UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

MAGISTRSKO DELO

ANALIZA PRILOŽNOSTI IN OVIR UVEDBE PODATKOVNEGA RUDARJENJA V PRAVNO INFORMACIJSKEM PORTALU IUS-INFO

Ljubljana, avgust 2018

MIHA KRONOVŠEK

IZJAVA O AVTORSTVU

Podpisani Miha Kronovšek, študent Ekonomske fakultete Univerze v Ljubljani, avtor predloženega dela z naslovom Analiza priložnosti in ovir uvedbe podatkovnega rudarjenja v pravno-informacijskem portalu IUS-INFO, pripravljenega v sodelovanju s svetovalcem prof. dr. Jurijem Jakličem

IZJAVLJAM

1. da sem predloženo delo pripravil 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 poskrbel/-a, 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 pridobil/-a vsa potrebna dovoljenja za uporabo podatkov in avtorskih del v predloženem delu in jih v njem jasno označil/-a;

7. da sem pri pripravi predloženega dela ravnal/-a 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 30.8.2018 Podpis študenta: _____

KAZALO

UVOD **1**

1 POSLOVNA INTELIGENCA IN PODATKOVNO RUDARJENJE **4**
1.1 Delitev tehnologij poslovne inteligence **4**
1.2 Podatkovno rudarjenje in učenje iz preteklih podatkov **5**
1.3 Glavne naloge podatkovnega rudarjenja **6**
1.3.1 Analiza asociacij 6
1.3.2 Analiza skupin 7
1.3.3 Napovedovanje 7
1.3.4 Zaznavanje odstopanj 7
1.4 Pregled prednosti, priložnosti, koristi, slabosti, nevarnosti, izzivov, ovir in omejitev podatkovnega rudarjenja **8**
1.4.1 Velikost podjetja, organizacijska pripravljenost in zahtevana znanja 8
1.4.2 Varnost, zasebnost in integriteta podatkov 8
1.4.3 Ustreznost in kakovost podatkov 9
1.4.4 Koristi in finančni učinki 10
1.5 Uporaba podatkovnega rudarjenja pri trženju in prodaji **10**
1.5.1 Ohranjanje in napovedovanje potencialnih kupcev 11
1.5.2 Navzkrižna prodaja in prodaja dražje različice 11
1.5.3 Oglaševanje 11
1.6 Primeri dosedanjih aplikacij **12**

2 PROCES PODATKOVNEGA RUDARJENJA **13**
2.1 Razumevanje področja obravnave **14**
2.2 Razumevanje podatkov **14**
2.3 Priprava podatkov **14**
2.4 Modeliranje **15**
2.5 Ocena rezultatov **16**
2.6 Uvedba v prakso **16**
2.7 Pogosto uporabljene metode in tehnike strojnega učenja **16**
2.7.1 Asociativna pravila 17
2.7.2 Globoko učenje 17
2.7.3 Najbližji sosed 17
2.7.4 Odločitvena drevesa 17
2.7.5 Razvrščanje v skupine 18
2.7.6 Regresijska analiza 18
2.8 Orodje RapidMiner 18

3 PRIMER PODATKOVNEGA RUDARJENJA NA PORTALU IUS-INFO **20**
3.1 O portalu, podjetju in dejavnosti **20**

3.2 Faze projekta **20**
3.2.1 Razumevanje področja obravnave 21
3.2.2 Razumevanje podatkov 22
3.2.2.1 Postopek izdelave relacijskega modela 23
3.2.3 Priprava podatkov 25
3.2.3.1 Kakovost podatkov 33
3.2.3.2 Izzivi pri integraciji 34
3.2.4 Modeliranje 36
3.2.4.1 Testno modeliranje 36
3.2.4.2 Model preprečitve odhodov kupcev 39
3.2.4.3 Model prioritetne lestvice potencialnih kupcev 43
3.2.5 Ocena rezultatov 45
3.2.5.1 Model preprečitve odhodov kupcev 45
3.2.5.2 Model prioritetne lestvice potencialnih kupcev 49
3.2.6 Uvedba v prakso 53
3.2.6.1 Model preprečitve odhodov kupcev 53
3.2.6.2 Model prioritetne lestvice potencialnih kupcev 54

4 DISKUSIJA **55**

SKLEP **58**

LITERATURA IN VIRI **60**

PRILOGA

KAZALO TABEL

Tabela 1: Kakovost podatkov 33
Tabela 2: Primerjava zmogljivosti odločitvenih dreves in nevronske mreže 42

KAZALO SLIK

Slika 1: CRISP DM metodologija 13
Slika 2: Magični kvadrant podatkovne znanosti in platform strojnega učenja 19
Slika 3: IUS-INFO relacijski model projekta podatkovnega rudarjenja 24
Slika 4: Proces orodja PDI 26
Slika 5: Izračun spremenljivke velikost podjetja 27
Slika 6: Izračun spremenljivke skupno število emailov in klicev 28
Slika 7: Izračun spremenljivk zapustil podjetje in akcijski naročnik 30
Slika 8: Izračun spremenljivk število plačljivih paketov in trajanje naročnin 31
Slika 9: Izračun 9 finančnih spremenljivk 32

Slika 10: Postopek formatiranja datuma 35
Slika 11: Testni model preprečitve odhodov kupcev 36
Slika 12: Povzetek testnega modela preprečitve odhodov kupcev 37
Slika 13: Grafikon dviga testnega modela preprečitve odhodov kupcev 37
Slika 14: Testni model prioritetne lestvice potencialih kupcev - grafikon profilov 38
Slika 15: Testni model prioritetne lestvice potencialih kupcev - pomembnost spremenljivk 39
Slika 16: Glavni proces izgradnje modela preprečitve odhodov kupcev v orodju RapidMiner 41
Slika 17: Merila kvalitete (nekaterih) spremenljivk odločitvenega drevesa in nevronske mreže 41
Slika 18: Glavni proces izgradnje modela prioritetne lestvice potencialnih kupcev v orodju RapidMiner 44
Slika 19: Merila kvalitete spremenljivk segmentacije 44
Slika 20: Korelacijska matrika uporabljenih spremenljivk segmentacije 45
Slika 21: Povzetek drevesa odločanja 46
Slika 22: Povzetek nevronske mreže 46
Slika 23: Končna rešitev drevesa odločanja 47
Slika 24: Končna rešitev nevronske mreže 47
Slika 25: Grafikon dviga drevesa odločanja 48
Slika 26: Grafikon dviga nevronske mreže 49
Slika 27: Opis najpomembnejših značilnosti segmentov brezplačnih uporabnikov 49
Slika 28: Profili brezplačnih uporabnikov 50
Slika 29: Toplotna karta spremenljivk brezplačnih uporabnikov 50
Slika 30: Profili plačljivih uporabnikov 51
Slika 31: Opis najpomembnejših značilnosti segmentov plačljivih uporabnikov 51
Slika 32: Toplotna karta spremenljivk plačljivih uporabnikov 52
Slika 33: Primerjava profilov brezplačnih in plačljivih uporabnikov 52

SEZNAM KRATIC

IS – informacijski sistem
angl. – angleško
SME – (angl. Small and Medium Enterprises); mala in srednje velika podjetja
mio – milijon
EU – Evropska unija
PDI – orodje Pentaho Data Integration
BI – (angl. Business Intelligence); poslovna inteligenca
BIS – (angl. Business Inteligence System); poslovno-inteligenčni sistem
GDPR – (angl. General Data Protection Regulation); Splošna uredba EU o varstvu podatkov
CRM – (angl. Customer Relationship Management); upravljanje odnosov s kupci

CRISP DM – (angl. Cross-Industry Standard Process for Data Mining); medsektorski proces podatkovnega rudarjenja

KDD – (angl. Knowledge Discovery in Database); metodologija odkrivanja znanja v zbirkah podatkov

SEMMA – (angl. Sampling, Exploration, Modification, Model and Verification); metodologija vzorčenja, raziskovanja, spreminjanja, modeliranja in preverjanja

ETL – (angl. Extract, Transorm, Load); proces črpanja, pretvorbe in nalaganja podatkov

KNN – (angl. k-Nearest Neighbor); tehnika k-najbljižjih sosedov

CART – (angl. Classification and Regression Trees); klasifikacijska in regresijska odločitvena drevesa

CHAID – (angl. Chi Square Automatic Interaction Detection); odločitvena drevesa, ki avtomatično zaznavajo interakcijo s pomočjo $\chi 2$

SZAT – sistem za avtomatizacijo trženja

ERP – (angl. Enterprise Resource Planning); celovita informacijska rešitev

GK – glavni ključ

SURS – Statistični urad Republike Slovenije

UVOD

Podjetja dandanes generirajo ogromne količine transakcijskih podatkov, ki predstavljajo koristi, če jih uspešno pretvorijo v informacije. Ocenjena količina informacij naj bi se na svetu podvojila skoraj vsakih 20 mesecev (Dwivedi, Kasliwal, & Soni, 2016, str. 1). Podatki so shranjeni v številnih informacijskih sistemih (v nadaljevanju IS), posledično so različnih oblik, časovnih serij ter vsebinsko težko razumljivi kot povezana celota. Upravljanje IS poteka bodisi v podjetju ali pri zunanjem izvajalcu, kar za sabo pusti določene prednosti in omejitve.

Mala in srednje velika podjetja (angl. *Small and Medium Enterprises,* v nadaljevanju SME) predstavljajo 99 % od ocenjenih 19,3 milijonov (v nadaljevanju mio) podjetij v Evropski uniji (v nadaljevanju EU) in ustvarjajo 65 mio zaposlitev, kar predstavlja dve tretjini vseh zaposlitev. Povprečno evropsko podjetje šteje zgolj 4 zaposlene, vključno z lastnikom ali managerjem. Iz njihove majhnosti lahko sklepamo, da ponavadi nimajo primerno usposobljenih kadrov za analitične potrebe. V primeru, da prepoznajo koristi uporabe podatkovnega rudarjenja (angl. *Data Mining)* in različnih virov podatkov, bodo v prihodnje verjetno bolj motivirana k investiranju virov v podatkovno analitiko (angl. *Data Analytics*), (Coleman Shirley, 2016, str. 850-851).

Kot vemo, za poslovno odločanje ponavadi niso primerni neposredno transakcijski, surovi podatki, ampak podatki pripravljeni posebej za ta namen. Če upoštevamo dejstva, da podatki naraščajo v obsegu (angl. *Data Volume*), raznolikosti formatov (angl. *Data Variety*), da se čas procesiranja le-teh razlikuje (angl. *Data Velocity*) ter da je v podatkih vse več negotovosti (angl. *Data Veracity*), lahko hitro pridemo do zaključka, da tradicionalne aplikacije za obdelavo ne izpolnjujejo naraščajočih zahtev. Rešitev predstavlja nov pristop, imenovan podatkovno rudarjenje (Dwivedi in drugi, 2016, str. 1).

Podatkovno rudarjenje je hitro rastoča disciplina, trenutno ena najpomembnejših paradigem naprednih inteligenčnih sistemov in sistemov za podporo odločanja. Lahko ga definiramo kot rudarjenje koristnih informacij in znanja iz baz podatkov, ki se odrazijo v finančnih in nefinančnih koristih. Je proces odkrivanja novih zanimivih vzorcev, povezav, razmerij in trendov iz ogromne količine podatkov. Ta metoda ni omejena na specifično dejavnost ali disciplino, uporablja se tako za poslovne kot neposlovne namene. Zahteva inteligentno tehnologijo, ustrezno poznavanje pristopa in željo po odkrivanju potencialnega skritega znanja iz množice strukturiranih in ali nestrukturiranih podatkov (Amani & Fadlalla, 2017, str. 2; Farooqi & Raza, 2012, str 1).

Ideja omenjenega pristopa je spraviti ogromno količino podatkov na skupni imenovalec ter s pomočjo algoritmov strojnega učenja (angl. *Machine Learning*) izdelati modele, ki dajejo prej nepoznane odgovore. Torej na podlagi dosedanjih podatkov iščemo dodano vrednost, ki se obrestuje v prihodnje.

Podatkovno rudarjenje prinaša potencialne velike koristi, konkurenčne prednosti za podjetje, ki odtehtajo investiran napor in ne bi bile dosegljive z drugačnim, bolj preprostim analitičnim pristopom. V mislih imamo prihranke časa oddelkov, predvsem oddelka prodaje in trženja, kjer je, poleg ostalih veščin, vse bolj potreben kvantitativni pristop. S pomočjo podatkovnega rudarjenja lahko organizacija tako poveča prihodke kot zmanjša stroške. Prihodke je možno povečati tako, da identificiramo kupce, ki nas zapuščajo, in s pravim pristopom zadržimo profitabilne. Prav tako lahko ugotovimo, kateri nekupci v naši podatkovni bazi so potencialni, da postanejo plačljivi uporabniki oz. je vanje vredno investirati različne vire. Ravno v nasprotni smeri na znižanje stroškov deluje identifikacija oseb, ki nas zapuščajo, in dopustitev, da neprofitabilni odidejo. Tudi neukvarjanje z nizko potencialnim segmentom vpliva na prihranke virov. Opisani ukrepi doprinesejo k višji profitabilnosti, saj imamo v podjetju na dolgi rok več dobrih kupcev, kar se odrazi v boljšem poslovnem izidu. 10 do 20 % kupcev ustvari 50 do 80 % dobička podjetja in s pomočjo podatkovnega rudarjenja je možno identificirati ustrezne v bazi vseh oseb. Podjetja, ki na trgu delujejo več let, se srečajo s padcem stopnje rasti, sočasno jim je na razpolago zgodovina transakcij oz. ogromna količina neizkoriščenega potenciala (Farooqi & Raza, 2012, str. 2; Gupta & Gupta, 2012, str. 5; Saarenvirta, 1998, str. 3-5).

Uporaba podatkovnega rudarjenja v praksi ni toliko razširjena oz. je verjetnost njegove uvedbe, zaradi same kompleksnosti, manjša. Potrebni so številni deležniki z raznolikim znanjem in pogledi z različnih zornih kotov. Ključen je sponzor projekta, to je oseba, ki zaradi svojega statusa v podjetju doprinese k realizaciji nalog, ne glede na ovire, ki se pojavijo na poti. Predpogoj za uresničitev zastavljenih ciljev so pravi podatki, zbrani na ustrezen način. Podjetje mora oceniti kvaliteto, vsebino, dostopnost in kvantiteto razpoložljivih podatkov (Saarenvirta, 1998, str 5). Zaradi enega ali več prej naštetih razlogov se lahko že na tej točki podatkovno rudarjenje ustavi. Če je podjetje še mlado, na razpolago verjetno nima velike količine podatkov o svojih kupcih. V primeru, da ni uspelo za vsako enoto pridobiti vseh vrednosti spremenljivk, lahko sklepamo, da je kvaliteta podatkov na nižji ravni. Če so bili podatki primarno zbrani za drugačen namen, je verjetnost njihove ustreznosti in posledično uporabnosti manjša. Pomembno vlogo igra tudi izbira orodja ali programskega jezika, v katerem se projekt tehnično izvede. Glavno vprašanje je, ali bomo s pomočjo določene izbire uspeli doseči poslovne zahteve. V primeru, da želimo popolno svobodo pri ustvarjanju, je primerna izbira programskega jezika Python, R ali sorodnega. Seveda je ta izbor pogojen z znanjem programiranja. V nasprotju danes na trgu obstajajo številna celovita orodja za podatkovno analitiko, kjer znanje programskih jezikov ni potrebno. Modeli se izdelujejo s pomočjo grafičnega vmesnika v obliki procesa. Za delujoč proces so potrebni vhodni podatki, vsaj nekaj gradnikov in nastavitve le-teh. Prav tako moramo omeniti potencialno investicijo v nakup komercialnega orodja in strošek izvajalca projekta (Dwivedi in drugi, 2016, str. 1-3).

Če potegnemo črto pod prej opisanimi dejavniki, ni vnaprej znano, kakšen je končni izplen podatkovnega rudarjenja, ko ga uporabimo v resničnem poslovnem okolju. Z drugimi

besedami povedano, ni mogoče zagotoviti, da bodo koristi odtehtale stroške oz. da bomo izpolnili zastavljene cilje, ni mogoče vedeti, ali bomo odkrili kaj novega, zanimivega in uporabnega ter v kolikem času bodo dobljeni rezultati in ugotovitve postali neuporabni oz. zastareli.

Zaradi prej naštetih razlogov bom v magistrski nalogi analiziral in predstavil prednosti, priložnosti in koristi na eni strani ter slabosti, nevarnosti, izzive, ovire in omejitve, ki jih podatkovno rudarjenje ponuja skozi vsako od faz, pred končno implementacijo v poslovno okolje.

Prvi cilj magistrskega dela je skozi analizo primera ugotoviti ključne dejavnike za končni uspeh projekta podatkovnega rudarjenja. Drugi cilj dela je preveriti možnost uporabe podatkovnega rudarjenja v podjetju LEXPERA d. o. o. s pomočjo izdelave v praksi delujočega modela preprečitve odhodov kupcev (angl. *Churn*) ter modela prioritetne lestvice potencialnih kupcev portala IUS-INFO. Doseganje zastavljenih ciljev in celoten empirični del temelji na primeru največjega hrvaškega pravno-informacijskega portala IUS-INFO (http://www.iusinfo.hr).

Za potrebe empiričnega dela in v sklopu projekta podatkovnega rudarjenja so bile izvedene 4 delavnice s ključnimi deležniki. Pred pričetkom projekta so bile podjetju predstavljene potencialne koristi, ki jih prinaša podatkovno rudarjenje, potrebni viri in njihova pričakovanja. Z viri so mišljeni razpoložljivi podatki in osebe, ki bodo sodelovale v projektu. Tehnična združljivost podatkov različnih sistemov je bila preverjena s pomočjo izdelave relacijskega modela v orodju Microsoft Access, po pridobitvi vzorčnih podatkov. Priprava podatkov za poslovno-inteligenčne potrebe je bila izvedena s pomočjo vrtilnih tabel orodja Microsoft Excel in orodja Pentaho Data Integration (v nadaljevanju PDI).

Preizkusno in dejansko modeliranje je bilo izvedeno v orodju RapidMiner. Za doseganje ciljev projekta so bili uporabljeni številni algoritmi strojnega učenja in primerjane zmogljivosti na obstoječih podatkih. Končne rešitve modelov vsebujejo na eni strani tehniko drevesa odločanja in nevronsko mrežo, kjer omenjena algoritma klasificirata enote, ter na drugi strani segmentacijo z metodo voditeljev, kjer je kot mera za razvrščanje uporabljena evklidska razdalja. Primerjava zmogljivosti odločitvenega drevesa in nevronske mreže je bila narejena s pomočjo grafikona dviga (angl. *Lift Chart*).

Na zadnji delavnici so bili celoten proces in končne ugotovitve predstavljeni sodelujočim na projektu. Magistrska naloga zajema tudi ukrepe na podlagi končnih rešitev modelov, do katerih sem med drugim prišel s pomočjo primerjave dejanskega stanja obdobja 6 mesecev po raziskavi z napovedmi modelov. Podajam interpretacijo modelov in lastna priporočila glede uporabe v poslovnem okolju.

Magistrsko delo je sestavljeno iz treh glavnih delov, dveh teoretičnih in empiričnega. Pred njimi se nahaja uvod s problematiko, namenom, cilji in metodami dela ter za njimi diskusija in sklep s povzetkom glavnih ugotovitev. Prvi teoretični del, imenovan Poslovna inteligenca in podatkovno rudarjenje, zajema: delitev tehnologij poslovne inteligence, podatkovno rudarjenje in učenje iz preteklih podatkov, glavne naloge podatkovnega rudarjenja, pregled prednosti, priložnosti, koristi, slabosti, nevarnosti, izzivov, ovir in omejitev podatkovnega rudarjenja, uporabo podatkovnega rudarjenja pri trženju in prodaji ter primere dosedanjih aplikacij. Drugi teoretični del, imenovan Proces podatkovnega rudarjenja, vsebuje: razumevanje področja obravnave, razumevanje podatkov, pripravo podatkov, modeliranje, oceno rezultatov, uvedbo v prakso, pogosto uporabljene metode in tehnike strojnega učenja ter nekaj podatkov o orodju RapidMiner. V empiričnem delu je predstavljen primer podatkovnega rudarjenja na portalu IUS-INFO, podjetje LEXPERA d. o. o. in njegova dejavnost. Podrobno je analizirana in predstavljena izvedba vsake faze: razumevanje področja obravnave, razumevanje podatkov, priprava podatkov, modeliranje, ocena rezultatov, uvedba v prakso ter njene glavne sestavine.

1 POSLOVNA INTELIGENCA IN PODATKOVNO RUDARJENJE

1.1 Delitev tehnologij poslovne inteligence

Poslovna inteligenca (angl. *Business Intelligence*, v nadaljevanju BI) ima številne definicije. Lahko jo opišemo kot nabor metodologij, procesov, arhitektur in tehnologij, s pomočjo katerih je zmožna pretvoriti surove podatke v smiselne in uporabne informacije. Prav tako lahko rečemo, da gre za skupen izraz, ki opisuje koncepte in metode, ki izboljšujejo poslovne odločitve oz. nudijo pomoč pri procesu odločanja (Williams, 2016, str. 52). BI je ključna za sprejemanje odločitev na vseh vodstvenih ravneh in vodi k večji učinkovitosti poslovanja, kar predstavlja konkurenčno prednost v zahtevnih tržnih pogojih. Delimo jo na več podpodročij (Klisarova-Belcheva, Ilieva, & Yankova, 2017, str. 298; Williams, 2016, str. 28):

- ad hoc poizvedovanje;
- kazalniki in nadzorne plošče;
- opozorila na elektronsko pošto;
- podatkovno rudarjenje;
- poizvedovanje in poročanje;
- statistične analize;
- večdimenzionalne analize.

Poslovno-inteligenčni sistemi (angl. *Business Inteligence Systems*, v nadaljevanju BIS) so posebna orodja za obdelovanje podatkov, poizvedovanje in poročanje. BIS za svoje učinkovito delovanje zahteva primerno tehnologijo. S tem mislimo na informacijsko-

tehnološko infrastrukturo, ki predstavlja njegovo ogrodje (Grublješič & Jaklič, 2015, str. 300).

1.2 Podatkovno rudarjenje in učenje iz preteklih podatkov

Podatkovno rudarjenje, interdisciplinarno podpodročje (Paiano & Pasanisi, 2017, str. 66), je hitro rastoča disciplina, ena najpomembnejših trenutnih paradigem naprednih inteligenčnih sistemov in sistemov za podporo odločanja. Lahko ga definiramo kot odkrivanje koristnih informacij in znanja iz baz podatkov (Verma & Nashine, 2012, str. 604). Je proces odkrivanja novih zanimivih vzorcev, povezav, razmerij in trendov v ogromni količini podatkov. Ta metoda ni omejena na specifično dejavnost ali disciplino, uporablja se tako za poslovne kot neposlovne namene. Zahteva inteligentno tehnologijo in željo po odkrivanju potencialnega skritega znanja v množici strukturiranih in ali nestrukturiranih podatkov (Amani & Fadlalla, 2017, str. 2; Farooqi & Raza, 2012, str. 1).

Glavni namen podatkovne analitike je uporaba naprednih analitičnih tehnik za velike in raznolike podatkovne nize (Dwivedi in drugi, 2016, str. 1). S pomočjo podatkovnega rudarjenja podjetja izkoriščajo potencial podatkov kot sredstva za doseganje finančnih in nefinančnih koristi. Tehnologija izboljša proces odločanja s pretvorbo podatkov v koristne in uporabne informacije, ki prinašajo konkurenčne prednosti (Bole, Popovič, Žabkar, Papa, & Jaklič, 2015, str. 1). S pristopom podatkovnega rudarjenja lažje identificirajo povezanost med kazalci uspešnosti, ocenijo verjetnost nastanka določenega dogodka in posledično zmanjšujejo intuitivne odločitve. Zelo pomembno je opozoriti, da gre samo za pristop, ki ne izloči potrebe po poznavanju posla, podatkov in analitičnih metod, ki so v ozadju orodja (Amani & Fadlalla, 2017, str. 2). Podatkovno rudarjenje ne najde vzorcev in znanja, ki jim je mogoče samodejno zaupati brez validacije (Patel, Karvekar, & Mehta, str. 710). Eden najpomembnejših delov projekta je ocena rezultatov. S tem mislimo na donosnost investicije (angl. *Return on Investment,* v nadaljevnaju ROI) ali podobnega kazalnika, ki bo sposoben izmeriti koristi uvedbe (Saarenvirta, 1998, str. 2). Podatkovno rudarjenje je najbolj učinkovito v organizacijah, ki lahko povežejo transakcije s točno določenimi kupci (Saarenvirta, 1998, str. 3).

Ponuja izzive in priložnosti na številnih področjih, med drugim v ekonomiji, astronomiji, kemiji, inženiringu, podnebnih študijah, geologiji, ekologiji, fiziki, biologiji, zdravstvu in računalništvu (Rawat & Rajamani, 2010, str. 1).

Podatkovna znanost (angl. *Data Science*) je kombinacija razumevanja podatkov, spretnosti shranjevanja in manipulacije podatkov ter razvoj in uporaba analitičnih metod. Že opisna analitika (angl. *Descriptive Analytics*) lahko zagotovi dobro predstavitev podatkov s pomočjo agregiranja, rezanja (angl. *Slice*) ali predstavitve v obliki kocke (angl. *Dice*). Ta metoda operira na preteklih podatkih in skuša pojasniti, kaj se je zgodilo. Poslovne potrebe so ponavadi bolj kompleksne, zato v naslednjem koraku pride v poštev napovedna analitika

(angl. *Predictive Analytics*). Le-ta uporablja analitične tehnike, s pomočjo katerih v dosedanjih podatkih odkrije zakonitosti, vzorce in razmerja ter naredi napovedi glede prihodnosti. Predpisujoča analitika (angl. *Prescriptive Anaytics*) gre še korak dlje in predlaga najboljšo rešitev za dan problem (Coleman Shirley, 2016, str. 851-852).

V empiričnem delu se dotaknem tudi spletnega rudarjenja (angl. *Web Mining*), ki je implikacija tehnik podatkovnega rudarjenja na podatkih, shranjenih v strežniškem dnevniku (angl. *Server Log*). Spletno rudarjenje se deli na 3 podpodročja, in sicer: rudarjenje spletne vsebine, rudarjenje spletne strukture in rudarjenje spletne uporabe. Osredotočam se na zadnje podpodročje, ki je definirano kot proces zbiranja informacij o uporabi spletnih strani. Namen rudarjenja spletne uporabe je odkriti vzorce uporabe, ki pripomorejo k boljšemu razumevanju in služenju potreb spletnih aplikacij (Rawat & Rajamani, 2010, str. 1-4).

Obstajajo razmere, v katerih so podatki negotovi in zanje obstaja večji potencial kot pri rudarjenju natančnih, transakcijskih podatkov. Negotovi podatki so povezani z verjetnostjo, da nastopijo v veliki bazi podatkov. Pri transakcijskih podatkih je verjetnost ponavadi 100 %. Zaradi omenjenega pogoja je odkrivanje zanimivih, uporabnih vzorcev pri negotovih podatkih težje in potencialna korist temu primerna (Leung, MacKinnon, & Jiang, 2017, str. 22). V empiričnem delu se ukvarjam izključno s transakcijskimi podatki.

