili v naslednjem obdobju in za katere je najbolj verjetno, da se bo to zgodilo. Spremembe so vidne tudi na področju upravljanja z marketinškimi kampanjami, kjer se dogaja premik iz vprašanja, katere kampanje so bile v preteklosti uspešne, na vprašanje, kateri je naslednji najboljši možni produkt za vsako posamezno stranko ali skupino strank. To so vprašanja, ki imajo za podjetje veliko vrednost in s pomočjo napredne analitike lahko sedaj na podlagi preteklih dogodkov napovedujemo dogajanja v prihodnosti.

Smo v času, ko imamo na razpolago vedno več podatkov o strankah in njihovih navadah. Zaradi količine podatkov pa raziskovanje vzorcev postaja bolj zapleteno, potrebna je boljša infrastruktura ter novi način shranjevanja nestrukturiranih podatkov. Podjetjem postaja jasno, da podatki sami po sebi nimajo neke vrednosti, vendar pa so informacije, ki nosijo v sebi, ko rudarimo po njih, neprecenljive.

Upravljanje z osipom strank je pojem, povezan s prehajanjem strank od enega ponudnika storitev k drugemu. Gre za prepoznavanje tistih strank, ki bi lahko imele namen menjave ponudnika. Na splošno pa se osip strank prikazuje kot izgubo strank. Ko ga identificiramo, lahko te stranke ciljamo s proaktivnimi marketinškimi kampanjami za njihovo zadržanje (Hadden et al., 2007).

Preprečevanje odhoda strank (zadrževanje strank) je postal ‘modni’ pojem v devetdesetih letih, predvsem zaradi del Reichhelda in Sasserja (1990), ki sta prva prikazala koristi zadrževanja strank (Portela & Menezes, 2011). V poslovnem svetu je dobro znano, da je pridobivanje novih strank dražje od ohranjanja obstoječih in da imajo zveste stranke svojo vrednost tudi na drugih področjih, kot je pozitivno nagovarjanje drugih (priporočila) in nižje stroške poslovanja, prav tako so manj občutljive na ceno storitve in skozi čas potrošijo več denarja (Reichheld & Sasser, 1990, Gençer et al., 2014). Poleg tega raziskovalci ugotavljajo, da je zadrževanje strank najboljša osnovna marketinška dejavnost za preživetje v neki industriji (Kim et al., 2004). Zato je izdelava in uporaba natančnih modelov, ki lahko napovejo osip strank, ključnega pomena pri kampanjah, namenjenih zadrževanju strank in maksimiranju dobička. Burez in Van den Poel (2007) sta na primer pokazala, da se dobički lahko podvojijo z uporabo njunih napovednih modelov osipa strank. Veliko dela je bilo zato vloženega v izboljševanje obstoječih metod in razvijanje novih za prepoznavanje strank, za katere je osip bolj verjeten.

Osip strank analizirajo v različnih industrijah. Največkrat je analiziran v telekomunikacijskem sektorju, pogosto tudi na področju bančništva in zavarovalništva, v trgovini na drobno, pri ponudnikih internetnih storitev, v storitveni industriji, založništvu, pri internetnih igrah… (Karnstedt et al., 2010, Hashmi et al., 2013).

Na splošno lahko modele osipa strank razdelimo v dve kategoriji, in sicer v pojasnjevalno in napovedno. V prvi skupini so modeli, ki pojasnjujejo vzorce osipa strank na podlagi različnih (mehkih) gradnikov, kot so marketinške dejavnosti podjetja, poznavanje strank ali dodatni koncepti, kot sta zadovoljstvo in zaznana kakovost storitve (Baumann et al., 2015). Namen napovednih modelov na drugi strani pa je, da z določeno verjetnostjo napovedujejo dogodke v prihodnosti. Ker pojasnjevalna moč ne pomeni tudi napovedne moči, je napovedna analitika potrebna za gradnjo izkustvenih (empiričnih) modelov, ki imajo napovedno moč. Napovedni modeli osipa strank imajo tako tri ključne vidike prepoznavanja strank z veliko verjetnostjo osipa, ki so napovedna natančnost, razumljivost in upravičenost (Verbeke et al., 2011).

V ta namen so bile razvite različne tehnike upravljanja z osipom strank. Najbolj znane tehnike za napovedovanje osipa so odločitvena drevesa, logistična regresija, nevronske mreže in metoda podpornih vektorjev. Raziskovalci pa uporabljajo tudi druge tehnike, kot so analiza preživetja, ansambelska metoda in analiza omrežja za napoved osipa strank.

Magistrska naloga temelji na pregledu širokega nabora literature in teorije na področju osipa strank tako v bančništvu kot tudi drugih dejavnostih. V nalogi iščem najprimernejšo metodo oziroma napovedni model za določanje verjetnosti odhoda posamezne stranke, ki bi se jo dalo implementirati v praksi.

Najprej na podlagi obstoječe literature izberem nekaj metod, ki so po mojem mnenju najprimernejše za določanje osipa strank, s poudarkom na osipu strank v bančništvu. Hiter pregled obstoječe literature na temo napovedovanja osipa strank na splošno pokaže, da so najpogostejše metode odločitvena drevesa, logistična regresija, nevronske mreže, naključni gozdovi in podporni vektorji. Med temi sem v drugem delu izbrala štiri metode, ki sem jih preizkusila na konkretnem primeru napovedovanja osipa strank v srednje veliki slovenski banki.

V empiričnem delu sem najprej pripravila bazo podatkov, ki je sestavljena iz učnega dela, v katerem sem model »naučila« napovedovati osip strank, ter testnega dela, kjer sem s pomočjo modela napovedovala osip.

Sledi izbor atributov, ki so najmočneje povezani z osipom strank. Po izboru atributov pa sem preizkusila različne napovedne modele, ki sem jih na koncu naloge primerjala med

seboj na podlagi karakterističnih krivulj in ostalih kazalnikov ter se na ta način odločila za model, ki v mojem primeru najbolje napoveduje osip strank.

Cilj magistrske naloge je s pomočjo pristopov podatkovne analitike napovedati osip strank z večjo zanesljivostjo kot naključnim (naivnim) izborom. V ta namen sem si zastavila dve raziskovalni vprašanji:

1. Kako dobro lahko s pomočjo transakcijskih in drugih podatkov o strankah napovemo, katera stranka bo zapustila banko.
2. Kateri izmed nabora napovednih modelov najbolje napoveduje osip strank.

Z nalogo sem poskušala dokazati, da je v velikem naboru podatkov, ki jih ima banka o svojih strankah, za določanje osipa strank ključen razmeroma majhen nabor atributov ter da so za določaneje osipa strank potrebna relativno kratkoročna gibanja ter da je napoved verjetnosti odhoda možna le za kratek čas v naprej.

Za dosego cilja sem tako poskušala:

- narediti pregled literature s področja osipa strank;
- narediti pregled različnih napovednih tehnik/modelov za osip strank;
- poiskati, kateri so kazalniki, ki napovedujejo skorajšnji odhod stranke;
- ter na bančnem primeru izgraditi optimalni napovedni model, ki bo omogočal zaznavanje potencialnih odhodov strank.

Omejitve naloge: primerjalni podatki z drugimi bankami na področju individualnih strank niso javno dostopni (varstvo potrošnikov) in tako ne morem ugotavljati, kje je izbrana banka v primerjavi s konkurenco v državi.

1 ZVESTOBA STRANK IN OSIP STRANK

1.1 Zvestoba strank

Stranke so v današnjem svetu vedno mobilnejše in pripravljene, ob primerni ponudbi zamenjajo ponudnika storitve ali produkta. Ponudniki storitev in produktov se zavedajo, da je veliko ceneje stranko obdržati, kot pa na trgu k sebi privabiti novo. Zadovoljstvo in zvestoba strank sta zato v zadnjih letih, ko je v svetu prihajalo do velikih sprememb, postala še toliko pomembnejša. Na trgu, kjer med sabo tekmuje veliko število ponudnikov, je zato zadovoljstvo strank še toliko pomembnejše. Vendar pa zadovoljstvo in zvestoba nista vedno neločljivo povezana. Povezava je lahko asimetrična. Zveste stranke so po navadi zadovoljne, zadovoljstvo stranke pa še ne pomeni njene zvestobe. Zadovoljstvo je

potreben korak do zvestobe, toda postaja manj pomemben, ko se zvestoba začne vzpostavljati tudi skozi druge mehanizme (Oliver, 1999).

Oliver (1997) zvestobo opisuje kot globoko pripadnost k ponovnemu nakupu produkta ali storitve, ki ji dajemo prednost. S tem je opredeljena kot ponavljajoče se nakupovanje iste znamke oziroma pri istem ponudniku skozi neko časovno obdobje, kljub dogajanju na trgu in marketinškim akcijam konkurence, ki bi lahko spremenila obnašanje stranke. Gre za stranko, ki še naprej kupuje pri nekem ponudniku, in ne želi drugega (proti vsem pričakovanjem in ne glede na stroške – največja oblika zvestobe).

Jacoby in Chestnut (1978) sta raziskala zvestobo s psihološkega vidika, pri čemer sta ugotovila, da definicija ponavljajočega se nakupa še ne pomeni nujno zvestobe, prav tako pa tudi občasni nakupi ne pomenijo nujno nezvestobe, saj je stranka lahko zvesta več znamkam ali ponudnikom. Za določanje prave zvestobe je zato treba raziskati še druge vidike, kot so strankine vrednote, vplive in namere znotraj tradicionalnega odnosa. Za to, da lahko rečemo, da je neka stranka res zvesta, mora veljati, da stranka to znamko oziroma ponudnika vrednoti višje od konkurence, to mora sovpadati z odnosom do znamke oziroma ponudnika in stranka mora dati prednost nakupu te znamke oziroma nakupu pri tem ponudniku pred nakupom pri konkurenci.

Zvestoba stranke se, po Oliverju (1997), stopnjuje, zato zanjo poznamo več faz:

- Kognitivna zvestoba je prva faza (zvestoba informaciji, kot je cena ali lastnost). Stranka daje neki znamki ali ponudniku prednost pred drugimi. Temelji zgolj na zaupanju znamki. Zvestoba je še zelo šibka. Če pa se zadovoljstvo stranke nadaljuje, to postane del strankine izkušnje, kar je začetek naslednje faze.
- Afektivna zvestoba (kupim, ker mi je všeč), ki se pojavi, ko je bila stranka večkrat zadovoljna z znamko oziroma ponudnikom. V tej fazi se pojavi zadovoljstvo kot prijetna izpolnitev pričakovanj. Stranka lahko v tej fazi, kljub zadovoljstvu, preide h konkurenci.
- Konotivna zvestoba (zavzemam se za nakup) predstavlja nagnjenost stranke k nakupu specifične znamke ali pri specifičnem ponudniku. Stranka ima željo po ponovnem nakupu, vendar pa ni gotovo, ali bo nakup tudi opravila.
- Zadnja faza je akcijska (dejanska) zvestoba, ko stranka nakup dejansko izvede, ne glede na ovire, ki so morebiti prisotne.

Pri zvestobi naletimo tudi na ovire. Stranka se ne obnaša vedno razumno. Lahko je zvesta več znamkam ali ponudnikom, lahko v celoti preneha uporabljati nek produkt ali storitev (prenehanje pitja, kajenja…) ali pa stranka spremeni svoje navade (ne sedi več doma pred televizijo, ampak se začne ukvarjati s športom) oziroma odraste (nek ponudnik se osredotoča samo na otroke, ne pa več na mladostnike).

Stranko lahko premamijo spodbude okolja, da preide k drugemu ponudniku ali drugi znamki. Zvestoba je lahko nerazumna (iracionalna), kar znajo tekmeci izkoriščati, da jih »ukradejo« drugim. Najlaže je na ta način speljati stranke, ki so kognitivno zveste, ker pri njih ne gre za dejansko zvestobo, ampak usmerjenost na ceno in koristi.

Za prepoznavanje zvestobe strank nekemu podjetju ali produktu moramo najprej postaviti merila, po katerih bomo zvestobo merili na ravni posamezne stranke, ter zagotoviti, da so ta nepristranska in dosledna ter omogočajo primerjavo med različnimi lokacijami oziroma, zaposlenimi. Temu sledi prikaz odgovorov strank o njihovem zadovoljstvu in ugotavljanje, kako zveste stranke ima podjetje. Na koncu pa je na podlagi ugotovitev treba pripraviti primerne strategije za izboljšanje zadovoljstva strank.

Stranke na podlagi zadovoljstva in zvestobe lahko razdelimo na štiri skupine (Jones & Sasser, 1995):

Slika 1: Zadovoljstvo, zvestoba in obnašanje posamezne stranke

	Zadovoljstvo	**Zvestoba**	**Obnašanje**
Apostol/Privrženec	Visoko	Visoka	Ostaja in priporoča
Terorist/prebežnik	Nizko do srednje	Nizka do srednja	Je odšel, odhaja in izraža nezadovoljstvo
Plačanec	Visoko	Nizka do srednja	Prihaja in odhaja, ni zvest
Talec	Nizko do srednje	Visoka	Ujet, ne more menjati ponudnika

Vir:T.O. Jones &W.E. Sasser, Why Satisfied Customers Defect ,1995.

Apostol in privrženec sta temelj podjetja. Njuna pričakovanja podjetje oziroma produkt popolnoma izpolnjuje in zato jima je po navadi lahko ustreči. Tisti privrženci, ki so nadpovprečno zadovoljni in katerih izkušnje presegajo pričakovanja, so apostoli, ki podjetje ali znamko priporočajo od ust do ust.

Prebežniki so stranke, ki so nevtralne ali nezadovoljne in ki imajo velik potencial, da bodo odšle (če še niso). Po navadi so to stranke, ki so bile zadovoljne, ko pa so imele problem ali pritožbo, podjetje te ni znalo pravilno ali pravočasno rešiti (pomen hitrega razreševanja pritožb). Tu so najnevarnejši teroristi, to so tiste stranke, ki so imele slabo izkušnjo in o tem na široko govorijo okolici.

Plačanci so stranke, ki so lahko povsem zadovoljne, vendar niso zveste. Iščejo boljše priložnosti na trgu in menjavajo ponudnike. Običajno pri enem ponudniku ne ostanejo dovolj dolgo, da bi bile zanj profitabilne.

Talci pa so ujeti. Tudi, če jim ponudnik nudi slabe produkte ali storitve, ne morejo oditi (monopolističen trg). In ker ne gredo in ne morejo nikamor, se podjetja največkrat niti ne trudijo, da bi njihovo situacijo izboljšala. Dostikrat se podjetja ne zavedajo, da se lahko njihova situacija hitro spremeni in takrat bodo plačali ceno z velikim in nepričakovanim osipom strank.

1.2 Osip strank

Osip strank (angl. *churn*) predstavlja število strank, ki v nekem obdobju zapustijo podjetje. Beseda *churn* pa ima sopomenke, kot so tresenje, premikanje, gibanje – gre za premikanje strank med ponudniki oziroma prenehanje uporabe nekega produkta ali storitve. Stopnjo osipa strank pa merimo kot delež strank, ki so odšle, v primerjavi s celotno bazo strank v začetku opazovanega obdobja. Obratno vrednost (1 – stopnja osipa strank) pa imenujemo stopnja preživetja.

Razlogi za osip strank so različni in ko govorimo o njih, moramo raziskati različne vidike razmerij s stranko (Jones & Sasser, 1995):

- Dolgoročno popolno zadovoljstvo strank je glavno vodilo do njene zvestobe in na visoko konkurenčnem trgu je to še pomembneje (na primer bančni sektor, telekomunikacije). Majhen padec zadovoljstva strank zelo zmanjša njihovo zvestobo in poveča verjetnost odhoda. Zato mora biti glavna prioriteta podjetja, da popolnoma zadosti potrebam strank.
- Tudi tam, kjer je konkurenčnost majhna, je zadovoljstvo strank ključnega pomena za zvestobo ponudniku ali znamki. Tu imamo opravka z dvema tipoma zvestobe strank: prava zvestoba in lažna zvestoba. O lažni zvestobi govorimo, ko je videti, da so stranke zveste, pa dejansko niso (ko je na primer konkurenca omejena, ko je zaradi cene menjava onemogočena, zelo dobri programi zvestobe). Te stranke ostajajo zveste toliko časa, dokler na trgu nimajo svobodne izbire. Zato je za podjetja vedno pomembno, da znajo določati stranke, ki so jim pomembne in zadostiti vsem njihovim potrebam.
- Slaba kakovost storitev pa ni edini krivec za osip strank niti ni nujno glavni krivec. Podjetje je lahko s svojo ponudbo privabilo napačne stranke (proces pridobivanja novih strank ima pomanjkljivosti) ali pa nima pravega mehanizma za zadržanje strank (najpogosteje po tem, ko so imele neko slabo izkušnjo, podjetje pa je ni znalo odpraviti sebi v prid).
- Različni vidiki nezadovoljstva stranke zahtevajo različne ukrepe, če želimo zmanjšati osip strank. Vedno pa je ključnega pomena zadovoljstvo, ki je sestavljeno iz štirih

delov: osnovne lastnosti produkta ali storitve so takšne, kot jih predstavimo stranki, nudena je osnovna pomoč pri uporabi, vpeljan je sistem reševanja pritožb in nudene so posebne storitve, ki zagotavljajo, da so potrebe posamezne stranke izpolnjene (in s tem dobijo občutek personalizirane ponudbe).

- Pri problemu osipa strank se ne smemo osredotočati zgolj na njihovo zadovoljstvo. S pomočjo raziskave zadovoljstva lahko dobimo pomembne podatke, ne morejo pa dati širine in globine za vodenje strategije podjetja in uvajanja inovacij (novi produkti oziroma storitve). Da bi se podjetje ubranilo konkurence in ohranjalo stranke, mora uporabiti širok nabor raznovrstnih metod, s katerimi lahko izboljša svoj ugled na trgu.

Stranke največkrat, po ugotovitvah različnih raziskav, odhajajo od ponudnika zaradi cene, odnosa do strank (kakovost storitve, reševanje pritožb), dostopnosti ponudnika (enostaven dostop, dobra lokacija, parkirna mesta), pomanjkanja stika s stranko in zagotavljanja njenih potreb. Sem seveda soadajo tudi naravni razlogi, kot je smrt, proti katerim pa smo nemočni.

Ker so stranke dragocene, je treba njihov osip meriti, izračunati verjetnost odhoda posamezne stranke in izvajati akcije, s katerimi osip preprečujemo. Poleg tega pa mora podjetje skrbeti, da živi tiste vrednote, ki jih oglašuje. Kajti le z dobrim prvim vtisom (sprejem stranke), doseganjem in preseganjem pričakovanj strank, neprestanim dodajanjem vrednosti storitvam ali produktom, ki jih podjetje nudi, kot tudi povečevanjem konkurenčne prednosti ter gradnje baze zvestih strank lahko plujemo med čermi osipa strank. Poleg tega mora podjetje neprestano spremljati, kaj se dogaja v okolici, poznati svoj ciljni trg, svoje močne in šibke strani, nuditi informacije ter izobraževati stranke, znati prisluhniti svojim strankam ter vedeti, katere stranke so pripravljeni izgubiti in zakaj.

2 MODELIRANJE OSIPA STRANK

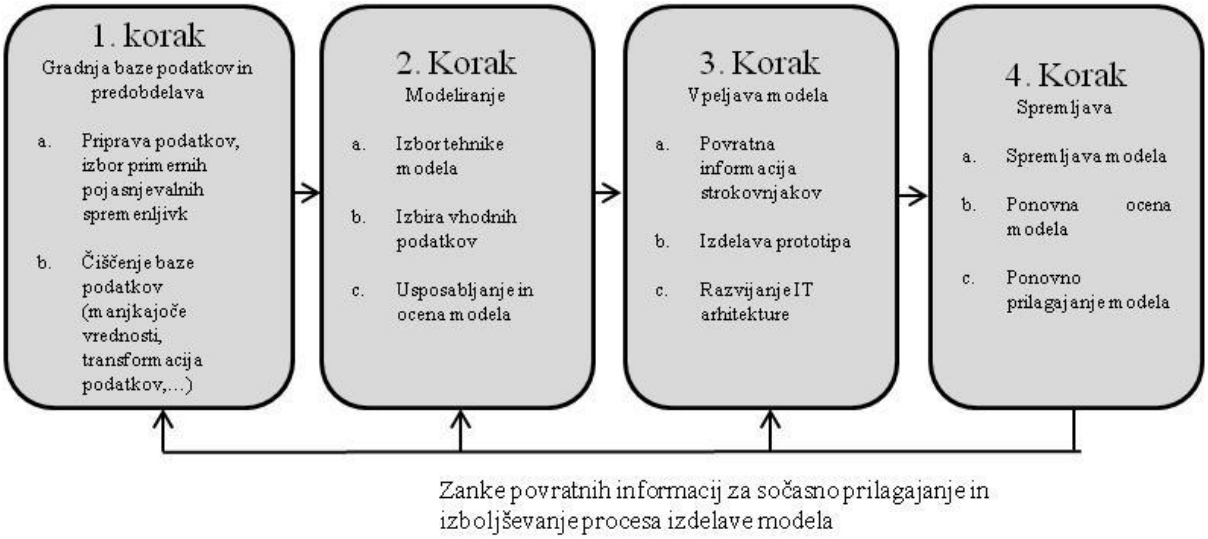
Zadrževanje strank naslavlja problem osipa strank, kjer osip strank opisuje odhod strank, medtem ko upravljanje z osipom strank opisuje trud podjetja, da bi prepoznalo in nadzorovalo problem osipa strank (Hung et al., 2006). Cilj podjetja je prepoznati tveganje osipa strank v zgodnjih fazah, da bi to lahko preprečili. Zlasti pri izjemno visokih letnih stopnjah osipa (20 do 40 %) so podjetja (predvsem se to dogaja v telekomunikacijskih storitvah) željna imeti uspešne programe za preprečevanje osipa strank z namenom maksimiranja svojih prihodkov (Kim et al., 2012). Zato večina ponudnikov telekomunikacijskih storitev uporablja napovedne modele osipa strank (Verbeke et al., 2012a). Preprost razlog za pozornost, ki jo je osip strank deležen je, da osip pomeni izgubo prihodka. Tudi na področju bančništva se takšni modeli pojavljajo vedno pogosteje.

Več študij je pokazalo, da je zadrževanje strank za podjetje precej bolj stroškovno učinkovito, kot privabljanje novih strank (Mozer et al., 2000) ter da privabljanje novih

strank stane petkrat do šestkrat več kot zadrževanje novih (Bhattacharya, 1998, Verbeke et al., 2012a) oziroma celo desetkrat več (Hadden, 2008). Zveste stranke pa so vredne tudi zato, ker so promotorji podjetja med ljudmi ter na dolgi rok nakupujejo več (in s tem generirajo večje prihodke), so cenovno manj občutljive na marketinške dejavnosti tekmecev in je poslovanje z njimi cenejše. Na drugi strani pa izguba stranke vodi v oportunitete stroške (Ganesh et.al., 2000). Zato lahko že majhno izboljšanje pri zadrževanju obstoječih strank znatno poveča dobičke podjetja (Van den Poel & Lariviere, 2004).

Običajni koraki pri modeliranju osipa strank so priprava podatkov in predobdelava, modeliranje, vpeljava modela in njegova spremljava, kot je opisano v Sliki 2.

Slika 2: Modeliranje napovedovanja osipa strank



Vir: W. Verbeke et al., New insights into churn prediction in the telecommunication sector: A profit driven data mining approach, 2012a.

2.1 Podatki in njihova predobdelava

Prvi korak pri razvoju modela osipa strank je ugotavljanje 'najprimernejših' podatkov za vključitev v model. Od kakovosti podatkov je odvisna moč in natančnost modela. Ker imajo različne kombinacije podatkov različno analitično moč, je pomembno, da se določijo tisti, ki so najprimernejši za analizo, ki jo izvajamo.

Različni tipi podatkov se lahko uporabljajo za napovedovanje osipa strank. To so socio-demografski podatki, podatki o imetništvu ter uporabi določenih storitev, zaračunani stroški, informacije družbenih omrežij. Zgodovinski podatki, na primer pogostost uporabe neke storitve, so lahko uporabljeni za uspešno prepoznavo odhajajočih strank dolgo preden

te pokažejo jasne znake odhoda in študije so pokazale, da se te metode dobro uporabljajo v reševanju problemov storitvenega sektorja (telekomunikacije, bančništvo, ponudniki internetnih storitev…) (Ng & Liu, 2001). Prav tako več raziskav predlaga trenutnost oziroma nedavnost (angl. *recency*), pogostost (angl. *frequency*) in finančne spremenljivke kot dobre pokazatelje obnašanja strank (Hadden, et al., 2007, Buckinx & Van den Poel, 2005). Trenutnostne spremenljivke se nanašajo na čas zadnje uporabe neke storitve ali produkta; pogostnostne spremenljivke se nanašajo na to, kako pogosto se neka storitev ali produkt uporablja, medtem ko so finančne spremenljivke povezane s celotno količino denarja, ki ga stranka porabi za neko storitev ali produkt v določenem obdobju. Mnoge raziskave kot dragocene informacije za napovedovanje osipa strank poudarjajo tudi socio-demografske značilnosti, kot so zadovoljstvo stranke, socialno mreženje in okoljske spremembe.

Pomembna stopnja v razvoju modela osipa strank je izbor spremenljivk, kjer gre za proces prepoznave področij, ki so za napoved najboljša oziroma najpomembnejša in hkrati pomagajo pri čiščenju in zmanjševanju podatkov z vključevanjem pomembnih lastnosti in izločevanjem odvečnih šumov in manj pomembnih lastnosti (Hadden et al., 2007). Predobdelava podatkov, kot je izbira vhodnih spremenljivk in vzorčenje, ima velik vpliv na končni rezultat, morda celo večji kot sam izbor tehnike modeliranja (Verbeke, 2012b).

Za namen gradnje in vrednotenja modela sklop podatkov po navadi razdelimo na dva dela: podatke za učenje (angl. *training set)* ter podatke za testiranje modela (angl. *test set*). V tem procesu je potrebno vzeti v obzir dejstvo, da ima večina nastavitev modela za napovedovanja osipa strank popačene razrede: veliko manj je tistih, ki odhajajo, kot tistih, ki ostajajo. To lahko nekaterim tehnikam razvrščanja povzroča težave pri učenju, katere stranke so potencialne za odhod. Zato je včasih treba imeti pretirano velik vzorec (angl. *oversampling*), da se izboljša proces učenja; to pomeni, da je razred, ki je v manjšini, v podatkih za učenje podvojen (Verbeke et al., 2011).

Poleg izbire spremenljivk je pomembna tudi količina podatkov, ki so potrebni za napoved. Na primer, ali je potrebno zbrati podatke o uporabi nekega produkta za celotno obdobje uporabe ali je dovolj, da vzamemo samo nekaj zadnjih obdobij? Jemanje podatkov za stranko za njeno celotno obdobje uporabe produkta lahko na eni strani pripelje do težav pri upravljanju s tako veliko količino podatkov, na drugi strani pa ne vodi nujno do večje natančnosti in uporabnosti modela. Nekateri avtorju predlagajo, da v analizo ni treba vključiti vseh strank, saj so za podjetje največkrat zanimive predvsem stranke, ki prinašajo največjo vrednost, ovrednoteno z denarnim tokom (Lu, 2002).

