UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

MAGISTRSKO DELO

VPELJEVANJE PODATKOVNE ZNANOSTI NA PRIMERU RAZVOJA IN UPORABE NAPOVEDNEGA MODELA

Ljubljana, junij 2020 ŽIGA KLJUN

IZJAVA O AVTORSTVU

Podpisani Žiga Kljun, študent Ekonomske fakultete Univerze v Ljubljani, avtor predloženega dela z naslovom Vpeljevanje podatkovne znanosti na primeru razvoja in uporabe napovednega modela, pripravljenega v sodelovanju s svetovalcem red. 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, 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 vsa potrebna dovoljenja za uporabo podatkov in avtorskih del v predloženem delu in jih v njem jasno označil;

7. da sem pri pripravi predloženega dela ravnal v skladu z etičnimi načeli in, kjer je to potrebno, za raziskavo pridobil 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 29.6.2020 Podpis študenta:_____

KAZALO

KAZALO TABEL

KAZALO SLIK

SEZNAM KRATIC

angl. - angleško

AUC – (ang. Area Under the Curve); Območje pod krivuljo

CRISP – (ang. Cross-industry standard process); Od dejavnosti neodvisen standardiziran proces

IT – (ang. Information Technology); Informacijska tehnologija

GIGO – (ang. Garbage In Garbage Out); Slabi vhodni podatki, slabi izhodni podatki

KPI – (ang. Key Performance Indicator); Ključni kazalnik učinkovitosti in uspešnosti

XGBoost – (ang. eXtreme Gradient Boosting); ime metode strojnega učenja

UVOD

Podatkovna znanost je veda, ki ponuja nabor različnih metod za obdelavo in interpretacijo ogromnih količin podatkov, pridobljenih iz različnih virov in na različne načine (Waller & Fawcett, 2013). V zadnjih treh desetletij je veda doživela velik razcvet. Ker podatkovna znanost omogoča, da iz podatkov dobimo uporabne informacije, ki jih pred tem nismo imeli, je porast števila podatkov vplival na to, da je zanimanje za podatkovno znanosti čedalje večje tako v javnih kot v privatnih organizacijah. Dandanes zato čedalje več procesov odločanja postaja bolj podatkovno usmerjenih. Senzorji in procesna orodja zaradi odprtokodnega načina postajajo dostopna ne le velikim, pač pa tudi srednjim in malim podjetjem. Spopadanje z velikimi podatki tako postaja v teh časih za podjetja, ki želijo ohranjati oziroma ustvarjati konkurenčno prednost, neizbežno (Vicario & Coleman, 2019).

Ker se je podatkovna znanost tako hitro razvila, danes na trgu obstaja primanjkljaj podatkovnih znanstvenikov; trenutni zaposleni pogosto nimajo dovolj znanj in veščin, da bi prevzeli njihove naloge. Podjetja se pogosto zatekajo k pristopu, v katerem razvijejo metodološki pristop k podatkovni znanosti, ki vključuje vloge, postopke in dobre prakse načina dela (Blais, 2019). Ta omogoča poslovnim, podatkovnim in drugim analitikom ter podatkovnim razvijalcem, da strokovno pravilno in učinkovito delujejo pri procesu podatkovne znanosti znotraj svojih ekip. Te skupine potem delajo na svojih ločenih projektih.