1.3 Glavne naloge podatkovnega rudarjenja

S pomočjo podatkovnega rudarjenja se rešujejo različni izzivi in temu primerne so tudi naloge. Določena raziskava se lahko osredotoča na analizo asociacij, kjer se odkrivajo razmerja med objekti. Analiza skupin poskuša enote čim bolj pametno razdeliti v segmente. Napovedovanje že v imenu ponuja razlago aktivnosti te naloge. Zadnja naloga podatkovnega rudarjenja je zaznavanje odstopanj, kjer se iščejo enote z neobičajnimi vrednostmi atributov.

1.3.1 Analiza asociacij

Pri analizi asociacij in afinitete gre za identifikacijo predhodno znanih in neznanih vzorcev, ki obstajajo v sklopu skupine številnih objektov. Kot skupina objektov so mišljeni obiskovalci spletne strani, izdelki in storitve, objavljena spletna vsebina, uporabniški računi, gospodinjstva itd. Med verjetno najbolj prepoznavne spada analiza nakupne košarice (angl. *Marker-basket Analysis*), (Saarenvirta, 1998, str. 3), katere glavni namen je, da prodajalec dobi informacijo v zvezi z nakupnim obnašanjem oz. katere artikle osebe pogosto nakupujejo skupaj (Kaur & Kang, 2016, str. 78). Odkrita razmerja prinašajo potencialno vrednost pri ostalih tehnikah podatkovnega rudarjenja (Aggarwal, 2014, str. 130), saj že vnaprej poznamo povezave med objekti. Z analizo asociacij lahko napovemo kombinacijo številnih atributov, kar daje svobodo v primerjavi z napovedovanjem, kjer so napovedi omejene na vnaprej določeno število razredov (Witten, Hall, & Frank, 2011, str. 72).

1.3.2 Analiza skupin

Analizo skupin lahko poimenujemo tudi segmentacija ali združevanje v grozde (angl. *Clustering*). Cilj te analize je poiskati enote, ki so si podobne in jih združiti v homogeno, povezano skupino. Na drugi strani želimo doseži čim večjo heterogenost, različnost med skupinami. Pri delitvi ne uporabljamo vnaprej definiranih razredov, ampak jo izvedemo na takšen način, da skupine izgledajo naravne (Witten in drugi, 2011, str. 138). V ozadju pristopa so hierarhične in porazdelitvene metode (Paiano & Pasanisi, 2017, str. 70). Enote določene skupine se po značilnostih močno razlikujejo od enot preostalih skupin. Ni jasnega pravila, na koliko skupin razdelimo populacijo, saj je to odvisno od primera do primera in vsebinske razlage posameznih segmentov. Definitivno morata biti vsaj 2, v praksi se podatke ponavadi razdeli na do 15 segmentov. Grozdenje je uporabljeno za celovit pogled na podatke in poslovanje podjetja (Farooqi & Raza, 2012, str. 3).

1.3.3 Napovedovanje

Omenjeno kategorijo lahko razdelimo na podkategoriji klasifikacija in napovedovanje, odvisno od tipa podatkov, na katerih operiramo. Gre za metodo, ki prepozna pomembne razrede podatkov ali napove prihodnji trend podatkov. Cilj je izdelati model, ki napove bodoče vedenje strank s klasifikacijo zapisov v razrede. Število razredov je vedno končno število, določeno vnaprej (Refaat, 2007, str. 7). Pogoj za uspešno uvrščanje enot v razrede je izpolnjevanje vnaprej določenih meril oziroma kriterijev. Za klasifikacijo so pogosto uporabljena odločitvena drevesa, Bayesov klasifikator, pravila za klasificiranje in metoda podpornih vektorjev. Napovedovanje je ocenjevanje bodočih vrednosti, kot primer lahko navedemo ceno delnic na borzi. Poleg odločitvenega drevesa in nevronske mreže je možno napovedovanje izvesti tudi z različnimi regresijskimi analizami (Farooqi & Raza, 2012, str. 3). Napovedni modeli zahtevajo fazo treniranja in fazo testiranja modela. Zelo pomembno je, da enote striktno ločimo in jih uporabimo bodisi v prvem ali drugem primeru (Aggarwal, 2014, str. 30). Zmožnost napovedovanja v prvi meri zavisi od kvalitete in kvantitete podatkov (Saarenvirta, 1998, str. 3).

1.3.4 Zaznavanje odstopanj

Podatek, ki značilno odstopa od običajnega, se imenuje osamelec (angl. *Outlier*). Razlog za njegovo odklonilno naravo je možno interpretirati s tem, da je nek drug mehanizem odgovoren za njegov nastanek, ki je lahko namerne ali nenamerne narave. S pomočjo podatkovnega rudarjenja je možno zaznati takšne enote (Kaur & Kang, 2016, str. 80). Analiza anomalij se izvede na podlagi primerjave aritmetične sredine, standardnega odklona ali mer nestanovitnosti (angl. *Volatility Measures*). Prav tako se primerjajo vrednosti atributov uporabnika ali skupine uporabnikov skozi daljše časovno oddobje, kjer je možno odkriti velike razlike (Shaw, Subramaniam, Tan, & Welge, 2001, str. 130). Algoritmi zaznavanja odstopanj so v praksi uporabljeni kot varnostni mehanizmi na različnih nivojih varnosti (Cao, 2009, str. 118).

1.4 Pregled prednosti, priložnosti, koristi, slabosti, nevarnosti, izzivov, ovir in omejitev podatkovnega rudarjenja

1.4.1 Velikost podjetja, organizacijska pripravljenost in zahtevana znanja

Mnoga velika podjetja prve korake v smeri podatkovne analitike naredijo s pomočjo partnerja oziroma zunanjega izvajalca, ki je strokovnjak področja z veliko mero znanja in izkušenj. Trgovska veriga Tesco je na primer leta 1995 pričela sodelovati s podjetjem Dunnhumby, ki je specializirano za področje zbiranja in analize podatkov. Skupaj so ustvarili program kartice zvestobe (angl. *Loyalty-Card Program*), ki je spodbudil kupce k večjemu zapravljanju. Podjetju je pomagal k boljšemu razumevanju, katere podatke o kupcih je potrebno zbirati, kakšne so nakupne navade in posledično so lahko pričeli s targetiranim oglaševanjem, kar je prineslo večjo prodajo. Končni rezultat je bil nakup prej omenjenega podjetja s strani Tesca (Coleman Shirley, 2016, str. 856).

Za uspešno izvedbo projekta podatkovnega rudarjenja je potrebno vključiti ljudi različnih znanj in pogledov. Najprej potrebujemo nekoga, ki je pobudnik oziroma sponzor projekta in ima moč, da se projekt pripelje do konca. Cilji projekta pridejo ponavadi od vrha, managementa, katerega del je lahko tudi sponzor. Na drugi strani potrebujemo tehnična znanja v zvezi z IS in bazami podatkov, kjer so shranjeni podatki. Nekje vmes, med managerskim in tehničnim pogledom, najdemo analitika, katerega naloga je pretvoriti razpoložljive podatke v modele, ki so skladni s cilji projekta. Pomemben faktor predstavlja pritisk konkurence, ki spodbudi organizacijo k inovativnim pristopom, kot je podatkovno rudarjenje. Deležniki imajo cilje in vizijo, vendar je pot do končnega uspeha zaradi spremenljivih zahtev tekom projekta nejasna. Skladno temu je priporočljiv prilagodljiv pristop z manj podrobnejšega planiranja in možnostjo sprotnega učenja in inoviranja (Bole in drugi, 2015, str. 2-3).

1.4.2 Varnost, zasebnost in integriteta podatkov

Podatki podjetja so ponavadi zasebni, namenjeni uporabi zgolj znotraj organizacije (Coleman Shirley, 2016, str. 854). Zaupnost podatkov definiramo kot tajnost podatkov, katerim skrijemo nabor vrednosti atributov pred nepooblaščenim uporabnikom (Wang, 2006, str. 366). Kot osebne podatke smatramo podatke identificiranih oseb. Vedenjski podatki se nanašajo na obnašanje ljudi v določenih okoliščinah (Cao, 2009, str. 115). Podatkovno rudarjenje vključuje tako osebne kot vedenjske podatke. Oseba, ki se strinja s koriščenjem njenih osebnih podatkov s strani organizacije, mora biti seznanjena z namenom koriščenja, zaščitnimi ukrepi zaupnosti in integritete, posledicami nedržanja dogovorjenih okvirov ter ostalimi pravicami, ki izvirajo iz tega razmerja (Witten in drugi, 2011, str. 34).

Splošna uredba EU o varstvu podatkov (angl. *General Data Protection Regulation*, v nadaljevanju GDPR), ki je stopila v veljavo 25. 5. 2018, določa nova pravila glede varstva osebnih podatkov na ravni EU, katere članica je tudi Hrvaška. Omenjena uredba ureja pravice posameznika in obveznosti upravljavcev ter obdelovalcev podatkov. Pravice

posameznika se nanašajo na osebne podatke posameznika, katerega osebni podatki se obdelujejo. Med temi pravicami so pravica do dostopa do osebnih podatkov ter pravice do popravka, pozabe, izbrisa, omejitve obdelave, ugovora in prenosljivosti podatkov (Informacijski pooblaščenec, 2016).

Določena je tudi obveznost upravljavcev, oseb, odgovornih za obdelavo podatkov glede zagotavljanja preglednih in lahko dostopnih informacij posameznikom, na katere se nanašajo osebni podatki, o obdelavi njihovih podatkov. V uredbi so podrobno opredeljene splošne obveznosti upravljavcev in oseb, ki osebne podatke obdelujejo v njihovem imenu – obdelovalcev. Med temi obveznostmi sta obveznost izvajanja ustreznih varnostnih ukrepov in obveznost uradnega obveščanja o kršitvah varstva osebnih podatkov. V skladu z uredbo bodo javni sektor ter podjetja, katerih temeljne dejavnosti zajemajo ali dejanja obdelave, ki pomenijo redno in sistematično obsežno spremljanje posameznikov, ali pa obsežno obdelavo posebnih vrst podatkov, morali imenovati uradno odgovorno osebo za varstvo podatkov. Upravljavci ne bodo več dolžni prijavljati zbirk osebnih podatkov v register zbirk osebnih podatkov, ostajajo pa dolžnosti sprejema katalogov. Te dolžnosti se krepijo in uvajajo tudi za pogodbene obdelovalce. Večji poudarek daje uredba tudi izvedbam analiz učinkov na varstvo osebnih podatkov. V primeru varnostnih incidentov, kot je npr. izguba osebnih podatkov, se uvaja obveznost poročanja nadzornemu organu in v določenih primerih tudi obveščanja vseh prizadetih posameznikov (Informacijski pooblaščenec, 2016).

GDPR na ravni EU na nek način omejuje podatkovno rudarjenje, saj je možno vključiti zgolj osebne podatke oseb, od katerih imamo eksplicitno dovoljenje za konkreten namen koriščenja.

1.4.3 Ustreznost in kakovost podatkov

Kakovost razumemo kot nivo, do katerega podatki ustrezajo merilom čistosti, popolnosti in zadostnosti (Bole in drugi, 2015, str. 3). Podatki podjetja so ponavadi zbrani za trenutne potrebe. Zadovoljiva kvaliteta za trenutne aktivnosti ni zagotovilo, da je kakovost na zadostnem nivoju za ostale namene (Coleman Shirley, 2016, str. 856). Izbira pravega vira podatkov je bistvenega pomena za celoten projekt podatkovnega rudarjenja. Popraviti je potrebno napačne zapise, nepomembna ali nasprotujoča si polja ter zastarele podatke (Verma & Nashine, 2012, str. 607). Prav tako ne želimo neravnovesja ali manjkajočih vrednosti (Cao, 2017, str. 64-65). Manjkajoče kategorične vrednosti lahko nadomestimo z najpogostejšo vrednostjo, manjkajoče številske pa s povrečno vrednostjo. Za pretvorbo numeričnih vrednosti v kategorične se uporabi diskretizacija (Goncalves, Barros, & Vieira, 2012, str. 3). Z omenjeno operacijo zmanjšamo število vrednosti določenega atributa tako, da jih združimo v intervalne vrednosti. S pomočjo normalizacije zmanjšamo razlike med podatki različnih virov in omogočimo njihovo primerljivost (Sullivan, 2012, str. 200-227).

1.4.4 Koristi in finančni učinki

Uvedba projekta podatkovnega rudarjenja podjetje prisili, da preveri kvaliteto, vsebino, dostopnost in količino razpoložljivih podatkov. Prav tako mora dokumentirati do sedaj uporabljene analitične metode in dosežene rezultate. Razumevanje poslovnega modela ter tehnične zmožnosti podjetja in poslovnih analitikov so med drugim sestavni del analize dosežkov. Že zaradi zgolj prej omenjenih koristi bi morala podjetja danes začeti s pilotnimi projekti podatkovnega rudarjenja (Saarenvirta, 1998, str. 5).

Vse več organizacij koristi v podatkih odkrito znanje kot razlog oz. povod za odločitve. Podjetja s samega vrha po omenjenem konceptu so bila v povrečju 5 % bolj produktivna in 6 % bolj profitabilna kot njihovi konkurenti (Coleman Shirley, 2016, str. 859). Algoritmi strojnega učenja so zmožni prioritizirati ali celo avtomatizirati odločanje. V poslovnem okolju je hitrost sprejema odločitve ključnega pomena, zato najbolj agilna podjetja analizirajo podatke IS v realnem času. Pri tem so identificirane prednostne naloge in potrebne aktivnosti za njihovo uresničitev (SAP, 2018).

Podatkovno rudarjenje prinaša visoke potencialne koristi, ki se na koncu odražajo tudi finančno. 10 in 20 % kupcev ustvari med 50 in 80 % dobička podjetja. S pomočjo podatkovnega rudarjenja lažje identificiramo te osebe (Saarenvirta, 1998, str. 3). Človeška sposobnost sočasnega procesiranja velikih in zapletenih zbirk podatkov se težko kosa z napredno podatkovno analitiko, ki deluje ravno na tej ravni in je gonilnik za boljše poslovne rezultate (SAP, 2018).

1.5 Uporaba podatkovnega rudarjenja pri trženju in prodaji

Pri trženju poznamo dva nasprotujoča si pristopa, masovnega (angl. *Mass Marketing*) in direktnega (angl. *Direct Marketing*). V prvem primeru uporabimo identično sporočilo za vse kupce, v drugem le-tega prilagajamo na ravni posameznika ali skupine posameznikov (Van Der Putten, 1999, str. 1). Pri direktnem, ki je v današnjem času prisoten v večji meri, je ključno upravljanje odnosov s kupci (angl. *Customer Relationship Management,* v nadaljevanju CRM). CRM definiramo kot proces pridobivanja, zadrževanja in rasti profitabilnih kupcev. CRM aktivnosti za sabo potegnejo stroške, zato je ključno vedeti, katere aktivnosti izvajati s katero osebo. Značilnosti kupcev in njihove nakupne navade so v veliki meri skrite ter posledično neizkoriščene (Shaw in drugi, 2001, str. 127). Dandanes nadaljnje poslovanje s stranko ni samoumevno, lojalnosti je čedalje manj, zato kupci izginjajo preko noči. Podatkovno rudarjenje je pristop, ki lahko okrepi odnose s strankami. Zgolj poznavanje dejanskih potreb in resničnih želja kupcem predstavlja vrednost in se odrazi v obliki zvestobe. Velikokrat imamo s terminom CRM v mislih že programsko opremo, ki podpira prej omenjeni proces in shranjuje vse relavantne informacije. CRM ni koncept niti projekt, ampak poslovna strategija, ki stremi k razumevanju trenutnih in potencialnih kupcev. Osredotočenost na kupca predstavlja pomembno sredstvo današnjega

časa (Farooqi & Raza, 2012, str. 2; Gupta & Gupta, 2012, str. 1-7). Področja koriščenja podatkovnega rudarjenja znotraj trženja in prodaje so številna.

1.5.1 Ohranjanje in napovedovanje potencialnih kupcev

Želja podjetja je, da zadrži kupce (angl. *Customer Retantion*), vendar je pomembno, katere. Fokus je na tistih, ki podjetju prinašajo sedanjo ali potencialno vrednost (Chorianopoulos, 2015, str. 94). Pravilnost izbire se obrestuje v obliki poslovnega izida. Na podlagi tistih, ki so organizacijo že zapustili, se izdela model. Uporabi se na sedanjih kupcih in tako ugotovi, kateri najverjetneje odhajajo v bodoče. Profitabilnih kupcev, tistih z visoko življenjsko vrednostjo in nizkim tveganjem, podjetje ne želi izgubiti (Saarenvirta, 1998, str. 3). Med drugim je omenjen model pogosto uporabljen pri prodaji na drobno, spletni prodaji in v finančni ter telekomunikacijski industriji (Klinkenberg & Hofmann, 2013, str. 7).

Napovedovanje potenciala nekupcev lahko izvedemo s pomočjo klasifikacije, in sicer ko oseba prvič vstopi v sistem, jo na podlagi njenih značilnosti uvrstimo v razred kupcev ali nekupcev. Uvrstitev je izvedena na podlagi dosedanjih kupcev oz. njihovih značilnosti. Drugo možnost preverbe potenciala ponuja segmentacija, ki osebe združi na podlagi istih značilnosti. Pristop podjetja do vsakega segmenta nato poteka na prilagojen način (Gupta & Gupta, 2012, str. 5).

1.5.2 Navzkrižna prodaja in prodaja dražje različice

Navzkrižna prodaja (angl. *Cross-selling*) in prodaja dražje različice (angl. *Up-selling*) se nanašata na promocijske aktivnosti, katerih cilj je povečati prodajo sorodnih ali tesno povezanih produktov, ki jih stranka koristi znotraj podjetja. Za potrebe navzkrižne prodaje je odlične rezultate pokazala analiza nakupne košarice, ki pove, kateri izdelki se ponavadi kupujejo skupaj. Zelo znan je primer nakupa piva in plenic. Za potrebe prodaje dražje različice lahko izvedemo model, ki segmentira kupce določenega izdelka po različnih verzijah le-tega. Znotraj segmentov ne najdražje različice iščemo kupce s podobnimi značilnostmi, kot so tisti iz najdražjega segmenta, in jim le-tega ponudimo (Patel in drugi, str. 712). S premium segmentom oz. izdelkom mislimo na tistega, ki je bolj profitabilen od obstoječega (Chorianopoulos, 2015, str. 94).

1.5.3 Oglaševanje

V tem primeru je govora o čim bolj targetiranem oglaševanju, bodisi preko tiskanih ali elektronskih medijev. V ozadju oglaševanja so stroški, zato je cilj imeti čim boljši odziv na vsako kampanjo. Dober odziv si zagotovimo s tem, da vemo, kakšne oblike ter vsebine naj bo sporočilo in komu ga pošiljati. Demografski podatki kupcev ponavadi ne korelirajo s podatki profitabilnosti, zato jih ne združujemo s podatki nakupnega obnašanja. Segmentacija na podlagi nakupnega obnašanja razdeli kupce skladno z njihovo dobičkonosnostjo. Demografska razdelitev znotraj omenjenih segmentov pomaga izbrati primerne kampanje in kanale oglaševanja (Saarenvirta, 1998, str. 4). Z naprednimi metodami je smotrno ugotoviti,

kateri kanal komunikacije je določenemu kupcu prva izbira in katero ponudbo bo verjetno sprejel (Gupta & Gupta, 2012, str. 4-5).

1.6 Primeri dosedanjih aplikacij

Ena od iranskih bank je na podlagi 30.000 strank izvedla model, ki napoveduje odziv na naslednjo trženjsko kampanjo oz. sprejem ponudbe s strani kupcev. Odziv brez modela je do sedaj v povprečju znašal zgolj 7 %, kar nakazuje na možnost izboljšave. Kot je znano, so s pošiljanjem povezani stroški in zato želimo čim boljši odziv oz. ponudbo pošljemo zgolj tistim osebam, ki jo bodo verjetno sprejele. S pomočjo modela, kjer so bili med drugim uporabljeni podatki o času od zadnjega nakupa stranke (angl. *Recency*), številu vseh nakupov stranke (angl. *Frequency*) in dosedanjih izdatkih stranke (angl. *Monetary*), dosežejo 57 % odziv, s kontaktiranjem zgolj 20 % populacije (Zhao & Cen, 2013, str. 177-178).

Telekomunikacijsko podjetje je želelo tradicionalno delitev kupcev na poslovne in fizične uporabnike ter njihove pripadajoče podskupine obogatiti s podrobnejšo segmentacijo obnašanja uporabnikov. Zbralo je podatke 6-mesečnega obdobja koriščenja storitev in s tem pridobilo stabilne, nevolatilne vzorce vedenja, namesto naključnih ali zastarelih. S pomočjo modela so stranke segmentirali v 5 skupin, ki jih je mogoče poimenovati na podlagi značilnosti njihovih pripadnikov: uporabniki SMS storitev, uporabniki mednarodnih storitev, profesionalni uporabniki, osnovni uporabniki in tipični pogovorni uporabniki. Segmentacijo so izvedli z namenom pridobitve segmentov na podlagi obnašanja in koriščenja teh vpogledov pri nagovarjanju strank. Poznavanje pravih želja in njihovo zadovoljevanje je privedlo k višjemu zadovoljstvu in zvestobi (Chorianopoulos, 2015, str. 332-359).

Neimenovana banka je izvedla model preprečitve odhodov kupcev segmenta z najvišjo vrednostjo. Po njihovem prepričanju je teh dobrih 20.000 strank ključnih za podjetje in jih ne želijo izgubiti. Podatki o uporabi plačilnih kartic so bili zbrani in analizirani za obdobje 1 leta. Na podlagi modela so identificirali 20 % strank srednjega in visokega tveganja, da zapustijo podjetje. Temu primerno so lahko ukrepali s številnimi manjšimi trženjskimi kampanjami (Chorianopoulos, 2015, str. 217-266).

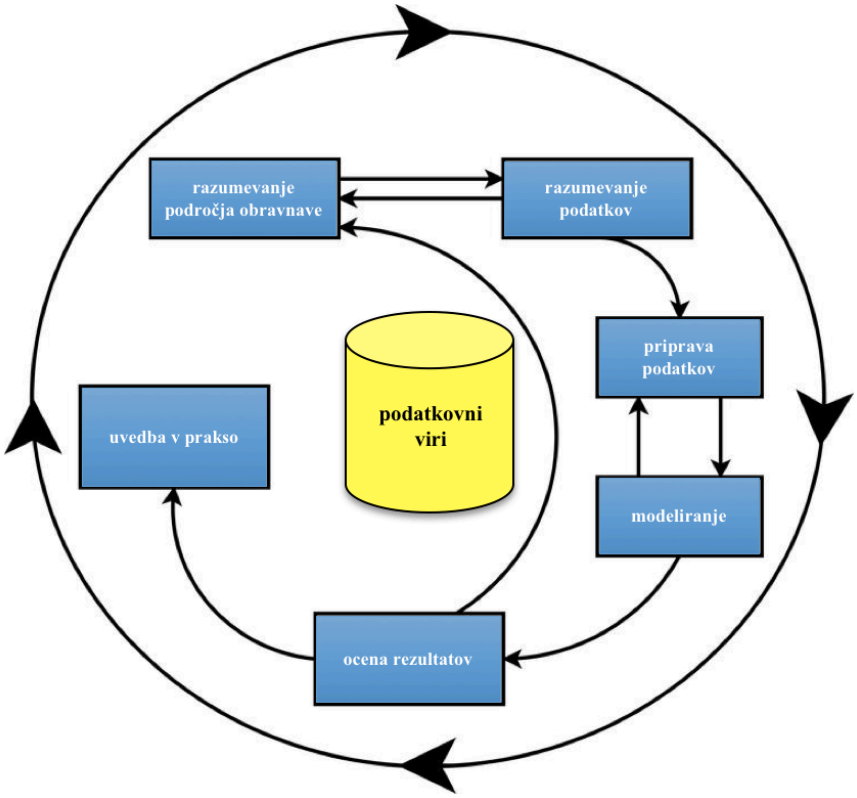
Podjetje Spotify nudi naročnikom oblačno storitev poslušanja glasbe, brez oglasnih sporočil. Kljub temu, da je njihova glavna dejavnost glasba, uporabljajo tehnike podatkovnega rudarjenja za iskanje povezav med skladbami. Za ta namen so kupili podjetje Echo Nest, specializirano v glasbeni analitiki. Rezultat nakupa se je pokazal v 42 % rasti prodaje v letu 2013 v primerjavi z letom poprej. Med drugim je podjetje sposobno napovedati, katera glasba bo verjetno postala popularna oz. bo poskočila na vrh lestvic (Coleman Shirley, 2016, str. 856).

2 PROCES PODATKOVNEGA RUDARJENJA

V empiričnem delu sledim IBM-ovemu medsektorskemu procesu podatkovnega rudarjenja (angl. *Cross-Industry Standard Process for Data Mining*, v nadaljevanju CRISP DM), ki je le ena od možno uporabljenih metodologij pri podatkovnem rudarjenju. Poleg CRISP DM, prve izbire v praksi, spadata v ožjo skupino najbolj uporabljenih pristopov še metodologija odkrivanja znanja v zbirkah podatkov (angl. *Knowledge Discovery in Database*, v nadaljevanju KDD) ter metodologija vzorčenja, raziskovanja, spreminjanja, modeliranja in preverjanja (angl. *Sampling, Exploration, Modification, Model and Verification*, v nadaljevanju SEMMA), (Rogalewicz & Sika, 2016, str. 100-101).

CRISP DM sestavlja 6 medsebojno odvisnih faz, prehodi med njimi so tako naprej kot nazaj, z namenom izboljšanja celote. CRISP DM kot metodologija vključuje opise delov projekta, opise nalog posameznih delov in razmerja med aktivnostmi. Gre za proces, ki omogoča celovit pogled na življenjski cikel podatkovnega rudarjenja (IBM, 2016, str. 1-2). V nadaljevanju sledijo podrobnejši opisi posameznih faz in njihovih ključnih elementov.

Slika 1: CRISP DM metodologija



Vir: IBM (1994).

2.1 Razumevanje področja obravnave

Projekt podatkovnega rudarjenja se prične z razumevanjem področja obravnave (angl. *Business Understanding*). Potrebno je analizirati poslovni model podjetja, identificirati priložnosti, definirati želen rezultat oz. kaj bo organizacija s projektom pridobila (Chorianopoulos, 2015, str. 115). V proces mora biti vključenih čim več ključnih deležnikov, med drugim sponzor projekta, ki je gonilna sila v primeru zapletov na poti. V začetku je potrebno jasno definirati cilje, potencialne izzive in težave ter razpoložljive vire. Z viri mislimo na finančne, človeške vire, trajanje izvajanja projekta ter razpoložljive podatke. S tem zmanjšamo verjetnost končnega neuspeha. Enega ali več primarnih ciljev je potrebno jasno izraziti v pričakovani poslovni koristi, na primer dvig prodaje najdražjega naročniškega paketa za 15 % v primerjavi s trenutno prodajo. Doseganje zastavljenih ciljev lahko merimo objektivno ali subjektivno. Potrebno se je vprašati, ali obstajajo varnostne ali pravne omejitve v zvezi z dostopom do podatkov ali v zvezi s končnimi rezultati. Prav tako moramo vedeti, na kakšen način bodo rezultati predstavljeni vodstvu podjetja, bodisi kot razlaga celotnega postopka ali zgolj končnih rezultatov modelov (IBM, 2016, str. 3-9).