2.2 Modeliranje: gradnja modela in njegovo ocenjevanje

Obstaja veliko naprednih modelov za napovedovanje osipa strank. Veliko raziskav poroča, da so najpogostejše tehnike odločitvena drevesa, nevronske mreže in logistična regresija (Hadden, et al., 2007, Ngai et al., 2009, Hashimi et.al., 2013). Na primer, Hashimi in njegovi sodelavci (2013) so pregledali raziskave, ki se ukvarjajo z osipom strank, v revijah na svetovnem spletu v obdobju od 2002 do 2013 in ugotovili, da so bila najpogosteje uporabljena odločitvena drevesa (v 25 % člankov), medtem ko so bile nevronske mreže in logistična regresija uporabljene v 15 % člankov. Druge priljubljene tehnike so še regresijska analiza, metoda grozdenja, metoda podpornih vektorjev, naključni gozdovi, genetski algoritmi in še mnogo drugih. Raziskave so pokazale, da pri osipu strank izbor metode pomembno vpliva na rezultat in da iz množice statističnih metod, metod rudarjenja po podatkih in metod strojnega učenja ni tako enostavno izbrati eno. Nekateri avtorji so ponudili smernice za podjetja, ki se začenjajo ukvarjati z napovednimi modeli. Neslin in njegovi sodelavci (2006) so kot dobre tehnike za začetek na primer predlagali logistično regresijo in odločitvena drevesa.

Čeprav je bilo modeliranje osipa strank že dobro raziskano, ni splošnega strinjanja glede delovanja tehnik napovednega modela osipa in obstajajo nasprotujoči si rezultati, ko modele med seboj primerjamo (Verbeke et al., 2011). Ker do nasprotovanj pogosto prihaja pri primerjavi različnih napovednih modelov osipa strank, ostaja problem, katero metodo je najbolje izbrati, nerazrešen (Verbeke et.al., 2012a). Nekatere smiselne matrike ocenjevanja modela, ki jih dobimo iz klasifikacijske matrike (angl. *confusion matrix*), so predstavljene v Tabeli 1. To so občutljivost, specifičnost in natančnost. Klasifikacijska natančnosti je prikazana v enačbi (1)

$$\mathrm{CA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

je odstotek opazovanj, ki so pravilno razvrščena. Mnogo avtorjev opozarja, da je klasifikacijska natančnost dvoumen kazalnik, predvsem v primeru skrajnih (ekstremnih) podatkov in je lahko zavajajoč, ker ocenjuje napačno pozitivne (angl. *false positives* – FP) in napačno negativne (angl. *false negatives* – FN) enako in ne upošteva asimetričnosti (angl. *skewness*) razvrstitve podatkov (Hassouna et.al.,2015, Verbeke et.al., 2012). Občutljivost se izračuna kot je navedeno v enačbi (2)

$$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

je delež dejansko pozitivnih, ki so pravilno razvrščeni, specifičnost, ki se izračuna kot je prikazano v enačbi (3)

$$\frac{\mathrm{TN}}{\mathrm{TN+FP}} \tag{3}$$

pa je delež dejansko negativnih, ki so pravilno razvrščeni. Še ena mera, ki se uporablja za ocenjevanje, je vrednost F1 (angl. *F1 score*), ki se izračuna na podlagi občutljivosti in specifičnosti.

Poleg teh kazalnikov se izračunavajo še nekateri, s katerimi ocenjujemo napovedno moč modela. To sta točnost, prikazana v enačbi (4), ki nam pove, kakšen je delež pravilnih napovedi v celotnem vzorcu. Kazalnik je dober pokazatelj, ko imamo uravnotežen model. V primeru neuravnoteženega modela (ko je, če govorimo o osipu, strank, ki ostajajo, mnogo več kot tistih, ki odhajajo), pa je neuporaben

$$\frac{\mathrm{TP + TN}}{\mathrm{TP + \ TN + FP + FN}} \tag{4}$$

in stopnja FP, prikazana v enačbi (5), ki nam pove, kolikokrat napovemo dogodek, pa do njega dejansko ne pride

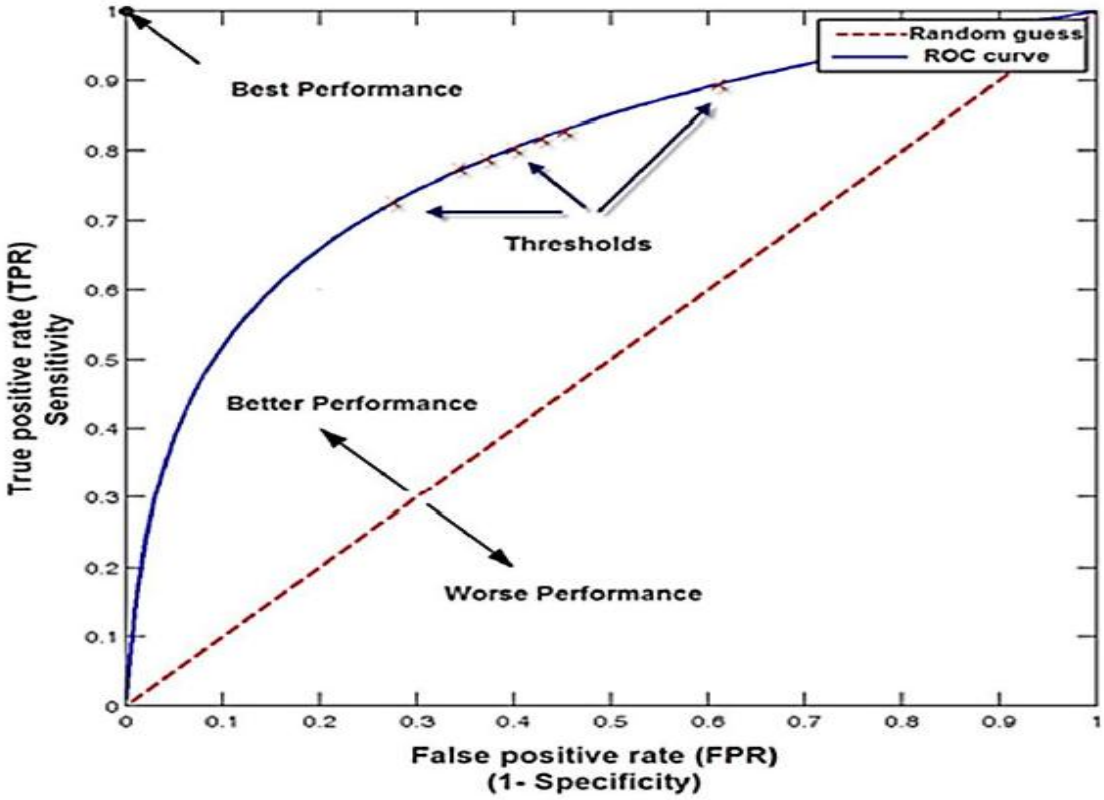
$$\frac{\mathrm{TN}}{\mathrm{FN + TN}} \tag{5}$$

Tabela 1: Prikaz klasifikacijske matrike

		Napovedano stanje	
		P	N
Dejansko stanje	P	Pravilno pozitivne (TP)	Napačno negativne (FN)
	N	Napačno pozitivne (FP)	Pravilno negativne (TN)

Vir: T. Fawcett, An Introduction to ROC Analysis, 2006.

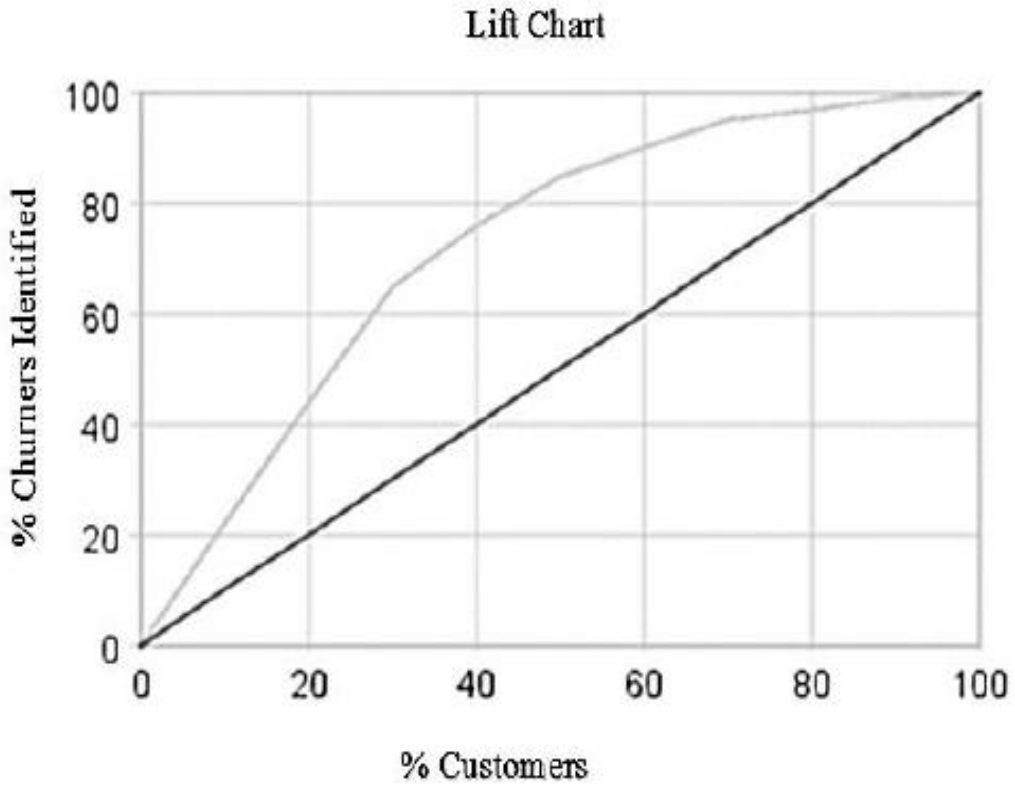
Slika 3: Primer krivulje ROC



Vir: M. Hassouna et. al, Customer Churn in Mobile Markets: A Comparison of Techniques, 2015.

Še en pogosto uporabljeni model za ocenjevanje napovedi osipa strank v literaturi je značilnostna krivulja signala (angl. *reciever operating characteristics* – ROC) in področje pod krivuljo (angl. *area under curve* – AUC). Krivulja ROC, kot je prikazana v Sliki 3, je grafični prikaz pravilno pozitivnih stopenj v primerjavi z nepravilno pozitivnimi. Popolna klasifikacija je točka (0,1), kjer imamo 0 % napačno pozitivnih in 100 % pravilno pozitivnih na grafu. Torej, bližje kot je krivulja ROC tej točki, boljši je model. Izračunavanje področja pod krivuljo ROC (AUC) nam omogoči primerjavo krivulj ROC med različnimi modeli. Vrednosti AUC so v območju med 0 in 1. Za modele z višjo vrednostjo AUC se šteje, da bolje delujejo.

Slika 4: Primer grafikona dviga



Vir: M. Hassouna et. al, Customer Churn in Mobile Markets: A Comparison of Techniques 2015.

Ocenjevalna matrika, ki je podobna krivulji ROC je grafikon dviga (angl. *lift chart*). Grafikon dviga razvršča stranke v decile glede na njihovo verjetnost osipa. Meri odstotek najvišje razvrščenih strank, ki dejansko odhajajo, z odstotkom vseh odhodov v celotni populaciji. Na Sliki 4 je prikazan primer grafikona dviga.

Nekatere druge priljubljene tehnike ocenjevanja in primerjanja modelov so Ginijev koeficient (ki je enak dvakratniku področja med krivuljo ROC in diagonalo) in Kolmogorov Smirnova statistika, ki nam da maksimalno razdaljo med krivuljo ROC in diagonalo na določeni presečni vrednosti (Verbeke et al., 2012a). Kot so poročali Verbeke in njegovi sodelavci (2012a), se modeli napovedovanja osipa strank običajno ocenjujejo na podlagi statističnih mer uspešnosti, kar povzroči neoptimalno izbiro modela. Zato so predlagali novo mero, ki je osredotočena na dobiček. Izračunava se jo kot dobiček, ki bi ga lahko dobili, če bi vključili optimalen delež strank z najvišjo verjetnostjo odhoda v kampanji za zadržanje strank. Mnogi drugi avtorji prav tako predlagajo mere ocenjevanja modela, ki imajo dobiček kot osnovo (Neslin et al., 2006, Glady et al., 2009, Correa Bahnsen et al., 2015).

3 TEORIJA IZBRANIH MODELOV

Obstaja več tehnik podatkovnega rudarjenja za napovedovanje osipa strank. V vseh primerih gre za metode iz družine metod nadzorovanega učenja (angl. *supervised learning*). V nadaljevanju opisujem nekatere, tradicionalno najpogosteje uporabljene metode: logistično regresijo, diskriminantno analizo, naivni Bayesovski model ter nekatere metode, ki so na svojem pomenu pridobile veljavo v zadnjem desetletju, predvsem po zaslugi velike natančnosti napovedi: metoda podpornih vektorjev ter različne metode ansamblov.

3.1 Logistična regresija

3.1.1 Delovanje modela

Regresijska analiza proučuje odnos med eno odvisno in eno ali več neodvisnimi spremenljivkami (pojasnjevalnimi spremenljivkami). Z njo izračunavamo vrednost odvisne spremenljivke. Uporabljamo jo, ko nas zanima, kolikšna je verjetnost, da bo neka stranka spadala v eno ali drugo skupino.

Glede na število neodvisnih spremenljivk ločimo regresije na bivariatno, kjer imamo opravka z dvema neodvisnima spremenljivkama, in multivariatno regresijo, kjer se uporabi več kot dve neodvisni spremenljivki.

Regresije ločimo tudi na podlagi vrste odvisne spremenljivke na navadno regresijo, kjer je odvisna spremenljivka numerična in ima normalno porazdelitev, in logistično regresijo, kjer je odvisna spremenljivka opisna (kategorična) in porazdeljena binomsko.

Še ena ločitev pa je glede na obliko, in sicer na linearno regresijo, kjer imamo linearno funkcijo in nelinearno regresijo, kjer imamo opravka s kvadratno, kubično, ekspotencialno funkcijo.

V nadaljevanju bom predstavila logistično regresijo, ki se jo pogosto uporablja v napovednih modelih. V številnih aplikacijah, logistična regresija ponuja dobro napovedno uspešnost, razumljive modele in razmerja logaritmov možnosti (angl. *log odds ratio*) za ocenjevanje pojasnjevalnih spremenljivk. Logistična regresija se uporablja za modeliranje dihotomnih izhodnih spremenljivk, na primer: je odšel/ni odšel.

V modelu logistične regresije so logaritmi možnosti rezultata modelirani kot binarna kombinacija napovedovalnih spremenljivk. Če je prisotna več kot ena neodvisna spremenljivka, je možno uporabiti logistično regresijo. Pri logistični regresiji določamo točko reza (angl. *cut off point*), ki nam predstavlja točko, kjer se populacija razdeli na dva

dela. Prav tako je, z uporabo funkcije verjetja (angl. *likelihood ratio*), Waldovega testa ali Score testa možno testirati, ali vključitev nove spremenljivke prinese kakšno dodatno informacijo. Tu ocenjujemo verjetnost, ali bosta napovedan rezultat in dejansko stanje enaka.

Modeliranje s pomočjo logistične regresije je dobro poznano, konceptualno enostavno in pogosto uporabljeno v marketingu (Van den Poel, 2003) ter se zato uporablja za napovedovanje obnašanja stranke. Mnogi avtorji priporočajo modeliranje z logistično regresijo za napovedovanje osipa strank.

Kriteriji za izvajanje logistične regresije so, da imamo na razpolago eno ali več neodvisnih spremenljivk, podatki morajo biti neodvisni, vrednosti, ki jih lahko zavzema odvisna spremenljivka, pa se ne smejo prekrivati (na primer »da«/»ne«) ter obstajati mora linearna zveza med neodvisnimi spremenljivkami in logaritemsko transformirano odvisno spremenljivko.

Model logistične regresije dobimo na podlagi funkcije iz enačbe (6)

$$\ln Yi = \alpha + \beta Xi + \varepsilon i \qquad (6)$$

kjer so:

- lnY – odvisna spremenljivka
- X – neodvisna spremenljivka
- β – ocena koeficienta, kar pomeni, da če se neodvisna spremenljivka spremi za 1, se odvisna spremeni za βXi (če je β pozitivna, je verjetnost dogodka večja, če je β negativna, je manjša)
- α – konstanta
- ε – slučajna napaka

3.1.2 Primeri uporabe v praksi

Athanassopoulos (2000) je uporabil potrditveno faktorsko analizo (angl. *confirmatory factor analysis*), analizo grozdenja (angl. *cluster analysis*) in logistično regresijo za napovedovanje osipa strank v bančnem sektorju (prebivalstvo in mala podjetja (angl. *retail*)). Z uporabo dveh logističnih regresij (eno za prebivalstvo in drugo za mala podjetja) je analiziral povezavo med odločitvijo, da stranka banko zamenja, in različnimi pojasnjevalnimi dejavniki, ki se nanašajo na zadovoljstvo z nudeno storitvijo in socio-demografski profil stranke in pokazal, da so tako stranke prebivalstva kot mali podjetniki v principu zvesti in se za menjavo h konkurenčni banki odločijo šele po nizu negativnih dogodkov, ki so jih bili v banki deležni.

Van den Poel (2003) je za nakupno obnašanje kupcev poštnih storitev predlagal konceptualni model segmentacijskih spremenljivk, obrnjen k stranki. Za to je uporabil podatke o preteklih nakupnih navadah (kateri produkt ali storitev je stranka kupila, kakšno količino, za kakšno ceno, s čim je plačala) in podatke z vprašalnika, na katerega so odgovarjala gospodinjstva (podatki o stranki, zadovoljstvo, nakupi). Njegov model logistične regresije je potrdil, da so vse običajno uporabljene spremenljivke: nedavnost (angl. *recency*), pogostost in vrednost, zelo pomembne pri napovedovanju, kdo bo v naslednjem obdobju izvedel nakup poštnih storitev. Pri tem je bila pogostost najpomembnejša. Prav tako je predlagal, da uporaba plačilnega sredstva, trajanje razmerja in nakupno obnašanje pri poštnih storitvah lahko izboljšajo napovedno moč modela. Model je ovrednotil z uporabo PCC (angl. *Pearson product-moment correlation coefficient*) in meril AUC.

Kim in Yoon (2004) sta uporabila model logistične regresije za določanje osipa naročnikov v telekomunikacijskih storitvah, ki temelji na teoriji diskretne odločitve (angl. *discrete choice theory*), to je na obnašanju stranke v situacijah, kjer se mora odločiti med končno množico možnosti.

Ahn, Han in Lee (2006) sta uporabila multinominalno logistično regresijo za napovedovanje osipa strank v korejskih telekomunikacijskih storitvah. Poleg uporabe zgolj informacij o navadah na nivoju posamezne stranke ter demografskih značilnosti stranke, sta vpeljala še eno multinominalno spremenljivko, ki prikazuje status stranke: aktivni uporabnik, ne-uporabnik in prekinjeno sodelovanje. Z namenom analiziranja vpliva statusa stranke sta izvedla tri logistične regresije. Rezultati kažejo, da nekatere determinante osipa vplivajo na dejanski osip, neposredno ali posredno (ali oboje) prek spremembe statusa stranke.

Lima, Mues in Baesens (2009) so pokazali, kako je znanje o domenah[1] (angl. *domain knowledge*) lahko vključeno v proces rudarjenja (angl. *data mining*) za napoved osipa strank. To se lahko naredi z oceno koeficientnih znakov (angl. *coefficient signs*) v modelu logistične regresije ter z analizo odločitvene tabele, izvožene iz odločitvenega drevesa ali na pravilu temelječih klasifikatorjev (angl. *rule-based classifier*).

Nie in sodelavci (2011) so uporabili dve močni metodi razvrščanja za izgradnjo napovednega modela osipa strank v finančnem sektorju: odločitveno drevo in logistično regresijo. Uporabili so podatke o kreditnih karticah, ki so jih povezali s strankinimi osebnimi podatki, osnovnimi podatki o kartici, informacijah o tveganju in transakcijskimi podatki. V njihovi analizi so prikazali, da se je logistična regresija odrezala bolje kot

[1] Znanje o domenah se nanaša na informacije o posameznih domenah ali podatkih, zbranih iz predhodnih sistemov ali dokumentacije, ali pridobljenih od domenskih strokovnjakov in se lahko uporablja za iskanje smiselnih informacij, ki se lahko uporabijo kot vodilo pri procesu raziskovanja.

odločitveno drevo. Poleg točnosti rezultatov analize so izdelali še merilo za izračun stroškov napačnega razvrščanja z uporabo dveh tipov napak in ekonomske smiselnosti.

Mnogi drugi avtorji so pokazali, da je logistična regresija boljša od drugih metod. Neslin in ostali (2004) so na primer zaključili, da logistično modeliranje prekaša celo bolj prefinjene tehnike, kot je na primer nevronska mreža. Tudi Owczarczuk (2010) je v svojih raziskavah modeliranja osipa strank na podlagi velike množice naročnikov v telekomunikacijskih storitvah prikazal, da so linearni modeli, še posebej logistična regresija, zelo dobra izbira, medtem ko so odločitvena drevesa nestabilna v višjih percentilih krivulj dviga, in ne priporoča njihove uporabe.

3.2 Analiza preživetja (angl. *survival analysis*)

Konceptualno se lahko problem osipa strank skrči na splošni problem binarnega razvrščanja: ali odide ali ne, in model osipa strank se lahko osredotoča na spremenjeno obnašanje stranke. Se pa daje prednost modelom, ki lahko ocenijo tveganje ali stopnjo nevarnosti (angl. *hazard rates*), ki se skozi čas spreminja, ker ti lahko izdelajo natančnejše napovedi kot modeli, ki ne dovoljujejo, da neka spremenljivka zavzame neko drugo vrednost v času (Van den Poel & Lariviere, 2004).

Eden od statističnih pristopov, ki omogoča takšno analizo, je analiza zgodovine dogodkov (angl. *event history analysis* – EHA). EHA imenujemo tudi analiza preživetja, trajanja in zanesljivosti. Analiza preživetja je rastoč, vendar pogosto premalo izkoriščen statistični pristop k preizkušanju teorij dinamike na različnih področjih znanosti (Box-Steffensmeier & Jones, 2004). EHA se ukvarja s časom do nekega dogodka oziroma s časom med dvema stanjema. Cilj EHA je ocena preživetvene (distribucijske) funkcije, primerjava preživetvene funkcije ter iskanje povezav med izidom (čas preživetja) in napovednimi spremenljivkami. Značilnost EHA je prisotnost cenzuriranih podatkov (informacija o času preživetja ni popolna, najpogosteje je desno cenzuriranje – stranka preživi dlje od napovedi za pojav dogodka). Sledenje osipa strank je dober primer preživetvenih podatkov. Preživetveni podatki imajo dve skupni lastnosti, s katerimi je težko ravnati s konvencionalnimi statističnimi metodami: cenzuriranje in časovno-odvisne spremenljivke/kovariate (Lu, 2002).

3.2.1 Delovanje modela

Preživetvena metoda je sestavljena iz več metod analize podatkov, kjer je izhodna spremenljivka čas do dogodka. Dogodek je lahko odhod stranke, nastop bolezni, smrt. Preživetveni čas do nastanka tega dogodka pa se meri v časovnih enotah (ura, teden, leto). V primeru, da za posamezno stranko izračunamo čas do dogodka, to stranko spremljamo določeno obdobje.

Odvisna spremenljivka v modelu je sestavljena iz dveh delov, to sta čas do nastopa dogodka in status dogodka (meri, ali se je dogodek zgodil ali ne). Vrednost preživetvene funkcije v kateremkoli časovnem trenutku (*t*) je prikazana kot delež tistih, ki še niso doživeli dogodka (niso odšli) v tem času. Druga pomembna funkcija je funkcija tveganja, ki opisuje tveganost nastanka dogodka (v našem primeru, da stranka odide) na nekem intervalnem pasu po času *t* pod pogojem, da je stranka preživela do časa *t*.

Funkcija tveganja je bolj intuitivna za uporabo v preživetveni analizi, ker poskuša ovrednotiti trenutno tveganje, da bo stranka odšla v času *t* ob predpostavki, da stranka preživi do časa *t*. Skupna predpostavka v preživetveni analizi je sorazmerna domneva tveganja. To predpostavko pa je potrebno preveriti, ker lahko neopažena heterogenost povzroči zavajajoče ocene. V tem primeru je treba uporabiti modele s heterogenostjo (kot na primer eksponentni gamma model, Weibull-gamma model, Weibullov model z latentnimi razredi).

3.2.2 Primeri uporabe v praksi

Več avtorjev za modeliranje osipa strank predlaga EHA. Bhattacharya (1998) in Bolton (1998) sta naredila enega prvih zapisov, ki se ukvarja z osipom strank z metodo EHA. Bhattacharya je v svojem delu analiziral model s stopnjami tveganja (angl. *hazard rate model*) na arhivskih podatkih, ki se nanašajo na članstvo v umetnostnem muzeju, in našel negativno povezavo med časom, ko je bilo treba obnoviti članstvo in tveganostjo odhoda. Poleg tega pa je ugotovil, da ima znižanje bonitete člana pozitiven vpliv na odhod stranke. Bolton (1998) je razvil in ocenil model trajanja (angl. *duration model*), ki se osredotoča na vpliv strankinega zadovoljstva v mobilnih telekomunikacijskih storitvah. Model je ocenjen kot levo-nagnjena, proporcionalno tvegana regresija (angl. *proportional hazards regression*) s presečnimi in časovnimi serijami podatkov, ki opisujejo uporabnikovo zaznavanje in obnašanje. Rezultati kažejo, da je ocena zadovoljstva kupcev, pridobljena pred kakršnokoli odločitvijo da odidejo ali ostanejo zvesti ponudniku, pozitivno povezana s trajanjem razmerja.

Hoggart in Griffin (2001) sta predstavila nelinearni Bayesovski delitveni model (angl. *Bayesian partition model* – BPM) za kovariate v modelu preživetja in sta model uporabila na velikem naboru bančnih podatkov. Primerjala sta rezultate generaliziranega linearnega modela kovariat in pokazala, da ima BPM boljšo napovedno moč. To sta ocenila z razdelitvijo podatkov v dva dela in naknadno primerjavo napovedi enega seta opazovanj pogojevano na drug set v vsakem modelu. V model so bile vključene demografske spremenljivke (spol in starost) ter podatki o uporabi produkta ali storitve (na primer število kreditov, število kreditnih kartic, tip zavarovanja kredita…).

Lu (2002) je prikazal, kako lahko tehnike preživetvene analize uporabimo v telekomunikacijski industriji za identifikacijo strank, ki imajo veliko tveganje za odhod in za določitev časa, kdaj bodo odšle. V svoji raziskavi je zajel le visoko vredne stranke. Uporabil je sklop demografskih podatkov (spol, zakonski status, prihodki gospodinjstva, kreditne kartice,…), podatke o uporabi storitev (povprečno tedensko število klicev, odstotkovna sprememba v minutah, razmerje med deležem domačih in tujih prihodkov) in kontaktne podatke stranke (klici stranke v klicni center in pošiljanje elektronske pošte stranki).