Da bi oblikovali najboljši metodološki pristop k podatkovni znanosti za zgled ostalim uporabnikom, pa podjetje potrebuje skupino izkušenih podatkovnih znanstvenikov z različnim naborom veščin (Davenport & Patil, 2012). Danes se veliko poudarja, da morajo imeti podatkovni znanstveniki zelo širok nabor znanj in veščin. To je pogosto lahko past, ki predstavlja pomankanje specializacije, kar posledično prinese težave v prihodnosti (Little Miss Data, 2018). Na primer pomankanje poslovnih veščin lahko pomeni, da namenimo ure in celo dneve za reševanje določenega problema, ki se na koncu izkaže za nepomembnega. Seveda pa to ne pomeni, da moramo imeti zgolj specializirana znanja. S pomankanjem širine in statističnega razumevanja lahko prihaja do subjektivnih odločanj, ki prinašajo slabe poslovne rezultate (Jagadish, 2015). Podatkovni znanstveniki morajo tako imeti pravo mero kombinacije različnih veščin ter dobrih tehničnih veščin. Ker takih ljudi podjetja pogosto nimajo, se ta zatekajo k zunanjim izvajalcem, ki podjetju zagotovijo skupino podatkovnih znanstvenikov, ki jim pomagajo postaviti za njih primeren metodološki pristop k podatkovni znanosti, na podlagi katerega podjetje v nadaljevanju razvija svoje rešitve. Podatkovni znanstveniki metodološki pristop predstavijo tako, da v praksi na primeru dokažejo delovanje koncepta (angl. proof of contept) (Blais, 2019). Njihova naloga je tudi, da pomagajo podjetju pri učenju, kako postati bolj podatkovno usmerjena organizacija. Tukaj imajo majhna podjetja pogosto prednost, saj so bolj agilna in lažje spreminjajo svojo strukturo

in način vodenja (Felipe, 2019), medtem ko je v večjih podjetij ta proces veliko težji in je od tako od vodstva kot od podatkovnih znanstvenikov odvisno, kako dobro jim bo to uspelo.

V današnjem času se pogosto zgodi, da podjetja pod vplivom »modne besedne zveze« zanemarjajo pomen razumevanja problematike in pomen razumevanja statističnih metod, ki se v procesu strojnega učenja izvajajo. Zato se velikokrat zatekajo k rešitvam črnih zabojev (angl. black box), kjer programi v ozadju naredijo kalkulacije, ki uporabniku niso vidne in podajo odgovor. Rešitev ima sicer prednost v zelo zapletenih primerih, kjer je metoda tako zapletena, da jo človek težko razume, in je bolje, da postopka ne spreminja (Vicario & Coleman, 2019). V večini primerov pa dolgoročno črni zaboj zaradi pomankanja razumevanja in nizke stopnje robustnosti in prilagodljivosti prinaša slabše rezultate.

V procesu izvajanja podatkovne znanosti ima poslovno razumevanje torej veliko vlogo. Za podatkovne znanstvenike je pomembno, da sodelujejo s poslovnimi analitiki, saj ti, čeprav morda nimajo takšnega tehničnega znanja, s svojim poznavanjem vsebinskih izzivov ključno pripomorejo k razumevanju problema in izboljšanju končnih rezultatov (Foster, 2013). Podjetja se morajo zavedati, da podatkovna znanost zahteva ekipni pristop in ne predstavlja trivialnega postopka, ki bo dobil nabor podatkov in povečal uspešnost poslovanja. Podatkovna znanost podjetjem omogoči, da lažje razumejo njihovo občinstvo, na podlagi tega izboljšajo svoje izdelke oziroma storitve ter tako šele dosežejo boljše poslovne rezultate. Kot je že Einstein izjavil: »če bi mi dali eno uro, da rešim svet, bi porabil 59 minut za definiranje problema in 1 minuto za reševanje«.

Da bi podjetje doseglo dolgoročno uspešnost na področju podatkovne znanosti, mora to torej upoštevati številne vidike, na podlagi katerih nato zastavi metodološki pristop, ki je kasneje vodič pri izvajanju nadaljnjih projektov na področju podatkovne znanosti. Od definiranega metodološkega pristopa je odvisno, kako učinkovito bo podjetje izvajalo nadaljnje projekte podatkovne znanosti in kako uspešno bo lahko reševalo izzive, ki se jim bodo pri tem pojavljali (Silipo, 2019).

Namen magistrskega dela je na podlagi primera oblikovati predlog učinkovitega pristopa k vpeljevanju podatkovne znanosti v podjetje, ki je na tem področju novo in si želi postati bolj podatkovno usmerjeno.