2.2 Razumevanje podatkov

Z razumevanjem podatkov (angl. *Data Understanding*) mislimo na podrobejši pregled razpoložljivih podatkov. Podatki se ponavadi nahajajo v različnih IS, v številnih bazah podatkov, njihova struktura, tip in format so redko enaki. Potrebno jih je analizirati in ugotoviti način njihove integracije. Najpogosteje se opis podatkov osredotoča na kvantiteto in kvaliteto. S kvantiteto mislimo na število vrstic (zapisov) in število stolpcev (atributov), s kvaliteto izrazimo stanje podatkov. Atributi oz. spremenljivke so vnaprej določen niz vrednosti, ki karakterizirajo določen zapis (Witten in drugi, 2011, str. 49). Podatki so zelo redko popolni, vsebujejo napake pri kodiranju, manjkajoče vrednosti ali druge nedoslednosti. V primeru enormnih zapisov je smotrn premislek o uporabi vzorčnih podatkov. S pomočjo razumevanja poslovne logike in skladno z vrednostmi statističnih koeficientov odkrijemo spremenljivke, ki predstavljajo vrednost za določen model (IBM, 2016, str. 11-15).

2.3 Priprava podatkov

Tretja faza je priprava podatkov (angl. *Data Preparation*), ki je ena najdolgotrajnejših in kritičnih z vidika uspešnosti projekta podatkovnega rudarjenja (Goncalves in drugi, 2012, str. 1-6). Pripravo podatkov razumemo kot proces črpanja, pretvorbe in nalaganja podatkov (angl. *Data Extraction, Transormation, Loading,* v nadaljevanju ETL), (Pan, Zhang & Qin, 2018, str. 275). Uspešna izvedba ETL-a je predpogoj, da bo model zgrajen pravilno in hitreje. Podatki so črpani iz homogenih ali heterogenih virov. Večina algoritmov strojnega učenja je sposobnih brati zgolj dvodimenzionalno tabelo. Operacijski podatki so ponavadi shranjeni v relacijskih bazah, ki imajo arhitekturo v obliki več povezanih tabel. Prvi korak je torej agregiranje podatkov in posledično zmanjšanje števila dimenzij do skladnosti z

zahtevami algoritmov (Refaat, 2007, str. 3). Generalno gledano poznamo 2 tipa spremenljivk, numeričnega in nominalnega. S prvim mislimo na številske zapise, medtem ko drugi shranjuje tekstovne nize. Poznamo tudi binarne atribute, ki zavzemajo izključno 2 številski vrednosti. Večina algoritmov strojnega učenja je sposobna operirati zgolj na številskih, vključno z binarnimi spremenljivkami. Temu primerno je potrebno v fazi priprave podatkov ustrezno kodirati nominalne vrednosti (Michalewicz, 2007, str. 51). Po predhodnih aktivnostih podatke naložimo v ciljno podatkovno bazo, natančneje v operativno bazo, podatkovni sklad ali podatkovno skladišče (Pan in drugi, 2018, str. 275). Podatkovno skladišče (angl. *Data Warehouse*) pripomore k učinkovitejši izvedbi podatkovnega rudarjenja, saj so podatki zbrani sumarno in pospešijo čas procesiranja algoritmov strojnega učenja (Gupta & Gupta, 2012, str. 2). Ocenjeno je, da ponavadi ta faza zahteva 50–70 % časa in napora izvajanja celotnega projekta. Dolgotrajnost je posledica velikosti baz podatkov, ki jih merimo z globino in širino. Širša baza vsebuje več stolpcev spremenljivk in globlja več enot opazovanja (Verma & Nashine, 2012, str. 605). Nekoliko časa prihranimo tako, da kvalitetno izvedemo predhodni fazi razumevanja posla in razumevanja podatkov. Izzivi in težave, ki jih je potrebno odpraviti (IBM, 2016, str. 17-21):

- manjkajoče vrednosti – v tem primeru lahko izločimo katerega od zapisov ali katero od spremenljivk ali napolnimo manjkajoče vrednosti z ocenjenimi;
- napake podatkov – potrebno je uporabiti logiko, lastnoročno odkriti in odpraviti napake ali izločiti določene spremenljivke;
- neskladnosti pri kodiranju podatkov – določiti je potrebno enotno shemo kodiranja in popraviti neskladnosti;
- manjkajoči ali slabi metapodatki – lastnoročno je potrebno preučiti nenavadna polja in ugotoviti pravilen pomen.

2.4 Modeliranje

Modeliranje (angl. *Modeling*) je učenje modela na podlagi situacije, kjer poznamo končni rezultat, in uporaba le-tega v drugi, kjer je ravno nasprotno (Verma & Nashine, 2012, str. 606). V fazi modeliranja se pričnejo kazati sadovi dosedanjega dela. Ustrezno pripravljeni podatki se modelirajo v številnih ponovitvah. Zgradimo več modelov, pustimo prvotne nastavitve parametrov, primerjamo natačnost modelov in kasneje prilagajamo parametre, dokler nismo zadovoljni. Optimalne nastavitve si zabeležimo, če se kdaj odločimo za avtomatizacijo ali ponovno izgradnjo z novimi podatki. Obstaja tudi možnost, da se vrnemo v fazo priprave podatkov in jih prilagodimo za potrebe izbranega modela. Cilj je odkriti model, ki daje čim bolj natančne in učinkovite rezultate. Različni pogledi so upoštevani pri določenem problemu ali izzivu, posledično za ta namen zgradimo številne modele. Na koncu jih je potrebno rangirati objektivno, kot je npr. natančnost modela, ter subjektivno, kot je npr. enostavnost uporabe ali predstavitve končnih rezultatov (IBM, 2016, str. 23-27). Zmogljivost napovednih modelov lahko primerjamo tudi s pomočjo grafikona dviga, ki pove, ali je napoved boljša od naključnega ugibanja. V tem primeru je želja, da s čim

manjšim deležem populacije dosežemo čim višji delež enot razreda, ki ga napovedujemo (RapidMiner, 2018d).

2.5 Ocena rezultatov

Na tej točki je večina projekta opravljenega. Potrebno je oceniti, ali rezultati prejšnje faze ustrezajo postavljenim merilom uspešnosti. Ocena rezultatov (angl. *Evaluation*) je lahko precej enostavna, in sicer na podlagi števila ustreznih napovedi izračunamo natačnost določenega modela ter jo primerjamo z ostalimi modeli (Nisbet, Miner, & Elder, 2009, str. 45). Pri napovednih modelih predstavlja dober rezultat nizka vrednost merila *False-Positive*, ki pove, koliko vrednosti pozitivnega razreda je bilo uvrščenih v napačen razred (Sheikh, 2013, str. 156). Poznamo dve vrsti rezultatov, končno izbrani model z vrednostmi napovedi ter kakršnekoli ugotovitve, ki izhajajo iz samih modelov ali iz celotnega procesa podatkovnega rudarjenja. Po opravljeni fazi modeliranja je razumevanje področja obravnave še bolj pomembno pri odločitvi, kako rezultate uporabiti v praksi in katere nadaljnje študije so upravičene. Sledi točka, kjer se moramo odločiti, ali bomo pridobljene rezultate uvedli v realno poslovno okolje ali se bomo vrnili v fazo modeliranja in poskušali izboljšati dosedanje ugotovitve. Po številnih vrnitvah v predhodno fazo je možno med drugim ugotoviti, da rezultati nimajo uporabne vrednosti in lahko zgolj definiramo potrebne podatke za v prihodnje (IBM, 2016, str. 29-31).

2.6 Uvedba v prakso

V zadnji fazi rezultati modeliranja zaživijo v praksi. Uvedbo (angl. *Deployment*) lahko vidimo kot proces uporabe novih vpogledov, ki pripomorejo k izboljšavam znotraj organizacije. Poznamo uvedbo, kjer gre za formalno integracijo rezultatov, pridobljenih v orodju za obdelavo podatkov s podatkovnim skladiščem, kamor se zapišejo. Drugo vrsto uvedbe predstavlja uporaba novih koristnih informacij, pridobljenih skozi celoten proces podatkovnega rudarjenja. Predstavitev rezultatov se prilagodi vsakemu tipu uporabnika, na eni strani je poudarek na tehničnih podrobnostih ter na drugi strani na vplivu na poslovanje. Sledi premislek, kdaj je zaradi spreminjajočih okoliščin potrebno ponovno oceniti učinkovitost uvedenega modela. Z drugimi besedami rečeno, kako hitro obstoječi model zastari. Čisto na koncu je zelo koristno narediti krajši intervju s ključnimi deležniki projekta in dokumentirati njihovo izkušnjo (IBM, 2016, str. 33-36).

2.7 Pogosto uporabljene metode in tehnike strojnega učenja

V praksi je razširjena uporaba številnih algoritmov strojnega učenja, vendar je izbira odvisna od naloge podatkovnega rudarjenja. Konkretno pri analizi asociacij pridejo v poštev asociativna pravila, pri analizi skupin tehnika najbližjega soseda in razvrščanje v skupine, pri napovedovanju globoko učenje, najbližji sosed, odločitvena drevesa in regresijska analiza ter pri zaznavanju odstopanj tehnika razvrščanja v skupine.

2.7.1 Asociativna pravila

S pomočjo asociativnih pravil (angl. *Association Rule Learning*) odkrivamo razmerja med spremenljivkami v velikih bazah podatkov. Kot primer lahko navedemo analizo transakcij v nakupovalnem središču, kjer se je izkazalo, da bodo kupci, ki kupijo skupaj tako računalnik kot monitor, verjetno kupili tudi brezprekinitveno napajanje. Omenjena informacija se lahko koristi pri targetiranem oglaševanju ali pri postavitvi izdelkov na police (Farooqi & Raza, 2012, str. 3). Pomembno je definirati termina podpore (angl. *Support*) in zaupanja (angl. *Confidence*), ki sta ključna parametra asociativnega pravila. Podpora je definirana kot delež transakcij T, ki vsebuje izdelka X in Y. Podpora = $(X \cup Y) / n$. Zaupanje je definirano kot delež izdelka X v transakcijah T, ki hkrati vsebuje izdelek Y. Zaupanje = $(X \cup Y) / X$ (Aggarwal, 2014, str. 131).

2.7.2 Globoko učenje

Globoko učenje (angl. *Deep Learning*) spada v skupino nelinearnih napovednih algoritmov, ki po strukturi delujejo kot biološke nevronske mreže (Verma & Nashine, 2012, str. 606). Namenjene so učenju in izkoriščanju zmožnosti interne predstavitve podatkov v svojih skritih plasteh (Aggarwal, 2014, str. 232-233). V splošnem poznamo 2 tipa nevronskih mrež (angl. *Neural Networks*), takšne z enosmernim tokom podatkov (angl. *Feed-forward Neural Network*) in takšne s povratnimi zankami (angl. *Recurrent Neural Network*). Določena mreža je sestavljena iz vhodne plasti, kjer so vrednosti izbranih atributov ene ali več skritih plasti, kjer se izvajajo izračuni, ter izhodne plasti, kamor prispejo vrednosti želenega števila izhodnih spremenljivk. Možne so številne arhitekture, odvisno od poslovnega problema (Michalewicz, 2007, str. 136-137). Nevronska mreža, ki se imenuje tudi povezovalni model, je zmožna opraviti katero koli nalogo razvrščanja (Aggarwal, 2014, str 206).

2.7.3 Najbližji sosed

Tehnika najbližjega soseda (angl. *Nearest Neighbour*) uvršča zapise na podlagi sosednjih zapisov (Verma & Nashine, 2012, str. 606). Spada v skupino lenih učencev, kjer model ni zgrajen vse do konca postopka. Enote, ki so uporabljene pri učenju algoritma, so uvrščene v razred šele, ko pridejo do algoritma podatki namenjeni testiranju. Metoda shrani vse instance, zato tudi spada v kategorijo učenja na podlagi instanc. Takšen pristop potrebuje ogromno število enot, da daje dobre rezultate. Ena glavnih prednosti omenjene metode je zmožnost modeliranja zelo kompleksnih prostorov. *K-Nearest Neighbor* (v nadaljevanju KNN) je najbolj poznan algoritem tega tipa (Sullivan, 2012, str. 469-470). KNN deluje tako, da najde najbližje k-enote, ki so bile uporabljene v fazi učenja glede na enoto, ki jo trenutno uvršča. Nato tej enoti dodeli najpogostejši razred med sosednjimi k-enotami ali povprečno vrednost v primeru regresije (Klinkenberg & Hofmann, 2013, str. 34).

2.7.4 Odločitvena drevesa

Odločitvena drevesa (angl. *Decision Trees*) so uporabljena tako za klasifikacijo kot napovedovanje vrednosti. Posledično je lahko odvisna spremenljivka binarna, nominalna,

intervalna ali zvezna (Refaat, 2007, str. 21). So strukture v obliki drevesa, ki predstavljajo nabor pogojev, na podlagi katerih je podatkovni niz uvrščen v ustrezno skupino. Uvrščanje nepoznanega primera poteka tako, da potuje po drevesu glede na vrednosti atributov, skozi ustrezna vozlišča. Ko je končni list dosežen, se primer razvrsti glede na razred dodeljen listu (Witten in drugi, 2011, str. 64). Poznamo klasifikacijska in regresijska drevesa (angl. *Classification and Regression Trees*, v nadaljevanju CART) in drevesa, ki avtomatično zaznavajo interakcijo s pomočjo χ^2 (angl. *Chi Square Automatic Interaction Detection*, v nadaljevanju CHAID), (Verma & Nashine, 2012, str. 606). Drevo odločanja hierarhično deli prostor, dokler ne doseže podprostora, ki je značilen za točno določen razred. Rezultate omenjene tehnike je zaradi vizualne predstavitve enostavno intepretirati (Aggarwal, 2014, str. 87).

2.7.5 Razvrščanje v skupine

Algoritem grozdenja omogoča izvajanje segmentacije podatkov oz. razvrščanje v skupine. Ena pogosteje uporabljenih je metoda voditeljev oz. algoritem *K-Means (Refaat, 2007, str. 25-26)*. Pri tej metodi je vsak segment predstavljen s središčem, ki se imenuje centroid. Določen centroid je izračunan kot povprečje vseh točk, ki so mu bile dodeljene. Pomanjkljivost metode voditeljev je določanje števila skupin vnaprej s strani uporabnika in jo deloma odpravlja algoritem *X-Means*, kjer uporabnik določi spodnjo in zgornjo mejo števila skupin. Cilj omenjenih tehnik je ustvariti uporabno in homogeno število segmentov (Chorianopoulos, 2015, str. 204). Pri obeh tehnikah je potrebno izbrati merilo, na podlagi katerega so vrednosti atributov enot izračunane in dodeljene v razrede. Pogosto uporabljena je evklidska razdalja (Paiano & Pasanisi, 2017, str. 71).

2.7.6 Regresijska analiza

Statistiki uporabljajo izraz regresija za proces napovedovanja številske vrednosti in termin regresijski model kot sopomenko linearnega modela. Linearni model je enostavno predstaviti v dvodimenzionalnem prostoru, kjer narišemo premico funkcije tako, da se v prostoru najbolje prilega nizu točk (Witten in drugi, 2011, str. 62). Regresija je proces iskanja vrednosti odvisne spremenljivke na podlagi vrednosti neodvisnih spremenljivk. Potrebno je ugotoviti, katere spremenljivke so statistično značilne in posledično vključene v model (Gupta & Gupta, 2012, str. 5). Model je najlažje razumljiv v smislu zmogljivosti, matematičnega izračuna, meril kvalitete in prileganja (Refaat, 2007, str. 15).

2.8 Orodje RapidMiner

Danes na trgu obstaja kopica odprtokodnih orodij za obdelovanje podatkov, kot npr. RapidMiner, Weka, KNIME, orodje R, OpenRefine, Solver, Orange itd. V večini primerov orodja vsebujejo številne gradnike, ki imajo v ozadju analitične tehnike za shranjevanje in manipulacijo podatkov. Nekatera ponujajo tudi možnost generiranja poročil ter vizualizacijo (Dwivedi in drugi, 2016, str. 1). RapidMiner, ki je uporabljen v empiričnem delu magistrske

naloge, je prijazen tistim uporabnikom, ki niso vešči programskih jezikov in imajo kljub temu željo po naprednem obdelovanju podatkov.

Zagotavlja integrirano okolje za strojno učenje, tekstovno rudarjenje (angl. *Text Mining*), napovedno analitiko in različne poslovne analize. Rešitev je možno uporabljati kot oblačno storitev ali kot namizno verzijo za osebne računalnike. V nekaterih raziskavah zadovoljstva uporabnikov iz let 2011, 2013 in 2014 se je RapidMiner uvrstil na sam vrh (Dwivedi in drugi, 2016, str. 1). Glavni razlog za njegovo izbiro se skriva v dejstvu, da ga je Gartner že 5. leto zapored uvrstil v kvadrant vodij na področju podatkovne znanosti in platform strojnega učenja (angl. *Gartner Magic Quadrant For Data Science And Machine Learning Platforms*), ("Magic Quadrant for Data Science and Machine-Learning Platforms", 2018).

Slika 2: Magični kvadrant podatkovne znanosti in platform strojnega učenja

Vir: Gartner (2018).

3 PRIMER PODATKOVNEGA RUDARJENJA NA PORTALU IUS-INFO

Podjetja na trgu imajo različno poslanstvo, cilje in vizijo ter temu primerno strategijo. Večina jih stremi k rasti, razvoju in doseganju poslovnega izida (Gurley, Peters, Collins, & Fifolt, 2015, str. 221-225). V začetku delovanja je ponavadi na razpolago malo število potencialnih kupcev, zato se je smotrno ukvarjati z vsemi. Skozi časovno obdobje se številka poveča do te mere, da je nujen razmislek, s kom se ukvarjati in koliko investirati v določeno razmerje, da bodo še vedno doseženi zastavljeni cilji (Saarenvirta, 1998, str. 2-5).

Portal IUS-INFO je zrasel do te mere, da je mesečni priliv novih uporabnikov znatno višji od razpoložljivih človeških virov, zato je izbira najbolj potencialnih oseb ključnega pomena. Prav tako podjetje ne želi, da odidejo dobre stranke. Pomoč pri obeh izzivih predstavlja podatkovno rudarjenje, ki med drugim omogoča napovedati potencial novih uporabnikov in identificirati tiste, ki najverjetneje zapuščajo podjetje.

3.1 O portalu, podjetju in dejavnosti

Portal IUS-INFO je največji na področju pravno-informacijskih storitev na Hrvaškem in šteje preko 40.000 registriranih uporabnikov. Je edinstven spletni pravno-informacijski sistem, ki vsebuje posodobljeno zakonodajo in sodno prakso v izvirni obliki in obliki kataloga ter ponuja možnost naprednega iskanja. Podjetje LEXPERA pravne in poslovne informacije d. o. o. je bilo ustanovljeno leta 2006 v Zagrebu in ima med drugim v lasti blagovno znamko IUS-INFO (LEXPERA, 2009a). Podjetje LEXPERA d. o. o. je del skupine LEXPERA SA, ki deluje v Luksemburgu, Sloveniji, Hrvaški in Turčiji (LEXPERA, 2016).

Portal IUS-INFO na hrvaškem trgu deluje od leta 2009. Hčerinsko podjetje v Sloveniji ima po standardni klasifikaciji dejavnosti, ki se uporablja za določanje dejavnosti poslovnih subjektov (SURS, 2008), kot glavno dejavnost navedeno 63.120 – obratovanje spletnih portalov.

3.2 Faze projekta

Celoten projekt podatkovnega rudarjenja poteka skladno s CRISP DM metodologijo, prične se z razumevanjem področja obravnave, sledi razumevanje podatkov, nato priprava podatkov, modeliranje, ocena rezultatov in na koncu uvedba v prakso.

Z namenom uspešne izvedbe projekta smo oblikovali projektno skupino, v kateri so sodelovale 3 osebe podjetja LEXPERA d. o. o., konkretno direktor, tehnični direktor in sistemski administrator, ter svetovalno podjetje FrodX d. o. o., ki upravlja njihov sistem za avtomatizacijo trženja (angl. *Marketing Automation System*, v nadaljevanju SZAT). V obdobju med 25. 8. 2017 in 27. 3. 2018 smo imeli uvodno srečanje, delavnico pred izvozom

želenih podatkov, delavnico po testnem modeliranju in zadnje srečanje po zaključku projekta.

Pred pričetkom projekta so bile podjetju predstavljene potencialne koristi, ki jih prinaša podatkovno rudarjenje, potrebni viri in njihova pričakovanja. Z viri so mišljeni razpoložljivi podatki, ki so na voljo znotraj interne organizacije, in osebe, ki bodo sodelovale v projektu. Dogovoril sem se, da bodo uporabljeni podatki ERP in CRM sistema, baze spletne strani in SZAT-a.

Tehnično združljivost podatkov različnih sistemov sem preveril s pomočjo izdelave relacijskega modela v orodju Microsoft Access, po pridobitvi vzorčnih podatkov. Na delavnici, pred pridobitvijo želenih dejanskih podatkov, sem se dogovoril za izvoz podatkov prej omenjenih sistemov v Excelove datoteke. Priprava podatkov za poslovno-inteligenčne potrebe je bila izvedena s pomočjo vrtilnih tabel orodja Microsoft Excel in orodja PDI.

Na srečanju z direktorjem po testnem modeliranju sem dobil odgovore na nekatera poslovna in tehnična vprašanja, ki so se dodatno pojavila tekom izdelave preizkusnih modelov. Preizkusno in dejansko modeliranje je bilo izvedeno v orodju RapidMiner. Na zadnji delavnici so bili celoten proces in končne ugotovitve predstavljeni sodelujočim na projektu. Prav tako je bil namenjen čas za vprašanja z diskusijo, ki je podala zanimive poglede na podjetju doslej nepoznan pristop.

3.2.1 Razumevanje področja obravnave

Projekt se prične z razumevanjem poslovnega modela pravno-informacijskega portala IUS-INFO. Spletni portal beleži 3 tipe obiskovalcev, in sicer neregistrirane uporabnike, registrirane brezplačne uporabnike in registrirane plačljive uporabnike. Analiza uporabe portala temelji izključno na registriranih uporabnikih, ki imajo unikatno uporabniško ime. Registrirani brezplačni uporabniki so mesečno omejeni s številom dokumentov, do katerih lahko dostopajo, in prav tako s številom točk, ki so potrebne za dostop do določenega dokumenta. Registrirani plačljivi uporabniki načeloma teh omejitev nimajo, odvisno od tipa naročnine. Obstajajo različne možnosti, splošna delitev je v 3 skupine: IUS-INFO BASIC, IUS-INFO CLASSIC in IUS-INFO PROFESSIONAL (LEXPERA, 2009b). Na eni strani torej obstajajo brezplačni registrirani uporabniki z omejenim dostopom, na drugi strani pa plačljivi uporabniki z neomejenim dostopom, ki naročnino sklenejo za obdobje vsaj 1 meseca.

Na podlagi poslovnega modela identificiram potencialne koristi podatkovnega rudarjenja na konkretnem primeru in definiram cilje projekta. Na eni strani želim ugotoviti, kateri brezplačni uporabniki so potencialni, da postanejo plačljivi naročniki. Na drugi strani želim ugotoviti, kateri trenutno plačljivi uporabniki verjetno nameravajo zapustiti podjetje. Cilja projekta sta izdelava v praksi delujočega modela preprečitve odhodov kupcev ter modela

prioritetne lestvice potencialnih kupcev portala IUS-INFO. Glavna cilja razčlenim na več manjših, ki omogočajo sistematično spremljanje izvajanja in doseganja:

- preučiti poslovni model portala IUS-INFO;
- ugotoviti zmožnost povezljivosti podatkov s pomočjo relacijskega modela na podlagi vzorčnih podatkov;
- ovrednotiti kakovost podatkov različnih sistemov na podlagi vzorčnih podatkov;
- zagotoviti kakovostne podatke za izgradnjo modela prioritetne lestvice in modela preprečitve odhodov kupcev;
- integrirati podatke različnih sistemov in pridobiti dvodimenzionalno tabelo, ki služi kot input posameznega modela;
- zgraditi, testirati in primerjati zmogljivost različnih metod in tehnik strojnega učenja pri posameznem modelu;
- interpretirati rezultate posameznega modela (uporabnost v praksi);
- identificirati potencialne plačljive uporabnike;
- identificirati plačljive uporabnike, ki nameravajo zapustiti podjetje;
- predlagati način uporabe modelov oz. praktične ukrepe.

3.2.2 Razumevanje podatkov

Poleg človeških virov, potrebnih za izvedbo projekta, so podatki eden ključnih faktorjev za končni uspeh. Govora je o razpoložljivih transakcijskih podatkih podjetja. Skupno prepričanje sodelujočih na projektu je želja po pridobitvi čim bolj širokega pogleda na obnašanje uporabnikov IUS-INFO portala z vidika podatkov. Glede na to, da interesantne vzorce tvorijo številni podatkovni viri kot celota, je le-te potrebno predhodno povezati (Farooqi & Raza, 2012, str. 4).

Dogovoril sem se, da bodo uporabljeni podatki ERP in CRM sistema, spletne strani (portalske baze in baze prijav) in SZAT-a. V tem primeru so analizirani: vidik koriščenja portala, odzivi na trženjske kampanje, interakcije v okviru CRM sistema in zgodovina koriščenja paketov s pripadajočim finančnim delom.

Najprej sem se želel prepričati, ali je tehnično možno integrirati podatke in jih koristiti kot celoto. Na podlagi vzorčnih podatkov v orodju Microsoft Access izdelam relacijski model, ki predstavlja načrt naslednje faze, priprave podatkov. Vzorčni podatki so izvoženi v Excelove tabele. Preden se lotim povezovanja 7 tabel (Portalske baze, Baze prijav, ERP – paketov, ERP – plačil, CRM – stikov, CRM – prodajnih aktivnosti in SZAT-a), je potreben razmislek o ustreznih primarnih ključih. Glavni ključ (angl. *Primary Key*, v nadaljevanju GK) mora izpolnjevati naslednje 3 kriterije (Harrington, 2016, str. 92-96): mora enolično identificirati vsako vrstico, nikoli ne sme biti prazno ali ničelno in se ne sme pogosto spreminjati, v idealnem primeru nikoli.

3.2.2.1 Postopek izdelave relacijskega modela

Najprej uvozim podatke osrednje tabele ter inkrementalno dodajam ostale tabele, jih povežem preko ključev, ki ne predstavljajo kršitev. Portalska baza, osrednja tabela ima atribut *UserName* za GK.