Van den Poel in Larivière (2004) sta uporabila Coxovo proporcionalno metodo tveganja za raziskavo osipa stran v finančnem sektorju (bančništvo in zavarovalništvo). Uporabila sta podatke o obnašanju strank (lastništvo produktov, tipi storitev, ki jih stranka uporablja (na primer telefonsko bančništvo), čas med posameznimi nakupi), demografske podatke o stranki (starost, spol, izobrazba, kraj bivanja) in makroekonomske napovedi za napoved osipa strank. Ugotovila sta, da imajo na zadrževanje strank velik vpliv demografske karakteristike in okoljske spremembe (na primer število poslovalnic), medtem ko ima obnašanje omejen vpliv na obnašanje o odhodu na glede na število produktov v lasti in pretečenem času med dvema nakupoma. Poleg tega sta identificirala še dve kritični obdobji za potencialni odhod: prvo je nekaj let po tem, ko nekdo postane stranka, drugo pa po več kot 20 letih zvestobe ponudniku.

Larivière in Van den Poel (2004) sta uporabila proporcionalno metodo tveganja za testiranje vpliva navzkrižne prodaje v trenutku, ko ima stranka veliko verjetnost odhoda, da bi ocenila, ali to lahko prepreči odhod. Proučevala sta odhod strank z varčevalnimi in investicijskimi produkti in ustvarila različne kategorije obnašanja strank, ki so odšle, z uvedbo dveh dimenzij: ročnost produkta in vključena tveganja med kapitalom in prihodki. Upoštevajoč te lastnosti produkta sta najprej proučila vpogled v časovnico odhodov s pomočjo ocen Kaplan-Meier in nato z analizo nadaljevala na najbolj alarmantni skupini strank, ki sta jo dobila s pomočjo predhodne raziskovalne analize. Izdelan je bil model tveganja za odkrivanje najprimernejših produktnih skupin za navzkrižno prodajo z namenom zmanjšanja števila odhodov. Hkrati je bil ocenjen multinominalni probit model (regresijski model, kjer ima odvisna spremenljivka le dve možni vrednosti, na primer bo ostal, bo odšel) za raziskavo želja stran glede na lastnosti vključenega produkta in za preverjanje, ali te sovpadajo z ugotovitvami analize preživetja.

Jamal in Bucklin (2006) sta razvila pristop k modelu tveganja, ki napoveduje osip strank med ponudniki televizijskih storitev. Raziskala sta naravo empiričnih oziroma izkustvenih povezav med osipom strank in dejavniki, kot so uporabniška izkušnja s storitvijo, čas za odpravo napake in pravičnost plačila. Pristop uporablja Weibullov model z latentnimi razredi tveganja s časovno-spremenljivimi kovariatami. Model vključuje heterogenost tako

v izhodiščnih verjetnostih tveganja kot tudi v parametrih odziva. Ugotavljata, da se napoved osipa strank bistveno izboljša, ko dodamo heterogenost.

Wong (2011) je uporabil analizo preživetja za modeliranje, koliko časa potrebuje stranka v brezžičnih telekomunikacijskih storitvah, da odide. Opazoval je dve glavni skupini lastnosti strank: demografske lastnosti (starost, lokacija, jezik) in značilnosti obnašanja (menjavanje paketa, status pogodbe, primernost paketa in menjavanje aparata). S pomočjo Coxove regresije je bil zgrajen model, ki ga v marketingu lahko uporabljajo za ugotavljanje segmentov strank, ki so občutljivi za odhod. Analiza funkcije tveganja predlaga marketingu, da morajo posebno pozornost posvečati mladim strankam in tistim iz določene regije. Nadalje je pokazala, da imajo stranke, ki so naredile spremembe v paketu, manjšo verjetnost odhoda, prav tako kot tiste, ki so kupile nov aparat.

Portela in Menzes (2011) sta uporabila analizo preživetja za napoved osipa strank pri fiksnih telekomunikacijskih storitvah. Model je bil ocenjen z veliko kovariatami, ki so vključevale osnovne informacije o stranki, demografske podatke, zgodovinske podatke o uporabi storitev, oznako, ali gre za potencialno stranko v odhajanju, izstavljene račune, naročnino…

Tradicionalne analize preprečevanja osipa strank se opirajo na logistične modele. Nekateri raziskovalci danes trdijo, da dajejo analize preživetja, skupaj z razvojem in izboljšavo statističnih metod, boljšo oceno življenjske vrednosti stranke ker ocenjujejo, kdaj bo stranka odšla, ne samo, ali bo odšla ali ne ter zagotavljajo boljšo uporabo časovno spreminjajočih se makroekonomski spremenljivk (Fu & Hongyuan, 2013). Po drugi strani pa so pomanjkljivosti analize preživetja, da uvedba tega modela ni tako enostavna, kot je uvedba binarnega modela.

Poleg priljubljenih regresijskih metod za napovedovanje osipa strank, logistične regresije in ključne – Coxove regresije, obstaja še več regresijskih metod za modeliranje osipa strank. Madden, Savage in Coble-Neal (1999) so uporabili binarni probit model za povezavo med verjetnostjo odhoda stranke od trenutnega ponudnika interneta na podlagi spremenljivk, kot so ekonomske spremenljivke, uporaba interneta, značilnosti ponudnika, socio-demografske lastnosti.

Kim, Lee in Johnson (2013) so predlagali model osipa strank, ki temelji na delni optimizaciji najmanjših kvadratov, ki izrecno upošteva stroške upravljanja obvladljivih marketinških spremenljivk. Metoda delne optimizacije najmanjših kvadratov (angl. *partial least square method*) je metoda multivariatnega napovednega pristopa, ki upošteva tako različne odgovore kot različne napovedne spremenljivke. Znano je, da je robustna pri podatkih, ki vsebujejo napake v meritvah in kolinearnosti (Kim et al., 2013).

3.3 Diskriminantna analiza

Še ena od metod rudarjenje po podatkih, s katero želimo poiskati linearno kombinacijo merjenih spremenljivk, da bomo imeli kot rezultat skupine, ki so si med seboj čim bolj različne, je diskriminanta analiza. Iščemo tiste podatke, ki nam najbolje pojasnijo razlike med vnaprej določenimi skupinami ter zmanjšajo napako pri uvrščanju enot v te skupine (zagotoviti čim manjše mešanje med enotami posameznih skupin). Primerna je za preučevanje odvisnosti, ko je odvisna spremenljivka opisna, neodvisne spremenljivke pa so številske.

Za pravilno delovanje modela mora biti izpolnjenih nekaj predpostavk:

- Število skupin mora zadostiti pogoju $k \geq 2$.
- Vsaka skupina mora vsebovati vsaj 2 enoti.
- Ne sme priti do multikolinearnosti (ni linearnih kombinacij preostalih spremenljivk).
- Variančno-kovariančna matrika $p \times p$ mora biti v vsaki populacijski skupini približno enaka.
- $p < n–2$, kjer je p število spremenljivk in n število enot v vzorcu.
- Spremenljivke morajo biti normalne ali ordinalne.
- Za ocenjevanje se predpostavlja, da so v vsaki skupini enot spremenljivke slučajno izbrane iz populacije z večrazsežnostno normalno porazdelitvijo spremenljivk.

Predpostavke so lahko kršene, če imamo zagotovljeno dovolj veliko število opazovanih enot ($n > 100$).

Kadar ima odvisna spremenljivka dve vrednosti, govorimo o diskriminantni analizi z dvema skupinama (na primer 0 = stranka ostaja, 1 = stranka odide), če pa je skupin več, pa govorimo o multipli diskriminantni analizi.

V primeru modeliranja osipa strank imamo opravka z diskriminantno skupino z dvema skupinama. Diskriminantno analizo tako lahko zapišemo kot je navedeno v enačbi (7):

$$D = a_1 y_1 + a_2 y_2 + \ldots + a_k y_k \tag{7}$$

kjer so:

- D je vrednost diskriminantne funkcije
- a_k je koeficient diskriminantne funkcije pri spremenljivki y_k
- y_k je k-ta neodvisna spremenljivka

Vrednosti D morajo biti oblikovane tako, da sta si porazdelitvi med seboj čim bolj oddaljeni, kar dosežemo z maksimiranjem razmerja

$$\frac{Variabilnost\ med\ skupinami}{Variabilnost\ znotraj\ skupin}$$

Nato pa je treba ugotoviti, katere spremenljivke največ prispevajo k razlikovanju med skupinami. To naredimo s koeficienti diskriminantne funkcije, kjer nam večja vrednost koeficienta pomeni večji relativni prispevek te spremenljivke k vrednosti diskriminantne funkcije.

3.4 Naivni Bayesovski model

Pri rudarjenju po podatkih lahko uporabljamo najrazličnejše algoritme, njihov cilj pa je enak, in to je čim boljša napovedna moč. Način, kako napovedujemo, pa lahko jemljemo iz različnih multidisciplinarnih tehnik.

Vse od 50-ih let prejšnjega stoletja se veliko uporablja naivni Bayesovski model, ki temelji na Bayesovem teoremu (opisuje verjetnost nastanka nekega dogodka na podlagi pogojev, ki so lahko povezani s tem dogodkom). Tako naivni Bayesovski model izhaja iz statistike in verjetnostne teorije. Algoritem, ki je v uporabi pri tej teoriji, daje vzvod verjetnostnim razmerjem med atributi in oznakam razredov. Model pa včasih nudi naivne domneve o neodvisnosti med atributi, kar pomeni, da ta včasih ni pravilno določena. Po drugi strani pa s svojo robustnostjo in enostavnostjo kompenzira omejitve, ki jih poda predpostavka neodvisnosti.

3.4.1 Delovanje modela

X predstavlja set atributov, znotraj seta pa so posamezni atributi, imenovani X_1, X_2, X_3, ..., X_i. Y pa je dogodek/izid. P(Y) predstavlja verjetnost dogodka, s katero napovedujemo, da se bo nek dogodek zgodil in jo imenujemo zato predhodna verjetnost. P(Y/X) predstavlja pogojeno verjetnost, ki nam pove verjetnost dogodka, ko poznamo vrednost X, imenujemo jo tudi izkustvena verjetnost. Ta nam predstavlja izid, ko spoznamo vrednosti vhodnih atributov (Kotu & Deshpanade, 2015), prikazano v enačbi (8) (Bayesov teorem):

$$P(Y/X) = \frac{P(Y) + P(X/Y)}{P(X)} \tag{8}$$

P(X/Y) je še ena pogojena verjetnost, definirana kot verjetnost, da nek atribut zavzame določeno mesto v oznaki razreda (angl. *class label*) in se jo izračunava iz seta podatkov. P(X) pa nam prikazuje verjetnost dokaza. Ker je enak za vsako vrednost znotraj razreda, ga lahko jemljemo kot konstanto, ki je prikazana v enačbi (9)

(9)

$$P(Y/X) = \frac{P(Y) * \prod_{i=1}^{n} P\,(X_i/Y)}{P(X)}$$

Model je uporaben v skoraj vseh programskih jezikih in njegova izvedba je hitra. Proces izgradnje je podoben tistemu pri odločitvenih drevesih in drugih metodah razvrščanja, lahko pa z njim delamo tako z nominalnimi in numeričnimi atributi.

3.4.2 Omejitve modela

Naivni Bayesovski model ima svoje omejitve. Glede na to, da je enostaven in robusten, mu manjkajoče vrednosti ne predstavljajo problema, če jih imamo v testnem delu podatkov. Če pa se manjkajoča vrednost pojavi v učljivem delu, posteriorna verjetnost zaradi enačbe modela dobi vrednost 0. Da bi se temu izognili, lahko manjkajoče vrednosti, umetno napolnimo z nizkimi privzetimi vrednostmi (Lapazov popravek).

Drugo omejitev predstavljajo manjkajoči podatki v neprekinjenih numeričnih spremenljivkah. Z diskretizacijo jih lahko spremenimo v nominalne (in izgubimo nekaj informacije) in uporabimo Lapazov popravek ali pa uporabimo funkcijo verjetnosti gostote.

Še ena omejitev modela pa je neodvisnost atributov, ki je ena glavnih predpostavk modela, saj ga lahko uporabljamo le za neodvisne atribute. V realnem svetu pa je ta pogoj težko doseči. Še vedno pa model deluje dobro tudi, če so atributi rahlo medsebojno povezani. Preveliki povezanosti med atributi v modelu se lahko izognemo, če že v pripravi podatkov izločimo močno povezane atribute.

3.5 Metoda podpornih vektorjev

Še ena tehnika za napovedovanje osipa strank je metoda podpornih vektorjev (angl. *support vector machines* – SVM). To je robusten in široko uporaben klasifikacijski algoritem v društvu algoritmov za strojno učenje (Gençer et al., 2014), ki temelji na tehnologiji nevronskih mrež z uporabo teorije statističnega učenja. V binarnem okviru razvrščanja, poskuša SVM najti linearno optimalno hiper-ravnino, tako da je stopnja ločitve med pozitivnimi in negativnimi primeri maksimirana (Coussement & Van den Poel, 2006). Pred vpeljavo SVM morajo biti optimizirani številni parametri, da se lahko zgradijo prvo razredni klasifikatorji. To je ključni korak, ko vpeljujemo SVM.

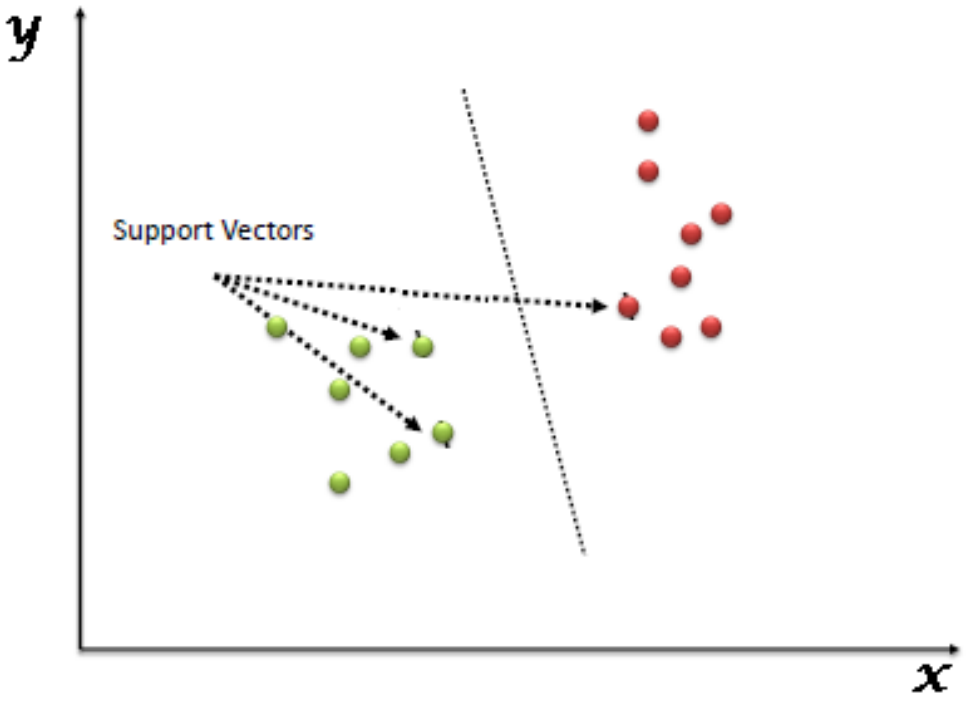
3.5.1 Delovanje modela

Metoda podpornih vektorjev je precej nova metoda strojnega učenja, ki sta jo leta 1995 predstavila Cortes in Vapnik. Med seboj povezuje več znanstvenih področij, statistiko, matematično teorijo optimizacije in računalniško znanje.

SVM se uporablja pri kategorizaciji besedil (rudarjenje po besedilih), kot razvrščevalec slik, za prepoznavo značilnosti pisav, v biologiji in drugih znanostih.

SVM je algoritem nadzorovanega strojnega učenja, ki se ga lahko uporablja tako za regresijo kot tudi za razvrščanje, za kar se ga uporablja pogosteje. Temelji na principu postavljanja meje nekemu območju, znotraj katerega so točke z enakimi lastnostmi. Vsak podatek pokaže kot točko v *n*-dimenzionalnem okolju, ker vrednost koordinate predstavlja posamezno točko. Nato pa s klasifikacijo iščemo tisto hiper-ravnino, kjer se dve skupini razlikujeta med seboj. Opravka imamo z ne-linearno verjetnostno binarno linearno klasifikacijo (angl. *non-probabilistic binary linear classifier*).

Slika 5: Hiper-ravnina z dvema različnima skupinama



Vir: S. Ray, Understanding Support Vector Machine algorithm from examples (along with code), 2015.

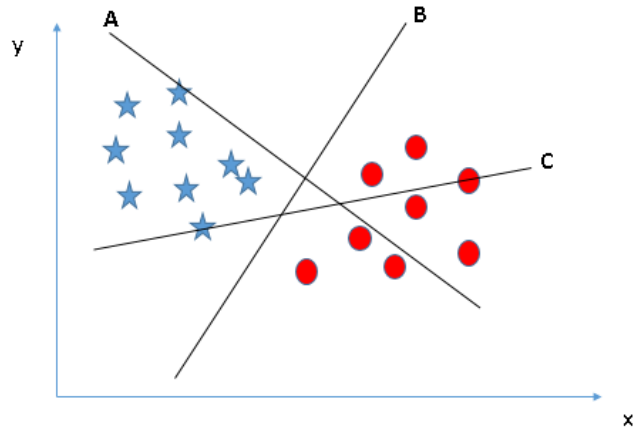
Ko na vzorcu za trening določimo skupine, na podatkih iz testnega vzorca le pogledamo, ali posamezna spremenljivka pade znotraj določene meje ali ne. Ko so meje določene, je večina učljivih podatkov odvečna, saj potrebujemo za določitev in postavitev meje le

jedrne točke. Te podatkovne točke podpirajo naše meje in se zato imenujejo podporni vektorji.

Hiper-ravnino pa lahko določimo na različne načine (Ray, 2015):

- Poišče se hiper-ravnina, ki najbolje loči dva razreda. V spodnjem primeru je to hiper-ravnina B.

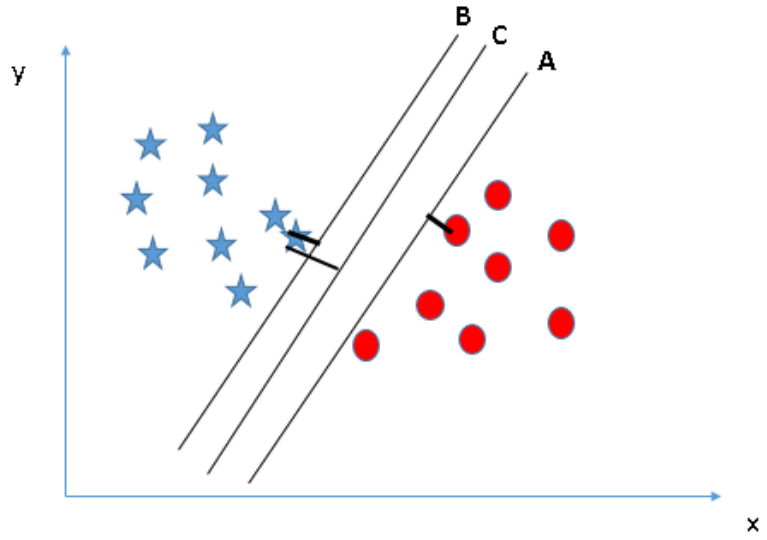
Slika 6: Iskane najboljše premice za ločevanje dveh razredov



Vir: S. Ray, Understanding Support Vector Machine algorithm from examples (along with code), 2015.

- Vse tri hiper-ravnine dobro ločujejo razreda, vendar pa je najboljša tista, ki maksimira razdaljo med najbližjima točkama vsakega razreda (imenovana razlika (Margin)). V spodnjem premeru je to hiper-ravnina C.

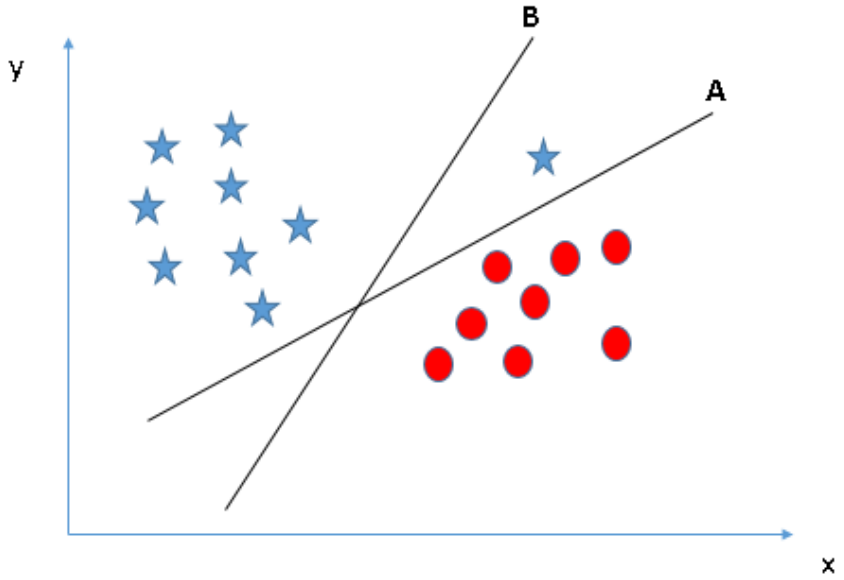
Slika 7: Maksimiranje razdalje med najbližjima točkama vsakega razreda



Vir: S. Ray, Understanding Support Vector Machine algorithm from examples (along with code), 2015.

- SVM izbere hiper-ravnino, ki natančno razvršča skupine preden maksimira maržo. Zato je pravilna hiper-ravnina A, ki ima vse točke pravilno razvrščene.

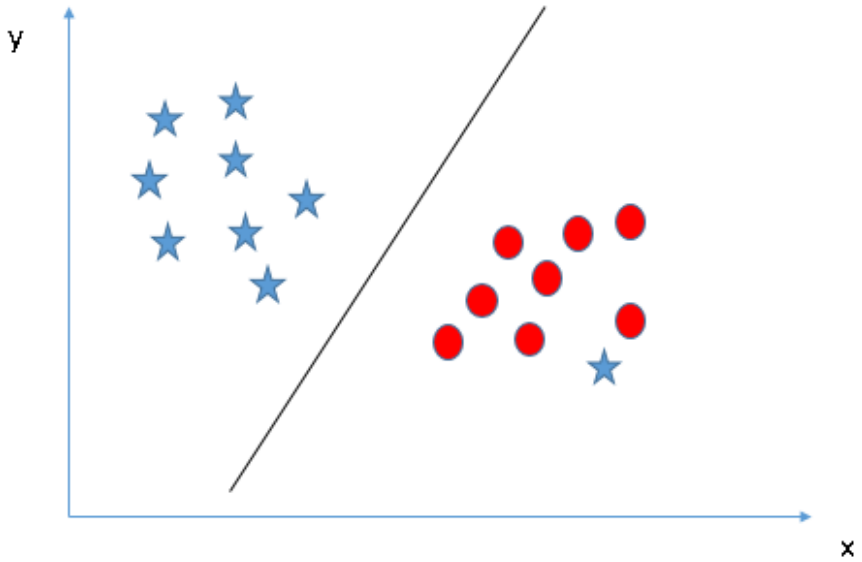
Slika 8: Pravilno razvrščanje točk v razrede



Vir: S. Ray, Understanding Support Vector Machine algorithm from examples (along with code), 2015.

- Tu skupine z ravno črto ne moremo ločiti na dve skupini. Vendar pa je en podatek v tem primeru osamelec in ga metoda SVM izloči ter poišče hiper-ravnino, ki ima maksimirano maržo.

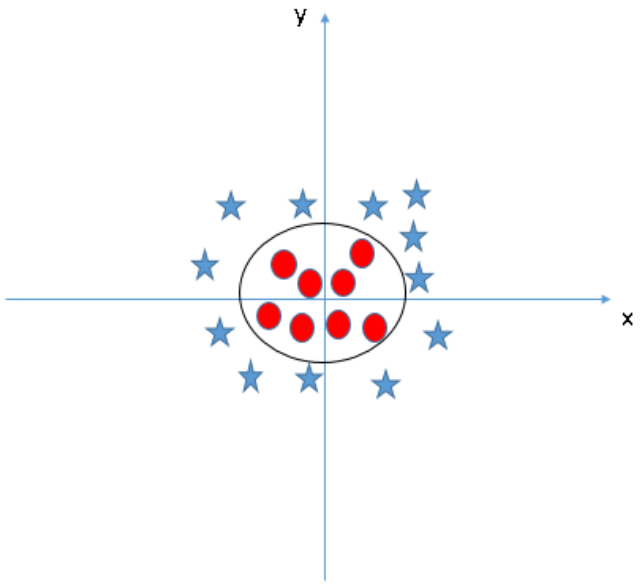
Slika 9: Maksimiranje marže, ko imamo med podatki osamelce



Vir: S. Ray, Understanding Support Vector Machine algorithm from examples (along with code), 2015.

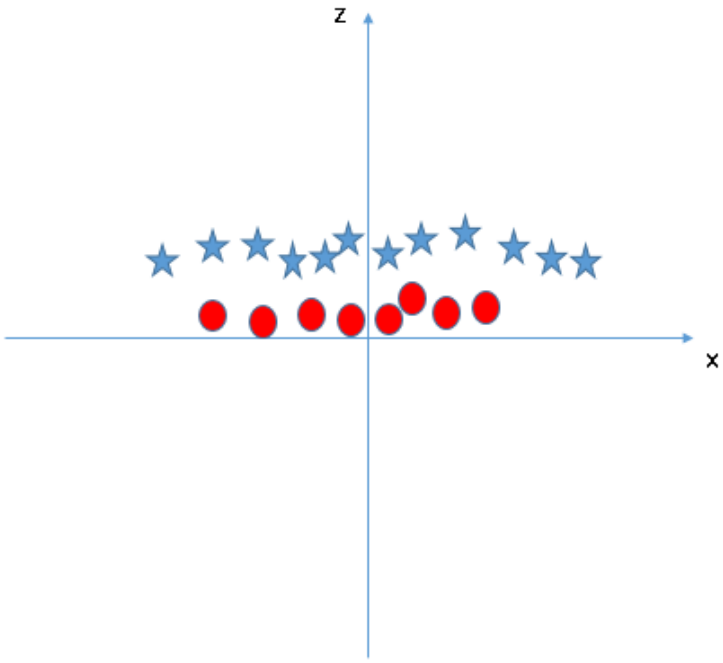
- Ko imamo opravka s tako razvrščenimi podatki, na primer v obliki kroga (Slika 10), SVM uvede dodatno funkcijo (z=x2+y2).

Slika 10: Podatki, razpršeni v obliki dveh krogov



Vir: S. Ray, Understanding Support Vector Machine algorithm from examples (along with code), 2015.

Slika 11: Transformacija podatkov z uvedbo kvadratne funkcije



Vir: S. Ray, Understanding Support Vector Machine algorithm from examples (along with code), 2015.

3.5.2 Omejitve modela

Pri uporabi metode SVM se srečamo z njenimi dobrimi in slabimi lastnostmi. Med dobre štejemo, da dobro deluje pri postavljanju mej ter je učinkovita v večdimenzionalnih prostorih. Prav tako se obnese v primerih, ko je število dimenzij večje, kot je število primerov v vzorcu.