Cilj magistrskega dela je analiza projekta vpeljave podatkovne znanosti v podjetje na primeru napovednega modela, s katero želimo identificirati potencialne težave in možne rešitve ter dejavnike, ki pozitivno vplivajo na uspešnost vpeljave podatkovne znanosti.

Teoretični del magistrskega dela obsega pregled obstoječe relevantne literature, ki je sestavljena iz dveh delov. Prvi vključuje področja podatkovne znanosti kot take in v povezavi s poslovno inteligenco in strojnim učenjem. V drugem delu sem se osredotočil na literaturo obstoječih metodoloških pristopov in praks, ki so aktualne v današnjem času.

Praktični del magistrskega dela vključuje analizo realnega primera vpeljevanja podatkovne znanosti v podjetje, ki vsebuje predstavitev in pregled vseh dejavnosti in izzivov znotraj projekta. V analizi sem sledil poteku projekta, vključno s predpripravo na projekt in predstavitvijo rezultatov naročnikom projekta. Na podlagi prej ugotovljenih teoretičnih ugotovitev sem primer tudi kritično ovrednotil. Pri tem sem upošteval, kako uspešno so bili doseženi zastavljeni cilji in povratne informacije, ki sem jih zbral tako v času projekta kot po njem.

Magistrsko delo je sestavljeno iz petih glavnih poglavij. V prvem poglavju sem se osredotočil na teoretični del podatkovne znanosti in njeno klasifikacijo, medtem ko sem v drugem poglavju predstavil teorijo metodološkega pristopa k podatkovni znanosti. V tretjem poglavju sem se posvetil praktičnemu pristopu, kjer sem predstavil metodološki pristop na realnem primeru in njegov projektni plan. V četrtem poglavju sledi predstavitev projekta in na koncu še diskusija, v kateri sem analiziral in kritično ovrednotil naš realen primer.

1 PODATKOVNA ZNANOST

1.1 Zgodovina

Čeprav je izraz podatkovna znanost relativno nov in se v veliki meri povezuje z računalništvom, področje primarno izvira iz statistike. Statistiki so se pomembnosti analize podatkov začeli zavedati že v šestdesetih letih 19. stoletja, vendar je do večjega skoka v popularnosti vede prišlo šele ob prelomu tisočletja v »pika-com« (angl. dot-com) dobi, ko so trdi diski postali precej cenejši, kot so bili do takrat (Dhar, 2012). Korporacije in vlade so to izkoristile in začele kupovati velike količine le-teh. Ena izmed posledic Parkinsonovega zakona je, da se podatki vedno širijo, da zapolnijo razpoložljivi prostor na disku. Interakcija »disk-podatki« je privedla do eksponentnega cikla med nakupom vedno večjega števila diskov in nabiranjem vedno večje količine podatkov. To je privedlo do izraza gmota podatkov (angl. big data), ki se uporablja za opisovanje tako velikih in zapletenih nizov podatkov, da je uporaba tradicionalnih orodij za podatkovne baze nad njimi prepočasna in neuporabna (Dhar, 2012).

V tem obdobju se je tudi bolj razširil izraz podatkovna znanost. Uradno je izraz prvič leta 2001 uporabil ameriški profesor računalništva in statistike na Univerzi v Purdue William S. Cleveland v knjigi z naslovom »Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics«. Izraz sta leto kasneje prevzela še Mednarodni svet za znanost (2002) in Univerza v Kolumbiji (2003), ko sta začela objavljati vsak svoj dnevnik podatkovne znanosti. Izraz podatkovna znanost danes velja za kombinacijo različnih orodij, algoritmov in načel strojnega učenja s ciljem odkriti skrite vzorce iz surovih podatkov.