V poštev pridejo samo prijave registriranih uporabnikov (polje *UserName* ni prazno), preostali zapisi Baze prijav so izbrisani pred uvozom v Access. Zapisom dodam atribut *ID*, ki predstavlja GK. Bazo prijav in Portalsko bazo povežem preko atributa *UserName*. Določen uporabnik ima lahko več prijav, določena prijava se nanaša na točno določenega uporabnika, povezava je 1:M.

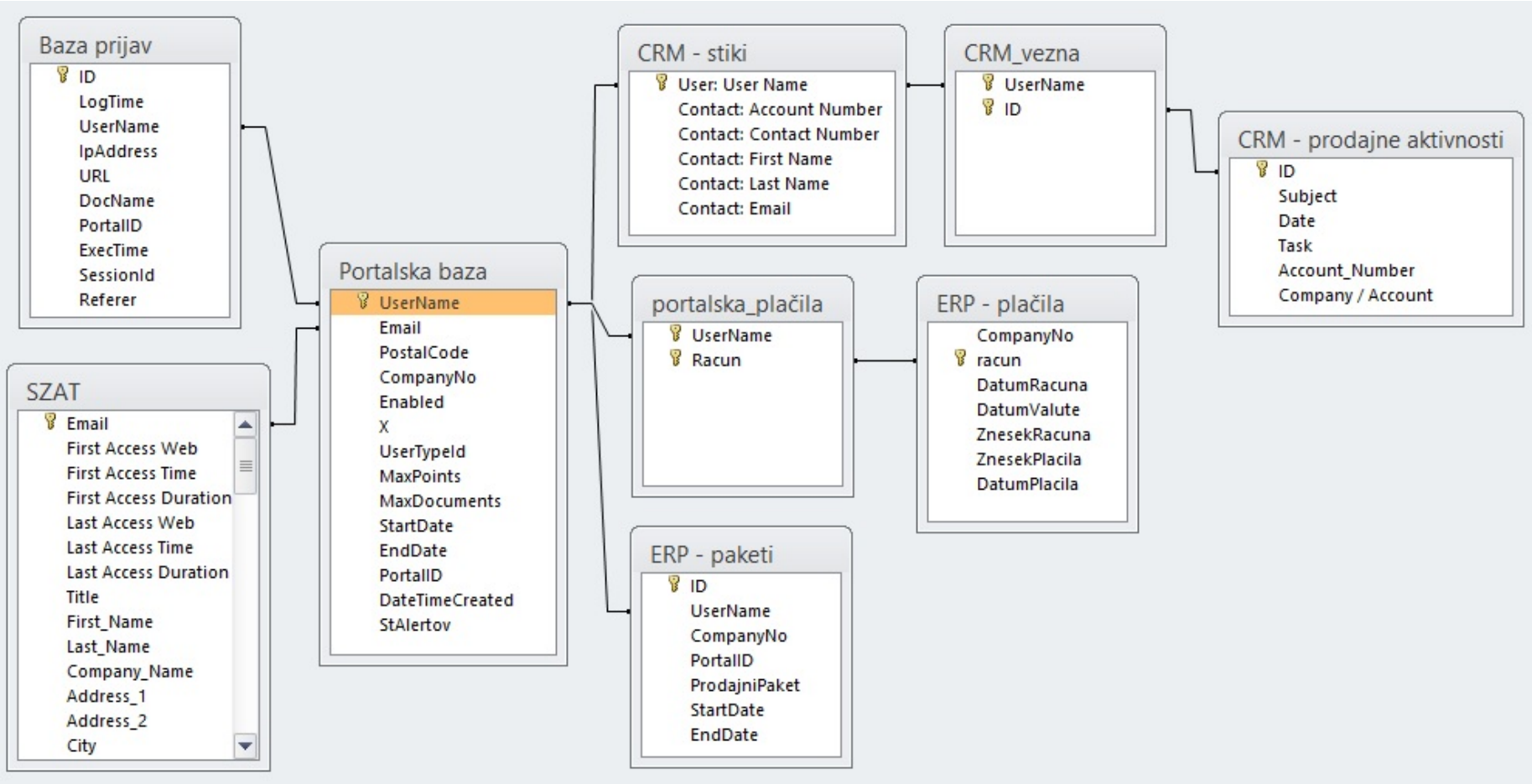
Prav tako je tabeli ERP – paketi dodan atribut *ID*, ki predstavlja GK. Uporabnik lahko skozi časovno obdobje koristi različne pakete, ki so pogojeni z datumom pričetka uporabe. Z osrednjo tabelo jo povežem preko atributa *UserName*, povezava je 1:M.

Racun predstavlja GK tabele ERP – plačila. Poveza s Portalsko bazo predstavlja razmerje M:M, in sicer ima več uporabnikov identično *CompanyNo* in več računov je izdanih na določeno *CompanyNo*. Potrebno je uvesti vezno tabelo. Skupni atribut obeh tabel je *CompanyNo*, kar po mojem mnenju predstavlja pomanjkljivost, in sicer ne vemo, kateri računi se nanašajo na katerega uporabnika, ampak zgolj na katero podjetje.

Glavni ključ tabele CRM – stiki predstavlja *UserName*, povezava s Portalsko bazo je 1:1, preko istoimenskega atributa. Tabeli CRM – prodajne aktivnosti je dodan atribut *ID*, ki predstavlja GK. Tabelo CRM – stiki in CRM – prodajne aktivnosti povežem s pomočjo vezne tabele, in sicer ima več uporabnikov določen *Account_Number* in določen *Account_Number* ima več prodajnih aktivnosti. Problem nastane v tabeli CRM – prodajne aktivnosti, saj omenjeni atribut nima znanih vrednosti za vse zapise.

Pri SZAT-u za GK izberem atribut *Email*. Bolj primerna izbira bi bila spremenljivka *UserName101* (Uporabniško ime IUS-INFO portala), ki ima prazne vrednosti pri zapisih vzorčnih podatkov. Tabelo Portalske baze in SZAT-a povežem preko atributa *Email*, povezava je 1:1.

Slika 3: IUS-INFO relacijski model projekta podatkovnega rudarjenja



Vir: lastno delo.

Po preverbi tehnične združljivosti podatkov sem s strani podjetja pridobil vsebinsko razlago vsake spremenljivke in analiziral, katere imajo potencialno vrednost pri izdelavi modela potencialnih kupcev ter katere pri modelu preprečitve odhodov kupcev.

Po tehtnem razmisleku sem za enoto preučevanja določil registriranega uporabnika IUS-INFO portala. Druga možnost je bilo podjetje uporabnika, vendar je bila zaradi narave podatkov prva izbira boljša. Večina spremenljivk se je odražala neposredno na obnašanje uporabnika. Tiste, ki so se nanašale na podjetje uporabnika, je bilo možno prevesti na uporabnika. V nasprotni smeri ne bi bilo smiselno.

Na podlagi analize potenciala spremenljivk in skladno z enoto preučevanja sem se dogovoril za izvoz dejanskih podatkov prej omenjenih 4 sistemov, prav tako v Excelove datoteke. Sledi kvantiteta podatkov, s čimer mislim na število vrstic zapisov in spremenljivk, ki jih vsebuje določen izvoz:

- portalska baza in baza prijav (logov) skupaj: 41.394 vrstic, 26 spremenljivk;
- CRM – stiki: 43.777 vrstic, 6 spremenljivk;
- CRM – prodajne aktivnosti: 53.471 vrstic, 8 spremenljivk;
- ERP – paketi: 78.055 vrstic, 6 spremenljivk;
- ERP – plačila: 29.293 vrstic, 7 spremenljivk;
- SZAT: 31.032 vrstic, 2 spremenljivki.

Podatki so iz časovnega obdobja med 1. 9. 2009 in 30. 9. 2017, z izjemo CRM podatkov, ki so na razpolago od 1. 1. 2014 do 30. 9. 2017. Na tej točki je potrebno poudariti, da so izvoženi podatki transakcijski, ki v večji meri niso neposredno uporabni za poslovno-inteligenčne potrebe. V naslednji fazi bodo prilagojeni temu namenu, pomoč pri pripravi predstavlja relacijski model.

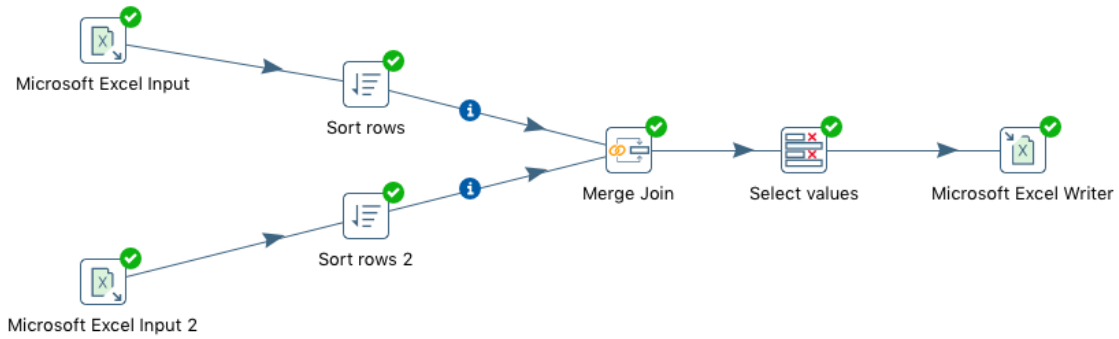
3.2.3 Priprava podatkov

Prvi cilj te faze je spraviti podatke 4 sistemov oz. 5 baz podatkov (Portalske baze, Baze prijav, ERP in CRM sistema ter SZAT-a) na skupni imenovalec ter ustvariti dvodimenzionalno tabelo, ki služi kot input naslednje faze, faze modeliranja. Drugi cilj je pretvoriti transakcijske podatke v podatke primerne za poslovno-inteligenčne potrebe. Določene spremenljivke je možno neposredno črpati iz obstoječih sistemov, preostale vrednosti je potrebno transformirati. S transformacijo mislim na izračunavanje vrednosti iz ene ali več spremenljivk transakcijskih podatkov. Izvozi podatkov se nahajajo v 5 Excelovih datotekah, v 6 različnih tabelah. Povezovanje, čiščenje in transformiranje podatkov je narejeno s pomočjo orodja PDI, imenovanega tudi Kettle, ter Microsoft Excela. Kettle je najbolj priljubljeno orodje med odprtokodnimi ETL rešitvami (Pentaho, 2015).

ETL postopek poteka tako, da osrednji tabeli inkrementalno dodajam spremenljivke ostalih tabel. Z osrednjo tabelo je mišljena Excelova tabela, kjer se nahajajo skupno izvoženi

podatki Portalske baze in Baze prijav, dimenzije 41.310 vrstic in 26 spremenljivk. V prvem koraku združim osrednjo tabelo in tabelo SZAT-a. Na sliki 4 je razviden zgolj eden od procesov orodja PDI, s pomočjo katerega uspešno integriramo podatke, ki se lahko nahajajo v številnih virih.

Slika 4: Proces orodja PDI



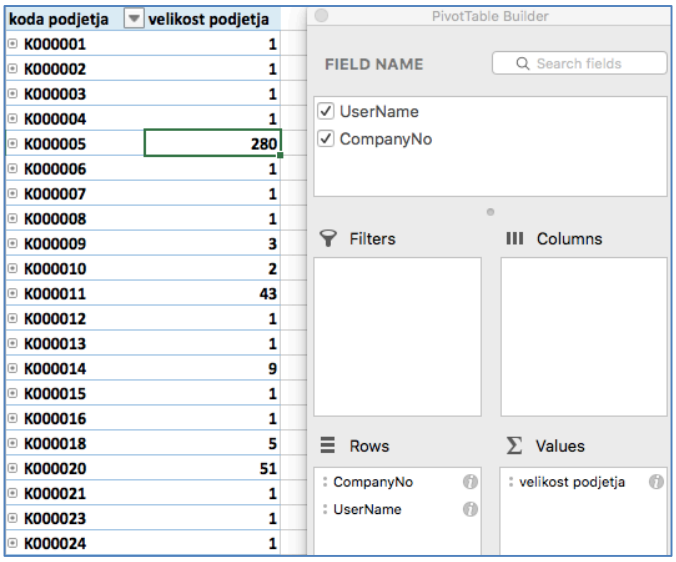
Vir: lastno delo.

Začetek procesa predstavljata 2 *Microsoft Excel Imput* gradnika (slika 4), ki prebereta podatke. Nato podatke vsake datoteke s pomočjo gradnika *Sort rows* razvrstim po spremenljivki *Email*, ki je v tem primeru ključ, po katerem združim podatke. Sledi združevanje, kjer uporabim LEFT OUTER JOIN, in sicer so pri omenjenem procesu ohranjeni vsi zapisi osrednje tabele. Po združevanju s *Select values* odstranim neželene spremenljivke. Končni rezultat predstavlja Excelova datoteka: 41.310 vrstic zapisov, 27 spremenljivk.

Sledi izračun dihotomne spremenljivke imenovane *Trenutno plačljiv uporabnik*, ki zavzema vrednosti 0 (NE) in 1 (DA). Najprej v Excelu filtriram in izločim uporabnike tipa 8 (administrator) ter tipa 9 (IUS-INFO partnerji), ki so interni uporabniki podjetja in pri raziskavi ne pridejo v poštev. Nato dodelim vrednost 0 vsem, ki imajo že zapadel datum odpovedi naročniškega razmerja. Preostale vrednosti izračunam s funkcijo *IF,* in sicer spadajo uporabniki tipa 4, 5, 6, 7 in 10 med plačljive ter tipa 2 in 3 med brezplačne uporabnike. Trenutno je 7.038 plačljivih in 34.272 brezplačnih uporabnikov.

Naslednja spremenljivka, ki jo računam, je *Velikost podjetja*. Z velikostjo je mišljeno število trenutno aktivnih uporabniških računov. Pod isto podjetje spadajo tisti računi, ki imajo enako kodo podjetja. Kodi tipa K000000 in K999999 pomenita, da ne poznamo pravega podjetja uporabnika, zato zanju izračun velikosti ni smiseln. S pomočjo vrtilne tabele preštejem število uporabnikov znotraj določene kode.

Slika 5: Izračun spremenljivke velikost podjetja



koda podjetja	velikost podjetja
K000001	1
K000002	1
K000003	1
K000004	1
K000005	280
K000006	1
K000007	1
K000008	1
K000009	3
K000010	2
K000011	43
K000012	1
K000013	1
K000014	9
K000015	1
K000016	1
K000018	5
K000020	51
K000021	1
K000023	1
K000024	1

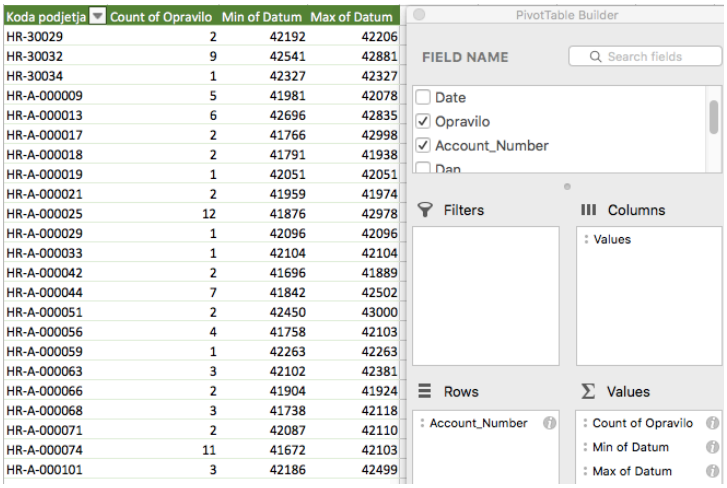
Vir: lastno delo.

Nato s pomočjo orodja PDI agregat velikosti dodelim uporabnikom, in sicer združim vrstice preko kode podjetja. Tako izračunana velikost podjetja posredno odraža velikost, in sicer večje kot je podjetje v resnici, večje število računov ima zakupljeno. Pomembna omejitev pri tej spremenljivki je to, da so znane kode podjetja samo za tiste, ki so bili do sedaj kupci. To pomeni 1.893 različnih podjetij oz. 10.472 uporabniških računov. Vsem ostalim dodelim vrednost 1 v primeru, da imajo osnovni paket (uporabnik tipa 4), ali vrednost 2 v primeru, da imajo premium paket (uporabnik tipa 5). Po posvetu z IUS-INFO vrednosti velikosti podjetij, ki imajo zakupljene številne pakete (skupno kvoto), delim s 5. Tem uporabnikom (tipa 7) po deljenju dodatno prištejem število 2 in tako dobim končne izračune velikosti.

V naslednjem koraku povežem tabeli CRM sistema, in sicer so v prvi tabeli podatki o kupcih (43.777 vrstic zapisov), v drugi pa podatki o prodajnih aktivnostih v zvezi s preteklimi, trenutnimi in potencialnimi kupci (53.471 vrstic zapisov). Spremenljivki, ki ju želim izračunati iz tabele prodajnih aktivnosti, sta *Skupno število sestankov* ter *Skupno število emailov in klicev*. Prav tako me zanima čas prve in zadnje interakcije s kupcem, ki ga v nadaljevanju uporabim za druge izračune.

Za izračun atributa *Skupno število emailov in klicev* (vrednost spremenljivke *Task* = 1) najprej izločim vse fizične sestanke (vrednost spremenljivke *Task* = 0) in zapise, ki nimajo znane vrednosti, niti spremenljivke *Account Number* niti spremenljivke *Contact Number,* ter tako dobim 50.444 zapisov aktivnosti. V naslednjem koraku filtriram aktivnosti po *Account Number* ter teh 43.099 zapisov agregiram s pomočjo vrtilne tabele.

Slika 6: Izračun spremenljivke skupno število emailov in klicev



Koda podjetja	Count of Opravilo	Min of Datum	Max of Datum
HR-30029	2	42192	42206
HR-30032	9	42541	42881
HR-30034	1	42327	42327
HR-A-000009	5	41981	42078
HR-A-000013	6	42696	42835
HR-A-000017	2	41766	42998
HR-A-000018	2	41791	41938
HR-A-000019	1	42051	42051
HR-A-000021	2	41959	41974
HR-A-000025	12	41876	42978
HR-A-000029	1	42096	42096
HR-A-000033	1	42104	42104
HR-A-000042	2	41696	41889
HR-A-000044	7	41842	42502
HR-A-000051	2	42450	43000
HR-A-000056	4	41758	42103
HR-A-000059	1	42263	42263
HR-A-000063	3	42102	42381
HR-A-000066	2	41904	41924
HR-A-000068	3	41738	42118
HR-A-000071	2	42087	42110
HR-A-000074	11	41672	42103
HR-A-000101	3	42186	42499

Vir: lastno delo.

Rezultat pove skupno število poslanih elektronskih sporočil in telefonskih klicev z določenim podjetjem. Beležim interakcije z 8.146 različnimi podjetji. Podobno naredim za preostale zapise, in sicer jih filtriram po *Contact Number* (*Account Number* zanj ni znan) ter teh 7.345 zapisov agregiram s pomočjo vrtilne tabele. Rezultat pove skupno število poslanih elektronskih sporočil in telefonskih klicev z določeno osebo. Beležim interakcije z 2.867 različnimi uporabniki. Poleg preštetja vrednosti spremenljivke *Opravilo* v polje *Values* vrtilne tabele dodam še funkcijo *MIN* in *MAX* (*Datum*) ter tako dobim datum prve in zadnje interakcije bodisi s podjetjem ali določeno osebo.

V predhodnem izpisu vrtilne tabele je možno razbrati nenavadne vrednosti datumov, ki v Excelu pomenijo izračunano vrednost glede na leto 1900, kjer je vrednost enaka 1. Datume pretvorim v lažje berljivo obliko z zamenjavo formata.

Določene zapise CRM tabel združim preko kode podjetja (*Account Number*), preostale preko kode uporabnika (*Contact Number*). Po prvem združevanju preko spremenljivke *Account Number* je tabeli CRM – stiki dodeljenih 13.748 zapisov od 11.013 različnih vrstic agregiranih podatkov prodajnih aktivnosti (*število interakcij, prva interakcija, zadnja interakcija*). Nato vzamem 30.029 vrstic prve tabele, katerim v prvem združevanju ni bila dodeljena nobena vrednost, in ponovno poizkusim preko spremenljivke *Contact Number*. Tokrat je dodatno dodeljenih 1.546 vrstic. V obeh primerih v PDI orodju uporabimo LEFT OUTER JOIN kot način združevanja, ki ohrani vse vrstice tabele CRM – stiki. Končni rezultat združitve CRM podatkov predstavlja 43.777 uporabnikov, od katerih je znanih 15.294 vrstic podatkov prodajnih aktivnosti.

Na koncu tabeli CRM sistema združim z osrednjo tabelo preko uporabniškega imena. Pri povezovanju najprej uporabim INNER JOIN, nova osrednja tabela šteje 40.585 vrstic

zapisov, kar je 725 manj, kot jih je pred združevanjem. Ugotavljam, da primanjkljaj izhaja iz dejstva, da teh uporabniških računov ni v CRM tabeli, ki pred združevanjem šteje 43.777 vrstic. Ker želimo ohraniti vse zapise osrednje tabele, ponovim postopek in tokrat uporabim LEFT OUTER JOIN. Rezultat predstavlja 41.310 vrstic zapisov.

Zadnjo manjkajočo spremenljivko modela prioritetne lestvice predstavlja *Število dni med registracijo in prvo interakcijo*. Izračunam jo enostavno z odštevanjem datumov, in sicer od datuma prve interakcije odštejem datum registracije ali datum začetka uporabe in dobim želeno vrednost atributa. Datum registracije in datum začetka uporabe je potrebno izračunati s pomočjo operacije *Text to Columns* in funkcije *DATE (leto; mesec; dan)*. Pri uporabnikih, ki imajo znan datum začetka uporabe, uporabim le-tega, saj je starejši kot datum registracije.

V model preprečitve odhodov poleg prej opisanih spremenljivk vključim tudi tiste, ki jih pridobimo iz ERP sistema, in sicer tabele paketov, kjer je zabeležena zgodovina vseh paketov uporabnikov, ter tabele plačil, kjer so zabeležena vsa plačila podjetja uporabnika. Tabela paketov v začetku šteje 78.055 vrstic 6 spremenljivk (*uporabniško ime, koda podjetja, portal ID, prodajni paket, začetek paketa, konec paketa*). Iz teh podatkov dodatno izračunam spremenljivko *datum začetka zadnjega paketa*, ki je potrebna za izračun spremenljivke *število dni med začetkom zadnjega paketa in zadnjo interakcijo,* dihotomno spremenljivko *zapustil podjetje* (0 – NE, 1 – DA), dihotomno spremenljivko *akcijski naročnik* (0 – NE, 1 – DA), spremenljivko *število plačljivih paketov*, spremenljivko *trajanje plačljivih naročnin,* spremenljivko *povprečno trajanje naročnine* in dihotomno spremenljivko *akcijski naročnik postane plačljiv* (0 – NE, 1 – DA). Določen uporabnik ima lahko v istem trenutku aktivnih več naročnin. Prav tako so možne menjave paketov še med aktivno naročnino. Ob izteku naročnine se paket spremeni v *REG-IIHR*, kar pomeni, da plačljivi uporabnik dobi pravice brezplačnega uporabnika. Ob ponovni sklenitvi se mu pravice dodelijo v skladu s tistimi, ki jih določa prodajni paket.

V izhodiščnih podatkih tabele ERP – paketi najprej pretvorim datum začetka in konca paketov iz tipa *Text* v ustrezen datum s pomočjo operacije *Text to Columns* in funkcije *DATE (leto; mesec; dan)*. 33.815 vrstic zapisov nima znanega konca trajanja paketa, saj gre za brezplačne uporabnike. Neznan datum je zapisan z vrednostjo 01/01/1753, ki jo izbrišem pri vseh zapisih.

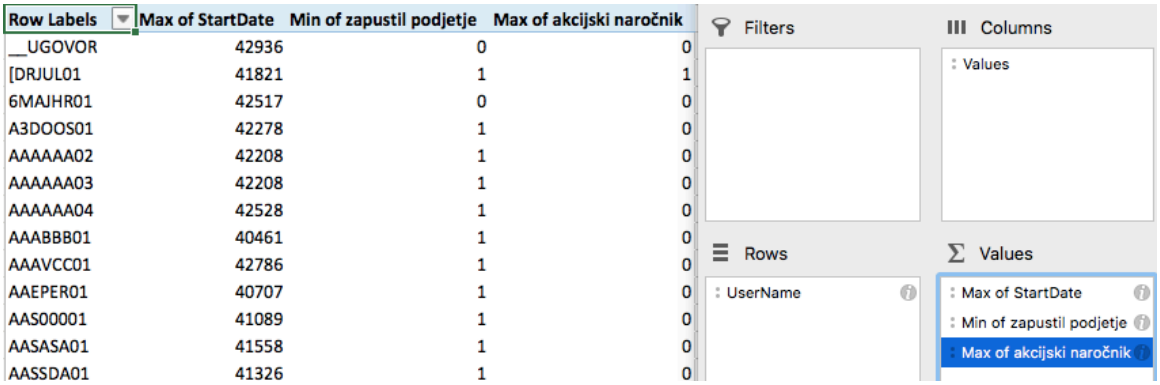
Datum začetka zadnjega paketa dobim s pomočjo vrtilne tabele in funkcije *MAX* (*start date*). Rezultat predstavlja 40.875 vrstic zapisov.

Nato računam spremenljivko *Zapustil podjetje,* in sicer za vsak paket, ki ga je skozi zgodovino imel uporabnik, določim vrednosti 0 (NE) ali 1 (DA). Najprej razvrstim uporabnike po paketih, tistim s paketom *REG-IIHR* dodelim vrednost 1. Pri preostalih 39.885 vrsticah zapisov uporabim funkcijo IF, in sicer če je datum konca paketa manjši od 1. 1. 2018, potem je uporabnik zapustil podjetje in mu dodelim vrednost 1, v nasprotnem

primeru dobi vrednost 0. Izpis šteje 7.291 vrstic plačljivih paketov in 70.764 brezplačnih. To seveda ne predstavlja končne rešitve, saj moram s pomočjo vrtilne tabele izbrati zgolj zadnjo naročnino v celotni verigi. Rezultat dosežem s pomočjo funkcije *MIN* (*zapustil podjetje*), in sicer ima zadnji ali več zadnjih aktivnih plačljivih paketov vrednost 0. Končni rezultat predstavlja 7.022 plačljivih in 34.288 brezplačnih uporabnikov, kar nam pove, da imajo določene osebe v danem trenutku aktivnih več plačljivih paketov.