Kar pa se tiče slabosti metode SVM, so te predvsem, da se ne obnese najbolje, ko imamo veliko količino podatkov (dolgotrajnost), prav tako pa ne deluje najbolje, ko imamo v podatkih šume (prekrivanje podatkov). Prav tako nam neposredno ne nudi ocene verjetnosti (do te pridemo šele s petkratno navzkrižno validacijo (angl. *five-fold cross-validation*).

3.6 Metoda ansambla (angl. *ensemble method*)

3.6.1 Delovanje modela

Pri modeliranju podatkov želimo zgraditi model, ki nam bo pojasnjeval razmerja med vhodnimi in izhodnimi spremenljivkami. In kot je to pri odločanju ljudi, ko do boljše odločitve običajno pridejo skupine, je tako tudi pri strojnem učenju. Ansambelske metode so po svoji vsebini učljive metode, ki svojo zmogljivost prikazujejo s tem, da med seboj

kombinirajo različne modele ter jih združijo v enega ter s tem dosegajo dobro učinkovitost oziroma izboljšanje rezultatov. Metode so vzporedne, zaradi česar so učinkovitejše v času treninga modela in testiranja (omejitev: potreben je dostop do več računalniških procesorjev). Ansambelske metode so postale pomembne pri problemih razvrščanja.

Ansambelska metoda je nadzorovan učljiv algoritem, ki med seboj kombinira različne napovedne modele, imenovane osnovni modeli (odločitvena drevesa, logistična regresija…), pri čemer so vhodni podatki za vsakega od modelov enaki, vsak model pa dobi svoj rezultat, ki se na koncu združi v enega. Znane metode, ki se uporabljajo pri ansambelski metodi so glasovanje (angl. *voting*), zlaganje (angl. *stacking*), nalaganje (angl. *bagging*) in povečevanje (angl. *boosting*) ali povečevanje z naklonom (angl. *gradiant boosting*)

Ansambelske metode lahko nudijo visoko natančne rezultate, sploh, kadar kombiniramo raznovrstne oziroma raznolike modele. Njihova pomanjkljivost pa je ta, da so težko razumljive in jih težko uporabimo tam, kjer je razumevanje modela ključno.

3.6.1.1 Glasovanje

Glasovanje je Bayesovski optimalni razvrščevalec (angl. *Bayes optimal classifier*). To je tehnika, ki združuje vse hipoteze, pri čemer je vsaki hipotezi dan glas, glede na verjetnost, da bodo učljivi podatki vzorčeni iz sistema, ki vsebuje to potrjeno hipotezo. Glas vsake hipoteze pa je pomnožen s predhodno verjetnostjo te hipoteze. Odločanje je tu podobno kot pri odločanju v nadzornem svetu ali med porotniki na sodišču, kjer je sprejetje skupne odločitve boljše od izbora posameznikove, ki je dostikrat pristranska.

Pogoj za učinkovitost te metode je, da se glasovi med seboj razlikujejo, to raznolikost pa lahko dobimo na dva načina, in sicer s spremembo algoritma ali pa nabora podatkov. Neposreden pristop je gradnja ansambla z učenjem različnih razvrščevalcev. Kot pri učenju z enim modelom, dobimo set podatkov, vendar pa v tem primeru učimo več različnih razvrščevalcev naenkrat (odločevalsko drevo, nevronske mreže…).

Prednost več razvrščevalcev pri ansambelski metodi je v tem, da se težko zgodi, da bi vsi naredili isto napako. Dokler vsako napako naredi le nekaj razvrščevalcev, dosežemo končno optimalno razvrstitev. Se pa pri tem pojavi slabost, da so različni algoritmi izračunavanja razvrščevalcev nagnjeni k delanju istih napak.

Pri glasovanju se lahko poslužimo še druge možnosti, da dobimo raznolikost podatkov, to je spremembe seta podatkov. Do te lahko pridemo z delitvijo osnovnega vzorca na več

manjših, kjer pa so posamezni sklopi lahko premajhni in so zato rezultati slabši. Da se temu izognemo, lahko uporabimo povečevanje podatkov (angl. *bootstrapping*[2])

3.6.1.2 Nalaganje

Nalaganje je imenovano tudi združevanje povečevanih podatkov (angl. *bootstrap aggregating*) je sestavljeno na način, da ima vsak model v ansamblu enako utež. Zgrajeni so na način, da se vsakemu osnovnemu modelu zamenjajo učljivi podatki (pridobljeni z metodo povečevanja podatkov) in na njihovi podlagi se izdelajo modeli, ki jih na koncu združimo (spreminjanje variance modela). S tem nestabilni modeli pridobijo na svoji stabilnosti.

3.6.1.3 Povečevanje

Povečevanje je postopno vključevanje in treniranje vsakega novega modela v ansambelski metodi z uteževanjem vseh učljivih zapisov. Nudi rešitev, ko imamo šibke modele (angl. *learners*) in jih združuje v enega močnega (ob minimiziranju pristranskosti variance). V nekaterih primerih so njegovi rezultati boljši od rezultatov nalaganja, vendar pa ta pogosteje vodi v preprileganje (angl. *over-fit*) učljivih podatkov.

Za začetek so vsi učljivi zapisi enako uteženi. Utež pa se uporabi za porazdelitev vzorčenja pri izbiri z nadomeščanjem. Neprimerno razvrščeni podatki so bolj uteženi in imajo tako večjo verjetnost, da bodo izbrani v naslednjem krogu, naslednji model pa bo vseboval več podatkov, ki jih je težko razvrstiti. V vsaki ponovitvi se uteži določa na novo in z veliko ponovitvami se razvrsti tako tiste podatke, ki jih razvrstimo enostavno, kot tudi tiste, ki jih težje razvrstimo. Največkrat se za to uporablja AdaBoost (prilagojeno povečevanje oziroma angl. *Adaptive Boositng*), ki je proces strojnega učenja in lahko izboljšuje druge učljive algoritme (s šibkimi členi). AdaBoost je občutljiv na šume v podatkih in osamelce. Velikokrat se ga opisuje kot direktno uporabnega razvrščevalca (angl. *out-of-the-box classifier*).

3.6.1.4 Povečevanje z naklonom

Čeprav teoretična utemeljitev povečevanja z naklonom sega že v pozna devetdeseta leta, pa so se tovrstne metode zaradi različnih razlogov v programski opremi za strojno učenje in podatkovno analitiko pojavile šele v sredini leta 2015. Trenutno so te metode med najobetavnejšimi metodami za nadzorovano učenje (angl. *supervised learning*) za

[2] Ime prihaja iz fraze »To lift himself up by his bootstraps. – Dvigniti samega sebe za zanke na čevljih.«, kar je nemogoče in nesmiselno. To je metoda, kjer z omogočanjem zamenjave podatkov v vzorcu in dopuščanjem ponavljajočih se podatkov, nov vzorec ni enak predhodnemu in je zato njegovo obnašanje drugačno, kar vodi k raznolikosti rezultatov.

strukturirane in razmeroma nizkodimenzionalne podatke, kamor sodijo tudi podatki iz moje naloge.

Povečevanje z naklonom si z ostalimi povečevalnimi metodami deli to, da gre v vseh primerih za ansambel šibkih modelov, združenih v enega močnega. Za razliko od ostalih povečevanj, pri povečevanju z naklonom ne utežujemo enot, temveč optimiziramo funkcijo izgube (angl. *loss function*), če imamo na primer, opravka z regresijo. Prvi model je lahko običajna regresija po metodi najmanjših kvadratov (lahko je celo povsem naivna napoved, da so vse vrednosti povprečne). To je šibek model, saj je v njem (običajno) precej napak. V naslednjem koraku ponovimo model, tokrat na ostankih iz prvega modela, in tako naprej.

V mojem primeru je napovedna spremenljivka diskretna, zato potrebujemo primeren osnovni (šibek) model. Pri povečevanju z naklonom se za osnovni model za diskretne spremenljivke najpogosteje jemlje kar običajna odločitvena drevesa fiksne velikosti.

3.6.1.5 Zlaganje

Zlaganje združuje napovedi več učljivih algoritmov na podlagi usposabljanja učljivega algoritma. Najprej naučimo vse algoritme s podatki, ki jih imamo na razpolago, temu pa sledi učenje združenega algoritma, s pomočjo katerega naredimo končno napoved. Z uporabo poljubnega združevalnega algoritma lahko metodo nalaganja uporabimo pri vseh zgoraj omenjenih ansambelskih tehnikah, dejansko pa se največkrat uporablja z enoplastno (angl. *single-layer*) logistično regresijo.

Zlaganje običajno daje boljše rezultate od posamičnega modela, kot učinkovito pa se je izkazalo tako pri modelih z nadzorovanim učenjem kot tudi pri modelih z nenadzorovanim učenjem.

3.6.1.6 Naključni gozdovi (angl. *random forrest*)

Tehnika naključnih gozdov spada med ansambelske metode za razvrščanje, ki deluje na podlagi izgradnje velikega števila odločitvenih dreves in kot rezultat da povprečno napoved (regresijo) posameznih dreves. Temelji na podobnem konceptu kot metoda nalaganja. Zanjo pa je potrebna velika računalniška moč za treniranje odločitvenih dreves znotraj gozda. Hitro deluje, če gledamo zgolj odločitvene štore oziroma štrclje (angl. *stumps*), če pa gremo globlje po strukturi drevesa, postane kompleksno. Za poenostavitev modela je priročna uporaba odločitvenih dreves s fiksno strukturo in izbiranjem naključnih lastnosti.

Zbirka dreves se imenuje gozd in razvrščevalci, ki so grajeni na tak način, so dobili ime naključni gozdovi. Imajo le tri dele. To so podatki, globina želenega odločitvenega drevesa

in število dreves, ki jih moramo narediti. Lastnosti posameznega drevesa so izbrane naključno, po navadi se en podatek lahko pojavlja večkrat, celo na isti veji. Odločitve so narejene na listih drevesa in temeljijo na učljivih podatkih. Razvrščevalec, ki ga dobimo, je en glas med mnogimi naključnimi drevesi. Metoda deluje dobro, ker ko imamo v gozdu veliko dreves, naključna drevesa izničijo šum v podatkih in s tem dobimo dobro napovedno moč modela.

Model se izvaja v štirih korakih (Kotu & Deshpanade, 2015). Imamo *n* učljivih podatkov in *m* atributov ter *k* število dreves v gozdu. Za vsako drevo naredimo sledeče:

1. Z opcijo nadomeščanja podatkov izberemo *n* zaporedni vzorec (enako kot pri metodi nalaganja).
2. Izberemo *D* (število atributov, ki se jih upošteva pri delitvi vozlišča), kjer je $D >> m$.
3. Začnemo z gradnjo odločitvenega drevesa, kjer za vsako vozlišče vzamemo poljubno število atributov *D* in ta korak ponovimo na vseh vozliščih.
4. Kot pri ansambelski metodi, večja je raznolikost osnovnih dreves, manjša je napaka ansambla.

3.7 Ostale tehnike podatkovnega rudarjenja

Ena najbolj priljubljenih tehnik, ki je postala pomembna v strukturi znanja, ki se uporablja za razvrščanje, je **odločitveno drevo**. Odločitvena drevesa predstavljajo ocenjevanje od zgoraj navzdol. Začne se pri korenskem vozlišču (angl. *rot node*) in gre navzdol s pomočjo ocen različnih spremenljivk, dokler ne pridemo do vozlišča lista (angl. *lean node*), kjer se izvede končno razvrstitev (Lima, Mues & Baesens, 2009). Razvoj odločitvenega drevesa je po navadi sestavljen iz gradnje drevesa (rekurzivna delitev učljivega dela, dokler večina elementov v vsakem delu ne vsebuje enake vrednosti) in obrezovanja drevesa (angl. *tree pruning*) (izbira in odstranjevanje vej, ki vsebujejo največjo ocenjeno napako). Odločitveno drevo raste rekurzivno z deljenjem učljivih zapisov v zaporedne čistejše podsklope, kjer mora minimalno število opazovanj pasti v vsak sklop, sicer se drevo obreže (Verbeke et al., 2012a). Odločitvena drevesa so postala zelo priljubljena zaradi enostavne predstavitve. Še ena prednost odločitvenega drevesa pa je, da lahko upravljajo s kovaratami merjenimi z različnimi merskimi stopnjami.

Eden glavnih problemov z odločitvenimi drevesi je njihova velika nestabilnost. Majhna sprememba v podatkih pogosto povzroči povsem drugačno delitev, kar ni optimalno, ko potrjujemo učljivi model. Ta problem pa lahko rešimo z uporabo tehnike naključnih gozdov (angl. *random forest*). Naključni gozdovi uporabljajo podmnožico *m* naključno izbranih napovednikov (angl. *predictor*) za rast vsakega drevesa na vzorcu s povečanim številom učljivih podatkov. Ko je generirano veliko število dreves, vsako drevo glasuje za najpriljubljenejši razred. Po združevanju teh glasov med različnimi drevesi, vsak primer

dobi napovedano oznako razreda. Tehnika naključnih gozdov je enostavna za vpeljavo, ker je treba določiti le dva prosta parametra: število naključno izbranih napovednikov *m* in število generiranih dreves. V marketingu se je postavljanje naključnih gozdov že pokazalo za boljše od drugih bolj tradicionalnih tehnik razvrščanja (Coussement & Van den Poel, 2006).

Nevronske mreže so struktura za obdelavo podatkov, ki ima sposobnost učenja, kar je koncept, ki temelji na bioloških možganih. Nevronske mreže so zgrajene iz mreže nevronov, med seboj povezanih s funkcijami in utežmi, ki morajo biti ocenjene z umestitvijo mreže v učljive podatke. Z uporabo učljive mreže na atribute stranke, je mogoče dobiti približek njegove posteriorne verjetnosti, da odide (Verbeke et al., 2012a). Nevronske mreže so drugačne od odločitvenih dreves in drugih tehnik razvrščanja, ker nudijo napoved z njeno verjetnostjo. Sčasoma so se pojavili različni pristopi k nevronski mreži (Hadden et al., 2007).

Analiza grozdenja je multivariatna statistična metoda, ki se ukvarja s prepoznavanjem skupin. Cilj analize grozdenja je razdelitev sklopa opazovanj v več skupin ali grozdov (angl. *cluster*), minimizirati variabilnost znotraj grozda in maksimirati variabilnost med grozdi. Obstaja veliko metod grozdenja, najbolj priljubljena pa je verjetno metoda *k*-voditeljev (angl. *k-means*). Ta nehierarhična metoda grozdenja stremi h *k*-razdelitvi opazovanj v *k*-grozde, kjer vsako opazovanje pripada v grozd z najbližjim povprečjem. Število grozdov in mere oddaljenosti morajo biti določene vnaprej.

Razvite so še številne druge tehnike napovedovanja osipa strank, kot so generični algoritmi, algoritmi generiranja pravil (angl. *rule induction classifiers*), ki imajo kot rezultat razumljiv set pravil »if – then« za napovedovanje manjšinskega razreda, medtem ko je večinski razred privzeto določen, diskriminantna analiza, ki je multivariatna metoda. Ta izvira iz linearne kombinacije merjenih spremenljivk, ki razlikujejo med vnaprej določenimi skupinami na način, da so stopnje napačnega razvrščanja minimiziranje. Nekateri avtorji predlagajo združevanje prednosti različnih metod za napovedovanje osipa strank. To je znano pod imenom ansambelska metoda.

3.8 Uporaba različnih napovednih modelov v praksi

Napovedne modele je v zadnjem desetletju v praksi uporabilo veliko avtorjev, ki so se ukvarjali z napovedovanjem osipa strank. Modele so, zaradi količine in dostopnosti podatkov največ izdelovali v sektorju telekomunikacij, sledijo pa bančništvo ter internetni ponudniki. V Tabeli 2 prikazujem pregled raziskav s tega področja, nekatere pomembnejše izsledke pa na koncu opišem.

Tabela 2: Pregled literature napovedi osipa strank

Avtor	Delo	Uporabljena statistična metoda	Sektor
Bhattacharya (1998)	Ko so stranke člani: zadrževanje strank in plačevanje članarine (angl. *When customers are members: customer retention in paid membership contexts*)	Analiza preživetja	Stranke umetnostnega muzeja
Bolton (1998)	Dinamični model trajanja razmerja s stranko pri istem ponudniku storitev: vloga zadovoljstva (angl. *A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction*)	Analiza preživetja	Telekomunikacijske storitve
Madden, Savage in Coble-Neal (1999)	Osip naročnikov pri avstralskem ponudniku internetnih storitev. Informacije, ekonomija in politika (angl. *Subscriber churn in the Australian ISP market. Information Economics and Policy*)	Binomialni probit model	Ponudniki internetnih storitev
Mozer et al. (2000)	Napovedovanje nezadovoljstva naročnikov in preprečevanje odhoda v brezžičnih telekomunikacijskih storitvah (angl. *Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry*)	Logistična regresija, odločitveno drevo, nevronske mreže	Telekomunikacijske storitve
Datta et al. (2006)	Avtomatiziran sistem modeliranja in napovedovanja na velikem obsegu (angl. *Authomated Cellular Modeling and Prediction on a Large Scale*)	Odločitvena drevesa, genetski algoritmi	Telekomunikacijske storitve
Athanassopoulos (2000)	Znamenja zadovoljstva strank v podporo segmentaciji trga in pojasnitvi razlogov za menjavo (angl. Customer satisfaction cues to support market segmentation and explain switching behavior)	Potrditvena faktorska analiza, grozdenje, logistična regresija	Finančne storitve (bančništvo)
Hoggart in Griffin (2001)	Bayesijski model delitve osipa strank (angl. A Bayesovski Partition Model for Customer Attrition)	Bayesovski model delitve v modelu preživetja	Bančništvo na drobno
Van den Poel in Lariviere (2004)	Analiza osipa strank v finančnih storitvah z uporabo proporcionalnih modelov tveganja (angl. Customer attrition analysis for financial services using proportional hazard models)	Analiza preživetja	Finančne storitve (bančništvo in zavarovalništvo)

se nadaljuje

Tabela 2: Pregled literature napovedi osipa strank (nad)

Larivière in Van den Poel (2004)	Preučevanje vloge lastnosti produkta pri preprečevanju osipa strank z uporabo analize preživetja in izbire. Primer finančnih storitev (angl. Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services)	Analiza preživetja, multinomialni probit model	Finančne storitve
Kim in Yoon (2004)	Determinante osipa naročnikov na korejskem trgu mobilne telefonije (angl. *Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market*)	Logistična regresija,	Telekomunikacijske storitve
Buckinx in Van den Poel (2005)	Analiza baze strank: delni osip po obnašanju zvestih strank v nepogodbenih maloprodajnih storitvah (angl. *Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting*)	Logistična regresija, nevronske mreže, naključni gozdovi	Maloprodajne storitve
Hung, Yen in Wang (2006)	Uporaba rudarjenja po podatkih v upravljanju z osipom strank v telekomunikacijah (angl. Applying data mining to telecom churn management)	Odločitveno drevo, nevronske mreže, grozdenje	Telekomunikacijske storitve
Qi et al. (2009)	Model ADTreesLogit za napovedovanje osipa strank (angl. ADTreesLogit model for customer churn prediction)	Logistična regresija, odločitveno drevo	Telekomunikacijske storitve
Wei in Chiu (2002)	Napovedovanje osipa strank s pomočjo podrobnih podatkov o klicih (angl. *Turning telecommunications call details to churn prediction*)	Odločitvena drevesa	Telekomunikacijske storitve
Ahn, Han in Lee (2006)	Analiza osipa strank: determinante osipa strank in mediacijski učinek delnega osipa strank v korejskih mobilnih telekomunikacijskih storitvah (angl. Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry)	Logistična regresija	Telekomunikacijske storitve
Jamal in Bucklin (2006),	Izboljšanje ugotavljanja in napovedovanja osipa strank: heterogeni pristop k modelu tveganja (angl. Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach)	Analiza preživetja	Ponudniki TV storitev

se nadaljuje

Tabela 2: Pregled literature napovedi osipa strank (nad)

Coussement in Van den Poel. (2006)	Napoved osipa strank v naročniških storitvah: uporaba metode podpornih vektorjev pri primerjavi dveh parametričnih izbornih tehnik (angl. Churn Prediction in Subscription Services: an Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques)	Metoda podpornih vektorjev, logistična regresija, naključni gozdovi	Naročniške storitve (časopisno naročništvo)
Burez in Van den Poel (2007)	CRM v podjetju s plačljivo televizijo: uporaba analitičnih modelov za zmanjševanje osipa strank s ciljnimi marketinškimi akcijami za naročniške storitve (angl. CRM at a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services)	Markova veriga, naključni gozdovi, logistična regresija	Naročniške storitve (plačljivi TV programi)
Glady, Baesens, in Croux (2009)	Modeliranje osipa strank z uporabo življenjske vrednosti stranke (angl. Modeling Churn Using Customer Lifetime Value)	Logistična regresija, odločitveno drevo, nevronske mreže	Finančne storitve
Lima, Mues in Baesens (2009)	Vpeljava domenskega znanja v rudarjenje po podatkih z uporabo odločitvenih table: študije primerov napovedi osipa strank (angl. Domain knowledge integration in data mining using decision tables: Case studies in churn prediction)	Logistična regresija, odločitveno drevo	Telekomunikacijske storitve
Abbasimehr et al. (2014)	Primerjalna anliza delovanja ansambelskega učenja pri napovedovanju osipa strank (angl. *A comparative assesment of the performance of ensemble learning in customer churn prediction*)	Ansambelska metoda	Primerjava učinkovitosti modelov na javno dostopnih podatkih neznanega izvora
Nie, Rowe, Zhang, Tian in Shi. (2011)	Napovedovanje osipa strank pri imetnikih kreditnih kartic z uporabo logistične regresije in odločitvenega drevesa (angl. *Credit card churn forecasting by logistic regression and decision tree*)	Logistična regresija, odločitveno drevo	Finančne storitve (bančništvo)
Verbeke, Martens, Mues, Baesens (2011)	Gradnje razumljivega napovednega modela osipa strank s tehniko naprednih indukcijskih pravil (angl. *Building comprehensible customer churn prediction models with advanced rule induction techniques*)	AntMiner+, ALBA (metoda podpornih vektorjev)	Telekomunikacijske storitve

se nadaljuje

Tabela 2: Pregled literature napovedi osipa strank (nad)

Portela in Menezes (2011)	Odkrivanje osipa strank: uporaba neprekinjenih modelov (angl. *Detecting Customer Defections: an application of continuous duration models*)	Analiza preživetja	Telekomunikacijske storitve
Wong, K.K. (2011)	Uporaba Coxove regresije za modeliranje časa do odhoda stranke v brezžičnih telekomunikacijskih storitvah (angl. *Using Cox regression to model customer time to churn in the wireless telecommunication industry*)	Analiza preživetja	Telekomunikacijske storitve
Huang, Kechadi in Buckley (2012)	Napovedovanje osipa strank v telekomunikacijskih storitvah (angl. *Customer churn prediction in telecommunications*)	Logistična regresija, linearni razvrščevalci, naivni Bayesovski model, odločitvena drevesa…	Telekomunikacijske storitve
Kim, Lee In Johnson (2013)	Optimizacija upravljanja z osipom strank z marketinškimi spremenljivkami, ki se jih lahko nadzoruje in pripadajočimi stroški (angl. Churn management optimization with controllable marketing variables and associated management costs)	Metoda najmanjših kvadratov	Telekomunikacijske storitve
Hassouna et.al. (2015)	Osip strank na trgu mobilnih storitev: primerjava tehnik (angl. Customer Churn in Mobile Markets: A Comparison of Techniques)	Logistična regresija, odločitveno drevo	Mobilne storitve
Verbraken et al. (2014)	Dobičkovno optimiziranje osipa strank s pomočjo Bayesovskih mrežnih metod (angl. *Profit optimizing chustomer churn with Bayesian network classifiers*)	Bayesovska mrežna metoda	Primerjava učinkovitosti modelov na umetno ustvarjenih podatkih
Gençer, Bilgin, Zan in Voyvodaoğlu (2014)	Ugotavljanje osipa strank in zadržanih strank mobilnih aplikacij z metodo strojnega učenja (angl. *Detection of Churned and Retained Users with Machine Learning Methods for Mobile Application*)	Metoda podpornih vektorjev	Mobilne aplikacije

se nadaljuje

Tabela 2: Pregled literature napovedi osipa strank (nad)

Baumann, Lessmann, Coussement in De Bock (2015).	Maksimiraj, kar je pomembno: napovedovanje osipa strank z metodo ansambla osredotočeno na odločitve (angl. *Maximize What Matters: Predicting Customer Churn with Decision-Centric Ensemble Selection*)	Metoda ansambla	Telekomunikacijske storitve

Eno od pionirskih del na področju napovedovanja osipa strank je delo Mozerja in drugi (2000), kjer avtorji raziskujejo tehnike s področja statističnega strojnega učenja za napovedovanje osipa strank. Tehnike vključujejo logistično regresijo, odločitvena drevesa, nevronske mreže in metodo povečevanja (algoritem AdaBoost). Poleg tega avtorji določajo, katere spodbude naj bodo ponujene naročnikom, da bi izboljšali stopnjo zadržanja in maksimirali dobičkonosnost lastniku. Njihov poskus temelji na naročniških podatkih telekomunikacijskih storitev, vključujoč informacije o njihovi uporabi storitev, stroških, napravah in zgodovini pritožb.

Datta in njegovi sodelavci (2001) so opisali avtomatiziran sistem modeliranja osipa naročnikov mobilnega omrežja. Uporabili so učljive metode, kot so odločitvena drevesa in genetski algoritmi za izbiro lastnosti, nevronske mreže pa za napovedovanje verjetnosti osipa.

Wei in Chiu (2002) sta predlagala, izdelala in poskusno ocenila tehniko napovedovanja osipa strank v telekomunikacijskih storitvah, ki napoveduje možen osip iz pogodbenih podatkov naročnikov in spremembe vzorca klicev, ki sta ga dobila iz podrobnosti klicev na podlagi odločitvenih dreves.

Buckinx in Van den Poel (2005) sta uporabila logistično regresijo, nevronske mreže in naključne gozdove za napoved delnega odhoda strank, ki so po obnašanju zveste, v nepogodbenem delu prodaje izdelkov za vsakdanjo rabo (angl. *fast moving consumer goods* – FCMG). Nista opazila pomembnih razlik v uspešnosti med različnimi tehnikami razvrščanja.

Hung in sodelavci (2006) so pokazali, da tehnika odločitvenih dreves in nevralnih mrež lahko poda natančne napovedi osipa strank v telekomunikacijskih storitvah z uporabo demografskih podatkov, računov, pogodbenega statusa, klicev in spremembah razmerja. Uporabili so metodo *k*-voditeljev grozdenja za segmentiranje strank v pet segmentov glede na njihovo višino računa, nato pa so ustvarili odločitveno drevo v vsakem segmentu, da so ocenili ali je obnašanje strank v različnih »vrednostnih« segmentih različno. Za segmentacijo so uporabili tudi nevronsko mrežo, ki so jo modelirali z odločitvenim drevesom.

Coussement in Van den Poel (2006) sta uporabila metodo podpornih vektorjev pri naročnikih časopisa, da sta zgradila model osipa strank z visoko napovedno močjo. Ti modeli so vezani na logistično regresijo in naključne gozdove in prikazujejo, da ko uporabimo optimalen postopek pri izbiri parametrov, je metoda podpornih vektorjev boljša od tradicionalne logistične regresije, naključni gozdovi pa se obnesejo boljše od obeh metod podpornih vektorjev.