S porastom količine podatkov so se podjetja začela zavedati, da morajo z velikimi

podatki nekaj narediti, saj jim samo shranjevanje ne prinaša velikih prednosti. Za to pa je potrebno veliko računalniške arhitekture, kar ne pomeni zgolj več fizičnih diskov za shranjevanje podatkov, pač pa tudi zmogljivejše procesne enote, ki bodo skrbele za obdelavo le-teh. Podjetja, kot so Google, Yahoo, in Amazon, so izumila novo računalniško arhitekturo, ki ji danes rečemo računalništvo v oblaku. Ta omogoča uporabnikom rabo računalniških virov na zahtevo, ne da bi za to potrebovali svojo lastno fizično opremo. Eno najpomembnejših odkritij na področju analize podatkov je tudi MapReduce, ki je kodificiran v programsko opremo, znano kot Hadoop (Dean & Ghemawat, 2010). MapReduce omogoča sočasno obdelavo velikih količin podatkov z delitvijo na manjše kose in vzporedno obdelavo na Hadoop strežnikih. Na koncu združi vse podatke iz več strežnikov in tako vrne konsolidiran izhod nazaj v aplikacijo.

Izkazalo se je sicer, da Hadoop ni najbolj enostaven za delo, saj zahteva napredne računalniške kapacitete. To je odprlo trg za ustvarjanje analitičnih orodij s preprostejšimi vmesniki, ki delujejo nad Hadoopom. Ta razred orodij se imenuje »orodja za masovno analitiko« (angl. Mass Analytics Tools) – to so orodja za analizo velike mase podatkov. Primeri za to so priporočilni sistemi, strojno učenje in kompleksna obdelava dogodkov (angl. complex event proccessing).

S pojavom orodij za analitiko in obdelavo gmot podatkov (angl. big data) je nastala potreba po ljudeh, ki razumejo orodja in z njimi opravijo analizo teh velikih gmot podatkov (Gutierrez, 2018). Tem ljudem pravimo podatkovni znanstveniki (angl. Data Scientist). Ti ljudje so sposobni izslediti nova analitična spoznanja, ki jih prej v svetu majhnih podatkov ni bilo mogoče izslediti (Vicario & Coleman, 2019). Obseg težav, ki jih rešujemo z analizo velikih podatkov, je pogosto takšen, da en posameznik ne more opraviti vseh potrebnih procesov in analiz, zato se podatkovna znanost najbolje izvaja v skupinah. Podjetja so se okoli leta 2010 začela zavedati o možnostih, ki jim jih omogočajo podatkovni znanstveniki in zanimanje za podatkovno znanost je močno naraslo in raste še danes.

1.2 Področja podatkovne znanosti

Ne obstaja samo ena delitev podatkovne znanosti. Podatkovno znanost lahko delimo na več načinov in različni avtorji so delitev opredelili različno. Mi bomo naslonili na delitev Ucrosove (2018) in razdelili podatkovno znanost na 8 glavnih disciplin. Gre za zelo širok pogled na podatkovno znanost, ki v okviru svojih področij vključuje tudi dele nekaterih drugih disciplin (Ucros, 2018).

1. **Podatkovni inženiring in podatkovno skladiščenje**: navezuje na pretvorbo podatkov v obliko, primerno za kasnejšo analizo. Proces vključuje upravljanje vira, strukturiranje, preverjanje kakovosti in shranjevanje podatkov, ki se kasneje uporabijo pri analizi.

2. **Podatkovno rudarjenje in statistična analiza**: navezuje se na uporabo statistike v

raziskovalnih analizah in napovednih modelih za odkrivanje vzorcev in podatkovnih trendov v obstoječem naboru podatkov. Namen te kategorije je razumeti poslovni problem, ga prenesti na podatke in na podlagi teh zgraditi poslovni model.

3. **Sistemska arhitektura**: to področje se ukvarja z načrtovanjem in implementacijo sistemske arhitekture, ki bo omogočala varno in zanesljivo delovanje ostalih procesov podatkovne znanosti. Lahko nudi tako rešitev v oblaku kot programje na mestu uporabe.