Ali je bil uporabnik registriran za določeno preizkusno obdobje, vidim s pomočjo paketa, ki se imenuje *AKC-IIHR*. Vsem, ki so ga imeli, dodelim vrednost 1 (DA), preostalim vrednost 0 (NE). Seznam šteje 7.559 vrstic akcijskih paketov in 70.496 preostalih. Nato s pomočjo vrtilne tabele in funkcije *MAX* (*akcijski uporabnik*) preverim zgodovino paketov uporabnika. Končni rezultat predstavlja 6.611 akcijskih uporabnikov, kar pove, da so določeni uporabniki preizkusili več kot 1 paket.

Slika 7: Izračun spremenljivk zapustil podjetje in akcijski naročnik



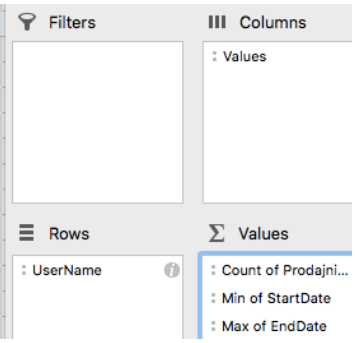
Row Labels	Max of StartDate	Min of zapustil podjetje	Max of akcijski naročnik
__UGOVOR	42936	0	0
[DRJUL01	41821	1	1
6MAJHR01	42517	0	0
A3DOOS01	42278	1	0
AAAAAA02	42208	1	0
AAAAAA03	42208	1	0
AAAAAA04	42528	1	0
AAABBB01	40461	1	0
AAAVCC01	42786	1	0
AAEPER01	40707	1	0
AAS00001	41089	1	0
AASASA01	41558	1	0
AASSDA01	41326	1	0

Vir: lastno delo.

Za izračun spremenljivke *Število plačljivih paketov* in *Trajanje plačljivih naročnin* je potrebno izločiti brezplačne in akcijske uporabnike. Dobim 32.326 vrstic zapisov, pri katerih s pomočjo vrtilne tabele preštejem skupno število paketov uporabnika, začetni datum prvega paketa ter končni datum zadnjega paketa. Podjetje ima 11.000 uporabnikov, ki so kadarkoli vzeli vsaj 1 naročnino, ter 29.875 preostalih. Od končnega datuma odštejem začetnega in dobim vrednosti atributa *Trajanje plačljivih naročnin.* Spremenljivko *Povprečno trajanje naročnine* predstavlja kvocient med spremenljivko *Trajanje plačljivih naročnin* in *Število plačljivih paketov*.

Slika 8: Izračun spremenljivk število plačljivih paketov in trajanje naročnin

Row Labels	Count of ProdajniPaket	Min of StartDate	Max of EndDate
__UGOVOR	5	42011	43300
[DRJUL01	1	41801	41820
6MAJHR01	3	41421	43246
A3DOOS01	2	41894	42277
ABAKUT01	1	41988	42004
ABBGRU01	1	40946	40968
ABCORP01	3	41417	42794
ABGRAD02	2	41623	42003
ABKDOO01	1	41892	41912
ACEKOV01	1	42613	42622
ACIVAN01	1	42755	42766
ADALOV01	1	40355	40374



Vir: lastno delo.

Sledi združevanje v orodju PDI, prvih treh ERP izračunanih spremenljivk (*začetek zadnjega paketa, zapustil podjetje* in *akcijski naročnik*), ki štejejo 42.485 vrstic zapisov, s preostalimi tremi (*št. plačljivih paketov, koliko dni je kupec* in *povprečno trajanje naročnine*), ki štejejo 11.359 vrstic zapisov. Ključ za združevanje je uporabniško ime, uporabim LEFT OUTER JOIN. Ko imam vseh 6 izračunanih spremenljivk združenih v 1 tabelo, lahko izračunam še zadnjo. Imenuje se *Akcijski naročnik postane plačljiv,* ki pove, ali je uporabnik registriran v okviru akcije kadarkoli postal plačljiv naročnik. Pogoja, ki ju mora izpolnjevati, sta vrednosti atributov *Akcijski naročnik* > 0 in *Število plačljivih paketov* > 0. Oba omenjena pogoja skupaj preverim s funkcijo *IF (AND(pogoj 1; pogoj 2); 1; 0)*. Ugotavljam, da jih od 6.611 preizkusnih uporabnikov kasneje postane 416 oz. 6,3 % plačljivih naročnikov.

Tabelo, ki se nanaša na spremenljivke ERP paketov, združim z osrednjo tabelo, uporabim LEFT OUTER JOIN, 40.875 od 41.310 vrsticam zapisov osrednje tabele so uspešno dodeljene vrednosti. Ključ za združevanje je uporabniško ime. Sedaj lahko izračunam spremenljivko *Število dni med začetkom zadnjega paketa in zadnjo interakcijo*, ki se izrazi kot *datum zadnje interakcije – datum začetka zadnjega paketa*.

Tabela plačil v začetku šteje 29.293 vrstic zapisov (izdanih računov) 7 spremenljivk (*koda podjetja, številka računa, datum računa, datum valute, datum plačila, znesek računa in znesek plačila*). Ponovno pretvorim datum računa, valute in plačila iz tipa *Text* v ustrezen datum s pomočjo operacije *Text to Columns* in funkcije *DATE (leto; mesec; dan)*. Znesek računa in plačila prav tako formatiram s prej omenjeno operacijo. Vsi zneski so v valuti hrvaška kuna (HRK). Ko je račun izdan, je le-ta bodisi plačan pravočasno (43 %), plačan z zamudo (48 %) ali pa ni bil poravnan (9 %). Iz izhodiščnih podatkov nato ustvarim dihotomne spremenljivke *Zamujeno plačilo* (0 – NE, 1 – DA), *Neplačilo* (0 – NE, 1 –DA) in *Plačani znesek je ustrezen* (0 – NE, 1 – DA) ter spremenljivki *Število dni zamude plačila* in *Manjkajoči/presežni znesek plačila*. Ali je bilo plačilo zamujeno, preverim s pomočjo datuma valute in datuma plačila. Kot neplačilo se smatra plačilo, ki nima znanega niti datuma niti zneska plačila. Ustreznost plačila se preveri s primerjanjem zneska računa in zneska

plačila. Pri zamujenem plačilu odštejem datum plačila in datum valute ter dobim število dni zamude. Pri neustreznih plačilih odštejem znesek plačila od zneska računa in dobim višino presežka ali primanjkljaja plačila. S pomočjo vrtilne tabele nato izračunam 9 agregiranih spremenljivk:

- *skupni plačani znesek podjetja uporabnika za vse pakete od začetka* – funkcija *SUM (plačila)*;
- *število izdanih računov na določeno podjetje* – funkcija *COUNT (številka računa)*;
- *število neplačanih računov podjetja uporabnika* – funkcija *SUM (neplačila)*;
- *število zamujenih plačil podjetja uporabnika* – funkcija *SUM (zamujena plačila)*;
- *povprečni znesek plačila podjetja uporabnika* – funkcija *AVERAGE (znesek plačila)*;
- *minimalno število dni zamude plačila* – funkcija *MIN (število dni zamude)*;
- *povprečno število dni zamude plačila* – funkcija *AVERAGE (število dni zamude)*;
- *maksimalno število dni zamude plačila* – funkcija *MAX (število dni zamude)*;
- *skupni manjkajoči/presežni znesek plačil podjetja uporabnika* – funkcija *SUM (manjkajoči/presežni znesek plačila)*.

Slika 9: Izračun 9 finančnih spremenljivk

Row Labels	Sum of ZnesekPlacila	Count of racun	Sum of neplačilo	Sum of zamujeno plačilo	Average of ZnesekPlacila	Min of Število dni zamude plačila	Average of Število dni zamude plačila	Max of Število dni zamude plačila	Sum of manjkajoči / presežni znesek plačila
K000001	14499	3	0	2	4833	0	55	153	7236
K000002	2600	3	1	2	1300	1	1,5	2	0
K000005	60187,54	37	2	32	1719,644	0	15,42857143	75	0
K000010	16560	36	4	32	517,5	2	53,5	119	0
K000011	2747,5	3	0	0	915,8333333	0	0	0	0
K000012	27780	3	0	2	9260	0	2,666666667	5	0
K000014	61462,5	6	2	0	15365,625	0	0	0	0
K000015	8095	4	2	0	4047,5	0	0	0	0
K000016	20906,06	36	1	2	597,316	0	4,771428571	157	0
K000018	71730,24	36	6	3	2391,008	0	5,766666667	160	0
K000020	36562,5	2	0	2	18281,25	16	18,5	21	0
K000021	18791,42	37	1	6	521,9838889	0	21,19444444	405	560
K000023	41223,84	27	1	8	1585,532308	0	5,769230769	75	781
K000026	23400	5	1	0	5850	0	0	0	0
K000029	14381,4	30	0	23	479,38	0	8,466666667	51	0
K000031	1651,21	5	0	5	330,242	1	62	122	0
K000032	4992,52	8	2	4	832,0866667	0	48,33333333	106	0
K000033	14160	36	4	32	442,5	6	57,59375	123	0
K000034	20897,05	37	1	13	580,4736111	0	7	115	0
K000035	26461,41	36	1	7	756,0402857	0	9,228571429	136	164,88
K000040	750	1	0	0	750	0	0	0	0
K000041	24849,55	43	5	11	653,9355263	0	12,26315789	136	0
K000042	19911,42	37	2	4	568,8977143	0	1,114285714	19	0
K000050	8970,08	19	3	10	560,63	0	4,5	35	0
K000052	14917,5	36	2	33	438,75	0	43,70588235	314	0

Vir: lastno delo.

Sledi izračun spremenljivk imenovanih *delež zamujenih plačil* in *delež neplačil*. Število zamujenih plačil delim s številom izdanih računov. Enaka formula velja pri deležu neplačil, tokrat je v števcu število neplačil.

V naslednjem koraku osrednji tabeli dodam 11 prej izračunanih spremenljivk, ki se nanašajo na plačila podjetij uporabnikov. Združevanje se naredi preko kode podjetja, in sicer so znani podatki za 2.155 podjetij. Vrednosti so uspešno dodeljene 9.123 različnim uporabnikom. Pri združevanju se uporabi LEFT OUTER JOIN, vseh 41.310 vrstic osrednje tabele je ohranjenih.
Za konec izračunam spremenljivko *Skupni plačani znesek / uporabnika* za vse pakete od začetka, in sicer delim višino skupno plačanega zneska podjetja z velikostjo podjetja oz.

številom njegovih uporabnikov. Ta spremenljivka izloči vpliv velikosti, vrednosti plačil so lažje primerljive med uporabniki.

Končni rezultat ETL postopka predstavlja tabela z 41.310 vrsticami zapisov 37 spremenljivk. To tabelo bom uporabil kot input modela prioritetne lestvice kupcev in modela preprečitve odhodov kupcev. Katere spremenljivke konkretno pridejo v poštev pri določenem modelu, ugotovim med modeliranjem, skladno s poslovno logiko in statistično značilnostjo.

3.2.3.1 Kakovost podatkov

V tem poglavju se osredotočim na analizo končnega rezultata faze priprave podatkov. Pri vsakem atributu analiziram delež manjkajočih in ekstremnih vrednosti ter vrednosti, ki sem jih dodatno napolnil ali odstranil, skladno s poslovno logiko podjetja. Prav tako je podan tip spremenljivke. Določeni atributi se nanašajo na podjetje, ki ima enega ali več registriranih uporabnikov, ali pa so vrednosti, ki se nanašajo primarno na podjetje, preračunane na uporabnika. Na ta način je dosežen celovit pogled na obnašanje uporabnika. V tabeli 1 je razvidna kakovost vseh spremenljivk, ki so bile pripravljene za poslovno-inteligenčne potrebe.

Tabela 1: Kakovost podatkov

Ime spremenljivke	**Tip spremenljivke**	**Delež manjkajočih vrednosti v %**	**Delež ekstremnih vrednosti v %**	**Delež napolnjenih ali odstranjenih vrednosti v %**
UserName	Tekst	0,00	0,00	0,04
CompanyNo	Tekst	0,00	74,64	0,00
Enabled	celo število	0,00	92,62	0,00
Login	celo število	0,00	81,22	0,00
StAlertovRegi	celo število	0,00	0,09	0,00
StAlertovZakoni	celo število	0,00	0,10	0,00
StAlertovPredpisi	celo število	0,00	0,01	0,00
StAlertovLokalniPredpisi	celo število	0,00	0,00	0,00
StAlertovAktualnaStevilka	celo število	0,00	0,09	0,00
StUspesnihPrijav	celo število	0,00	13,95	0,00
StNeuspesnihPrijav	celo število	0,00	5,55	0,00
StDniOdZadnjeUporabe	celo število	28,37	0,00	0,00
SteviloOgledovDokum	celo število	28,37	0,00	0,00
AvgExecTime	celo število	86,17	0,01	0,00
Lead Score	celo število	0,00	0,00	40,28
Plačljiv uporabnik	celo število	0,00	17,00	0,00

se nadaljuje

Tabela 1: Kakovost podatkov (nad.)

Ime spremenljivke	**Tip spremenljivke**	**Delež manjkajočih vrednosti v %**	**Delež ekstremnih vrednosti v %**	**Delež napolnjenih ali odstranjenih vrednosti v %**
Velikost podjetja	celo število	0,00	77,11	74,64
Število interakcij	celo število	0,00	0,00	66,23
Število dni med registracijo in prvo interakcijo	celo število	66,73	0,00	0,00
Akcijski naročnik	celo število	0,00	16,00	84,00
Število plačljivih paketov	celo število	0,00	0,00	0,00
Koliko dni je kupec	celo število	0,00	0,00	73,37
Povprečno trajanje naročnine	racionalno število	0,00	0,00	73,37
Akcijski uporabnik postal plačljiv	celo število	0,00	1,01	98,99
Število dni med začetkom zadnjega paketa in zadnjo interakcijo	celo število	66,77	0,00	0,00
Skupni plačani znesek podjetja	racionalno število	0,00	0,00	78,23
Število izdanih računov	celo število	0,00	0,00	77,92
Število neplačanih računov	celo število	0,00	0,00	77,92
Število zamujenih plačil	celo število	0,00	0,00	78,23
Povprečni plačani znesek računa	racionalno število	0,00	0,00	78,23
Minimalno število dni zamude plačila	celo število	0,00	0,00	78,23
Povprečno število dni zamude plačila	racionalno število	0,00	0,00	78,23
Maksimalno število dni zamude plačila	celo število	0,00	0,00	78,23
Skupni manjkajoči/presežni znesek plačil	racionalno število	0,00	0,00	78,23
Delež neplačil	racionalno število	0,00	0,31	77,92
Delež zamujenih plačil	racionalno število	0,00	0,85	77,92
Skupni plačani znesek/ uporabnika	racionalno število	0,00	0,00	77,92

Vir: lastno delo.

3.2.3.2 Izzivi pri integraciji

Skozi proces priprave podatkov sem naletel na določene izzive. Sledi opis težav in rešitev. Težave pri združevanju osrednje tabele in tabele SZAT-a:

- za 607 uporabnikov osrednje tabele ni znanih vrednosti spremenljivke email, zato teh zapisov ni mogoče združiti z zapisi tabele SZAT-a;
- nekateri elektronski naslovi osrednje tabele se pričnejo s presledkom, kar je potrebno odpraviti pred razvrščanjem zapisov po omenjeni spremenljivki. V nasprotnem primeru vrstni red ni konsistenten, kar je ključen pogoj za uspešno združevanje v orodju PDI.

Težave pri združevanju tebele CRM – stiki in CRM – prodajne aktivnosti ter pri kasnejši integraciji z osrednjo tabelo:

- ni znanih vrednosti zgolj 1 ključa v tabeli prodajnih aktivnosti, zato moram zapise ustrezno razvrstiti, ločiti in povezati preko spremenljivke *Account Number* ter nato *Contact Number*;
- pogosto zamenjani vrednosti spremenljivke *Account Number* in *Contact Number*. S tem mislim, da se zapisi pojavijo v nasprotnem stolpcu;
- ni znane nobene ustrezne vrednosti ključa, preko katere bi povezali 2.829 zapisov fizičnih sestankov tabele CRM – prodajne aktivnosti s tabelo CRM – stiki in tako ne morem izračunati želenega atributa *Skupno število sestankov*;
- nekonsistentnost glavnih ključev na relaciji tabele portalske baze in tabele CRM – stiki. V 165 primerih uporabniško ime ni identično v obeh primerih. Vrednosti je potrebno ročno poiskati v tabeli CRM – stiki in jih prenesti v osrednjo tabelo;
- formatiranje datuma (slika 10) posamezne prodajne aktivnosti, ki je v izvirnih podatkih tipa *Text* ter ga je potrebno pretvoriti v datum, da lahko delam izračune. Postopek formatiranja se prične z razčlenitvijo tekstovnega zapisa na 3 dele oziroma v 3 stolpce, in sicer na dan, mesec ter leto. To dosežem s pomočjo operacije *Text to Columns*. Kot pravilo, kje naj omenjena operacija razbije zapis, navedem piko. V zadnjem koraku s pomočjo funkcije *DATE (leto; mesec; dan)* izračunam pravilno delujoč datum. Omenjeni postopek po korakih od leve proti desni:

Slika 10: Postopek formatiranja datuma

Date	Dan	Mesec	Leto	Datum
21.7.2015	21	7	2015	21/07/2015
7.7.2015	7	7	2015	07/07/2015
13.7.2016	13	7	2016	13/07/2016
13.7.2016	13	7	2016	13/07/2016
14.7.2016	14	7	2016	14/07/2016
20.6.2016	20	6	2016	20/06/2016
20.6.2016	20	6	2016	20/06/2016

Vir: lastno delo.

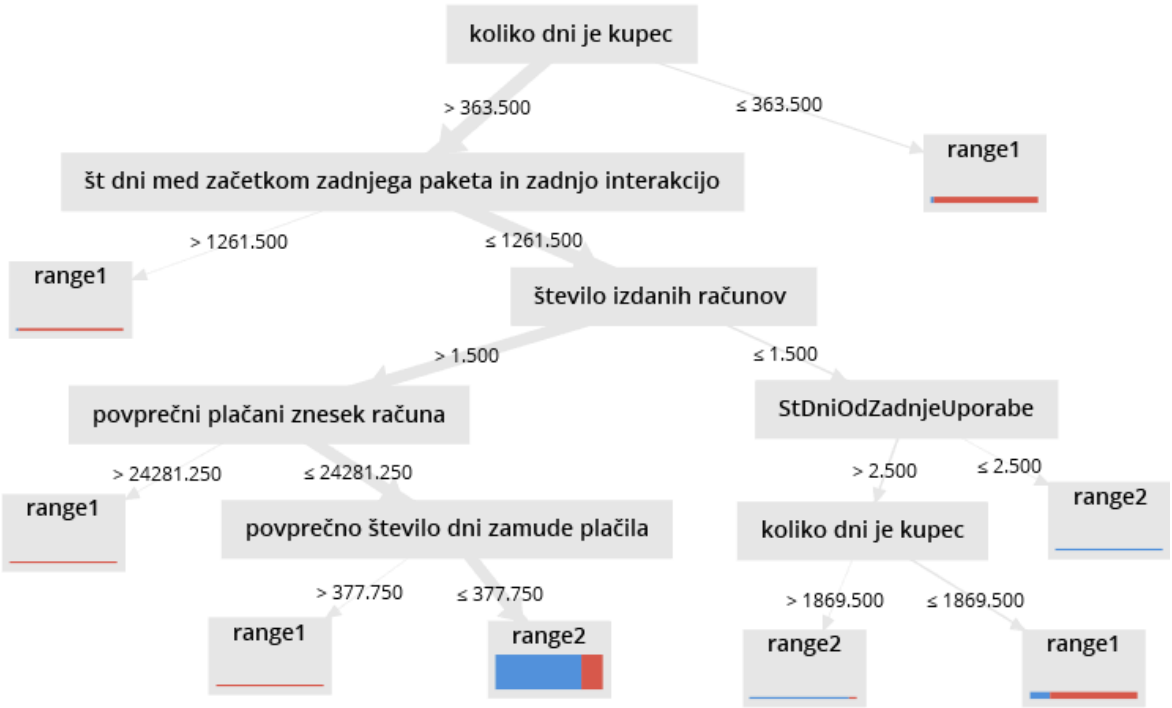
3.2.4 Modeliranje

Po uspešni pripravi podatkov se lotim modeliranja, kjer najprej izdelam testna modela preprečitve odhodov kupcev in prioritetne lestvice potencialnih kupcev. S tem želim preveriti, ali bodo modeli potencialno zaživeli v praksi in je smiselno investirati čas v njihovo nadaljnjo izboljšavo ali se je na tej točki bolje ustaviti in vrniti v predhodno fazo.

3.2.4.1 Testno modeliranje

Pri testnem modelu preprečitve odhodov kupcev je za enoto preučevanja določen registrirani uporabnik portala IUS-INFO z vsaj eno plačljivo naročnino (11.000 enot). Tehnika odločitvenega drevesa predstavlja model preprečitve odhodov (slika 11). Najpomembnejše spremenljivke so na vrhu narobe obrnjenega drevesa. Doseči želim čim bolj čiste liste (konce) drevesa, kjer naj bodo v čim večji meri pripadniki zgolj enega razreda. V tem primeru lahko opišem pogoje, pod katerimi kupci praviloma ostanejo v podjetju, ter pogoje, pod katerimi le-tega zapuščajo.

Slika 11: Testni model preprečitve odhodov kupcev



Vir: lastno delo.

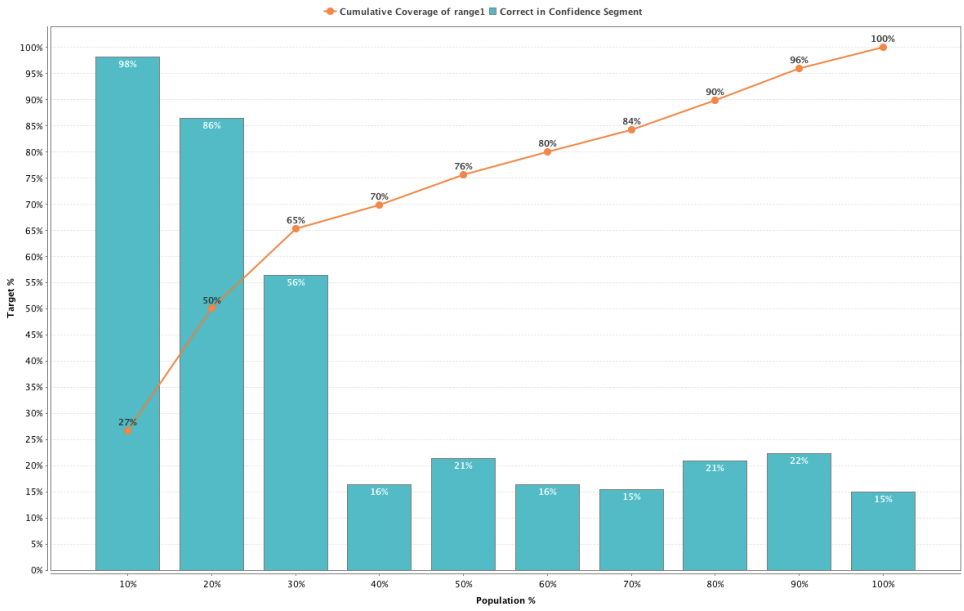
Natančnost testnega modela znaša 84 %, občutljivost 96,04 % in preciznost 81,64 %. V model je vključenih 37 spremenljivk, najpomembnejše so razvidne na sliki 12, spodaj levo. Po napovedi testnega modela namerava praviloma 20 % kupcev (range 1) zapustiti podjetje.

Slika 12: Povzetek testnega modela preprečitve odhodov kupcev

Vir: lastno delo.

Sledi ponazoritev testnega modela s pomočjo grafikona dviga (slika 13). Model identificira 50 % oseb, ki praviloma zapuščajo podjetje že z 20 % populacije. Vrednost zaupanja v tem primeru znaša 86 %. Brez modela bi bilo potrebnih 50 % enot za doseganje istega deleža.

Slika 13: Grafikon dviga testnega modela preprečitve odhodov kupcev

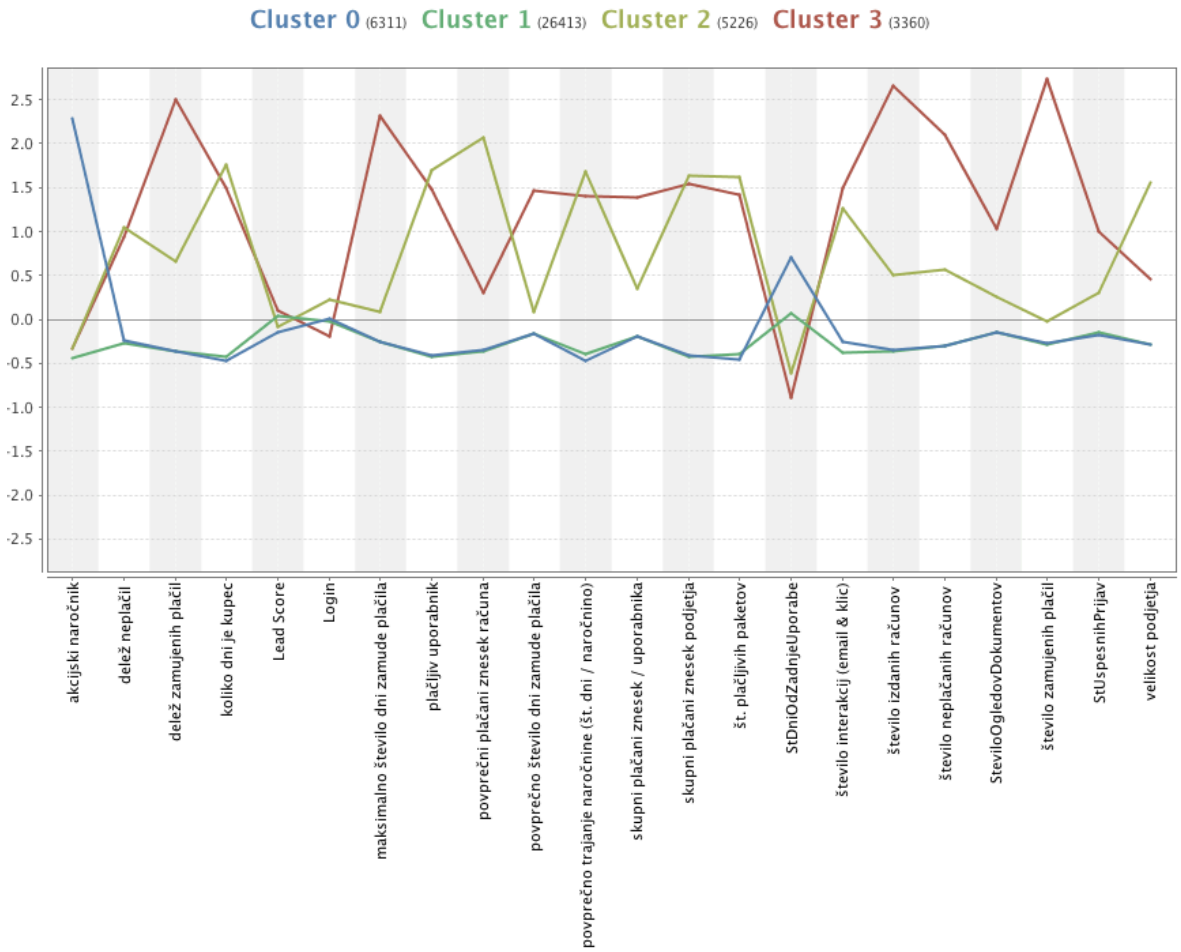


Vir: lastno delo.