Burez in Van den Poel (2007) sta uporabila Markovo verigo (angl. *Mark chain*), naključne gozdove in logistično regresijo pri napovedovanju osipa naročnikov plačljive televizije. Opisala sta tudi strategijo preprečevanja osipa v poskusu na terenu.

Qi in sodelavci (2009) so združili modificiran alternativen algoritem odločitvenega drevesa za napovedovanje osipa strank v telekomunikacijski industriji.

Verbeke in sodelavci (2011) so uporabili dve novejši metodi rudarjenja po podatkih za napoved osipa strank: AntMiner+, ki deluje na principu optimizacije kolonije mravelj in dovoljuje vključevanje domenskega znanja z uvedbo monotonih omejitev na končni sklop pravil, in pristop aktivnega učenja (angl. *Active Learning Based Approach* – ALBA), ki združuje visoko napovedno moč modela nelinearnih podpornih vektorjev z razumljivostjo formata postavljanja pravil.

Baumann in sodelavci (2014) so razvili okvir za model osipa strank in poročajo, da njihov model, osredotočen na odločitve, ki temelji na ansambelski izbiri, deluje bistveno bolje kot nekatere predhodne dobro poznane metode (na primer model logistične regresije). Poudarili so, da je problem prevladujočih pristopov k napovedovanju osipa strank, ki temelji na nadzorovanem učenju, v pomembnih omejitvah: analitiku v marketingu ne dovoljuje, da bi med izdelavo modela načrtoval cilje in omejitve kampanj. Njihova metoda vzvodov pa iz obstoječe statistične metode zgradi knjižnico kandidatov ter jo nato preusmeri na poslovni vidik pri iskanju podmnožice modelov, ki so najprimernejši za reševanje odločitvenega problema.

Abbasimehr in sodelavci (2014) so sistematično primerjali delovanje štirih ansambelskih metod in njihovi rezultati so pokazali da je vključevanje ansambelskega učenja prineslo znatno izboljšanje za posamično učenje baze v smislu treh indikatorjev uspešnosti, to so AUC, občutljivost in specifičnost. Zaključili so, da so ansambelske metode lahko dobri kandidati za napovedovanje osipa strank.

Verbraken, Verbeke in Baesens (2014) so ugotovili, da ne glede na to, da obstaja mnogo tehnik napovedovanja osipa strank, je bilo malo pozornosti posvečene uporabi razvrščanja po Bayesovski mrežni metodi (angl. *Bayesian network classifiers*), Njihova študija nudi

primerjavo napovedne moči različnih Bayesovskih algoritmov iz naivnih Bayesovskih razvrščevalcev v splošne Bayesovske mrežne razvrščevalce.

4 PRIPRAVA PODATKOV IN PROCES RUDARJENJA

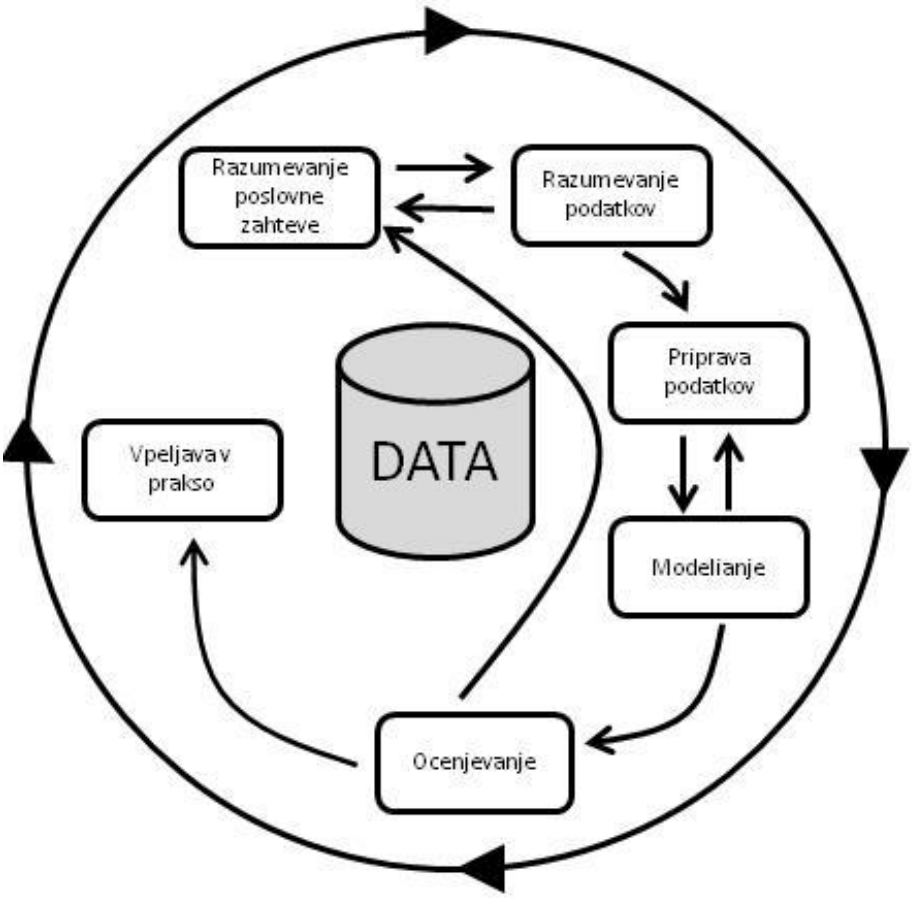
Rudarjenje po podatkih je odkrivanje različnih zanimivih povezav med podatki, ki so shranjeni v podatkovnih skladiščih ter različnih bazah podatkov. S pomočjo rudarjenja iščemo vzorce, povezave, spremembe, odstopanja... Zaradi široke dostopnosti do podatkov prihaja do potrebe, da iz njih pridobimo koristne informacije za marketinške analize, podporo pri odločanju in podjetja se tega vedno bolj zavedajo ter to uporabljajo (Fayyad, U.M. et al., 1996).

Pri rudarjenju po podatkih gre za iskanje znanja v podatkovnih bazah. To pa poteka v več fazah (Kotu & Deshpanade, 2015):

- razumevanje problema, ki ga bomo raziskovali (poznavanje ciljev naloge in poslovnih zahtev ter njihov prenos na področje podatkovnega rudarjenja);
- priprava podatkov (izbira podatkov, čiščenje, združevanje, transformacija,…; izvaja se večkrat tekom celotnega procesa);
- razvoj modela (uporaba različnih napovednih modelov, iskanje najprimernejšega, velikokrat je treba narediti korak nazaj na pripravo podatkov);
- ocenjevanje modela na testnih podatkih (pregled modela in ugotavljanje, ali dosega želene cilje ter odločitev o njegovi nadaljnji uporabi) in
- vpeljava v realni svet (uvedba v produkcijo, uporaba v praksi.

Eden najbolj znanih modelov rudarjenja po podatkih je CRISP-DM (angl. *Cross Industry Standard Process for Data Mining*), ki je bil razvit v sodelovanju več podjetij (Chapman et al., 2000).

Slika 12: Okvir CRISP-DM



Vir: P. Chapman et al., CRISP-DM 1.0: Step-by-step Data Mining guide, 2000.

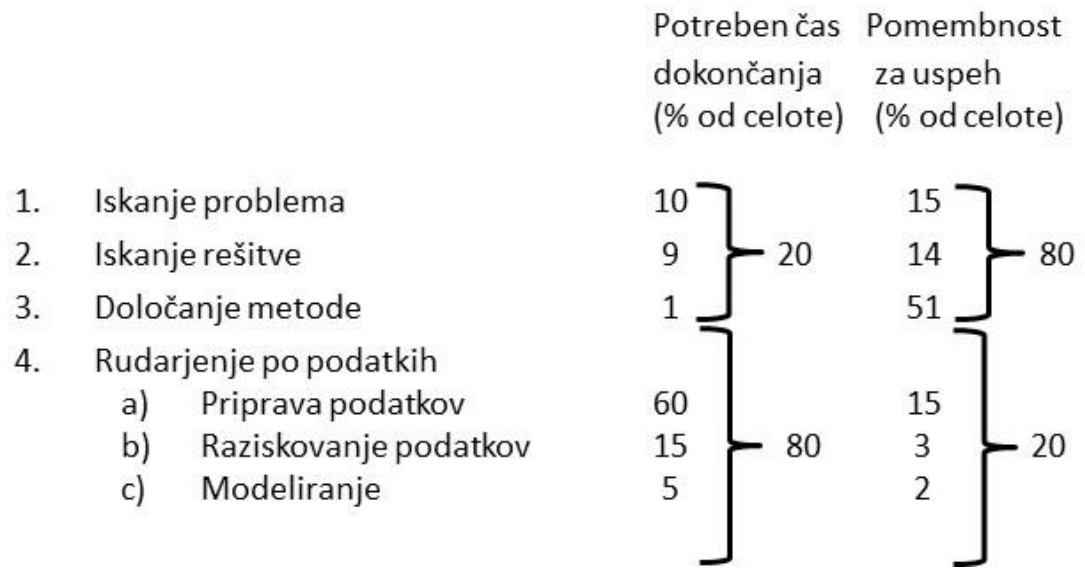
Iz Slike 12 je razvidno, da proces rudarjenja po podatkih ni linearen, ampak se v njem pojavljajo zanke, kjer iz ene faze lahko skačemo nazaj na vse predhodne. Najdolgotrajnejši proces pri rudarjenju po podatkih je priprava podatkov. Postavitev pravilnega poslovnega vprašanja, pridobitev poglobljenega poslovnega znanja o problemu, priprava podatkov in nato razvoj, so ključni za uspešnost procesa rudarjenja po podatkih.

4.1 Razumevanje poslovne zahteve in podatkov

Ker potreba po raziskovanju podatkov nastane iz nečesa, s čimer se spopadamo v realnem svetu moramo, preden se lotimo katerega koli naslednjega koraka v rudarjenju po podatkih, poznati problem, ki ga bomo raziskovali in cilje, ki jih želimo doseči. Proces rudarjenje po podatkih se namreč začne z opredelitvijo problema, ugotovitvijo, kako sovpada z našim poslovanjem, ter katere podatke potrebujemo za njegovo rešitev. Opredelitev problema je najpomembnejši del procesa, kajti brez tega koraka ne moremo najti niti primernih podatkov in posledično kasneje pravega algoritma/modela (Shearer, 2000).

Slika 13 prikazuje, kako je s časovnim okvirom priprave in obdelave podatkov in kaj se dogaja s pomembnostjo posamezne faze. Vidimo lahko, da je pravilno identificiran in opredeljen problem ključ do uspeha celotnega projekta. Sama poraba časa pa je največja v delu pravilne priprave podatkov in uporabe metode raziskovanja. Uvedba v produkcijo pa ne spada v del raziskovalnega projekta, čeprav je to »dokaz o uspehu«. Uvedba namreč zahteva postopkovne in organizacijske spremembe v podjetju.

Slika 13: Postopek priprave podatkov s prikazom porabljenega časa in pomembnosti



Vir: D. Pyle, Data Preparation for Data Mining, 1999.

Največ nerazumevanja in nerealnih pričakovanj izhaja iz opredelitve problema, saj pri rudarjenju po podatkih ne gre za avtomatično iskanje povezav med njimi, kjer se najdejo rešitve in odgovori na vprašanja, za katere še vedeli nismo, da obstajajo. Učinkovito raziskovanje podatkov je pristop, ki zahteva identificiranje poslovnega problema ter nato iskanje podatkov, ki bodo našla odgovor nanj (Pyle, 1999).

Sama opredelitev problema ni vedno enostavna, kajti dogaja se, da tisto, kar smo mislili, da nam v našem poslovnem svetu predstavlja problem, to sploh ni. Z drobljenjem osnovnega vprašanja na manjše gradnike (lahko s pomočjo miselnih vzorcev) se problemu lahko posvetimo z metodo podatkovnega rudarjenja ter na koncu ugotovimo, ali je to dejansko naš osnovni problem ali ga imamo kje drugje. Zato pravilna opredelitev problema že sama po sebi nakazuje tudi smer rešitve.

Ker se v procesu rudarjenja po podatkih srečujemo z različnimi vzorci, je potrebno dobro poznavanje problema raziskovanja, raziskovane vsebine ter podatkov, saj lahko le tako

ločimo napačne vzorce (anomalije) od pravilnih. Rudarjenje po podatkih nam tako da novo znanje, s katerim lahko nadgradimo obstoječega (Lidwell et al., 2010).

K poznavanju problema raziskovanja pa sodi tudi dobro poznavanje in razumevanje podatkov, ki jih bomo uporabili pri raziskavi. To pomeni poznavanje, od kje naši podatki prihajajo, kako so skladiščeni, transformirani, kdaj so nastali ter katere podatke je potrebno pridobiti iz drugih virov. Pri tem je potrebno poznati omejitve podatkov, kot so njihova kakovost, količina, dosegljivost. V tem delu je potrebno pogledati, ali s podatki, ki so nam na voljo, lahko najdemo odgovor na poslovni problem. Model je lahko dober samo toliko, kot so dobri podatki, ki jih imamo na razpolago (»smeti notri, smeti ven«).

Ko pripravljamo podatke, se srečamo z naslednjimi pojmi:

- Set podatkov – zbirka podatkov, ki ima svojo strukturo.
- Posamezno opazovanje (angl. *data point, record*) – posamezen podatek znotraj seta podatkov.
- Atribut – lastnost posameznega nabora podatkov. Lahko je kategorični, numerični, alfanumerični, časovni (status stranke, ročnost kredita).
- Oznaka (angl. *label*) – poseben atribut, ki se ga napove na podlagi ostalih vhodnih atributov (obrestni pribitek)
- Identifikator – dodajajo vsebino posameznim opazovanjem (na primer številka ali ime stranke, številka računa, kartice). Uporabljamo ga za povezovanje različnih setov podatkov med sabo – povezovalni ključ (številka stranke, pod katero jo vodimo v centralnem registru).

Pri naboru in pregledu podatkov pa moramo imeti v mislih tudi medsebojno povezanost med podatki (korelacijo med njimi). In ker povezava med podatki še ne pomeni nujno tudi vzročno-posledične zveze, je poznavanje podatkov in določanje pravilnega problema, ki ga želimo z rudarjenjem po podatkih razrešiti, še toliko pomembneje.

4.2 Priprava podatkov

Priprava podatkov je ključnega pomena in v celotnem procesu rudarjenja po podatkih vzame največ časa, po ocenah ga za ta korak porabimo 60 % do 80 %. Pravilna priprava podatkov pripravi oboje, tistega, ki rudari po podatkih in podatke. Priprava podatkov pomeni, da bo model pravilno zgrajen, priprava tistega, ki bo rudaril po podatkih pa, da bo zgrajen pravi model (Pyle, 1999).

Podatki so redko že na začetku v obliki, kot jo potrebujemo za rudarjenje po podatkih. Zato jih je te treba transformirati v zahtevano strukturo. Ob tem se srečamo z manjkajočimi podatki, ali nepravilnimi podatki.

Priprava podatkov se začne s pregledom podatkov, kjer dobimo občutek, s čim razpolagamo in razumevanje, kaj imamo. Gre za pregled osnovnih statistik (povprečja, mediane, standardni odkloni) in vizualizacijo podatkov, pregledamo, kaj se nam dogaja z ekstremnimi vrednostmi ter pogledamo povezave med posameznimi spremenljivkami.

Srečamo se tudi s kakovostjo podatkov, kjer moramo zagotoviti določeno stopnjo kakovosti, če želimo, da bo obdelava podatkov prinesla želene rezultate. Poleg tega imamo opravka tudi z manjkajočimi vrednostmi, kjer moramo najprej ugotoviti, zakaj do njih prihaja, nato pa lahko z njimi upravljamo na več načinov (lahko jih nadomestimo z neko umetno vrednostjo, te podatke lahko izpustimo iz analize). Potreben je pregled, kje se nam pojavljajo osamelci, saj nam ti lahko predstavljajo napake v podatkih, lahko pa gre zgolj za velika, vendar upravičena, odstopanja.

Znotraj podatkov pa imamo različne tipe spremenljivk. Lahko so numerične (številke), alfanumerične (črke in številke) ali kategorične oziroma ordinalne (opisne). Od izbranega algoritma pa je odvisno, s kakšnimi omejitvami se bomo srečali pri pripravi podatkov in kako bo treba transformirati podatke, ki so nam na voljo. Kategorične podatke lahko spremenimo v numerične, numerične lahko standardiziramo (vrednosti med 0 in 1), zmanjšamo število atributov z metodo glavnih komponent (angl. *principal componenet analysis*).

Ker je na razpolago veliko število raznovrstnih podatkov, moramo za poenostavitev modela in lažje rudarjenje po podatkih, včasih njihovo število zmanjšati s pomočjo metode izbire lastnosti. Iz osnovne baze podatkov nato izberemo vzorec, ki reprezentativno predstavlja osnovno bazo. To nam pohitri postopek in hkrati bistveno ne poslabša rezultata. Pred gradnjo modela podatke razdelimo na dva dela, učno in testno bazo.

Cilj priprave podatkov je, da dobimo bazo podatkov, ki je primerna za modeliranje in hkrati dobro predstavlja osnovno bazo podatkov. Ob tem pa moramo znati odgovoriti na vprašanja, kaj v bazi imamo, kje so pasti v podatkih, na kaj moramo biti pozorni in ali bomo z njo dobili odgovor na postavljeni problem. Ko vse to imamo, lahko začnemo z naslednjim korakom v procesu.

4.3 Izdelava modela

Model nam predstavlja abstrakten prikaz podatkov in povezave med njimi. Z izgradnjo modela želimo dobiti odgovor na predhodno izpostavljen problem. Pred tem pa naredimo analizo, s katero se še bolje spoznamo z vsebino podatkov in njihovimi lastnostmi, kar nam omogoča lažjo izgradnjo modela in boljše razumevanje ter upravljanje z rezultatom, ki ga dobimo. Samo modeliranje pa nam omogoča, da informacije, skrite v podatkih, transformiramo v obliko, ki je razumljiva človeku.

Pri modeliranju je potrebno upoštevati deset zlatih pravil (Pyle, 1999):

1. Izbira jasno definiranih problemov, ki bodo prinesli otipljive koristi.
2. Določitev zahtevane rešitve.
3. Določitev, kako bo rešitev uporabljena.
4. Čim boljše razumevanje problema in podatkov.
5. Problem naj vodi modeliranje (to je izbiro podatkov, izbiro orodja…).
6. Določitev domnev in razprava o njih.
7. Izboljševanje modela s ponavljanjem.
8. Izdelava preprostejšega in razumljivega modela.
9. Določitev nestabilnosti modela (to so področja, kjer majhna sprememba v vhodnih podatkih povzroči veliko spremembo v izhodnih).
10. Določitev negotovosti v modelu (tista področja v podatkih, kjer v modelu prihaja do slabega zaupanja v njegovo napovedno moč).

Model je abstrakten prikaz, kakšen vpliv imajo spremenljivke ena na drugo, kako sprememba ene vpliva na ostale ter s kakšno gotovostjo do prihaja sprememb.

Danes v svetu obstaja veliko različnih algoritmov in modelov ter veliko računalniških programov, s katerimi modele izdelujemo. Modele same pa lahko razvrstimo v različne kategorije: razvrščanje, regresija, analiza asociacij, grozdenja in odkrivanje osamelcev (Kotu & Deshpanade, 2015), kjer ima vsaka od njih na razpolago več algoritmov za reševanje problema.

Modele lahko razvrščamo na aktivne in pasivne. Pasivni so tisti, v katerih se v času nič ne spreminja, nam pa dajejo razumljiv odgovor na »zakaj se nekaj zgodi«. Za razliko od aktivnih modelov, ki se spreminjajo vsakič, ko jim dajemo nove vhodne podatke.

Naslednja delitev je na pojasnjevalne in napovedne modele. Pojasnjevalni modeli nam dajo odgovor na vprašanje, kako in zakaj stvari delujejo in nam dajo dovolj vsebine, da se na njihovi podlagi lahko odločamo. Napovedni modeli pa nam nudijo neko verjetnost nastanka dogodka.

Še ena delitev pa je na statične in učljive modele. Statični so tisti, ki bazirajo na zgodovinskih podatkih, ki jih ne posodabljamo. Ko so enkrat narejeni, se ne spreminjajo. Učljivi modeli pa so dinamični in vsebujejo številne sidrne točke (angl. *set points*), delovanje modela pa neprestano ocenjuje vstopne podatke in spreminja obnašanje modela s spreminjanjem parametrov, da sistem dovolj dobro ohranja sidrne točke. Na podlagi preteklih izkušenj spreminja svoje napovedi. Medtem, ko so podatki za statični model lahko ročno pripravljeni, saj se jih pripravlja le enkrat, pa mora biti za učljivi model priprava podatkov avtomatizirana.

4.4 Vpeljava modela v prakso

Ko je model zgrajen in testiran, je pripravljen, da ga vpeljemo v uporabo v produkcijski sistem podjetja. Tu se srečamo z nekaj fazami.

Prva faza je preverjanje primernosti modela za vpeljavo v produkcijo. Ko preverimo, da se model na testnih podatkih dobro obnaša in smo, če je le mogoče, izvedli nekajmesečno pilotsko testiranje v realnem svetu, sledi odločitev o tehnični uvedbi modela v računalniški sistem podjetja. Za te namene imajo moderni računalniški programi za modeliranje izhodne podatke v obliki datoteke PMML (angl. *Predictive Model Markup Language*), ki je berljiva ostalim računalniškim sistemom in se jo tako lahko neposredno uporabi kot vhodni podatek. Če bodo napovedi temeljile na podatkih v realnem času, je treba poskrbeti, da je odzivnost sistema dovolj velika (kompromis med hitrostjo delovanja in natančnostjo napovedi). Ko imamo model enkrat vpeljan v produkcijo, pa je treba bdeti nad njim ter ga dopolnjevati oziroma posodabljati, da zagotavljamo reprezentativnost in pravilnost delovanja.

5 PRIPRAVA IN PREVERJANJE PODATKOV

5.1 Predstavitev podjetja

Na finančnem trgu v Sloveniji vlada velika konkurenca in vsak ponudnik poskuša na svoj način zadržati ter privabiti nove kupce. Ker se ljudje za zamenjavo banke ne odločamo vsak dan, je razmerje običajno dolgotrajno, še posebej, ko je stranka na banko vezana z dolgoročnim kreditom. Na odločitev, katero banko bo stranka uporabljala, vpliva ugled banke, občutek varnosti, cena in nabor ponujenih storitev, bližina doma ali službe, mreža bankomatov, poleg tega pa tudi priporočila prijateljev ali družine (socialna mreža).

V tej nalogi obravnavam osip strank iz banke s statističnega vidika glede na socio-demografske podatke o stranki ter imetništvu produktov in ne iz vedenjskega ali cenovnega vidika. Najprej na kratko predstavim banko, nato pa opredelim proces priprave podatkov ter način izdelave modelov za napovedovanje osipa stran ter preverim njihovo napovedno moč.

Osip strank v tem primeru pomeni, da stranka preneha poslovati z banko. Kot prenehanje pa štejem, da zaključi svoje poslovanje na način, da zapre vse svoje produkte, ki jih je pri banki imela in to na lastno željo.

Izbrana banka je srednje velika slovenska banka, ki fizičnim in pravnim osebam ponuja širok nabor storitev dnevnega bančništva. Cilj banke je strankam na trgu ponujati konkurenčne produkte po primerljivi ceni ter s tem skrb za poprodajne procese in čim

boljše poznavanje stran in prepoznavanje njihovih potreb. Banka redno spremlja cenovno politiko konkurentov na trgu, prav tako tudi novosti doma in v tujini.

5.2 Priprava podatkov za podatkovno rudarjenje in preverjanje modelov

Pred pričetkom rudarjenja po podatkih in izdelave modelov sem morala pripraviti podatke o primernih strankah ter njihovih produktih za določeno časovno obdobje ter jih shraniti v primerno obliko za nadaljnje delo. Delo je vzelo kar precej časa, saj je bilo treba zbrati primerne podatke, jih pretvoriti v primerno obliko, združevati ali razdruževati. Zaradi poznavanja problema osipa strank in razumevanja podatkov, ki so mi bili na razpolago, je bil izbor nekoliko lažji.

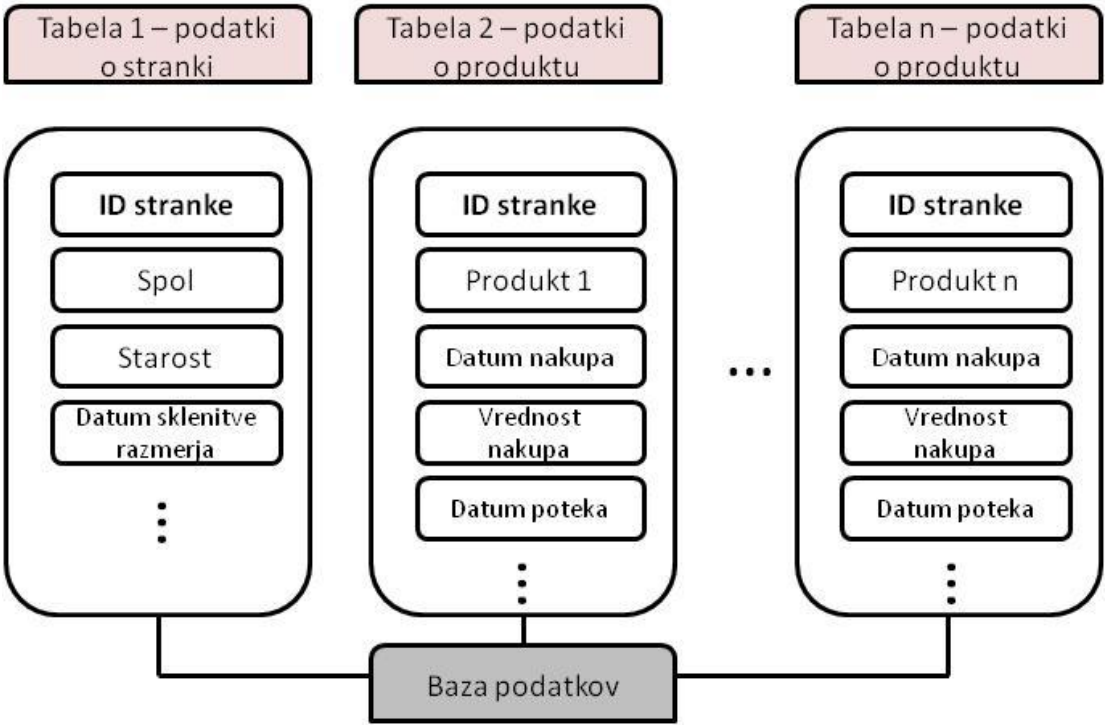
V nadaljevanju bom predstavila način dela in korake, kako sem se ga lotila, ter dobre in slabe strani, na katere sem naletela pri pripravi.