Pri testnem modelu prioritetne lestvice potencialih kupcev je za enoto preučevanja določen registrirani uporabnik portala IUS-INFO (41.310 enot). Segmentacija predstavlja testni model prioritetne lestvice potencialnih kupcev. Algoritem *Clustering X-Means* predlaga delitev uporabnikov v 4 skupine. V segmentacijo je vključenih 25 od 35 spremenljivk. Izbrane spremenljivke nimajo manjkajočih zapisov. Z vključitvijo številnih spremenljivk sem želel doseči celovit pogled na podatke oz. njihovo delitev v homogene skupine.

Grafikon profilov (slika 14) kaže značilnosti različnih segmentov. Grafikon pomembnosti spremenljivk (slika 15) kaže ključne atribute določenega segmenta. V segmentu 1 (*cluster 0,* slika 14) in segmentu 2 (*cluster 1,* slika 14) so brezplačni uporabniki, v preostalih 2 plačljivi uporabniki. Rezultati segmentacije odražajo, da gre za manjše razlike, tako med segmentoma nekupcev kot tudi kupcev. Razvrstitev enot v 4 skupine daje bolj nazorno sliko kot zgolj v skupino kupcev in nekupcev. Določene karakteristike so značilne za 2 od 4 segmentov, nekatere so izrazite zgolj za enega.

Slika 14: Testni model prioritetne lestvice potencialih kupcev – grafikon profilov



Vir: lastno delo.

Slika 15: Testni model prioritetne lestvice potencialih kupcev – pomembnost spremenljivk

Vir: lastno delo.

Po pogovoru z IUS-INFO sem prišel do zaključka, da je vredno nadaljevati in izboljšati testne modele. Na eni strani je poleg odločitvenega drevesa potrebno preizkusiti še kakšno od tehnik strojnega učenja pri modelu preprečitve odhodov kupcev ter na drugi strani podrobneje analizirati segmente nekupcev. Potrebna sta odgovora na vprašanji, kateri kupci praviloma nameravajo zapustiti podjetje in kateri brezplačni uporabniki so po obnašanju podobni trenutno plačljivim ter posledično najbolj potencialni.

3.2.4.2 Model preprečitve odhodov kupcev

Enako kot v primeru testnega modela preprečitve odhodov kupcev je tudi sedaj za enoto preučevanja določen registrirani uporabnik portala IUS-INFO z vsaj eno plačljivo naročnino. Številsko to pomeni 10.985 vključenih enot v model. Med možnim naborom 37 spremenljivk jih je bilo na podlagi pregleda korelacij in smiselnosti s poslovno logiko izbranih 25, tako v primeru uporabljene tehnike odločitvenega drevesa kot nevronske mreže.

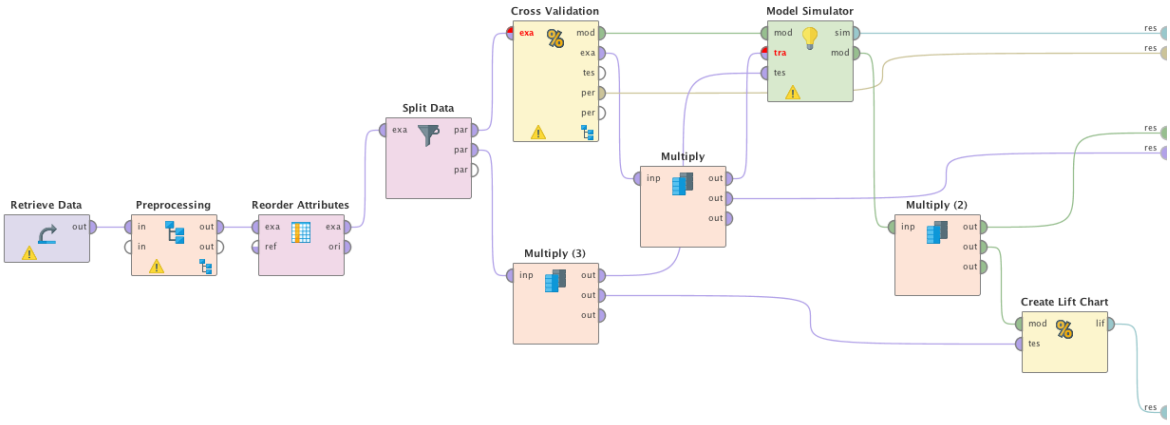
Faza modeliranja se prične v orodju Excel, kjer so zbrani vsi podatki predhodne faze. Tabela šteje 41.310 vrstic zapisov, 37 spremenljivk. Potrebno je filtrirati uporabnike po številu plačljivih paketov padajoče. Obdržim jih 11.000, ki imajo vrednost vsaj 1, preostale enote pri tem modelu ne pridejo v poštev. Prav tako izločim 15 podvojenih vrstic zapisov z identičnim uporabniškim imenom. Nato napolnim manjkajoče vrednosti 5 atributov, ki so vključeni v model:

- *število dni od zadnje uporabe* – 2.233 od 10.985 zapisov enot s povprečno vrednostjo 535;
- *število ogledov dokumentov* – 2.233 od 10.985 zapisov enot s povprečno vrednostjo 1.143;
- *avgExecTime* (povprečni čas odpiranja strani v sekundah) – 6.922 od 10.985 zapisov enot s povprečno vrednostjo 171;
- *število dni med registracijo in prvo interakcijo* – 343 od 10.985 zapisov s povprečno vrednostjo 251;
- *število dni med začetkom zadnjega paketa in zadnjo interakcijo* – 345 od 10.985 zapisov s povprečno vrednostjo 409.

Sledi shranjevanje in pričetek dela v orodju RapidMiner. S pomočjo čarovnika uvozim Excelovo tabelo podatkov ter jo shranim v lokalni repozitorij. S funkcionalnostjo *Automodel* hitro in enostavno izvedem primerjavo različnih algoritmov strojnega učenja ter ugotovim, katere tehnike je vredno v nadaljevanju uporabiti in izboljšati. *Automodel* prav tako pomaga identificirati spremenljivke, ki pridejo v poštev. Največ potenciala modela preprečitve odhoda kažeta odločitveno drevo in nevronska mreža. Za omenjena algoritma preizkusim različne nastavitve parametrov.

Glavni proces izgradnje modela (slika 16) se v orodju RapidMiner prične z branjem podatkov (gradnik *Retrieve Data*), ki sem jih že prej naložil ter lokalno shranil. Sledi podproces *Preprocessing*, kjer nastavim določene parametre v zvezi z napovednim modelom, nato atribute uredim po abecednem vrstnem redu (gradnik *Reorder Attributes*), razdelim podatke na 2 dela (gradnik *Split Data*): 70 % enot nadaljuje v fazo učenja (gradnik *Cross Validation*) in 30 % v fazo testiranja modela (gradnik *Model Simulator*) ter primerjavo napovedi s pomočjo grafikona dviga (gradnik *Create Lift chart*). Gradnik *Cross Validation* vsebuje v prvem primeru tehniko odločitvenega drevesa ter v drugem nevronsko mrežo. Proces se zaključi z rezultati, izhodi gradnikov *Model Simulator*, *Cross Validation* in *Create Lift chart*.

Slika 16: Glavni proces izgradnje modela preprečitve odhodov kupcev v orodju RapidMiner



Vir: lastno delo.

V praksi ni nujno, da bodo vse spremenljivke prispevale k napovedi. Nevključitev nekaterih lahko pospeši fazo učenja algoritma ali celo izboljša zmogljivost modela kot celote. Vendar je potrebno imeti čim bolj objektivna merila, na podlagi katerih je določena njihova usoda. Korelacija (angl. *Correlation*) pove povezanost določene spremenljivke z vrednostjo, ki se napoveduje. Premajhna ali prevelika povezanost ponavadi ni zaželena. 100 % identifikacija (angl. *ID-ness*) enolično določa atribut, vse njegove vrednosti so različne. Stabilnost (angl. *Stability*) odraža mero podobnosti med zapisi določene spremenljivke. Stolpec z manjkajočimi vrednostmi (angl. *Missing*) pove število manjkajočih zapisov pri določenem atributu (RapidMiner, 2018a).

Slika 17: Merila kvalitete (nekaterih) spremenljivk odločitvenega drevesa in nevronske mreže

Selected	Status	Quality	Name	Correlation ↓	ID-ness	Stability	Missing
☑	●		koliko dni je kupec	21.65%	0.00%	4.92%	0.00%
☑	●		StDniOdZadnjeUporabe	20.04%	0.00%	20.35%	0.00%
☑	●		povprečno trajanje naročnine (št. dni / naročnino)	17.95%	0.00%	4.90%	0.00%
☑	●		skupni plačani znesek podjetja	15.25%	0.00%	20.84%	0.00%
☑	●		skupni plačani znesek / uporabnika	9.24%	0.00%	20.84%	0.00%
☑	●		število izdanih računov	8.91%	0.00%	19.72%	0.00%

Vir: lastno delo.

Z namenom pridobitve čim bolj zmogljivega modela sem preizkusil različne globine algoritma drevesa odločanja in jih primerjal z že najzmogljivejšo nevronsko mrežo (tabela 2). Potrebno je dodati, da so bile nastavitve ostalih parametrov pri ugotavljanju najustreznejše globine odločitvenega drevesa nespremenjene.

Tabela 2: Primerjava zmogljivosti odločitvenih dreves in nevronske mreže

Globina drevesa / nevronska mreža	**7**	**8**	**9**	**10**	**11**	**Nevronska mreža**
Dvig (angl. *Lift*) v %	249,05	252,58	253,20	253,23	253,23	252,69
Delež natančnosti (angl. *Accuracy*)	82,72	83,07	83,19	83,22	83,22	94,62
Delež občutljivosti (angl. *Sensitivity*)	58,30	58,37	58,54	58,65	58,65	92,04
Delež preciznosti (angl. *Precision*)	91,94	93,25	93,48	93,49	93,49	93,29
Delež napačnih klasifikacij (angl. *Classification Error*)	17,28	16,93	16,81	16,78	16,78	5,38
Delež napovedi: oseba ostane - zapusti podjetje	80 - 20	80 - 20	80 - 20	80 - 20	80 - 20	70 - 30
Delež zaupanja v dvig pri 20 % populacije	91,21	96,36	97,27	97,27	96,97	98,18
Delež zaupanja v dvig pri 30 % populacije	46,97	41,82	40,91	40,91	40,30	96,06

Vir: lastno delo.

V tabeli 2 je možno razbrati, da je najzmogljivejše odločitveno drevo globine 10. Vrednosti nekaterih meril pri globini 11 so identične, vendar je boljša izbira manj razvejanega drevesa, ki ga je lažje interpretirati. Kriterij, na podlagi katerega se vozlišča drevesa delijo in posledično na koncu enote razvrščajo v ustrezno skupino, je *gain ratio*, ki meri delež dodatno pridobljene informacije z vključitvijo dodatnega vozlišča v model. Določena delitev vozlišča se zgodi ob predpogoju, da je s tem pridobljeno vsaj 10 % dodatne informacije. Prav tako mora biti izpolnjen pogoj prisotnosti minimalno 20 enot pri delitvi ter vsaj 10 enot v končnem listu drevesa. Stopnja zaupanja pri pesimističnem izračunu napake obrezovanja znaša 0,25 (RapidMiner, 2018b).

Najzmogljivejšo nevronsko mrežo sestavljajo 3 skrite plasti s po 16, 8 in 4 nevroni. Učenje algoritma poteka v 500 ponovitvah. Pri vsaki interakciji se uteži nevronov spremenijo za 0,3

(angl. *Learning Rate*). S pomočjo zagona (angl. *Momentum*), katerega vrednost znaša 0,2, se prepreči, da algoritem obtiči v lokalnem minimumu ali maksimumu (RapidMiner, 2018e).

3.2.4.3 Model prioritetne lestvice potencialnih kupcev

Enako kot v primeru testnega modela prioritetne lestvice potencialnih kupcev je tudi sedaj za enoto preučevanja določen registrirani uporabnik portala IUS-INFO. Številsko to pomeni 39.735 vključenih enot v model. Med možnim naborom 36 spremenljivk so bile na podlagi pregleda korelacij in smiselnosti s poslovno logiko izbrane 4, tako v primeru segmentacije brezplačnih uporabnikov kot segmentacije plačljivih uporabnikov. Zagotovljeni so enaki pogoji za analizo in primerjavo profilov.

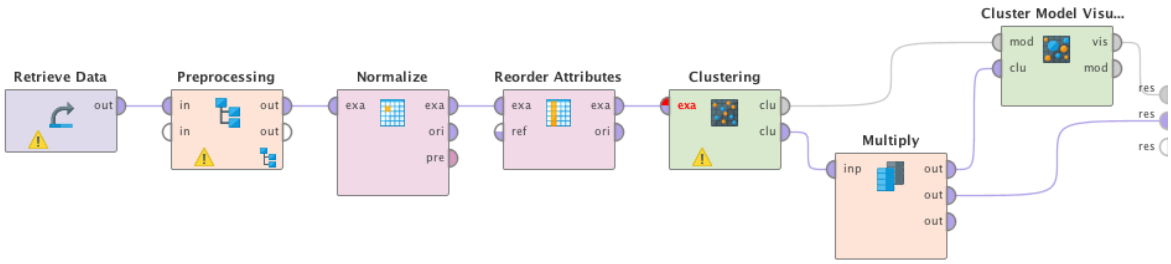
Faza modeliranja se prične v orodju Excel, kjer so zbrani vsi podatki predhodne faze. Tabela šteje 41.310 vrstic zapisov, 36 spremenljivk. Najprej filtriram uporabnike po tipu uporabnika naraščajoče. 1.500 oseb, ki so tipa *Knjižnica,* po posvetu z IUS-INFO ne pride v poštev pri modelu prioritetne lestvice potencialnih kupcev. Nato napolnim manjkajoče vrednosti 4 atributov, ki so vključeni v model:

- *število dni od zadnje uporabe* – 10.669 od 39.735 zapisov enot s povprečno vrednostjo 1.062;
- *število ogledov dokumentov* – 10.669 od 39.735 zapisov enot s povprečno vrednostjo 349;
- *število dni med registracijo in prvo interakcijo* – 27.357 od 39.735 zapisov s povprečno vrednostjo 336;
- *število dni med začetkom zadnjega paketa in zadnjo interakcijo* – 27.380 od 39.735 zapisov s povprečno vrednostjo 500.

Enote razdelim v 2 ločeni tabeli glede na to, ali so trenutno plačljivi ali brezplačni uporabniki portala IUS-INFO. Sledi shranjevanje in pričetek dela v orodju RapidMiner. S pomočjo čarovnika uvozim Excelovi tabeli brezplačnih in plačljivih uporabnikov ter ju ločeno shranim v lokalni repozitorij. S funkcionalnostjo *Automodel* hitro in enostavno izvedem segmentacijo, najprej pri nekupcih, nato še pri kupcih. Ugotovim, katere spremenljivke pridejo v poštev ter dobim dobro izhodišče za izboljšavo modelov.

Glavna procesa pri izgradnji obeh modelov sta v orodju RapidMiner identična, razlikujeta se v vhodnih podatkih in številu skupin, v katere so enote razdeljene. Proces (slika 18) se prične z branjem podatkov (gradnik *Retrieve Data*), ki sem jih že prej naložil in lokalno shranil v orodje. Sledi podproces *Preprocessing*, kjer nastavim določene parametre v zvezi s segmentacijo, nato podatke normaliziram (gradnik *Normalize*), atribute uredim po abecednem vrstnem redu (gradnik *Reorder Attributes*), izvedem delitev v skupine (gradnik *Clustering*) in na koncu rezultate vizualno predstavim (gradnik *Cluster Model Visualizer*).

Slika 18: Glavni proces izgradnje modela prioritetne lestvice potencialnih kupcev v orodju RapidMiner



Vir: lastno delo.

Najprej se lotim segmentacije brezplačnih uporabnikov, v katero vključim maksimalno število smiselnih spremenljivk, kar v tem primeru pomeni 4. Omenjene spremenljivke vnaprej ne pogojujejo mej med plačljivimi in brezplačnimi uporabniki, imajo zadostno variabilnost ter nobene manjkajoče vrednosti (slika 19). Stopnjo povezanosti med atributi je možno razbrati na sliki 20, korelacije so na eni strani šibke in pozitivne ter na drugi šibke in negativne. Preveriti želim, ali je kateri od profilov nekupcev po obnašanju podoben profilu kupcev ter posledično potencialen za nadaljnje trženjske aktivnosti. Zato pri segmentaciji plačljivih uporabnikov v model vključim identične spremenljivke, tako je zagotovljena primerljivost profilov.

Slika 19: Merila kvalitete spremenljivk segmentacije

Select...	Status	Quality	Name	Correlat...	ID-ness	Stability	Missing
✓	●		StDniOdZadnjeUporabe	?	0.00%	29.54%	0.00%
✓	●		SteviloOgledovDokumentov	?	0.00%	29.51%	0.00%
✓	●		Lead Score	?	0.00%	41.08%	0.00%
✓	●		število interakcij (email & klic)	?	0.00%	79.44%	0.00%

Vir: lastno delo.

Slika 20: Korelacijska matrika uporabljenih spremenljivk segmentacije

Attributes	Lead S...	StDniO...	število ...	Stevilo...
Lead Score	1	-0.183	-0.041	0.021
StDniOdZadnjeUporabe	-0.183	1	-0.044	-0.133
število interakcij (email & klic)	-0.041	-0.044	1	0.167
SteviloOgledovDokumentov	0.021	-0.133	0.167	1

Vir: lastno delo.

Pri segmentaciji brezplačnih in plačljivih naročnikov portala IUS-INFO je uporabljen algoritem *Clustering k-Means*, v prvem primeru gre za delitev v 3 skupine, v drugem zgolj v 2. Evklidska razdalja predstavlja merilo, na podlagi katerega je enota dodeljena ustrezni skupini. Algoritem najprej določi dobro izhodišče, nato računa razdalje med točkami oz. vključenimi spremenljivkami (RapidMiner, 2018c). Glede na končne nastavitve parametrov so v obeh primerih v modelu prisotni izključno številski atributi.

3.2.5 Ocena rezultatov

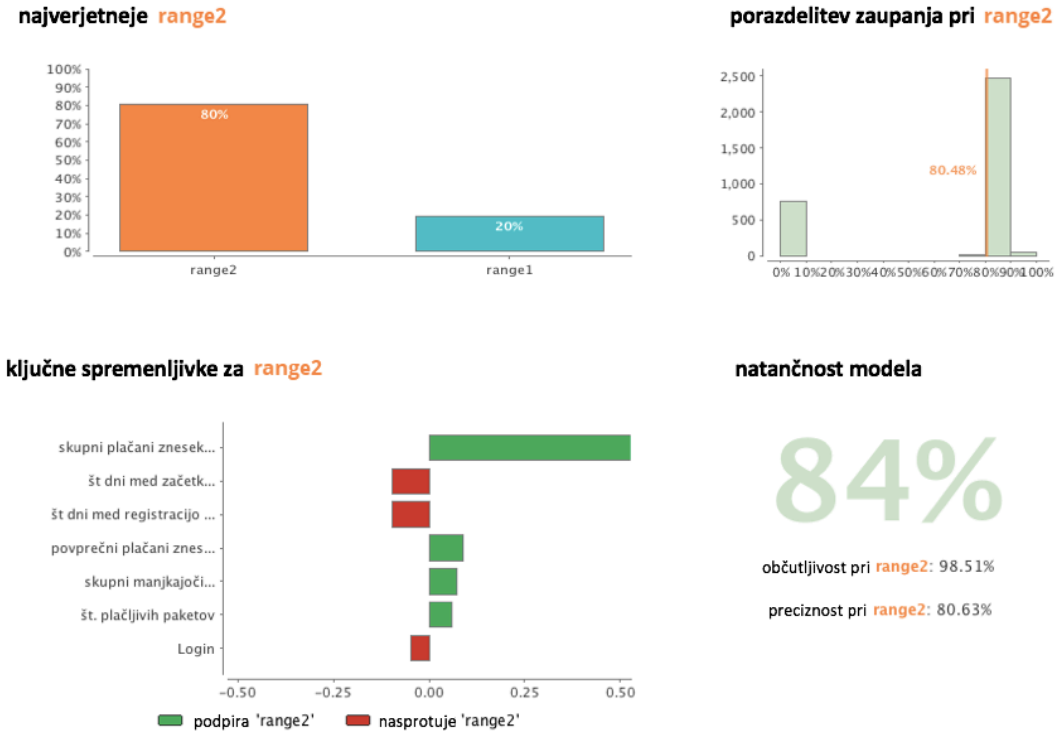
V prejšnjem poglavju sem predstavil pristop k doseganju glavnih ciljev projekta, izdelavi modela preprečitve odhodov kupcev in modela prioritetne lestvice potencialnih kupcev ter primerjal zmogljivosti algoritmov strojnega učenja na razpoložljivih podatkih. V tem poglavju se osredotočam na oceno najboljših rešitev.

3.2.5.1 Model preprečitve odhodov kupcev

Pri modelu preprečitve odhodov kupcev sta možni 2 rešitvi, na eni strani odločitveno drevo, kjer je lažje interpretirati dobljene rezultate in predlagati ukrepe, ter na drugi strani nevronska mreža, ki daje boljše napovedi. Med vsemi drevesi odločanja je najboljše tisto z 10 nivoji. V Tabeli 2 je razvidno, da ima omenjeno drevo v 2 od 8 primerov boljšo zmogljivost v primerjavi z najboljšo nevronsko mrežo. Oba modela spadata v kategorijo napovedne analitike.

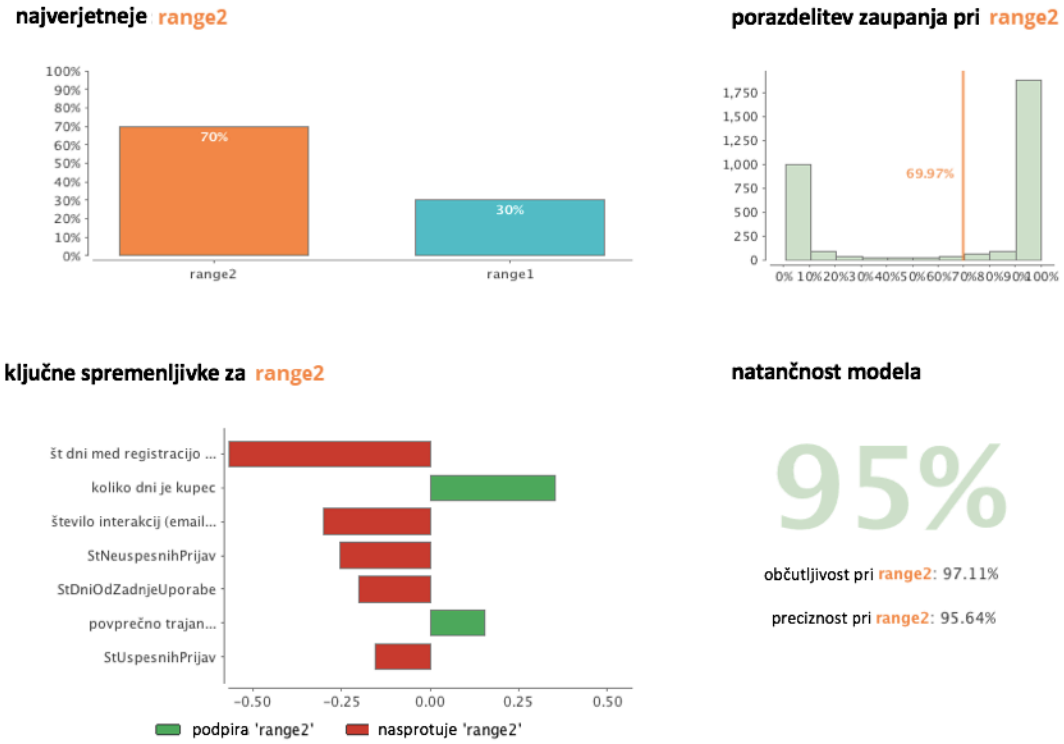
Na slikah 21 in 22 sta povzetka najboljšega drevesa odločanja ter najzmogljivejše nevronske mreže. Povzetek vključuje levo zgoraj delež napovedi po razredih (range1 so osebe, ki praviloma zapuščajo podjetje), desno zgoraj porazdelitev zaupanja, levo spodaj najpomembnejše spremenljivke ter desno spodaj natančnost, občutljivost in preciznost modela.

Slika 21: Povzetek drevesa odločanja



Vir: lastno delo.

Slika 22: Povzetek nevronske mreže



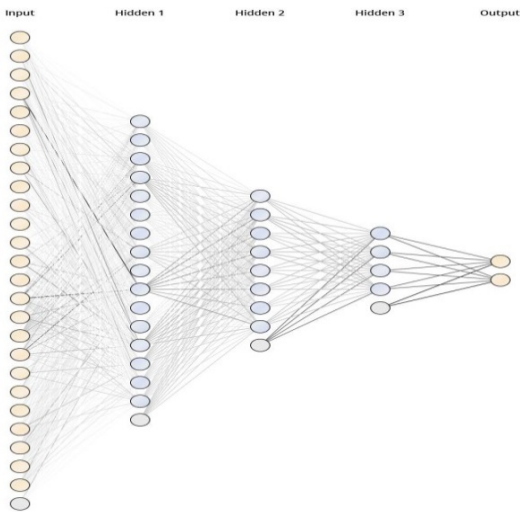
Vir: lastno delo.

S slike 23 je razvidno, da je v veliki meri dosežena čistost listov drevesa odločanja, kar je primarni cilj te tehnike. Pomembnost spremenljivk modela se zmanjšuje od vrha proti dnu narobe obrnjenega drevesa. Na podlagi vključenih atributov so bile enote klasificirane v kategorijo kupcev in nekupcev, znani so pogoji, pod katerimi osebe praviloma zapuščajo (range1) ali ostajajo v podjetju (range2).

Slika 23: Končna rešitev drevesa odločanja

Vir: lastno delo.