V izbrani banki so podatki shranjeni v podatkovnem skladišču, določeni izvedeni in agregirani podatki pa tudi v posameznih podatkovnih skladiščih, prilagojenih za lažje razumevanje in uporabo. Podatki o strankah, ki sem jih uporabila, so v več tabelah, kjer vrstice predstavljajo stranke, stolpci pa njihove atribute. Med seboj so povezljivi preko enotnega ključa stranke. Podatke sem tako izvozila iz več različnih tabel ter jih naknadno povezala med seboj, tako da mi je vsaka stranka z vsemi pripadajočimi atributi predstavljala svojo vrstico.

Predmet moje raziskave so bile stranke, ki so prostovoljno odšle iz banke ter naključen nabor strank, ki so še vedno stranke ter iskanje vzorca med njimi. Omejila sem se samo na fizične osebe, saj med njimi in pravnimi osebami veljajo drugačne zakonitosti, nabor fizičnih oseb in njihovih produktov pa je večji kot pri pravnih osebah. Izbrala sem 4 različna časovna obdobja v katerih so bile vse izbrane stranke tudi stranke banke in primerjala njihovo obnašanje skozi čas. Za obdobje odhoda pa sem vzela mesec julij 2015.

Zaradi poznavanja problema sem za večino podatkov vedela, da jih v modelu potrebujem, nekatere pa sem dodatno pripravila, če bodo morda uporabni pri pojasnjevanju ostalih. Nekatere atribute sem izločila, ker so podatki napačni, pomanjkljivi ali pa se pojavljajo tako redko, da njihovo vključevanje ne bi bilo smiselno. Podatki, ki sem jih uporabljala o strankah so bili anonimizirani, tako da iz njih ni mogoče ugotoviti, za katero stranko gre (osebni identifikacijski podatki, številke pogodb, imena so bili izločeni, ker za analizo niso pomembni).

Slika 14: Povezava podatkov strank v enotno tabelo v podatkovnem skladišču izbrane banke



5.2.1 Priprava nabora podatkov

Pri pripravi vhodnih podatkov sem se srečala z različnimi tipi podatkov ter z manjkajočimi podatki spol, starost, začetek razmerja. Manjkajoče številske podatke sem nadomestila z ničlami; logika nadomeščanja je v tem, da je v bazi ob odsotnem produktu manjkajoča vrednost, logično pa je stanje na produktu nič. Zaradi narave podatkov je manjkajočih vrednosti na posameznih produktih precej, saj imajo stranke v lasti manj produktov, kot jih je povprečno v ponudbi banke. Atribute, ki sem jih uporabila, lahko razdelim na enostavne atribute oziroma atribute prvega reda in izpeljane atribute, to je atribute drugega reda.

Enostavni atributi oziroma atributi prvega reda so tisti, ki sem jih lahko neposredno uporabila v modelu in za njih transformacija ni bila potrebna. To so atributi kot je starost, spol, indikatorji produktov. Ponekod je bila potrebna enostavna transformacija, da sem na primer iz datuma izločila samo leto.

Izpeljane atribute drugega reda sem dobila tako, da sem dva ali več atributov prvega reda med seboj seštela ali izvedla dodatne indikatorje (na primer ali stranka nek produkt ima ali ga nima). Namen transformacije je bilo moje mišljenje, da bi določeni produkti združeni lahko bolje pojasnili obnašanje stranke oziroma bili z logičnega vidika bolj smiselni (na primer združitev več tipov depozitov ali kreditov). Z dodajanjem izvedenih atributov sem želela izboljšati natančnost modela.

Poleg dodajanja atributov pa sem v nekaterih primerih morala izračunati nove atribute na podlagi monotonih spremenljivk. Tako sem dobila trajanje razmerja z banko kot razliko med datumom zadnjega opazovanega obdobja in datumom sklenitve razmerja. Ta podatek se mi je zdel pomemben pri ugotavljanju, ali je trajanje razmerja pomemben dejavnik pri napovedovanju odhoda stranke.

Skupaj sem za analizo izbrala več kot 50 atributov in 1.585 strank, ki sem jih vse združila v eno tabelo. Zbrane atribute lahko razdelim na demografske (starost stranke, kraj izbrane poslovalnice), podatke o produktih (osebni računi, krediti, depoziti, elektronsko bačništvo) ter transakcijske podatke (število prilivnih in odlivnih transakcij po posameznem produktu, stanja).

Tabela 3: Spremenljivke, uporabljene v modelu

Ime spremenljivke	**Tip**	**Pomen**
Client Number	Integer	Povezovalna sprmenljivka
Demografske spremenljivke		
Age	Integer	Starost
Class	Integer	Segment
Gender	Integer	Spol
Razmerje	Integer	Trajanje razmerja z banko
Produktne spremenljivke		
Asset Balance at End of Month	Nominal	Končno negativno stanje na računu
Avg Asset Balance	Nominal	Povprečno negativno stanje na računu
Avg Ledger Balance	Nominal	Povprečno stanje na računu
Avg Liability Balance	Nominal	Povprečno pozitivno stanje na računu
Avg Not Allowed Balance	Nominal	Povprečno nedovoljeno negativno stanje na računu
Card	Binominal	Imetništvo kreditne kartice
Liability Balance at End of Month	Nominal	Končno pozitivno stanje na računu
Life Insurance	Nominal	Življenjsko zavarovanje
Housing_loan_yes	Binominal	Imetništvo stanovanjskega kredita
No of Days from Last Transaction	Nominal	Število dni od zadnje transakcije
No of Deposits	Nominal	Število pologov
No of Liability Balance Days	Nominal	Število ni pozitivnega stanja
No of Not Allowed Balance Days	Nominal	Število dvi v nedovoljenem negativnem stanju
No of Withdrawals	Nominal	Število dvigov
Not Allowed Balance at End of Month	Nominal	Končno nedovoljeno negativno stanje na računu
Overd. Amount	Nominal	Znesek limita na računu
Ledger Balance at End of Month	Nominal	Končno stanje na računu
Overd. Type	Polynominal	Tip limita

se nadaljuje

Tabela 3: Spremenljivke, uporabljene v modelu (nad)

Overd. Renewal Indicator	Polynominal	Obnova limita
Pck Bonus	Nominal	Bonus, prejet na paketu
Pck no. prod	Nominal	Število produktov v paketu
Product Account Type	Polynominal	Tip računa
Propety Insurance	Nominal	Premoženjsko zavarovanje
Total Amount of Credit Operations	Nominal	Znesek prilivnih transakcij
Total Amount of Debit Operations	Nominal	Znesek odlivnih transakcij
Total Amount of Deposits	Nominal	Znesek dvigov
Total Amount of Withdrawals	Nominal	Znesek pologov
Časovne spremenljivke		
Overd. End Date	Date time	Potek limita
Last Login Date Serv	Date time	Zadnji vpogled v EB
Contract Signature Date Acct.	Date time	Datum podpisa pogodbe o otvoritvi računa
Contract Signature Date Serv	Date time	Datum podpisa pogodbe o EB
Pck Start Date	Date time	Začetek veljavnosti paketa
Life_Max_Cont_closing	Date time	Potek zadnje pogodbe o živ. zavarovanju
Life_Min_Cont_closing	Date time	Potek prve pogodbe o živ. zavarovanju
Overd. Start Date	Date time	Začetek limita
Izvedene in ostale spremenljivke		
Churner	Binominal	Identifikator, ali gre za stranko, ki je odšla ali ne
letna_sprememba_račun	Nominal	Sprememba poslovanja po računu
Depoziti_stanje	Nominal	Seštevek depozitov
VR_promet_skupaj	Nominal	Seštevek transakcij na varčevalnem računu
VR_stanje_skupaj	Nominal	Seštevek stanj na varčevalnih računih
mesečna_sprememba_račun	Nominal	Sprememba poslovanja po računu
Loans_outstanding_skupaj	Nominal	Izpostava po vseh kreditih
Loans_Anuiteta_skupaj	Nominal	Mesečna obveznost vseh kreditov
ER	Integer	Opremljenost stranke
Client Value	Nominal	Vrednost stranke
RI	Integer	Redni priliv
Base Acct Activity Level	Polynominal	Aktivnost računa
STL_yes	Binominal	Imetništvo kratkoročnega kredita

5.2.2 Pregled in čiščenje podatkov

Po izdelavi osnovnega nabora podatkov sem podatke morala pogledati, da sem preverila neskladnosti ter jih očistila.

Prvi korak je bil iskanje neskladnosti in njihovo čiščenje. Tu sem morala predvsem preveriti skladnost in enotnost zapisa datuma, poenotiti uporabo decimalnih pik in vejic, nekatere podatke preveriti na viru podatkov oziroma v definiciji nabora. Prav tako sem

preverjala pojav velikih odstopanj (doline in vrhovi). Nekatere najdene nepravilnosti sem v tabeli ročno popravila, za nekatere pa sem spremenila parametre izvoza podatkov. V tem koraku sem izvedla tudi čiščenje tistih atributov, ki so se izkazali za nepravilne oziroma sem imela na njih premalo zapisov za nadaljnjo smiselno obravnavo. Zaradi redkosti pojavljanja sem tako izpustila nekatere podatke o imetništvu finančnih inštrumentov. Prav tako sem iz nabora izključila umrle stranke ter tiste, ki banke niso zapustile na lastno željo.

Temu koraku je sledila transformacija podatkov, ker sem imela med atributi tako nominalne kot numerične vrednosti. Ker nekateri modeli z nominalnimi atributi ne znajo upravljati (na primer logistična regresija), sem jih pretvorila v numerične.

Kot zadnji korak pa sem morala določene manjkajoče vrednosti nadomestiti z neko vrednostjo, ki jo modeli prepoznajo. Za nominalne vrednosti sem uporabila vrednost »Ni podatka«, pri numeričnih pa sem manjkajočo vrednost nadomestila z »0«, saj je to, da vrednosti v naboru ni bilo, pomenilo, da stranka tega produkta nima, torej za njo ta produkt nima vrednosti (na primer nima kredita ali nima varčevanja).

5.2.3 Časovni nabor podatkov

Ker je bil cilj moje analize določitev najustreznejšega modela za napovedovanje osipa strank, sem potrebovala podatke za različna obdobja, da sem lahko preverila, kaj se je s stranko dogajalo v času in kateri atributi so najpomembnejši za napoved.

Tako sem v časovni vrsti izbrala trenutek, ko je stranka banko zapustila ter njene podatke dva meseca pred to odločitvijo (po interni raziskavi naj bi se stranke za menjavo banke odločale impulzivno in ko se odločijo, menjavo izvedejo hitro, v manj kot mesecu dni). Poleg tega pa sem preverila še, kaj se je s strankami dogajalo pol leta in leto pred odhodom. Pri tem sem zagotovila, da sem v nabor vzela stranke, ki so bile v celotnem obdobju stranke banke.

6 OPIS PROCESOV MODELIRANJA IN PREDSTAVITEV REZULTATOV

6.1 Orodje za modeliranje

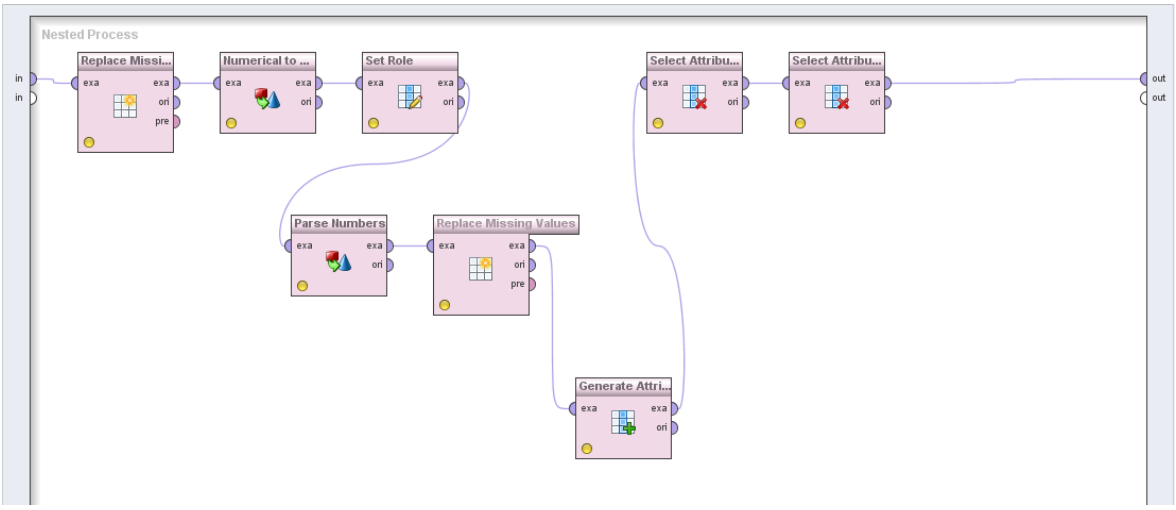
Preverjanje delovanja različnih napovednih modelov sem izvedla s programskim paketom RapidMiner Studio (v nadaljevanju RapidMiner). Programski paket je enostaven za modeliranje in ima grafični uporabniški vmesnik za modeliranje kompleksnih obdelav podatkov. Podatke za uvoz v RapidMiner sem pripravila v Excelovih tabelah, prav tako sem izvozila podatke iz programa.

Najprej sem v programu izvedla osnovni proces obdelave podatkov, ki mi je kasneje služil kot osnovni nabor za preverjanje napovednih modelov. Za preverjanje sem uporabila enotne podatke, ker sem na ta način zagotovila dosledno preverjanje napovedne moči posameznega modela.

Proces priprave vhodnih podatkov za uporabo v modelih je potekal po naslednjih korakih:

1. Z ukazom »Preprocessing« sem transformirala nekatere atribute. Numerične podatke sem spremenila v binominalne, tako so na primer stranke, ki so odšle, dobile vrednost 1 (odšla) oziroma 0 (ostaja), spol sem spremenila v 0 (moški) in 1 (ženske).
2. Atributu »odhajajoča stranka« sem spremenila vlogo, da je postal oznaka (angl. *label*).
3. Z ukazom »Parse Number« sem nominalne atribute spremenila v numerične in v naslednjem koraku nadomestila manjkajoče vrednosti z ničlami.
4. Z ukazom »Generate Attibutes« sem izračunala izveden atribute drugega reda.
5. V naslednjem koraku sem izbrala izločila atribute, ki sem jih uporabila pri izračunu izvedenih atributov, da med njimi ni prihajalo do korelacije.
6. V zadnjem koraku pa sem izbrala samo numerične atribute, ker je bil to najlažji način, da odstranim opisne atribute, ki jih v modelu nisem želela uporabiti
7. Za določene algoritme (naivni Bayesovski model, naključna drevesa) sem atribute diskretizirala tako, da sem oblikovala enako številčne, a različno široke razrede. Razlog za to je, da sem odstranila vpliv ekstremnih vrednosti ter zadovoljila predpostavko enakih predhodnih verjetnosti.

Slika 15: Proces programa RapidMiner za pripravo vhodnih podatkov v modelu



Tako pripravljene podatke sem v nadaljevanju raziskave uporabila v vsakem modelu, s katerim sem preverjala njegovo napovedno moč.

6.2 Preverjanje naivnega Bayesijsovskega modela

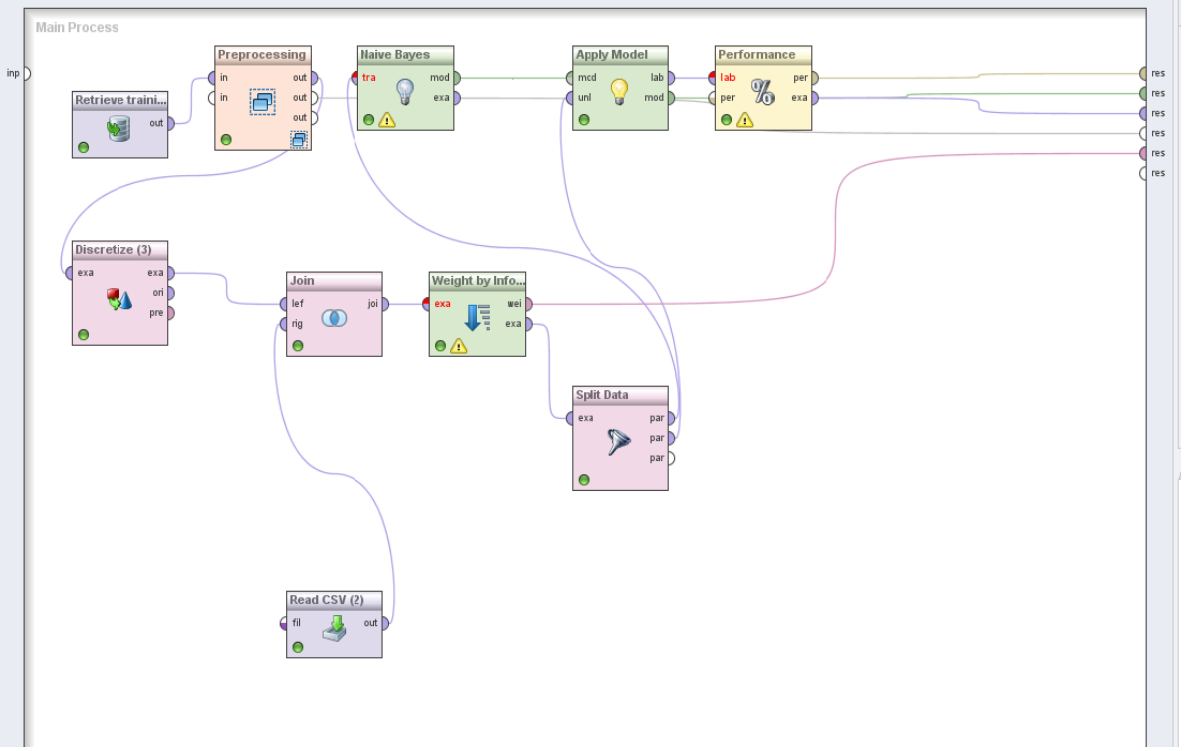
6.2.1 Priprava modela s programom RapidMiner

Za prvi model preverjanja napovedne moči sem raziskala naivni Bayesijsovski model. Izbrala sem ga zaradi narave modela, ker opisuje verjetnost nastanka nekega dogodka na podlagi pogojev, ki so lahko povezani s tem dogodkom, hkrati pa je dovolj robusten, da ga manjkajoče vrednosti ne motijo.

Opis izdelave modela in razlaga parametrov izračuna:

1. Z operatorjem »Retrieve« naložimo osnovne pripravljene tabele s podatki (v mojem primeru v obliki tabele Excel v obliki zapisa .csv).
2. Dodala sem operator »Preprocessing«, s katerim sem v modelu transformirala osnovne podatke iz nabora (kot opisujem v točki 7.1.).
3. Operator »Discretize« mi je spremenil numerične atribute v nominalne z uporabo košev (angl. *bins*). Velikost košare sem pustila 400, kot mi je program sam predlagal, saj s spreminjanjem velikosti nisem dobila boljšega rezultata.
4. Ker sem naknadno ugotovila, da potrebujem še nekaj dodatnih podatkov, sem z operatorjem »Join« in »Read« v model dodala nove podatke. Pri združevanju sem uporabila 'levo združevanje' kar pomeni, da mi je nove podatke pripelo k obstoječim, pri tem pa mi v nabor ni dalo nobene nove vrstice. Tudi dodatni podatki so bili pripravljeni na enak način kot osnovni in pripeljani v obliki Excelove datoteke v obliki zapisa .csv.
5. Z operatorjem »Weight by Operation Gain Ratio« sem utežila atribute glede na to, koliko informacije v modelu prinašajo. Višjo vrednost ima utež, pomembnejši je določen atribut. Razmerje uteži dodanih informacij (angl. *information gain ratio*) se včasih uporablja samo namesto uteži dodanih informacij, ker z njim izničimo vpliv pristranskosti upoštevanja atributov z velikimi številom različnih vrednosti. Uporabila sem ga z zavedanjem, da lahko ob tem atributi z majhno informacijsko vrednostjo lahko dobijo neupravičeno prednost. Atribute sem razvrstila po velikosti od največje vrednosti do najmanjše, glede na dodano vrednost, ki jo prinašajo.
6. Z ukazom »Split Data« sem bazo razdelila na dva dela, in sicer na učljivi in na testni del v razmerju 70:30.
7. Z ukazom »Naive Bayes« sem pognala naivni Bayesovski model.
8. Zadnji operator »Performance« pa mi je služil za izračun ocene modela. Za prikaz sem izbrala prikaz točnosti (angl. *accuracy*), področje AUC in dvig (angl. *lift*).

Slika 16: Proces programa RapidMiner za izdelavo naivnega Bayesijsovskega modela in njegovo preverjanje



6.2.2 Predstavitev rezultatov

Spremenljivke, ki naivni Bayesovski model najbolje pojasnjujejo, so opremljenost stranke s produkti, njena letna vrednost, redni priliv, ki ga mesečno prejema, izpostava na kreditih in mesečni obrok kredita ter nedovoljeno negativno stanje na računu.

Rezultat modela je klasifikacijska matrika za binarne razvrščevalce, testirana na podlagi testnega vzorca (479 strank ali 30 % strank iz moje baze podatkov). Ker napovedujemo odhod strank, nam vrednost »true true« pove, koliko strank je banko dejansko zapustilo, vrednost »true false« pa, koliko jih je ostalo. Na drug strani vrednost »predicted false« pove, koliko smo jih napovedali napačno, vrednost »predicted true« pa, koliko smo jih napovedali pravilno.

Slika 17: Klasifikacijska matrika naivnega Bayesovskega modela

accuracy: 68.48% lift: 162.72% (positive class: true)

	true false	true true	class precision
pred. false	214	27	88.80%
pred. true	124	114	47.90%
class recall	63.31%	80.85%	

Cilj za moj model je bil, da bi bilo čim več strank v presečišču vrednosti »true true« in »predicted true«, kar mi predstavlja stranke, za katere sem napovedala, da bodo odšle in tudi so odšle (angl. *true positive* – TP). Iz rezultatov je razvidno, da sem z modelom pravilno napovedala 114 strank (TP). Prav tako so bile pravilno napovedane stranke, ki so na presečišču vrednosti »predicted false« in »true false«, kajti to so stranke, za katere sem napovedala, da bodo ostale, in tudi so (angl. *true negatives* – TN). Teh strank je 214.

Naslednji dve skupini strank pa sta tisti, ki sem jih z modelom napačno napovedala. Za to skupino strank sem želela, da bi bila čim manjša. Prva skupina napačno napovedanih strank so napačno pozitivni (angl. *false positives* – FN), to so tisti, za katere sem napovedala, da bodo odšli, pa so ostali. Teh je v modelu 27. Druga skupina pa so napačno negativni (angl. *false negatives* – FN), kjer sem z modelom napovedala, da bodo ostali, pa so odšli. Teh je 124.

Na podlagi te matrike sem dobila tudi mero točnosti (angl. *accuracy*), ki znaša 68,5 %. To mi pove, da je model pravilno napovedal skoraj 70 % vseh strank, kar pomeni precej dobro napovedno moč. Stopnja napačne razvrstitve oziroma stopnja napake (angl. *misclassification rate* ali *error rate*) pa prikazuje obratni faktor, ki je v tem modelu 30 %.

Stopnja TP, imenovana tudi priklic oziroma občutljivost (angl. *recall* ali *sensitivity*) pove, kako pogosto sem napovedala odhod in je do njega prišlo. V modelu znaša 47,9 %. Stopnja FP, ki pove, kolikokrat sem napovedala odhod, pa do njega ni prišlo, znaša 11,2 %.

Poleg teh so pomembni še trije kazalci napovedne moči modela. To so specifičnost (angl. *specificity*), ki nam pove, kolikokrat smo med tistimi strankami, ki so ostale, tudi napovedali, da bodo ostale. V mojem primeru je specifičnost 88,8 %. Naslednji kazalec je natančnost (angl. *precision*), ki pove, kolikokrat smo med napovedanimi strankami za odhod te pravilno napovedali. V modelu se je to zgodilo v 80,9 %. Tretji kazalec pa je razširjenost (angl. *prevalence*), ki meri, kako pogosto se v našem vzorcu pojavi odhod stranke in znaša 49,7 %.

Ker imamo opravka z naivnim Bayesovskim modelom, nam priklic in specifičnost predstavljata pogojne verjetnosti, razširjenost nam predstavlja predhodno verjetnost (angl.

prior probability), pozitivno in negativno napovedane vrednosti pa predstavljajo kasnejše verjetnosti (angl. *posterior probability*).

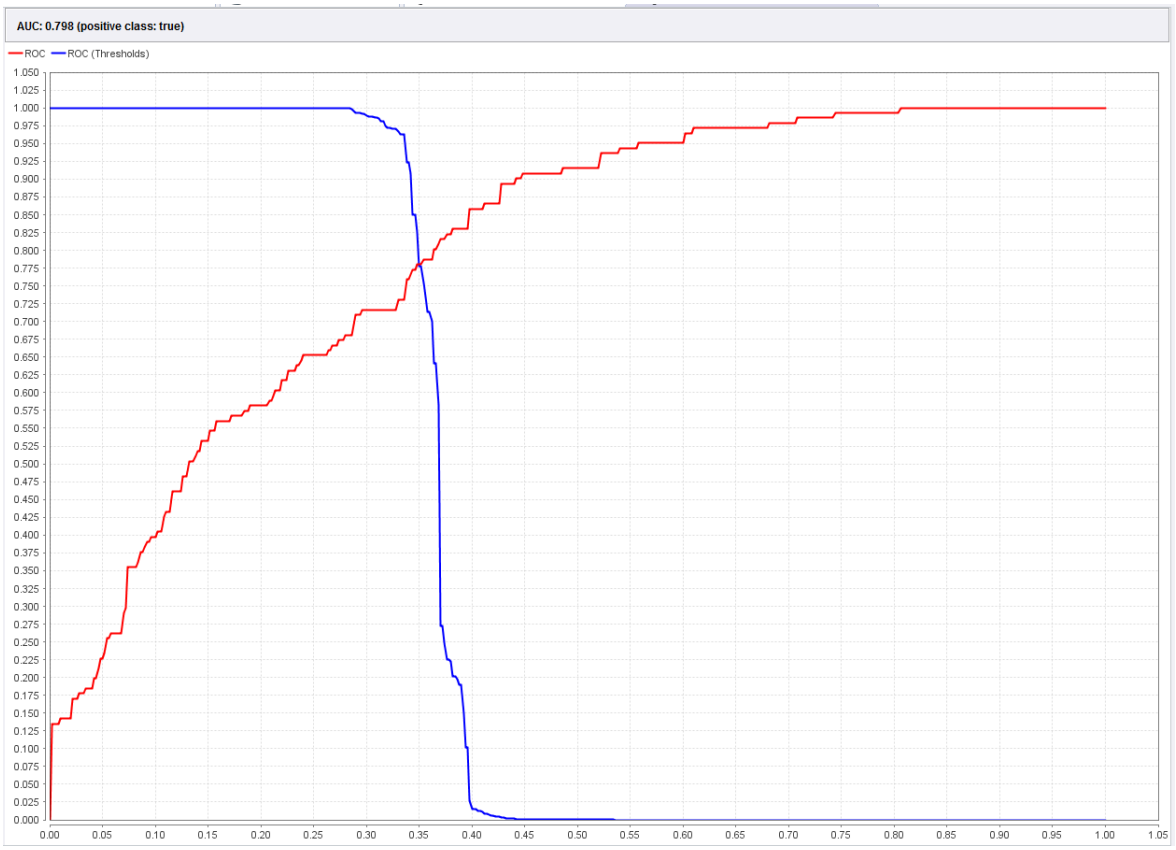
Dodatno lahko napovedno moč preverjamo še z nekaterimi kazalci kot je Cohenova Kappa, ki pove, kako dobro razvrščevalec napoveduje v primerjavi z naključnim izborom. Če je v modelu Kappa visoka, to pomeni, da je velika razlika med natančnostjo in tem, kako pogosto bi se motil, če bi vedno napovedoval razred z najvišjim številom zadetkov. V tem primeru je izračunana Kappa 0,37, kar oceni model na srednje dober.

Dvig pa je mera učinkovitosti razvrščevalnega modela ki se izračuna kot razmerje med dobljenimi rezultati z modelom in brez njega. Izračunamo ga kot odstotek pravilno napovedanih pozitivnih deljeno z odstotkom naključno pravilno napovedanih pozitivnih. V tem modelu znaša 162,7 %, kar pomeni, da z modelom napovem 63 % več resnično pozitivnih kot bi jih z naključnim izborom.

ROC krivulja, prikazana v Sliki 18, je primerna za prikazovanje binarnih razvrščevalcev, v mojem primeru, ali stranka odide ali ostane. V grafu na osi Y prikazujem stopnji TP oziroma občutljivost, na osi X pa stopnjo FP oziroma specifičnost. Vsaka točka na krivulji ROC prikazuje razmerje med občutljivostjo in specifičnostjo. Bolj je krivulja oddaljena od navidezne diagonale, ki poteka od izhodiščne točke (0, 0) proti skrajni točki (1, 1) in nam prikazuje naključni izbor (verjetnost 50 %), boljša je napovedna moč našega modela.

V primeru mojega naivnega Bayesijsovskega modela je krivulja lepo upognjena in izračun AUC (področja pod krivuljo ROC), ki nam pove, kako dobro moj razvrščevalec deli množico podatkov na dva dela in predstavlja odstotek pokritosti kvadrata s krivuljo ROC je 80 %, kar pomeni, da je napovedno model dober. V primeru AUC bi 100 % pomenilo popolnega razvrščevalca, 50 % pa naključni izbor.

Slika 18: Krivulja ROC in AUC za naivni Bayesovski model



6.3 Preverjanje modela logistične regresije

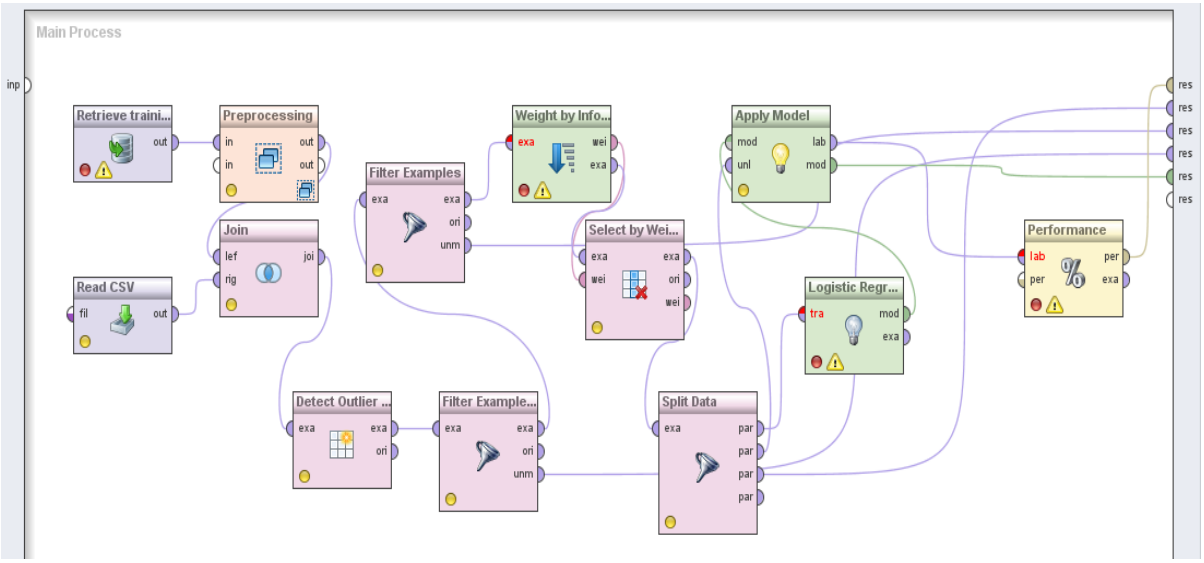
6.3.1 Priprava modela s programom RapidMiner

Opis izdelave modela in razlaga parametrov izračuna:

1. Delovanje operatorjev »Retrieve«, »Preprocessing«, »Join« in »Read« sem opisala v točki 6.2.1.
2. Z operatorjem »Detect Outlier« sem poiskala osamelce. Program jih izračunava na podlagi metode *k* najbližjih sosedov. Ta primerja lokalno gostoto (angl. *density*) objekta od lokalne gostote sosednjih objektov in tako identificira točke, ki imajo znatno nižjo gostoto od svojih sosedov. Na ta način je bilo iz osnovnega nabora izločena ena tretjina vrstic.
3. Sledita dva operatorja »Filter Examples«, s katerima sem počistila tiste manjkajoče atribute, ki sem jih dodala z novo pripeto datoteko.
4. Operator »Weight by Operation Gain Ratio« sem opisala v točki 62.1., z operatorjem »Select by Weight« pa sem izbrala samo tiste atribute, katerih utež je zadoščala kriteriju, da so večji od 0,1. Kriterij sem izbrala zato, da sem v modelu obdržala samo pomembnejše atribute, kar je izboljšalo napovedno moč modela.

5. Ukazom »Split Data« je bazo razdelil na dva dela, na učljivi in na testni del v razmerju 70:30.
6. Z ukazoma »Apply Model« in »Logistic Regresion (Evolutionary)« sem pognala model logistične regresije. Model ima, v primerjavi z »Logistic Regresion« to prednost, da mu lahko nastavimo, kolikokrat naj se model generira, preden se ustavi ter po koliko ponovitvah, ki ne izboljšajo modela, naj se ustavi. Poleg tega lahko nastavimo obliko funkcije (kernel type) in konstanto kompleksnosti, s katero določimo velikost tolerance napačne klasifikacije. Z uporabo modela »Evolutionarly« je bil rezultat veliko boljši kot ob uporabi enostavnega modela.
7. Zadnji operator »Performance« pa sem tudi tu uporabila za izračun ocene modela. Za prikaz sem izbrala prikaz točnosti (angl. *accuracy*), krivuljo ROC, AUC in dvig (angl. *lift*).

Slika 19: Proces ptograma RapidMiner za izdelavo modela logistične regresije in njegovo preverjanje



6.3.2 Predstavitev rezultatov

Rezultat modela prikazujem v obliki klasifikacijske matrike, testirane na podlagi testnega vzorca (332 strank ali 30 % strank iz moje baze podatkov po odstranitvi osamelcev).

Slika 20: Klasifikacijska matrika modela logistične regresije

accuracy: 75.00% lift: 229.64% (positive class: true)

	true false	true true	class precision
pred. false	228	71	76.25%
pred. true	12	21	63.64%
class recall	95.00%	22.83%	

Želela sem pravilno napovedati čim več dejanskih odhodov strank (TP). Iz rezultatov je razvidno, da sem z modelom pravilno napovedala odhod 21 strank (TP). Prav tako so bile pravilno napovedane stranke, za katere sem napovedala, da bodo ostale, in tudi so (TN). Teh strank je 228.

Z modelom sem napačno napovedala napačno pozitivne (FN), ki jih je v modelu 12 ter napačno negativne (FN), kjer sem z modelom napovedala, da bodo ostali, pa so odšli. Teh je 71.

Na podlagi te matrike sem dobila tudi mero točnosti, ki znaša 75 %. To mi pove, da je model pravilno napovedal 75 % vseh strank, kar pomeni dobro napovedno moč. Stopnja napačne razvrstitve oziroma stopnja napake, ki je v tem modelu 25 % pa mi prikazuje obratni faktor.

Stopnja TP (priklic oziroma občutljivost) mi pove, kako pogosto sem v modelu napovedala odhod in je do njega prišlo. V modelu znaša 63,6 %. Stopnja FP, ki pove, kolikokrat sem napovedala odhod, pa do njega ni prišlo, znaša 23 %.

Poleg teh so pomembni še trije kazalci napovedne moči modela. To so specifičnost (kolikokrat smo med strankami, ki so ostale, napovedali, da bodo ostale), ki v tem modelu znaša 76,3 %. Naslednja kazalca sta natančnost (kolikokrat smo med napovedanimi strankami za odhod te pravilno napovedali), ki znaša 22,8 % in razširjenost (kako pogosto se v našem vzorcu pojavi odhod stranke), ki v modelu znaša 9,9 %.

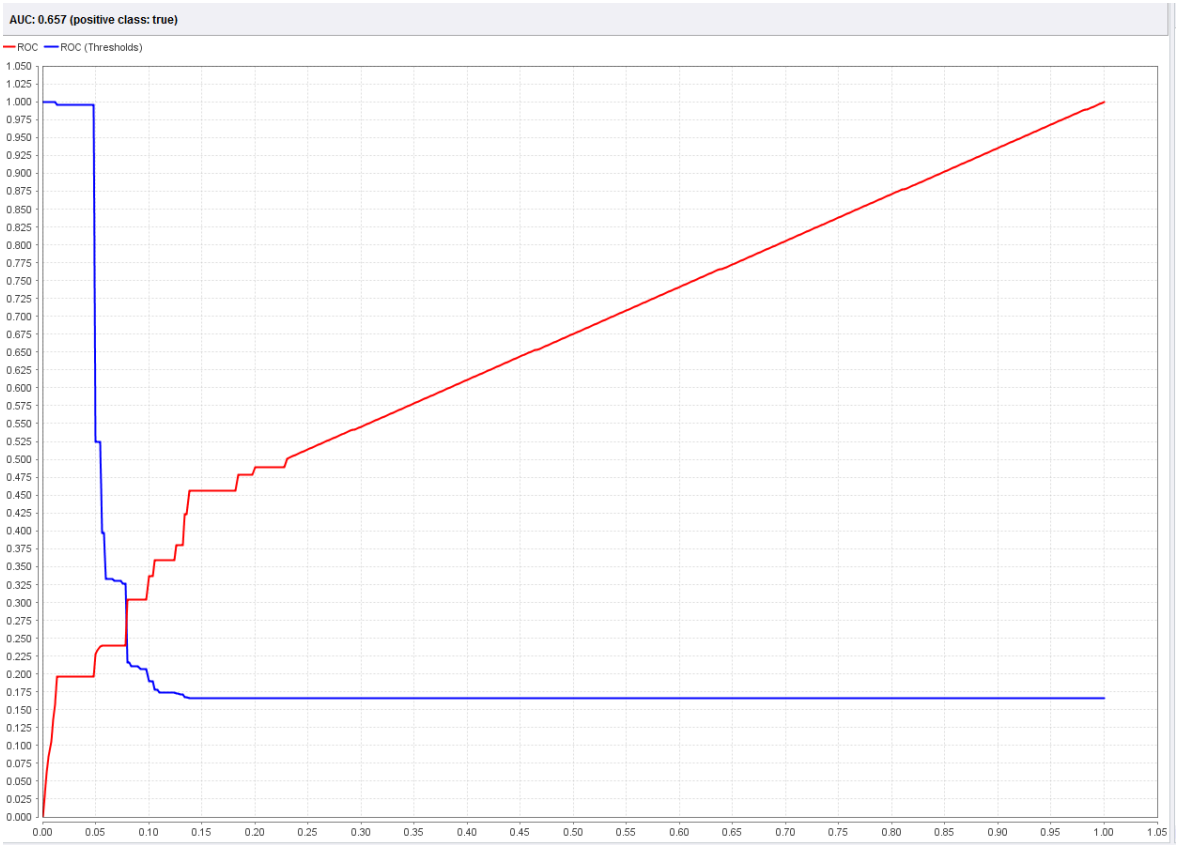
Kazalec Kappa, ki pove, kako dobro razvrščevalec napoveduje v primerjavi z naključnim izborom, je v modelu logistične regresije 0,22, kar oceni model za relativno slabega.

Kappa je, glede na nizko napovedno moč modela, vseeno visoka, kar je posledica majhnega števila strank, ki odhajajo.

Dvig pa je mera učinkovitosti razvrščevalnega modela. V opisanem modelu znaša 229,6 %, kar pomeni, da z modelom napovem 130 % več resnično pozitivnih kot bi jih z naključnim izborom.

Krivulja ROC, prikazana v Sliki 21, prikazuje, da model kot tak ni najboljši, saj je njena usločenost slaba in je skoraj poravnana z navidezno črto, ki poteka od točke (0,0) do (1,1), ki predstavlja naključni izbor, kar napove, da s tem modelom ne dobim dosti boljših rezultatov od naključnega izbora. Izračun AUC (področja pod krivuljo ROC), to potrjuje, saj znaša le 66 %, kar pomeni, da je napovedna moč modela šibka.

Slika 21: Krivulja ROC in AUC za model logistične regresije



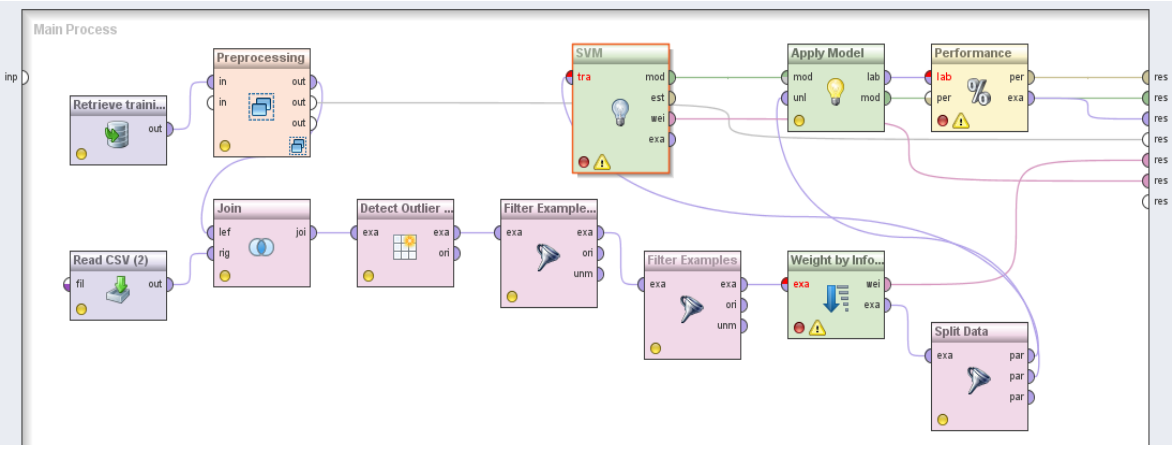
6.4 Preverjanje metode podpornih vektorjev

6.4.1 Priprava modela s programom RapidMiner

Opis izdelave modela in razlaga parametrov izračuna:

1. Delovanje operatorjev »Retrieve«, »Preprocessing«, »Join« in »Read« ter »Operation Gain Ratio« sem opisala v točki 6.2.1., operator »Detect Outlier« ter oba »Filter Examples« pa v točki 6.3.1.
2. Z ukazom »Split Data« sem bazo razdelila na dva dela, na učljivi in na testni del v razmerju 70:30.
3. Z ukazom »SVM« in »Apply Model« sem pognala model podpornih vektorjev.
4. Zadnji operator »Performance« pa mi je tudi tu služil za izračun ocene modela. Za prikaz sem izbrala prikaz točnosti (angl. *accuracy*), krivuljo ROC, AUC in dvig.

Slika 22: Proces programa RapidMiner za izdelavo metode podpornih vektorjev in njeno preverjanje



6.4.2 Predstavitev rezultatov

Spremenljivke, ki moj model podpornih vektorjev najbolje pojasnjujejo, so nedovoljeno negativno stanje na računu, opremljenost stranke s produkti v letu pred odhodom, število prilivov na osebni račun in stanje na osebnem računu ob koncu meseca. Vsi našteti vplivajo na model v pozitivni smeri. Spremenljivke, ki na model najbolj vplivajo v negativni smeri, pa so opremljenost s produkti tik pred odhodom, letna vrednost stranke, število odlivov z osebnega računa ter izpostava na kreditu in mesečni obrok kredita.

Slika 23: Klasifikacijska matrika metode podpornih vektorjev

accuracy: 73.19% lift: 206.21% (positive class: true)

	true false	true true	class precision
pred. false	231	80	74.28%
pred. true	9	12	57.14%
class recall	96.25%	13.04%	

Rezultat modela sem tudi tokrat prikazala v obliki klasifikacijske matrike, testirane na podlagi testnega vzorca (332 strank ali 30 % strank iz moje baze podatkov po odstranitvi osamelcev).

Želela sem pravilno napovedati čim več dejanskih odhodov strank (TP). Iz rezultatov je razvidno, da sem z modelom pravilno napovedala odhod 12 strank (TP). Prav tako so bile pravilno napovedane stranke, za katere sem napovedala, da bodo ostale, in tudi so (TN). Teh strank je 231.

Z modelom podpornih vektorjev sem napačno napovedala napačno pozitivne (FN), ki jih je v modelu 9 ter napačno negativne (FN), kjer sem z modelom napovedala, da bodo ostali, pa so odšli. Teh je 80.

Na podlagi te matrike sem dobila tudi mero točnosti, ki znaša 73,2 %, kar pomeni, da je bilo toliko strank z modelom pravilno napovedanih, kar mi pove, da ima model dobro napovedno moč. Stopnja napačne razvrstitve oziroma stopnja, ki je v tem modelu 26,8 % pa mi prikazuje obratni faktor.

Stopnja TP (priklic oziroma občutljivost) mi pove, kako pogosto sem v modelu napovedala odhod in je do njega prišlo. V modelu znaša 57,1 %. Stopnja FP, ki pove, kolikokrat sem napovedala odhod, pa do njega ni prišlo, znaša 25,7 %.

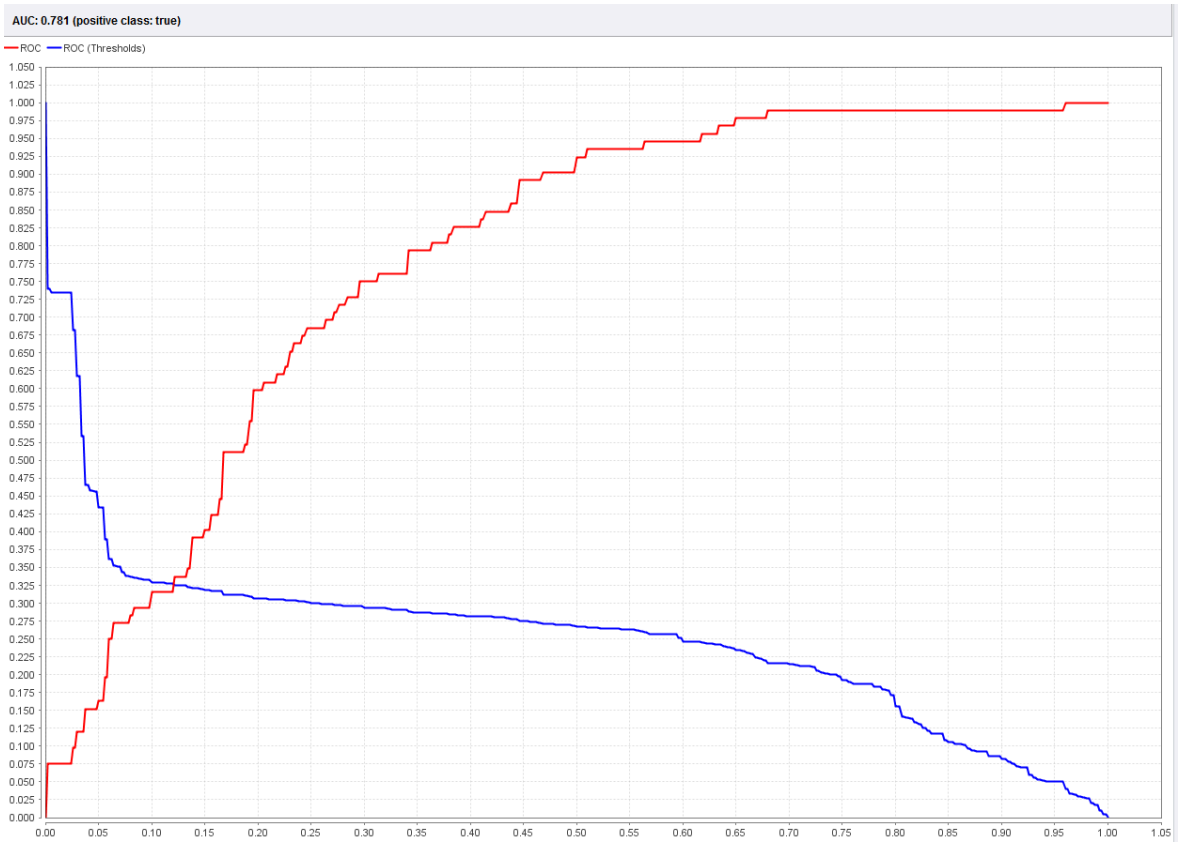
Poleg teh so pomembni še trije kazalci napovedne moči modela. To so specifičnost (kolikokrat smo med strankami, ki so ostale, napovedali, da bodo ostale), ki v tem modelu znaša 74,3 %. Naslednja kazalca sta natančnost (kolikokrat smo med napovedanimi strankami za odhod te pravilno napovedali), ki znaša 13 % in razširjenost (kako pogosto se v našem vzorcu pojavi odhod stranke), ki v modelu znaša 6,3 %.

Kazalec Kappa, ki pove, kako dobro razvrščevalec napoveduje v primerjavi z naključnim izborom, je v modelu logistične regresije 0,21, kar oceni model za relativno slabega. Kappa je, glede na nizko napovedno moč modela, tudi tu relativno visoka, kar je ponovno posledica majhnega števila strank, ki odhajajo.

Dvig pa je mera učinkovitosti razvrščevalnega modela. V opisanem modelu znaša 206,2 %, kar pomeni, da z modelom napovem 106 % več resnično pozitivnih, kot bi jih z naključnim izborom.

Krivulja ROC, prikazana v Sliki 24, prikazuje, da je model relativno dober, saj je krivulja lepo navzgor usločena in z modelom lahko napovemo precej boljše rezultate kot z naključnim izborom. Izračun AUC (področja pod krivuljo ROC), to potrjuje, saj znaša 78 %, kar pomeni, da je napovedna moč modela dobra. Še vedno pa je očitno, da bi, če bi želeli imeti res dober model, to AUC moral pokrivati večji del kvadrata po sabo.

Slika 24: Krivulja ROC in AUC za metodo podpornih vektorjev



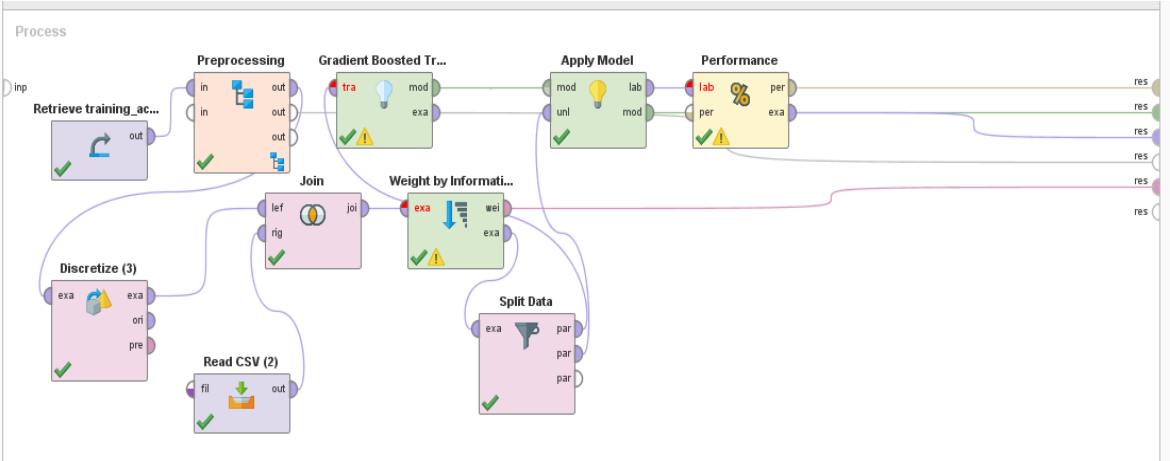
6.5 Preverjanje modela s povečevanjem naklona

6.5.1 Priprava modela s programom RapidMiner

Opis izdelave modela in razlaga parametrov izračuna je enaka kot za naivni Bayesovski model, z razliko, da sem tu za preračun modela izbrala model s povečevanjem naklona, ki je primeren za strukturirane in razmeroma nizkodimenzionalne podatke, kamor sodijo tudi podatke iz moje naloge. Ker je v mojem primeru napovedna spremenljivka diskretna, potrebujem primeren osnovni (šibek) model, v tem primeru jemljem običajna odločitvena

drevesa fiksne velikosti. Vzela sem 20 dreves z največjo globino 5. S povečevanjem števila dreves in njihove globine se je napovedna sposobnost modela zmanjševala.

Slika 25: Proces programa RapidMiner za izdelavo metode s povečevanjem naklona in njeno preverjanje

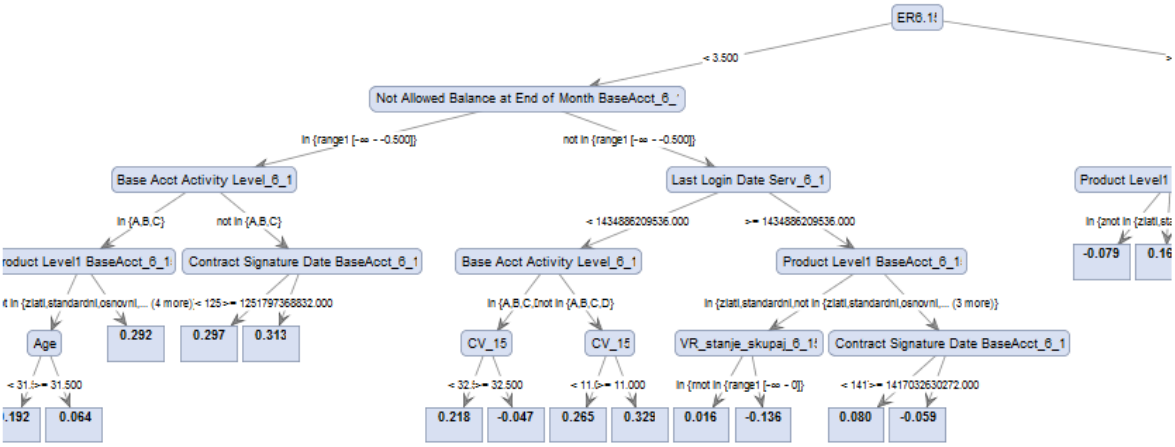


6.5.2 Predstavitev rezultatov

Algoritem dreves odločanja s povečevanjem naklona gradi serijo dreves odločanja, pri čemer v prvem koraku napoveduje dejansko opazovane vrednosti (v mojem primeru stranke, ki so odšle), v vsakem nadaljnjem koraku pa poskuša napovedati to, kar je napovedal predhodni model, pri čemer se čedalje večjo težo daje tistim napovedim, ki so se v predhodnih korakih izkazale kot pravilne.