Slika 24: Končna rešitev nevronske mreže

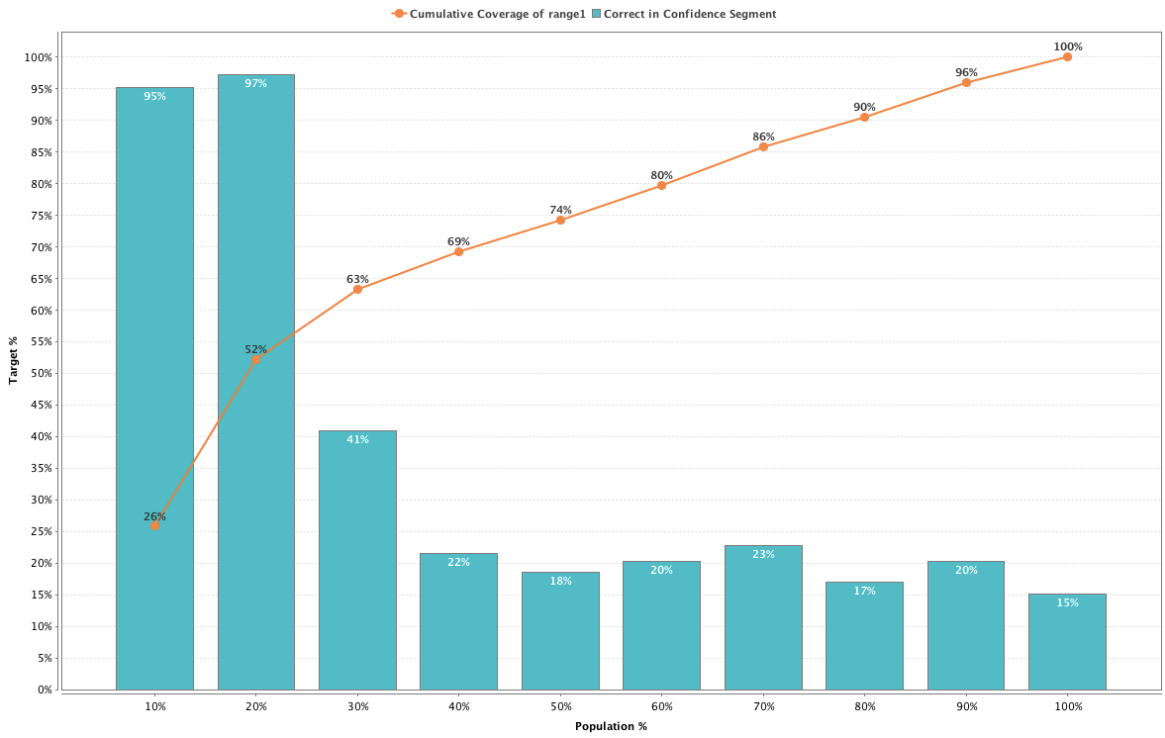


Vir: lastno delo.

Na sliki 24 je nevronska mreža, ki šteje 25 vhodnih spremenljivk (angl. *Input*), ima 3 skrite plasti (angl. *Hidden Layer 1–3*) ter 2 izhodni napovedni vrednosti (angl. *Output*). Prav tako so razvidne vse povezave med vhodi, nevroni in izhodi. Izhoda povesta, ali je bila oseba klasificirana v skupino, ki verjetno ostaja v podjetju, ali v skupino, ki jo verjetno zapušča.

Uporabnost odločitvenega drevesa (slika 25) je do 20 % populacije oz. 1.538 oseb. Pri tej meji znaša vrednost zaupanja 97 %. Povprečna napoved z modelom je 2,6-krat boljša od naključnega ugibanja. Model identificira 52 % pripadnikov razreda, ki praviloma zapuščajo podjetje (cumulative coverage of range1) že z 20 % populacije. Brez modela bi bilo za doseganje enakega deleža potrebnih 52 % enot.

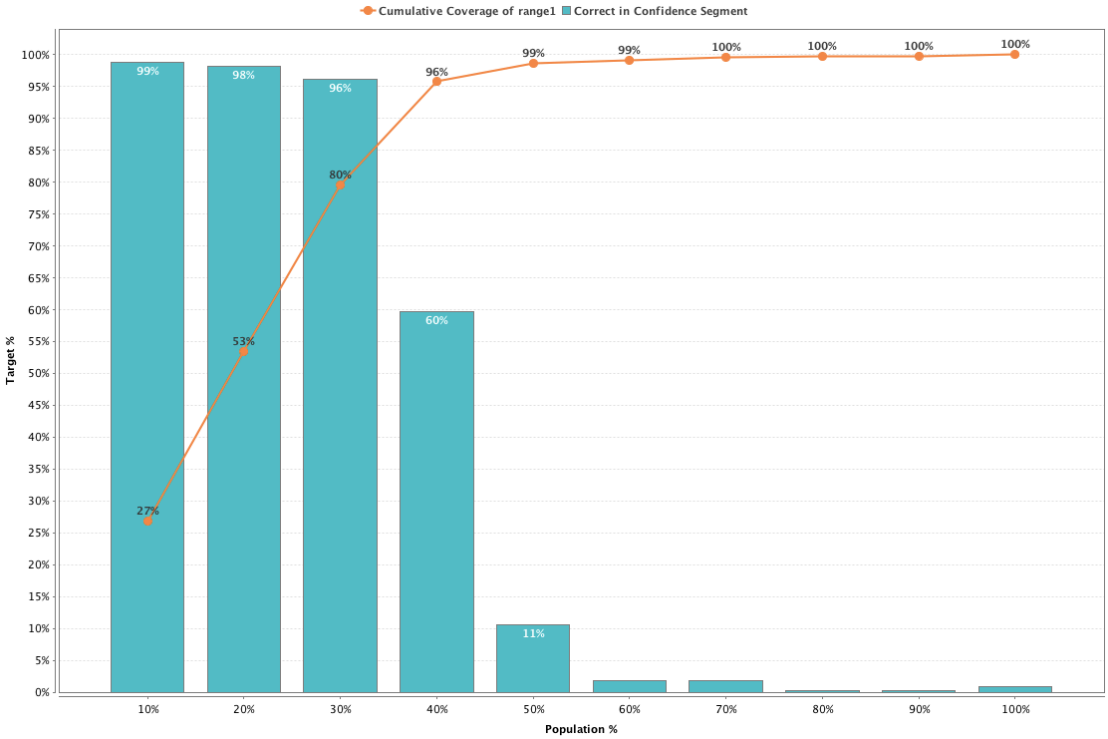
Slika 25: Grafikon dviga drevesa odločanja



Vir: lastno delo.

Uporabnost nevronske mreže (slika 26) je do 30 % populacije oz. 2.307 oseb. Pri tej meji znaša vrednost zaupanja 96 %. Povprečna napoved z modelom je 2,67-krat boljša od naključnega ugibanja. Model identificira 80 % pripadnikov razreda, ki praviloma zapuščajo podjetje (cumulative coverage of range1) že z 30 % populacije. Brez modela bi bilo za doseganje enakega deleža potrebnih 80 % enot.

Slika 26: Grafikon dviga nevronske mreže

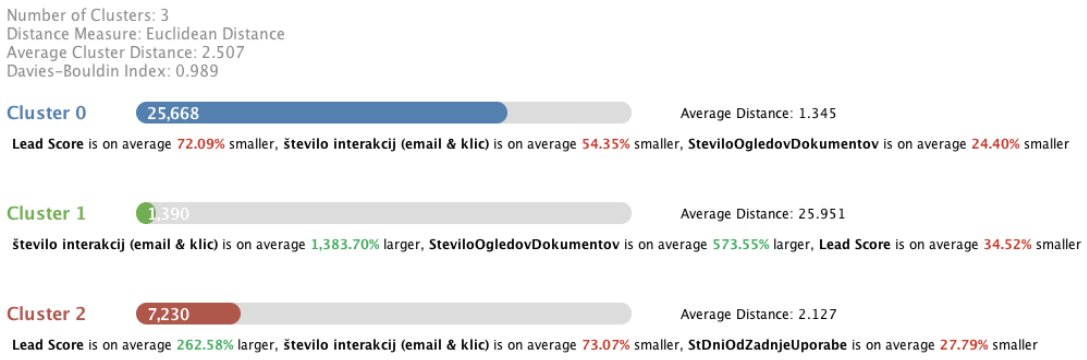


Vir: lastno delo.

3.2.5.2 Model prioritetne lestvice potencialnih kupcev

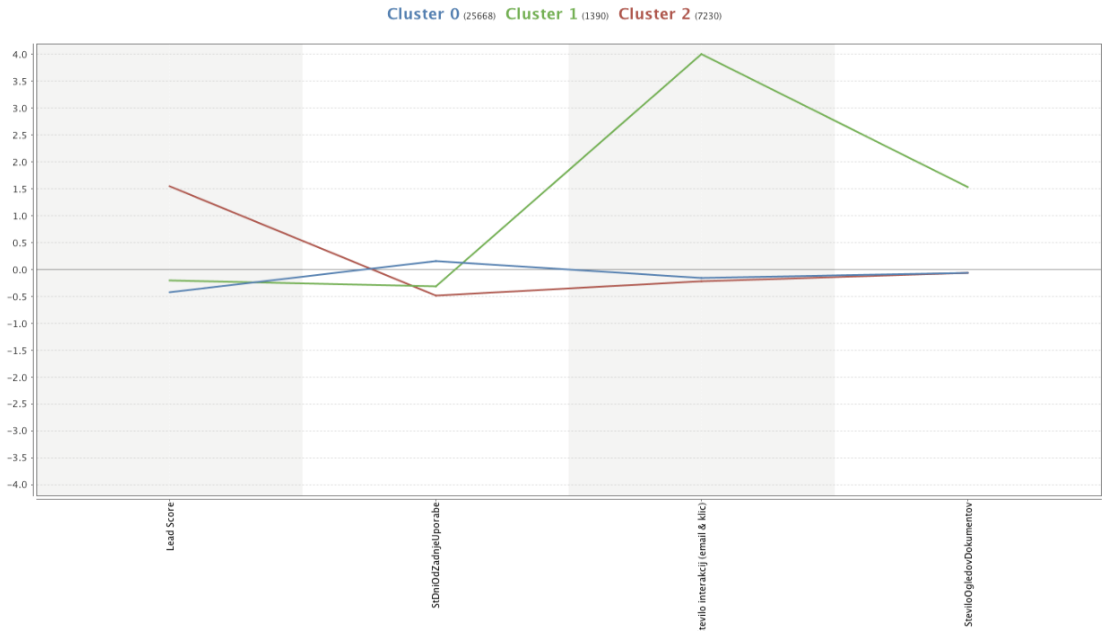
Odločim se za delitev brezplačnih uporabnikov v 3 skupine (slika 28): aktivne nekupce (*Cluster 2*), ki jih je 7.230, bivše kupce (*Cluster 1*), ki jih je 1.390, in neaktivne nekupce (*Cluster 0*), ki jih je 25.668. Delitev v 3 segmente daje najbolj raznolike profile, ki jih je glede na značilnosti pripadnikov možno poimenovati brez težav. Prav tako se skupine med seboj najbolj razlikujejo oz. je dosežena heterogenost. Znotraj posameznega segmenta je največja homogenost oz. so si pripadniki med sabo zelo podobni.

Slika 27: Opis najpomembnejših značilnosti segmentov brezplačnih uporabnikov



Vir: lastno delo.

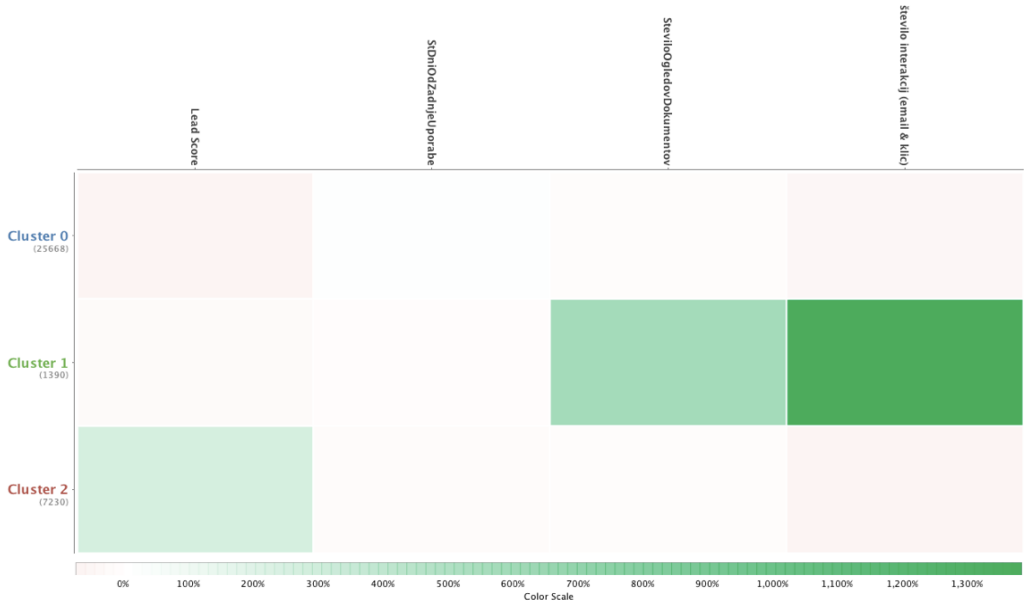
Slika 28: Profili brezplačnih uporabnikov



Vir: lastno delo.

Metodologija predstavitve s pomočjo toplotne karte (slika 29) uporablja barve za predstavitev različnih vrednosti določenega atributa. Termična analogija lahko vsebuje različne odtenke določene barve ali predstavitev z večjim številom barv (Sullivan, 2012, str. 166). Izrazitejši odtenki odražajo višje vrednosti določene spremenljivke.

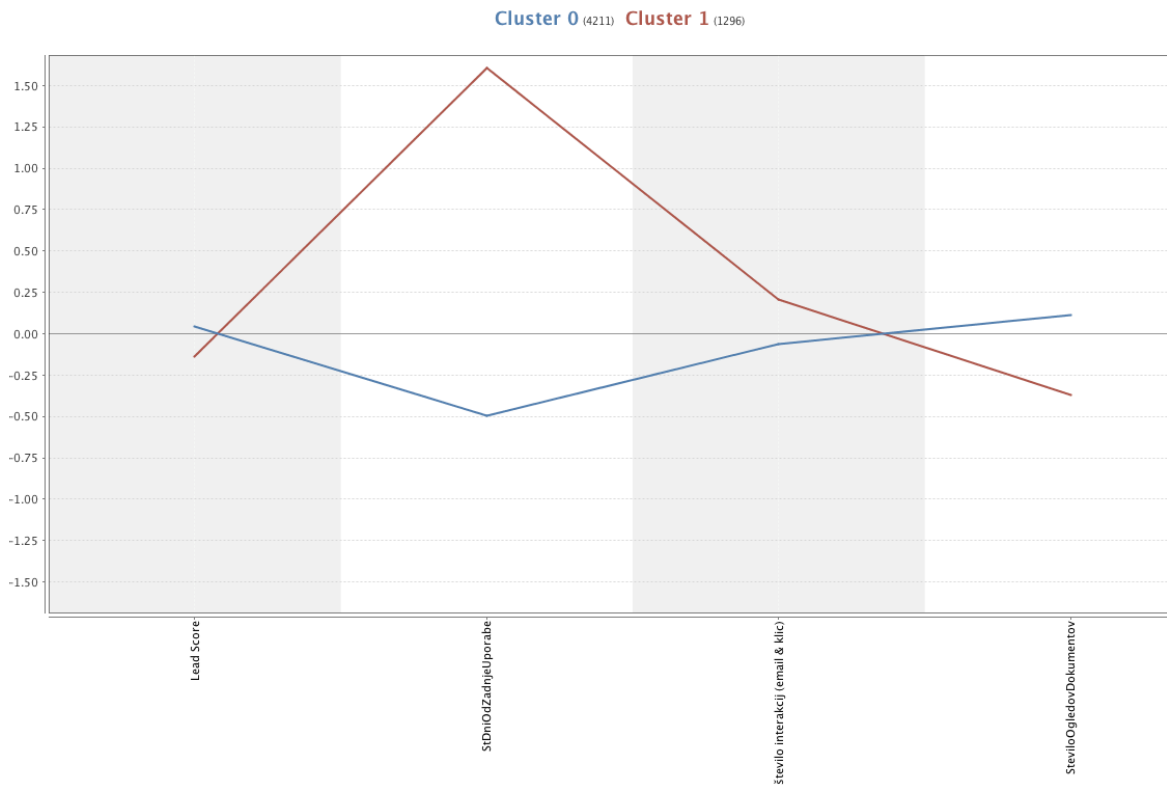
Slika 29: Toplotna karta spremenljivk brezplačnih uporabnikov



Vir: lastno delo.

Plačljive uporabnike razdelim v 2 skupini (slika 30), in sicer na bolj aktivne (*Cluster 0*), ki jih je 4.211, ter manj aktivne (*Cluster 1*), ki jih je 1.296. Segmentacija z zgolj 2 skupinama predstavlja v konkretnem primeru najboljši rezultat.

Slika 30: Profili plačljivih uporabnikov

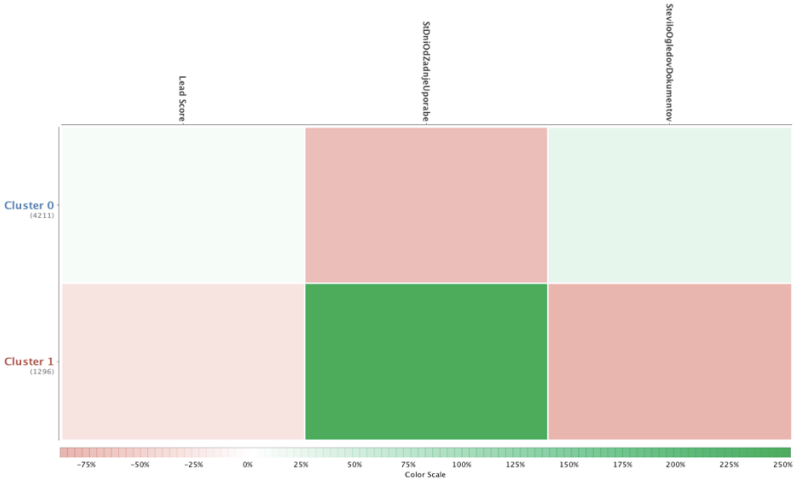


Vir: lastno delo.

Slika 31: Opis najpomembnejših značilnosti segmentov plačljivih uporabnikov

Vir: lastno delo.

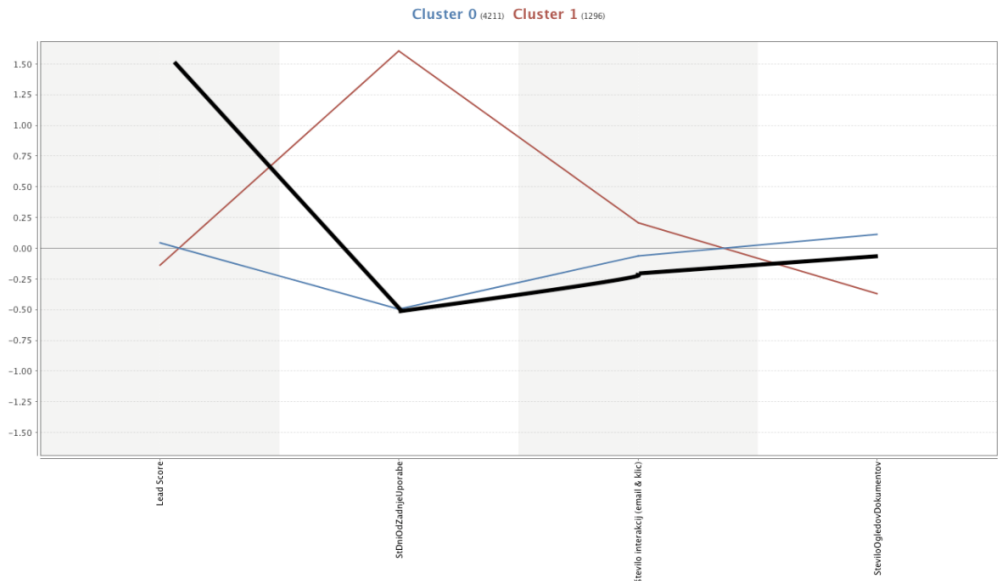
Slika 32: Toplotna karta spremenljivk plačljivih uporabnikov



Vir: lastno delo.

Kot je bilo rečeno, je bil cilj segmentacije odkriti potencialne brezplačne uporabnike. Po tem, ko sem v koordinatni sistem profilov plačljivih uporabnikov vrisal profil segmenta aktivnih nekupcev (odebeljena črna črta na sliki 33), sem prišel do zanimive ugotovitve. Skupina aktivnih nekupcev si je po obnašanju v veliki meri podobna s segmentom aktivnih kupcev (*Cluster 0*). Izjema je spremenljivka *Lead Score*, ki pri nekupcih znaša precej več, ujemanje v preostalih 3 atributih je očitno. Potrebno je povedati, da to ni v nasprotju s poslovno logiko, saj so najbolj potencialni brezplačni uporabniki po pričakovanjih bolj odzivni na trženjske aktivnosti, ki se merijo s pomočjo te spremenljivke.

Slika 33: Primerjava profilov brezplačnih in plačljivih uporabnikov



Vir: lastno delo.

3.2.6 Uvedba v prakso

Po koncu modeliranja in pridobljenih najboljših rezultatih sem celoten proces podatkovnega rudarjenja in rezultate predstavil sodelujočim na projektu ter nekaterim ostalim članom podjetja LEXPERA d. o. o. Želja je bila, da predlagam možne načine aplikacije v prakso. Ker je obstajal dvom, katero rešitev izbrati kot končno pri modelu preprečitve odhodov kupcev, sem dobil s strani podjetja dejanske podatke za prvih 6 mesecev glede na končne podatke, ki so bili uporabljeni v raziskavi. Dejansko stanje za to obdobje se bolje ujema z napovedmi odločitvenega drevesa oz. z ugotovitvami, ki jih ponuja odločitveno drevo. V nadaljevanju se osredotočam na ugotovitve posameznega modela in predloge ukrepov na podlagi ugotovitev ter prav tako izpostavim potencialne omejitve.

3.2.6.1 Model preprečitve odhodov kupcev

Na podlagi najzmogljivejšega odločitvenega drevesa je možno razbrati nove in koristne informacije o delovanju portala IUS-INFO. Tičejo se trajanja naročnine, števila zakupljenih paketov, trajanja od zadnje interakcije s podjetjem ter višine določenega plačila. Glavne ugotovitve so:

- za osebo, ki sklene naročniško razmerje za manj kot 1 leto, je velika verjetnost, da bo zapustila podjetje;
- podjetje z vsaj 4 zakupljenimi uporabniškimi računi praviloma lažje odpove naročnino 1 ali več uporabnikov kot manjše podjetje;
- v primeru, da je minilo več kot 1278 dni med začetkom zadnjega paketa in zadnjo interakcijo s kupcem, bo ta verjetno zapustil podjetje. Predpogoj predstavlja trajanje naročniškega razmerja z vrednostjo višjo od 363,5 dni;
- oseba s povprečno visokimi zneski plačil bo praviloma zapustila podjetje. V tem primeru sta predpostavki, da znaša skupno plačani znesek na uporabnika več kot 213 HRK ter da ni bilo stika v zadnjih 1278 dneh.

Menim, da je na podlagi ugotovitev odločitvenega drevesa možno izboljšati poslovanje, zato predlagam sledeče ukrepe:

- podjetje naj poskuša z uporabnikom že v začetku skleniti naročniško razmerje za obdobje vsaj 1 leta in eventuelno ponudi možnost plačila na obroke. S tem pridobi zgolj bolj pripadne in zveste kupce;
- s plačljivimi naročniki je potrebno komunicirati bolj pogosto ter jim tako izkazati pomembnost. V mislih imam interakcije preko SZAT-a;
- visok znesek računa je bolje razčleniti na večje število manjših plačil;
- za velika podjetja je potrebno pripraviti posebno ponudbo oz. jim ponuditi popust v primeru, da želijo skozi naročniško obdobje zmanjšati število uporabniških računov.

Ob tem je pomembno izpostaviti, da so pri enem ali več predlaganih ukrepih prisotne omejitve. Mišljeno je sklepanje naročnine za obdobje minimalno 1 leta, kar lahko pomeni potencialno izgubo prihodkov s strani tistih, ki vidijo korist IUS-INFO portala zgolj kratkoročno. Prav tako lahko preveč komunikacije povzroči ravno nasproten efekt od želenega, in sicer da se uporabnik počuti bombardiranega s sporočili in ravno zato odpove svojo naročnino.

3.2.6.2 Model prioritetne lestvice potencialnih kupcev

Na podlagi primerjave profilov brezplačnih in plačljivih uporabnikov je možno sklepati, da je segment 7.230 aktivnih nekupcev najbolj potencialen za nadaljnje trženjske aktivnosti. Dodaten argument predstavlja podatek, da 6.241 od 7.230 pripadnikov prej omenjenega segmenta še ni imelo breplačnega preizkusnega obdobja ter da jih 5.788 od 7.230 ni bilo niti enkrat kontaktiranih preko elektronske pošte ali klica.

Ker prej opisani segment kaže zadostno mero potenciala, predlagam, da se stanje preveri v praksi. Postopek izgleda tako, da podjetje najprej naključno izbere 100 oseb segmenta potencialnih nekupcev, jih kontaktira preko elektronske pošte ali telefonskega klica in preveri njihov dejanski interes za plačljivo naročnino ali brezplačno preizkusno obodobje. V primeru, da se na podlagi vzorca potrdi zadostna mera potenciala, se tokrat podjetje osredotoči na celoten segment. Izvede trženjsko kampanjo ali več le-teh po korakih. Osebam, ki se odzovejo, se ponudi plačljiv izdelek. Uporabnikom, ki ponudbo zavrnejo oz. v začetku niso pripravljeni plačati, se ponudi brezplačno preizkusno obdobje s kasnejšo plačljivo vezavo. Kot je znano, ravno osebe tega segmenta najbolj intenzivno uporabljajo portal IUS-INFO ter je posledično večja verjetnost, da postanejo plačljivi uporabniki, brez omejitev dostopa na mesečni ravni.

Segmentacija je bila izvedena na najboljši možni način, vendar so bile prisotne določene omejitve:

- prvo omejitev predstavlja dejstvo, da sta bila segmentiranje uporabnikov in primerjava profilov brezplačnih s plačljivimi segmenti narejena na zgolj 4 spremenljivkah. Ostali atributi niso primerni, saj njihove vrednosti že vnaprej izrazito razlikujejo kupce in nekupce;
- druga pomembna omejitev je raven kakovosti podatkov. S kakovostjo mislim na napolnitev določenih manjkajočih zapisov spremenljivke *Število dni od zadnje uporabe* in *Število ogledov dokumentov* s povprečnimi. 1.879 uporabnikov z napolnjeno vrednostjo vsaj ene spremenljivke pade ravno v segment potencialnih nekupcev. Če so te enote izločene, ostane še vedno 4.362 potencialnih oseb;
- tretja omejitev se posredno navezuje na prvo, gre za ažurnost podatkov. Vključene spremenljivke merijo aktivnosti uporabnika v zadnjih 3 mesecih uporabe. Podatki uporabnikov se skozi čas verjetno razlikujejo. Za konkretno raziskavo so na razpolago do 1. 1. 2018.

4 DISKUSIJA

Za začetek nekaj besed o tem, kako ustvariti interes za projekt, kot je podatkovno rudarjenje, v podjetju, kjer je omenjeni pristop neznanka, še manj znane pa so vse koristi, ki jih prinaša. Potrebno je oceniti organizacijo oz. študijo primera, ki bi jo lahko zanje potencialno izvedli. Dodana vrednost uporabe takšnega pristopa mora biti izrazita. Podatkovno rudarjenje se uporablja v številnih panogah, vendar ni primerno za vsak poslovni model. Sledi vprašanje, kje ga je vredno uporabiti. Zelo primerne so tiste dejavnosti, kjer je poslovanje s kupcem zgolj elektronsko in se nakup potencialno zgodi večkrat. Kot primer takšnega poslovnega modela navajam spletno trgovino s športno prehrano. V tem primeru je človeški faktor pri samem nakupnem procesu, ki ga je zahtevno vključiti v model, minimalen ali ničen. Prav tako je dokaj enostavno meriti obnašanje uporabnika na spletu. V dejavnostih, pri katerih se stranki izdelek ali storitev proda izključno enkrat, uporaba podatkovnega rudarjenja ni najbolj smiselna. Pomemben je čas obstoja podjetja na trgu, stopnja razvitosti in predvsem, koliko ter kakšne kvalitete so razpoložljivi podatki. Organizacija, ki deluje šele prvo leto, verjetno nima na voljo dovolj podatkov. Lahko sklepamo, da z leti raste obseg, kakovost podatkov pa je vprašljiva.

Moj pristop k pridobitvi študije primera iz realnega poslovnega okolja je bil takšen, da sem našel svetovalno podjetje FrodX, ki se ukvarja s sorodno dejavnostjo. Za stranke ima številna slovenska ter nekaj tujih organizacij in v bodoče namerava nuditi tudi bolj napredne analize, kot je podatkovno rudarjenje. Torej imeli smo skupni interes, možnost za pridobitev ustrezne študije sem si povečal, saj je bilo zaupanje na relaciji agencije in njenih strank že vzpostavljeno. Nekatere stranke so bile za noviteto, kot je uvedba podatkovnega rudarjenja, v njihov poslovni ekosistem primerne že v danem trenutku, nekatere bodo v bodoče. Zgolj eno takšnih je bilo podjetje LEXPERA d. o. o., ki se je odločilo za analizo uporabe njihovega pravno-informacijskega portala IUS-INFO na Hrvaškem.

Kot je bilo že rečeno, je bila z namenom uspešne uvedbe podatkovnega rudarjenja v pravno-informacijski portal IUS-INFO ustvarjena projektna skupina. Vključene osebe so imele raznolike poglede in znanja, od zelo tehničnega do vodstvenega pogleda ter vidika trženja. Ključna oseba pri projektu je bil sponzor, direktor podjetja, ki je zaradi svojega položaja omogočal napredek, ne glede na vse izzive, prepreke in zastoje. Pri predstavitvi rezultatov se je izkazalo, da je v projektni skupini mogoče od začetka manjkala še oseba oddelka prodaje, ki je dnevno ravno največ v stiku s strankami. Ideje, ki so bile izpostavljene s strani prodajnega oddelka, so bile posredno že vključene v model preprečitve odhodov kupcev in model prioritetne lestvice potencialnih kupcev. Vendar ne bi škodila dodatna perspektiva vse od začetka izvedbe. Zastoji oz. čakanje na podatke se odražajo v trajanju projekta. Prvo srečanje je bilo konec avgusta 2017 ter zadnje konec marca 2018, kar znaša skupaj 7 mesecev. Transakcijski podatki, potrebni za izvedbo, so bili na razpolago šele v začetku januarja 2018. Po mojem mnenju je bil glede na obseg projekta interval izvajanja (pre)dolg. Na podlagi trajanja je možno sklepati, da projekt ni bil med najpomembnejšimi prioritetami

podjetja. Kar ni presenetljivo, saj je šlo za primer prve takšne pilotne študije, z izrazito nejasnim končnim izkupičkom.

Drugi razlog za 7-mesečno trajanje je možno utemeljiti z dejstvom, da v začetku ni bil definiran datum zaključka. Dobro so bili opredeljeni cilji in razpoložljivi podatki, ki so se nahajali v 4 sistemih. Prav tako se ni podrobneje načrtovala izvedba vsake od projektnih faz. Za analizo je bila na razpolago ogromna količina podatkov, zato se je za lažje razumevanje izdelal relacijski model, ki je dobro identificiral potencialne težave pri integraciji in pripravi podatkov. Razumevanje podatkov je bilo med drugim enostavnejše tudi zaradi natančno podanega metaopisa spremenljivk s strani LEXPERE in FrodX-a. Primarni namen zbiranja razpoložljivih podatkov je bil drugačen, kot bi ustrezal BI potrebam, zato je bila kvaliteta podatkov generalno gledano na nižji ravni. Predvsem CRM podatki so vsebovali številne manjkajoče zapise, ki jih je bilo potrebno odpraviti v fazi priprave. Projekt predstavlja dobro izhodišče za zbiranje ustreznih in kvalitetnejših BI podatkov za v bodoče, saj sem ugotovil trenutne pomanjkljivosti. Kot glavno bi izpostavil nezmožnost uporabe zajetnih podatkov glede fizičnih sestankov, ki so bili zabeleženi, vendar niso imeli ustreznega ključa, preko katerega bi jih dodelil ustrezni enoti preučevanja.

V fazi priprave podatkov, kjer so se na novo računali številni atributi na osnovi obstoječih transakcijskih podatkov, so bile vrstice ponavadi najprej izračunane s pomočjo vrtilnih tabel orodja MS Excel, nato združene z orodjem PDI, preko ustreznih ključev. Še pred tem so se v Excelu s pomočjo filtrov in razvrščanja odstranili neuporabni zapisi ali so se ročno odpravile napake, narejene že pri vnosu v posamezni IS. Za izračunavanje so bile med drugim uporabljene funkcije MIN, MAX, SUM, COUNT, AVERAGE. Na takšen način se je zmanjšal obseg podatkov in vsebina prilagodila BI namenu. Kljub glavnim prednostim Excela, kot sta enostavnost uporabe ter preglednost, se orodje na določene trenutke ni pokazalo kot dovolj robustno. Pri uporabi večjega števila funkcij naenkrat je izvajanje zamrznilo ali se izrazito podaljšalo, zato so bile spremenljivke računane postopoma. Kar nakazuje na možnost uporabe vzorčnih transakcijskih podatkov, potrebo po zmogljivejši procesorski moči ali celo uporabo drugega orodja, v primeru enormnih zapisov.

Testno in dejansko modeliranje je bilo izvedeno s pomočjo orodja RapidMiner, ki je ravno v letu 2018 uvedlo noviteto znotraj rešitve, in sicer modul *Automodel*, ki pripomore k hitrejšemu ter enostavnejšemu modeliranju. Trditev se je potrdila tudi na študiji primera IUS-INFO, saj je omenjena funkcionalnost zelo dobro identificirala ključne spremenljivke modela glede na izbrano nalogo podatkovnega rudarjenja, primerjala zmogljivosti številnih algoritmov na vhodnih podatkih, nudila možnost primerjave končnega rezultata glede na spremembe vrednosti vhodnih atributov in v ozadju ustvarila proces z vsemi gradniki. Namesto preizkušanja in optimizacije številnih algoritmov strojnega učenja v iskanju najboljše rešitve sem se lahko hitro osredotočil na zgolj peščico le-teh. Po mojem prepričanju je takšna bližnjica prinesla velike prihranke časa, tako v primeru izdelave modela preprečitve odhodov kupcev kot modela prioritetne lestvice kupcev. Prav tako je orodje nudilo pomoč

pri odločitvi glede nastavitev posameznih parametrov, saj obstajajo podatki, deleži uporabe konkretnih nastavitev gradnikov na ravni celotne skupnosti oz. vseh uporabnikov.

Kljub temu, da so bile pri projektu nekatere pomanjkljivosti v primerjavi z uporabljeno CRISP DM metodologijo ter številni izzivi, sem prišel do novih, zanimivih in predvsem uporabnih ugotovitev. Kot je bilo že rečeno, je pri modelu preprečitve odhodov kupcev obstajala dilema glede izbora najboljše rešitve med nevronsko mrežo in odločitvenim drevesom. V zakup so bile vzete tako želje glede možne vizualne predstavitve vseh korakov do rezultatov kot dejansko ujemanje napovedi s podatki obdobja naslednjih 6 mesecev glede na končne podatke raziskave. Najboljšo rešitev je tako odražalo drevo odločanja z 10 nivoji.

Glede na 84 % natančnost napovednega modela preprečitve odhodov kupcev in povprečno 2,6-krat boljšo napoved z modelom od naključnega ugibanja, lahko trdim, da je model zanesljiv in bi neuporaba v poslovnem okolju predstavljala zgrešeno poslovno priložnost. Visoka natančnost modela je posledica vključitve 25 najbolj relavantnih spremenljivk, ki pokrivajo vse 4 izhodiščne vire podatkov. Torej, da osebe praviloma ne bodo odhajale, mora podjetje v začetku poizkusiti skleniti pogodbo za obdobje daljše od enega leta, s strankami komunicirati bolj pogosto, nuditi možnost plačila na obroke ter pripraviti prilagojeno ponudbo za organizacije, ki imajo vsaj 4 uporabniške naročnine. Hkrati je potrebno biti zmeren, saj lahko preveč komunikacije prav tako povzroči odhode. Aplikacija ugotovitev bi verjetno zahtevala določene prilagoditve obstoječih poslovnih procesov oz. poslovnih pravil.

Pri modelu prioritetne lestvice potencialnih kupcev je bila zaradi narave podatkov izbrana drugačna pot. Namesto napovednega modela se je izvedla segmentacija, ki je zmogljiva tudi na manjšem številu atributov ali enot preučevanja. Ločeno sem razvrstil brezplačne in plačljive uporabnike v smiselno število skupin ter iskal podobnosti med profili nekupcev in kupcev. Kljub večjim številom omejitev kot v primeru prvega modela sem prišel do zadovoljivih rezultatov, ki bi jih bilo prav tako vredno preizkusiti v praksi. Pri identifikaciji segmenta 4.362 najbolj potencialnih oseb, da postanejo kupci, sem v zakup vzel dejstvo, da je bila segmentacija izvedena na podlagi 4 najustreznejših spremenljivk, da kakovost teh atributov ni bila popolna ter da je obstajalo tveganje zastarelosti podatkov. Glavni razlog za visok potencial segmenta se odraža v dejstvu, da je obnašanje pripadnikov skoraj identično kot pri osebah segmenta aktivnih plačljivih uporabnikov ter da jih večina še ni bila nikoli kontaktiranih in prav tako ni imela brezplačnega preizkusnega obdobja plačljivega paketa. Zaradi naštetih omejitev bi bilo pametno postopno preveriti dejanski potencial uporabnikov, in sicer najprej na podlagi vzorca ter šele nato na ravni celotnega segmenta, če bi vzorec potrjeval zadostno mero nakupnega potenciala. S takšnim pristopom se zmanjša tveganje neuspešne ali napačne uvedbe.

SKLEP

Podjetja dnevno ustvarijo ogromne količine podatkov, ki jih je zahtevno analizirati. Pristopi, ki so v preteklosti dajali rezultate, niso več zadostni. Podatkovno rudarjenje je eden od možnih odgovorov na naraščajoče analitične potrebe. Pri odkrivanju znanja v podatkih se pojavijo številne priložnosti, izzivi, omejitve in predvsem potencialne visoke koristi, ki se ob ustrezno vloženem trudu odrazijo tudi v finančnem smislu.

Skozi analizo primera sem ugotovil ključne dejavnike za končni uspeh projekta podatkovnega rudarjenja, ki so: ustrezen pristop oz. izbrana metodologija podatkovnega rudarjenja, jasno definirani in merljivi cilji, izdelan projektni plan z roki dokončanja posamezne faze, sestava projektne skupine z raznolikim znanjem, sponzor projekta z vplivom v podjetju, zadostna količina ter zadovoljiva kakovost razpoložljivih podatkov, podan metaopis oz. vsebinska razlaga podatkov, ustrezno znanje in orodja za pripravo ter modeliranje podatkov, realno ovrednotenje in predstavitev dobljenih rezultatov.

Potrdil sem možnost uporabe podatkovnega rudarjenja v podjetju LEXPERA d. o. o. s pomočjo izdelave v praksi delujočega modela preprečitve odhodov kupcev ter modela prioritetne lestvice potencialnih kupcev portala IUS-INFO. Oba modela kažeta visoko verjetnost delovanja v poslovnem svetu. V zakup so bile vzete določene omejitve, na podlagi katerih sem prišel do končnih rezultatov.

Ključen prispevek magistrske naloge je po korakih, natančno predstavljen pristop k študiji primera podatkovnega rudarjenja v panogi pravnih storitev oz. na primeru portala IUS-INFO. Prav tako je predstavljena aktualna literatura področja, ki nudi teoretično oporo praktičnemu primeru. Pridobljeno znanje je možno uporabiti na podobnih primerih v isti dejavnosti, sorodnih ali širše; v vseh podjetjih doma in v tujini, ki imajo podoben poslovni model, torej stranke plačujejo naročnino za koriščenje vsebin spletnega portala. Njihovo poslovanje v večji meri poteka elektronsko, zato je možna sledljivost obnašanja kupcev ter izdelava modelov, ki nudijo boljšo pomoč pri odločanju kot naključno ugibanje.

V primeru, da bi bili na razpolago tudi podatki, kot je število zaposlenih v podjetju določenega kupca ter panoga dejavnosti podjetja stranke, bi pri modelu preprečitve odhodov kupcev verjetno dosegel še višjo natančnost modela. Na drugi strani, pri modelu prioritetne lestvice kupcev, ne bi škodili podatki oz. atributi zbrani na ravni kategorij, namesto na skupni ravni. Kot primer navajam število ogledov spletne strani po kategorijah, namesto skupnega števila ogledov spletne strani v zadnjem četrtletju uporabe.

Možnosti za nadaljnje delo so številne. Podjetje LEXPERA d. o. o. lahko po uporabi rezultatov na hrvaškem trgu pridobljeno znanje izkoristi na preostalih trgih oz. znotraj celotne skupine. Pred tem bi bilo možno izvesti analizo napovedi ter priporočil modelov v primerjavi z rezultati, ki bi jih dala uporaba na Hrvaškem. V bodoče bi se lahko analizirali

učinki pristopa podatkovnega rudarjenja glede na podobnosti in specifičnosti ostalih trgov, ki jih pokriva LEXPERA SA.

Zaradi kombinacije zrelosti algoritmov strojnega učenja, zadostne zmogljivosti in cenovne dostopnosti računalnikov ter programske opreme in ugodnih oblačnih storitev je pričakovati, da bo v bodoče umetna inteligenca, vključno s podatkovnim rudarjenjem, gonilna sila poslovnih odločitev. V nekaj prihodnjih letih si ni možno predstavljati podjetja, ki ne bo izkoriščalo potenciala v do sedaj zbranih podatkih.

LITERATURA IN VIRI

1. Aggarwal, C. C. (2014). *Data Classification: Algorithms and Applications*. Hoboken: Chapman and Hall/CRC.
2. Amani, F. A., & Fadlalla, A. M. (2017). Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems, 24*, 32–58.
3. Bole, U., Popovič, A., Žabkar, J., Papa, G., & Jaklič, J. (2015). A case analysis of embryonic data mining success. *International Journal of Information Management, 35*, 253–259.
4. Cao, L. (2009). *Data Mining for Business Applications*. New York: Springer.
5. Cao, L. (2017). Data Science: Challenges and Directions. *Communications of the ACM, 60*(8), 59–68.
6. Chorianopoulos, A. (2015). *Effective CRM Using Predictive Analytics*. Chichester, West Sussex: Wiley.
7. Coleman Shirley, Y. (2016). Data-Mining Opportunities for Small and Medium Enterprises with Official Statistics in the UK. *Journal of Official Statistics, 32*(4), 849–865.
8. Dwivedi, S., Kasliwal, P., & Soni, P. S. (2016). Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime). *Symposium on Colossal Data Analysis and Networking* (1–8). Indore: IEEE.
9. Farooqi, M. R., & Raza, K. (2011). A Comprehensive Study of CRM through Data Mining Techniques. *In Proceedings of the National Conference* (1–5). New Delhi: NCCIST.
10. Gartner. (2018). *Magic Quadrant for Data Science and Machine-Learning Platforms*. Pridobljeno 16. aprila 2018 iz https://www.gartner.com/doc/reprints?id=1-4RMUF5K&ct=180222&st=sb
11. Goncalves, J. P. M., Barros, R. S. M., & Vieira, D. C. L. (2012). On the Use of Data Mining Tools for Data Preparation in Classification Problems. *IEEE/ACIS 11th International Conference on Computer & Information Science* (173–178). Shanghai: IEEE.
12. Grublješič, T., & Jaklič, J. (2015). Business Intelligence Acceptance: The Prominence of Organizational Factors. *Information Systems Management, 32*(4), 299–315.
13. Gupta, A. K., & Gupta, C. (2012). Analyzing Customer Behavior using Data Mining Techniques: Optimizing Relationships with Customer. *Management Insight, 6*(1), 92–98.
14. Gurley, D. K., Peters, G. B., Collins, L., & Fifolt, M. (2015). Mission, vision, values, and goals: An exploration of key organizational statements and daily practice in schools. *Journal of Educational Change, 16*(2), 217–242.
15. Harrington, J. L. (2016). *Relational Database Design and Implementation* (4 izd.). Amsterdam: Morgan Kaufmann.
16. IBM. (2016). *IBM SPSS Modeler CRISP-DM Guide.* Pridobljeno 20. decembra 2017 iz https://www.ibm.com/support/knowledgecenter/en/SS3RA7_18.0.0/modeler_crispdm_ddita/modeler_crispdm_ddita-gentopic1.html
17. Informacijski pooblaščenec. (2016). *Reforma evropskega zakonodajnega okvira za varstvo osebnih podatkov.* Pridobljeno 21. maja 2018 iz https://www.ip-rs.si/zakonodaja/reforma-evropskega-zakonodajnega-okvira-za-varstvo-osebnih-podatkov/

18. IUS-INFO. (2009). *Postanite korisnik portala IUS INFO.* Pridobljeno 3. maja 2018 iz http://www.iusinfo.hr/Services/Registration.aspx?Id=0
19. Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science, 85*, 78–85.
20. Klinkenberg, R., & Hofmann, M. (2013). *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Boca Raton: Chapman and Hall/CRC.
21. Klisarova-Belcheva, S., Ilieva, G., & Yankova, T. (2017). Business Intelligence and Analytics – Contemporary System Model. *Trakia Journal of Sciences, 15*(1), 298–304.
22. Leung, C., MacKinnon, R., & Jiang, F. (2017). Finding efficiencies in frequent pattern mining from big uncertain data. *World Wide Web, 20*(3), 571.
23. LEXPERA. (2009a). *Osnovni podaci*. Pridobljeno 2. maja 2018 iz: http://www.lexpera.hr/Misc/Content.aspx?Id=OSCom
24. LEXPERA. (2009b). *Postanite korisnik portala IUS INFO!* Pridobljeno 2. maja 2018 iz: http://www.lexpera.hr/Services/Registration.aspx?Id=0
25. LEXPERA. (2016). *LEXPERA SA*. Pridobljeno 2. maja 2018 iz: http://www.lexpera.com/index-3.html
26. Michalewicz, Z. (2007). *Adaptive Business Intelligence*. Berlin: Springer.
27. Nisbet, R., Miner, G., & Elder, J. F. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. Amsterdam: Academic Press.
28. Paiano, R., & Pasanisi, S. (2017). A New Challenge for Information Mining. *Broad Research in Artificial Intelligence & Neuroscience, 8*(2), 63–80.
29. Pan, B., Zhang, G., & Qin, X. (2018). Design and realization of an ETL method in business intelligence project. *IEEE 3rd International Conference on Cloud Computing and Big Data Analysis* (275–279). Chengdu: IEEE.
30. Patel, M., Karvekar, S., & Mehta, Z. (2014). Customer Behavior Model Using Data Mining. *International Journal of Advanced Technology in Engineering and Science, 2*(1), 709–716.
31. Pentaho. (2015). *Pentaho Data Integration (Kettle) Tutorial.* Pridobljeno 4. januarja 2018 iz https://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+%28Kettle%29+Tutorial
32. RapidMiner. (2018a). *Auto Model*. Pridobljeno 9. maja 2018 iz https://docs.rapidminer.com/latest/studio/auto-model/
33. RapidMiner. (2018b). *Decision Tree*. Pridobljeno 8. junija 2018 iz https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_decision_tree.html
34. RapidMiner. (2018c). *k-Means*. Pridobljeno 8. junija 2018 iz https://docs.rapidminer.com/latest/studio/operators/modeling/segmentation/k_means.html
35. RapidMiner. (2018d). *Lift Chart (Simple)*. Pridobljeno 9. maja 2018 iz https://docs.rapidminer.com/latest/studio/operators/validation/visual/lift_chart.html
36. RapidMiner. (2018e). *Neural Net*. Pridobljeno 8. junija 2018 iz https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/neural_nets/neural_net.html
37. Rawat, S. S., & Rajamani, L. (2010). Discovering potential user browsing behaviors using custom-built apriori algorithm. *International journal of computer science & information Technology, 2*(4), 28–37.
38. Refaat, M. (2007). *Data Preparation for Data Mining Using SAS.* Amsterdam: Morgan Kaufmann.

39. Rogalewicz, M., & Sika, R. (2016). Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. *Management & Production Engineering Review, 7*(4), 97–108.
40. Saarenvirta, G. (1998). Data mining to improve profitability. *CMA Magazine, 72*(2), 1–8.
41. SAP. (2018). *The benefits of machine learning in business*. Pridobljeno 13. junija 2018 iz https://www.sap.com/products/leonardo/machine-learning/what-is-machine-learning.html
42. Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems, 31*, 127–137.
43. Sheikh, N. M. (2013). *Implementing Analytics: A Blueprint for Design, Development, and Adoption*. Burlington: Morgan Kaufmann.
44. Sullivan, R. (2012). *Introduction to Data Mining for the Life Sciences*. New York: Humana Press.
45. Statistični urad Republike Slovenije. (2008). *Kaj je SKD?* Pridobljeno 8. junija 2018 iz http://www.stat.si/SKD
46. Van Der Putten, P. (1999). *Data mining in direct marketing databases*. Singapore: World Scientific Publishers.
47. Verma, D., & Nashine, R. (2012). Data Mining: Next Generation Challenges and FutureDirections. *International Journal of Modeling and Optimization, 2*(5), 603-608.
48. Wang, J. (2006). *Encyclopedia of data warehousing and mining* (2 izd.). New York: Hershey.
49. Williams, S. (2016). *Business Intelligence Strategy and Big Data Analytics: A General Management Perspective*. Cambridge: Morgan Kaufmann.
50. Witten, I. H., Hall, M. A., & Frank, E. (2011). *Data mining: Practical Machine Learning Tools and Techniques*. Amsterdam: Morgan Kaufmann.
51. Zhao, Y., & Cen, Y. (2013). *Data Mining Applications with R*. Amsterdam: Academic Press.

PRILOGA

Priloga: Slovar uporabljenih izrazov

slovenski izraz	**angleški izraz**
mala in srednje velika podjetja	Small and Medium Enterprises
podatkovno rudarjenje	Data Mining
podatkovna analitika	Data Analytics
količina podatkov	Data Volume
raznolikost podatkovnih formatov	Data Variety
način procesiranja podatkov	Data Velocity
negotovost v podatkih	Data Veracity
strojno učenje	Machine Learning
kupec zapusti podjetje	Churn
poslovna inteligenca	Business Intelligence
poslovno-inteligenčni sistemi	Business Inteligence Systems
donosnost investicije	Return on Investment
podatkovna znanost	Data Science
opisna analitika	Descriptive Analytics
rezanje podatkov	Slice
predstavitev podatkov v obliki kocke	Dice
napovedna analitika	Predictive Analytics
predpisujoča analitika	Prescriptive Anaytics
spletno rudarjenje	Web Mining
strežniški dnevnik	Server Log
analiza nakupne košarice	Marker-basket Analysis
segmentacija	Clustering
osamelec	Outlier
mere nestanovitnosti	Volatility Measures
program kartice zvestobe	Loyalty-card Program
Splošna uredba EU o varstvu podatkov	General Data Protection Regulation
masovno trženje	Mass Marketing
direktno trženje	Direct Marketing
upravljanje odnosov s strankami	Customer Relationship Management
ohranjanje kupcev	Customer Retantion
navzkrižna prodaja	Cross-selling
prodaja dražje različice	Up-selling
čas od zadnjega nakupa stranke	Recency
število vseh nakupov stranke	Frequency
dosedanji izdatki stranke	Monetary
medsektorski proces podatkovnega rudarjenja	Cross-industry Standard Process for Data Mining
metodologija odkrivanja znanja v zbirkah podatkov	Knowledge Discovery in Database
metodologija vzorčenja, raziskovanja, spreminjanja, modeliranja in preverjanja	Sampling, Exploration, Modification, Model and Verification
razumevanje področja obravnave	Business Understanding
razumevanje podatkov	Data Understanding
priprava podatkov	Data Preparation
proces črpanja, pretvorbe in nalaganja podatkov	Data Extraction, Transormation, Loading
podatkovno skladišče	Data Warehouse
modeliranje podatkov	Modeling

ocena rezultatov modeliranja	Evaluation
uvedba rezultatov v prakso	Deployment
asociativna pravila	Association Rule Learning
podpora	Support
zaupanje	Confidence
globoko učenje	Deep Learning
nevronske mreže	Neural Networks
nevronska mreža z enosmernim tokom podatkov	Feed-forward Neural Network
nevronska mreža s povratnimi zankami podatkov	Recurrent Neural Network
tehnika najbližjega soseda	Nearest Neighbour
odločitvena drevesa	Decision Trees
klasifikacijska in regresijska drevesa	Classification and Regression Trees
drevesa, ki avtomatično zaznavajo interakcijo s pomočjo $\chi 2$ porazdelitve	Chi Square Automatic Interaction Detection
tekstovno rudarjenje	Text Mining
Gartnerjev magični kvadrant na področju podatkovne znanosti in platform strojnega učenja	Gartner Magic Quadrant For Data Science And Machine Learning Platforms
sistem za avtomatizacijo trženja	Marketing Automation System
glavni ključ	Primary Key
korelacija	Correlation
identifikacija	ID-ness
stabilnost	Stability
manjkajoči zapis	Missing
sprememba uteži pri nevronih	Learning Rate
zagon nevronske mreže	Momentum
grafikon dviga	Lift Chart
natančnost modela	Accuracy
občutljivost modela	Sensitivity
preciznosti modela	Precision
napačne klasifikacije modela	Classification Error
vhodni podatki (nevronske mreže)	Input
skrite plasti (nevronske mreže)	Hidden Layers
izhodne vrednosti (nevronske mreže)	Output
funkcija, ki prešteje vrednosti	Count
funkcija, ki sešteje vrednosti	Sum
funkcija, ki vrne povprečno vrednost	Average
funkcija, ki vrne minimalno vrednost	Min
funkcija, ki vrne maksimalno vrednost	Max