Slika 26 prikazuje enega izmed odločitvenih dreves v modelu. V njem je prvi, najmočnejši napovedni parameter redni priliv, kjer ena veja prikazuje stranke, ki imajo priliv manjši od 3.500 EUR. Ta veja se nadalje deli na stranke, ki imajo na računu dovoljeno stanje in na tiste, pri katerih je bilo opaženo nedovoljeno stanje. Slednje nam na naslednji veji deli glede na uporabljene produkte teh strank, v naslednjem koraku glede na tip produkta ter v eni od zadnjih vej glede na letno izračunano vrednost stranke. Ponderji na koncu nam kažejo, kolikšen je delež napačno razporejenih strank – negativne vrednosti pomenijo, da veja napoveduje premalo strank, ki ostanejo, pozitivne pa premalo strank, ki odhajajo. V poznejših drevesih se delež napačno razvrščenih hitro zmanjšuje, v idealnem primeru bi bili v zadnjem drevesu vsi ponderji nič, kar bi pomenilo, da nam je model uspel pravilno razvrstiti vse enote.

Slika 26: Primer enega od odločitvenih dreves modela



Slika 27: Klasifikacijska matrika modela s povečevanjem naklona

accuracy: 86.43% lift: 272.31% (positive class: true)

	true false	true true	class precision
pred. false	313	40	88.67%
pred. true	25	101	80.16%
class recall	92.60%	71.63%	

Rezultat modela tudi tokrat prikažem v obliki klasifikacijske matrike, testirane na podlagi testnega vzorca 479 strank ali 30 % strank iz moje baze podatkov).

Želela sem pravilno napovedati čim več dejanskih odhodov strank (TP). Iz rezultatov je razvidno, da sem z modelom pravilno napovedala odhod 101 strank (TP). Prav tako so bile pravilno napovedane stranke, za katere sem napovedala, da bodo ostale, in tudi so (TN). Teh strank je 313.

Z modelom podpornih vektorjev sem napačno napovedala napačno pozitivne (FN), ki jih je v modelu 25 ter napačno negativne (FN), kjer sem z modelom napovedala, da bodo ostali, pa so odšli. Teh je 40. Glede na število pravilno napovedanih strank je napaka napovedi relativno majhna.

Na podlagi te matrike sem dobila tudi mero točnosti, ki znaša 86,4 %, kar pomeni, da je bilo toliko strank z modelom pravilno napovedanih, kar mi pove, da je model točen in ima zelo dobro napovedno moč. Stopnja napačne razvrstitve oziroma stopnja napake, ki je v tem modelu 13,6 %, pa mi prikazuje obratni faktor.

Stopnja TP (priklic oziroma občutljivost) mi pove, kako pogosto sem v modelu napovedala odhod in je do njega prišlo. V modelu znaša 80,2 %. Stopnja FP, ki pove, kolikokrat sem napovedala odhod, pa do njega ni prišlo, znaša 11,3 %.

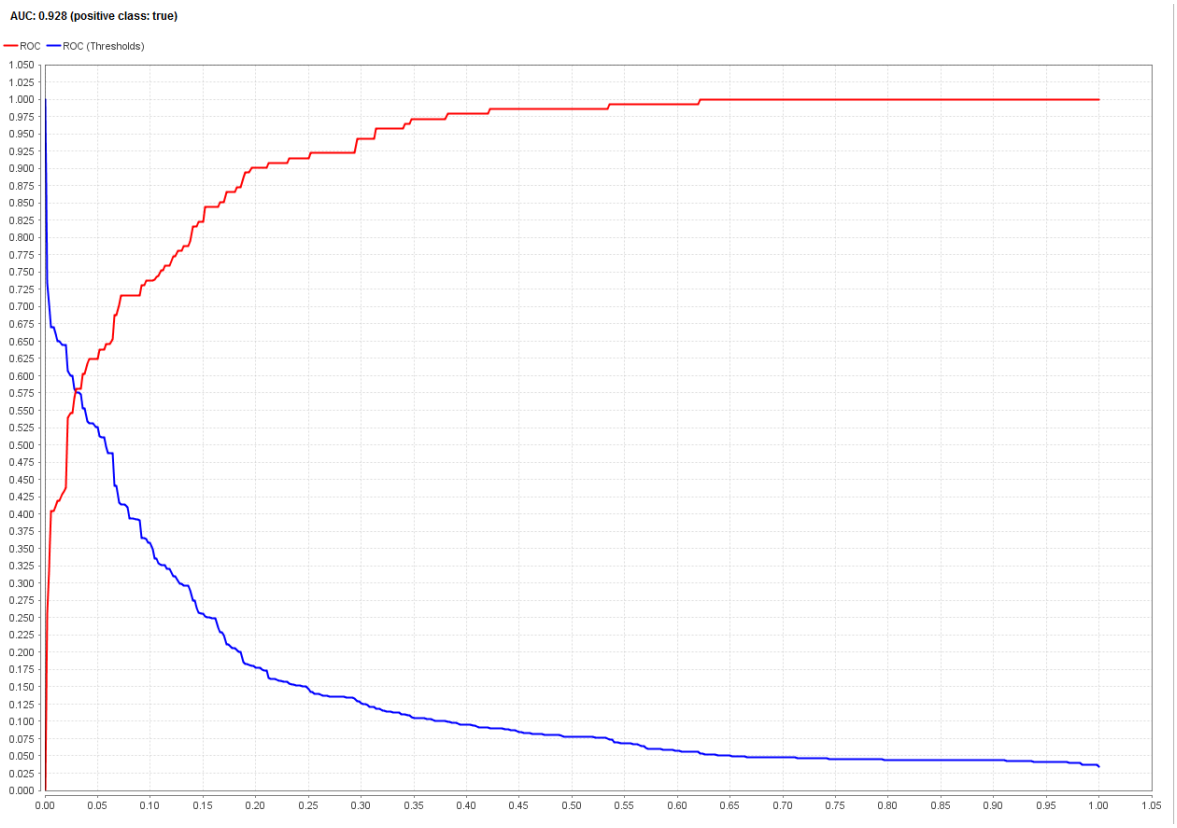
Poleg teh so pomembni še trije kazalci napovedne moči modela. To so specifičnost (kolikokrat smo med strankami, ki so ostale tudi napovedali, da bodo ostale), ki v obravnavanem modelu znaša 88,7 %. Naslednja kazalca sta natančnost (kolikokrat smo med napovedanimi strankami za odhod te pravilno napovedali), ki znaša 71,6 % in razširjenost (kako pogosto se v našem vzorcu pojavi odhod stranke), ki v modelu znaša 26,3 %. Tako specifičnost kot natančnost sta v modelu visoka, kar pomeni, da ta model dobro pojasnjuje tako odhode strank kot tudi stranke, ki ostajajo zveste.

Kazalec Kappa, ki pove, kako dobro razvrščevalec napoveduje v primerjavi z naključnim izborom, je v modelu s povečevanjem naklona 0,66, kar oceni model za dobrega. Vrednost Kappa je v tem modelu najvišja med vsemi proučevanimi modeli, kar pomeni, da ta model med izbranimi, najnatančneje razvršča stranke med tiste ki odhajajo in tistte, ki ostajajo.

Dvig pa je mera učinkovitosti razvrščevalnega modela. V opisanem modelu znaša 272,3 %, kar pomeni, da z modelom napovem 172 % več resnično pozitivnih kot bi jih z naključnim izborom. Tudi ta kazalec je v modelu zelo dober, saj se lepo vidi, da bi z uporabo zgolj naključnega izbora pojasnila mnogo manj odhodov strank.

Krivulja ROC, prikazana v Sliki 28, prikazuje, da je model zelo dober, saj je krivulja močno navzgor usločena in tako lahko tudi za ta kazalec trdim, da z modelom napovem veliko boljše rezultate kot z zgolj naključnim izborom. Izračun AUC (področja pod krivuljo ROC) to potrjuje, saj znaša 92 %, kar pomeni, da je napovedna moč modela odlična (100 % je največji možen AUC in to bi pomenilo popoln model).

Slika 28: Krivulja ROC in AUC za model s povečevanjem naklona



6.6 Medsebojna primerjava modelov in izbira najboljšega

Modele sem medsebojno primerjala glede na predhodno proučevane kazalnike njihove napovedne moči. Na podlagi kazalnikov sta se kot najboljša izkazala model s povečevanjem naklona in naivni Bayesovski model, najslabše pa metoda podpornih vektorjev. Tudi model logistične regresije, ki je sicer najpogosteje uporabljan model in zelo razširjen, se v mojem primeru ni izkazal kot dober.

Tabela :4 Primerjava modelov glede na izračunane kazalnike

Mere / Modeli	**Naivno Bayesovski model**	**Model logistične regresije**	**Metoda podpornih vektorjev**	**Model s povečevanjem naklona**
Točnost	68,5	75,0	73,2	86,4
Občutljivost	47,9	63,6	57,1	80,2
Stopnja FP	11,2	23,0	25,7	11,3
Specifičnost	88,8	76,3	74,3	88,7
Natančnost	80,9	22,8	13,0	71,6

se nadaljuje

Tabela 4: Primerjava modelov glede na izračunane kazalnike (nad)

Razširjenost	49,7	9,9	6,3	26,3
Kappa	0,37	0,22	0,21	0,66
Lift	162,7	229,6	206,2	272,3
AUC	79,8	65,7	78,1	92,8

Ko primerjam izbrane modele po posameznih kazalnikih, se metoda s povečevanjem naklona izkaže za metodo z največjo točnostjo, ker smo z njo napovedali največji odstotek dejanskih odhodov. Prav tako je to model z največjo občutljivostjo, kar pomeni, da smo največkrat napovedali odhod, kjer je do njega dejansko tudi prišlo, ima največjo Kappo, torej bi v primeru, ko bi se namesto uporabe katerekoli metode, odločila za naključni izbor, prav z neuporabo te metode naredilila največjo napako. Poleg tega ima metoda s povečevanjem naklona najvišji dvig, kar prav tako pomeni, da s tem modelom napovem veliko več pravih odhodov kot bi jih z naključnim izborom. Model je najboljši tudi v pokrivanju površine pod krivuljo ROC, kjer se s pokritostjo 92 % zelo približam popolnemu modelu).

Naivno Bayesovski model se je izkazal za malenkost boljšega od modela s povečevanjem naklona pri treh kazalnikih. Bolje se je izkazal pri specifičnosti, s katero merimo, koliko strank, za katere je bilo napovedano, da bodo ostale, je tudi dejansko ostalo. Prav tako je bil ta model najboljši pri stopnji FP, ki nam pove, koliko strankam smo napovedali odhod, pa do njega dejansko ni prišlo, ter pri natančnosti, ki nam pove, koliko napovedanih strank za odhod je dejansko odšlo. Imel je tudi najvišjo stopnjo razširjenosti, ki nam pove pogostost pojavljanja dogodka, v mojem primeru odhoda strank strank, v vzorcu.

Model s povečevanjem naklona, ki je trenutno eden najobetavnejših metod za strojno učenje, kjer gradimo drevo za drevesom na podlagi ostankov, je v mojem primeru najnatančneje napovedoval odhode strank, saj tudi na kazalnikih, kjer sicer ni bil najboljši, ni dosti zaostajal za naivnim Bayesovskim modelom.

Ostala modela, model logistične regresije in metoda podpornih vektorjev, se na nobenem kazalcu nista izkazala kot najboljša, zato lahko iz tega sklepam, da sta, na podlagi mojih vhodnih podatkov, slabša od prvih dveh. Je pa zanimivo, da sta oba slabša modela imela med seboj precej primerljive rezultate.

Kot naslednji korak v proučevanju pa bi bilo zame zanimivo oba najboljša modela preveriti na celotni bazi strank ter narediti napoved za prihodnost ter ugotoviti, kako se model izkaže v dejanskem poslovnem življenju. Zagotovo bi izvedba na celotni bazi zahtevala določene prilagoditve, tako na strani modela kot tudi uporabe podatkov, vendar menim, da

bi z njim lahko pokazala dodano vrednost uporabe napovednih modelov pri upravljanju s strankami.

SKLEP

Ko so trgi začeli postajati vedno bolj zasičeni, so podjetja začela spoznavati, da so njihova največja vrednost stranke, še posebej zato, ker povečani prihodki prihajajo predvsem od obstoječih strank in ne toliko od novih. Podjetja se vedno bolj zavedajo pomembnosti povečevanja zvestobe strank. Podjetja si zato prizadevajo ohranjati obstoječe stranke in napovedovati osip strank, da bi ga preprečili. Stranke poskušajo pritegniti k sebi in prikleniti nase na raznovrstne načine, preko ugodnih paketnih ponudb, popustov na količinske nakupe, popustov ali ugodnosti na zvestobo.

V sedanjem času imajo podjetja večinoma velike baze podatkov o svojih strankah, ki vsebujejo tako podatke o obnašanju in nakupnih navadah kot tudi socio-demografske značilnosti strank in to za več let nazaj. Prav tako pa imajo v svojih sistemih lahko zapisane tudi mehkejše podatke o strankah, ki jim pridejo prav predvsem pri neposrednem kontaktu s stranko. Študije so pokazale, da so vsi ti podatki lahko zelo uporabni pri napovedovanju osipa strank. Zato podjetja iščejo modele, ki jim lahko pomagajo čim natančneje določiti potencialne stranke za odhod ter z različnimi marketinškimi metodami na to dejanski odhod preprečiti.

Obstaja veliko poznanih tehnik, ki se uporabljajo v napovedovanju osipa uporabljajo. Najširše uporabljeni sta rudarjenje po podatkih in statistična analiza. Konvencionalni statistični modeli (logistična regresija, analiza preživetja, nevronske mreže, naključni gozdovi, genetični algoritmi, metoda podpornih vektorjev…) zelo dobro napovedujejo osip strank. Za napoved osipa strank so v uporabi tudi druge tehnike, kot so analiza preživetja, ansambelska metoda in analiza omrežja. Večina tehnik omogoča razvrščanje strank med tiste, ki so odhajajo in tiste, ki ostajajo, težko pa določijo, kdaj naj bi stranka odšla ali kako dolgo bo stranka ostala pri nekem ponudniku. Vsi ti glavni pristopi v napovedovanju osipa strank obravnavajo vsako stranko posebej in vsaki stranki napovedo njeno specifično verjetnost dogodka.

Poleg izbora natančne tehnike, so pomembni sestavni deli napovednega modela razumljivost, upravičenost in enostavnost uporabe v praksi, saj le model, ki ga uporabniki razumejo, lahko zaživi v poslovnem svetu. Raziskovalci so pokazali, da je pri napovedovanju osipa strank poleg nabora in kvalitete podatkov pomemben tudi izbor najustreznejše metode in je zato nujno iskati nove in boljše tehnike ter pred končno odločitvijo preizkusiti več modelov.

V nalogi sem prikazala nabor različnih napovednih modelov (tako tistih, z dolgo tradicijo, kot novejših, ki zahtevajo močnejšo računalniško podporo) in pokazala, da različni napovedni modeli na istih vhodnih podatkih dajejo zelo različne rezultate. S tem sem dokazala, da je izbor modela zelo pomemben in da je pomembno, da pri izdelavi napovednega modela osipa strank za neko podjetje preizkusimo več metod in se na koncu odločimo za najboljšega.

V nalogi sem preizkusila štiri različne napovedne modele in jih primerjala med seboj po različnih kazalnikih, ki sem jih izračunala s programskim orodjem RapidMiner. Izkazalo se je, da imata največjo napovedno moč model s povečevanjem naklona, ki se je izkazal kot najboljši skoraj na vseh proučevanih kazalnikih, ter mi je z njim uspelo pojasniti oziroma napovedati največ dejanskih odhodov strank. Sledi mu naivni Bayesovski model, ki je prav tako pokazal dobro napovedno moč, saj njegovi kazalniki ne zaostajajo veliko za tistimi, ki jih sem jih dobila z najbolje ocenjenim modelom. Preostala proučevana modela, logistična regresija in metoda podpornih vektorjev, pa sta se slabše odrezala, saj njuna napovedna moč ni visoka.

V mojem primeru se je izkazalo, da so novejši pristopi k modeliranju osipa strank boljši, kot je na primer pogosto uporabljena in široko znana linearna regresija. Novejši modeli so boljši v prvi meri zato, ker ne zahtevajo izpolnjevanja predpostavk (na primer o porazdelitvi spremenljivk), na katerih temeljijo tradicionalni modeli in ki v praksi skoraj nikoli niso izpolnjene. Poleg tega pa so modeli, ki na primer uporabljajo povečevanje naklona, samonastavljivi (angl. autotuning) – že v model je vgrajeno to, da se možni parametri sami nastavijo na optimalne vrednosti.

Kot naslednji korak v proučevanju pa želim oba najboljša modela preizkusiti v praksi na celotni bazi strank ter narediti napoved osipa strank in preveriti, kako se teoretična modela izkažeta v dejanskem poslovnem življenju in ali podjetju lahko prinedeta dodano vrednost, ki bi se izrazila v manjšem osipu strank ter se tem posledično povečanih prihodkih.

LITERATURA IN VIRI

1. Abbasimehr, H., Setak, M., Tarokh, M. (2014*).* A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction. *International Arab Journal of Information Technology*, *11*(6).
2. Ahn, J.H., Han, S.P. & Lee, Y.S. (2006*).* Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy, 30*(10–11), 552–568.
3. Athanassopoulos, A.D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of Business Research, 47*(3), 191–207.
4. Baumann, A., Lessmann, S., Coussement, K., & De Bock, K. W. (2015). Maximize What Matters: Predicting Customer Churn with Decision-Centric Ensemble Selection. *ECIS 2015 Completed Research Papers. Paper 15.* Najdeno 3. julija 2016 na spletnem naslovu http://aisel.aisnet.org/ecis2015_cr/15
5. Bhattacharya, C.B. (1998). When customers are members: customer retention in paid membership contexts. *Journal of the Academy of Marketing Science, 26*(1), 31–44.
6. Bolton, R.N. (1998). A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction, *Marketing Science, 17*(1), 45–65.
7. Box-Steffensmeier J. M., & Jones, B. S. (2004). *Event History Modeling. [A Guide for Social Scientists]*. Cambridge: Cambridge University Press.
8. 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.
9. Burez, J. & Van den Poel, D. (2007). CRM At a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications, 32*(2), 277–288.
10. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step–by–step Data Mining guide. *SPSS Inc*. Najdeno 15. junija 2016 na spletnem naslovu ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP–DM.pdf
11. Correa Bahnsen, A., Aouada, D. & Ottersten, B. (2015). A novel cost-sensitive framework for customer churn predictive modeling. *Decision Analytics, 2*(5).
12. Coussement, K., & Van den Poel, D. (2006). 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.
13. Datta, P., Masand, B., Mani, D.R., & Li, B. (2000). Automated Cellular Modeling and Prediction on a Large Scale. *Artificial Intelligence Review*. *14*(6), 485–502.

14. Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, *27*(8): 861 – 874.
15. Fayyad U.M., Piatetsky-Shapiro G., Smyth P. & Zthurusamy R. (eds.). (1996). *Advances in Knowledge Discovery and Data Mining.* Cambridge MA: AAAI/MIT Press.
16. Fu, L. & Wang, H. (2013). Estimate Attrition Using Survival Analysis. *Causality Actuarial Society Spring Meeting, Canada*. Najdeno 3. julija 2016 na spletnem naslovu https://cas.confex.com/cas/spring13/webprogram/Handout/Paper1619/Estimate%20Attrition%20Using%20Survival%20Analysis%20Spring%202013.pdf
17. Ganesh, J., Arnold, M., & Reynolds, K. (2000). Understanding the customer base of service providers: An examination of the differences between switchers and stayers. *Journal of Marketing, 64*(3), 65–87.
18. Gençer, M., Bilgin G., Zan Ö. & Voyvodaoğlu T. (2014). Detection of Churned and Retained Users with Machine Learning Methods for Mobile Application. Poglavje Design, User Experience, and Usability: User Experience Design for Diverse Interaction Platforms and Environments. *Springer International Publishing Switzerland, 8518*, 234–245.
19. Glady, N., Baesens, B. & Croux, C. (2009) Modeling Churn Using Customer Lifetime Value. *European Journal for Operational Research, 197*(1), 402–411.
20. Hadden, J. (2008). *A Customer Profiling Methodology for Churn Prediction*. Cranfield: Ph.D. Cranfield University.
21. Hadden, J., Tiwari, A., Roy, R. & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research, 34*(10).
22. Hashmi, N., Butt, N.A. & Iqbal, M. (2013). Customer churn prediction in telecommunication. *International Journal of Computer Science Issue, 10*(5).
23. Hassouna, M., Tarhini, A., Elyas, T. & Abou Trab M.S. (2015). Customer Churn in Mobile Markets: A Comparison of Techniques. *International Business Research, 8*(6), 224 – 237.
24. Hoggart, C.J. & Griffin, J.E. (2001). A Bayesian Partition Model for Customer Attrition. In: Bayesian Methods with Applications to Science, Policy and Official Statistics. *Selected Papers from the Sixth World Meeting of the International Society for Bayesian Analysis*, 223–232.
25. Huang, B., Kechadi, M.T. & Buckley, B. (2012) Customer churn prediction in telecommunications. *Expert Systems with Applications, 39*, 1414–1425.
26. Hung, S., Yen, D.C. & Wang, H. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications, 31*, 515–524.
27. Jacoby J., Chesnut R.W. (1978). *Brand Loyality.* New York: John Wiley & Sons.
28. Jamal, Z. & Bucklin, R. E. (2006). Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach. Journal of Interactive Marketing, *20*, 16–29.

29. Jones T.O. & Sasser W.E. (1995). Why Satisfied Customers Defect. Harward Business Review, Nov.–Dec. Najdeno 25. Julija 2016 na spletnem naslovu https://hbr.org/1995/11/why-satisfied-customers-defect
30. Karnstedt, M., Hennessy, T., Chan, J., Basuchowdhuri, P., Hayes, P. & Strufe T. (2010). Churn in Social Network. B. Furht (ur.), *Handbook of Social Network Technologies and Applications, Springer US,* 185–220.
31. Kim, H. & Yoon, C. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy, 28*, 751–65.
32. Kim, M., Park, M. & Jeong, D. (2004) The effects of customer satisfaction and switching barrier on customer loyalty in Korean telecommunications services. *Telecommunications Policy, 28* (2), 145–159.
33. Kim, Y. S., Lee, H. & Johnson, J. D. (2013). Churn management optimization with controllable marketing variables and associated management costs. *Expert Systems with Applications, 40*(6), 2198 – 2207.
34. Kotu V., Deshpanade B. (2015). *Predictive Analytics and Data Mining.* Burlington: Morgan Kaufmann Publishers Inc.
35. Larivière, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications, 27*(2), 277–285.
36. Lidwell, W., Holden, K. & Butler, J. (2010). *Universal Principles of Design, Revised and Updated: 125 ways to Enhance Usability, Inflence Perception, Increase Appeal, Make Better Design Decisions, and Teach Through Design*. Beverly, MA: Rockport Publishers.
37. Lima, E., Mues, C. & Baesens, B. (2009). Domain knowledge integration in data mining using decision tables: Case studies in churn prediction. *Journal of Operational Research Society, 60*(8), 1096–1106.
38. Lu, J. (2002). *Predicting customer churn in the telecommunications industry – An application of survival analysis modeling using SAS*. SAS User Group International (SUGI27) Online Proceedings, 114–127.
39. Madden, G., Savage, S., J. & Coble-Neal, G. (1999). Subscriber churn in the Australian ISP market. *Information Economics and Policy, Elsevier, 11*(2), 195–207.
40. Mozer, M.C., Wolniewicz, R., Grimes, D.B., Johnson, E. & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Trans Neural Netw., 11*(3), 690–696.
41. Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. (2006). Defection detection: Improving predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204–211.
42. Ng, K. & Liu, H. (2000). Customer Retention via Data Mining. *Artificial Intelligence Review, 14*(6), 569–590.

43. Ngai, E.W.T., Xiu, Li & Chau, D.C.K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification, *Expert Systems with Applications, 36*, 2592 – 2602.
44. Nie, G., Rowe, W., Zhang, L., Tian, Y. & Shi, Y. (2011). Credit card churnforecasting by logistic regression and decision tree. *Expert Systems with Applications, 38*(11), 15273–15285.
45. Oliver R.L. (1999). Whence Consumer Loyalty. *Journal of Marketing, 63* (Special Issue), 33–44.
46. Oliver R.L. (1997). *Satisfaction: A Behavioral Perspective on the Consumer*. New York: Irwin/McGraw–Hill.
47. Owczarczuk, M. (2010). Churn models for prepaid customers in the cellular telecommunication industry using large data marts. *Expert Systems with Applications, 37*, 4710–4712.
48. Portela, S. & Menezes, R. (2011). Detecting Customer Defections: an application of continuous duration models. *Journal of Global Strategic Management, 5*(1), 22–30.
49. Pyle D., (1999). *Data Preparation for Data Mining*, Burlington: Morgan Kaufmann Publishers Inc.
50. Qi, J., Zhang, L., Liu, Y., Li L, Zhou, Y., Shen, Y., Liang, L. & Li, H. (2009) AD TreesLogit model for customer churn prediction. *Annals of Operations Research, 168*(1), 247–265.
51. Ray, S. (2015) *Understanding Support Vector Machine algorithm from examples (along with code).* Najdeno 25. junija 2016 na spletnem naslovu https://www.analyticsvidhya.com/blog/2015/10/understaing–support–vector–machine–example–code/
52. Reichheld, F.F. & Sasser, W.E. (1990.) Zero Defections: Quality Comes to Services. *Harvard Business Review, 68*(5), 105–111.
53. Shearer, C. (2000). The CRISP–DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing, 5*(4), 13–22.
54. Van den Poel, D. (2003). Predicting Mail–Order Repeat Buying. Which Variables Matter? *Review of Business and Economics, Katholieke Universiteit Leuven, Faculteit Economie en Bedrijfswetenschappen, 0*(3), 371–404.
55. Van den Poel, D. & Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*(1), 196–217.
56. Verbeke, W. (2012 b). *Profit Driven Data Mining in Massive Customer Networks: New Insights and Algorithms*, Leuven: Ph.D., Katholieke Universiteit Leuven.
57. Verbeke, W., Dejaeger K., Martens D., Hur J. & Baesens B. (2012 a). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European journal of Operational research, 218*, 211–229.
58. Verbeke, W., Martens, D., Mues, D. & Baesens B. (2011) Building comprehensible customer churn prediction models with advanced rule induction techniques, *Expert*

Systems with Applications, 38(3). Najdeno 25. junija 2016 na internetni strain http://dx.doi.org/10.1016/j.eswa.2010.08.023

59. Verbraken, T., Verbeke, W. & Baesens, B. (2014) Profit optimizing customer churn with Bayesian network classifiers. *Intelligent Data Analysis–Business Analytics and Intelligent Optimization, 18*(1), 3–24.
60. Wei, C.P. & Chiu, I. T. (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with App lications, 23*, 103–112.
61. Wong, K.K. (2011) Using Cox regression to model customer time to churn in the wireless telecommunication industry. *Journal of Targeting, Measurement and Analysis for marketing, 19*(1), 37–43.