4. **Upravljanje s podatkovnimi bazami**: področje je zadolženo za načrtovanje, postavitev in vzdrževanje podatkovnih baz, ki podpirajo velike količine kompleksnih podatkovnih transakcij, namenjenih specifično določenim skupinam in storitvam

5. **Poslovna inteligenca**: ukvarja se z izdelovanjem prilagojenih analitičnih rešitev, upravljanjem nadzornih plošč, poročanjem sponzorjem in izboljševanjem procesa poročanja in analize na splošno (Su & Chiong, 2010).

6. **Strojno učenje**: gre za najbolj modno področje v podatkovni znanosti in predstavlja bolj kompleksno obliko podatkovnega rudarjenja. Vključuje analizo in pripravo podatkov, na katerih se nato z uporabo statističnega algoritma zgradi model, ki se na podlagi obstoječih podatkov uči in lahko na podlagi novega znanja napoveduje prihodnje rezultate oziroma nam pomaga pri odločanju (Patel in drugi, 2010). Čeprav za mnogo ljudi strojno učenje predstavlja jedro podatkovne znanosti, pa je pomembno, da se ne zanemari drugih vej, ki so prav tako pomembni deli te vede.

7. **Vizualizacija in predstavitev podatkov**: pomaga nam, da z uporabo vizualizacije, lažje razumemo sicer suhoparne podatke. Veda izhaja iz dela poslovne inteligence.

8. **Podatkovni analitik na specifičnem področju**: gre za področje podatkovne znanosti, kjer ima tehnično znanje nekoliko manjši pomen; osredotoča se na razumevanje in poznavanje specifičnega vsebinskega področja, na katerega se navezuje projekt podatkovne znanosti.

1.3 Podatkovna znanost in poslovna inteligenca

Čeprav podatkovna inteligenca danes predstavlja področje podatkovne znanosti, je veda tako obsežna, da jo lahko obravnavamo tudi samostojno. V zgodovini podatkovne analitike se je s preoblikovanjem podatkov v smiselne in koristne informacije primarno ukvarjala poslovna inteligenca (angl. business intelligence) (Su & Chiong, 2010). Šele z naraščanjem števila podatkov je prišla velika potreba po novi disciplini – podatkovni znanosti. Čeprav se poslovna inteligenca in podatkovna znanost v marsičem zelo prepletata, je razlikovanje obeh dveh ključnega pomena, da lahko izkoristimo potenciala obeh disciplin.

Da bi razumeli razlike, je najprej dobro, da poznamo definicije obeh disciplin.

Podatkovna znanost temelji na »podatkovnem pristopu« (angl. data-driven), kjer se skupaj uporabljajo številne interdisciplinarne znanosti za pridobivanje pomena in izvlečkov iz razpoložljivih poslovnih podatkov, ki so običajno zelo obsežni in kompleksni. Poslovna inteligenca na drugi strani, se osredotoča na spremljanje trenutnega stanja poslovnih podatkov z namenom razumevanja uspešnosti podjetja v preteklosti (Larson, 2019). V grobem lahko povzamemo, da se poslovna inteligenca ukvarja primarno z razlago preteklih podatkov in je osredotočena na opisno analitiko, medtem ko podatkovna znanost analizira pretekle podatke (trende ali vzorce), na podlagi katerih nato napoveduje prihodnost in je osredotočena na priporočilno in napovedno analitiko.

1.3.1 Skupne lastnosti

Tako podatkovna znanost kot poslovna inteligenca se osredotočata na podatke, s katerimi bi podjetju zagotovili večji dobiček, zadržali stranke, osvojili nov trg ipd. Končni uporabniki so običajno vodstveni delavci in menedžerji, ki pa pogosto nimajo niti časa, niti namena, da bi podrobno poznali tehnično ozadje podatkovne analize in si želijo zgolj orodja, ki jim bodo hitro in natančno pomagala pri njihovem odločanju in sprejemanju kritičnih potez v omejenem času, ki ga imajo na voljo (Larson, 2019). Zato so tako v primeru podatkovne znanosti kot poslovne inteligence za to zadolženi posebej usposobljeni strokovnjaki na izbranem področju, ki obogatijo in interpretirajo podatke ter rezultate pretvorijo v človeku prijazno in razumljivo obliko.

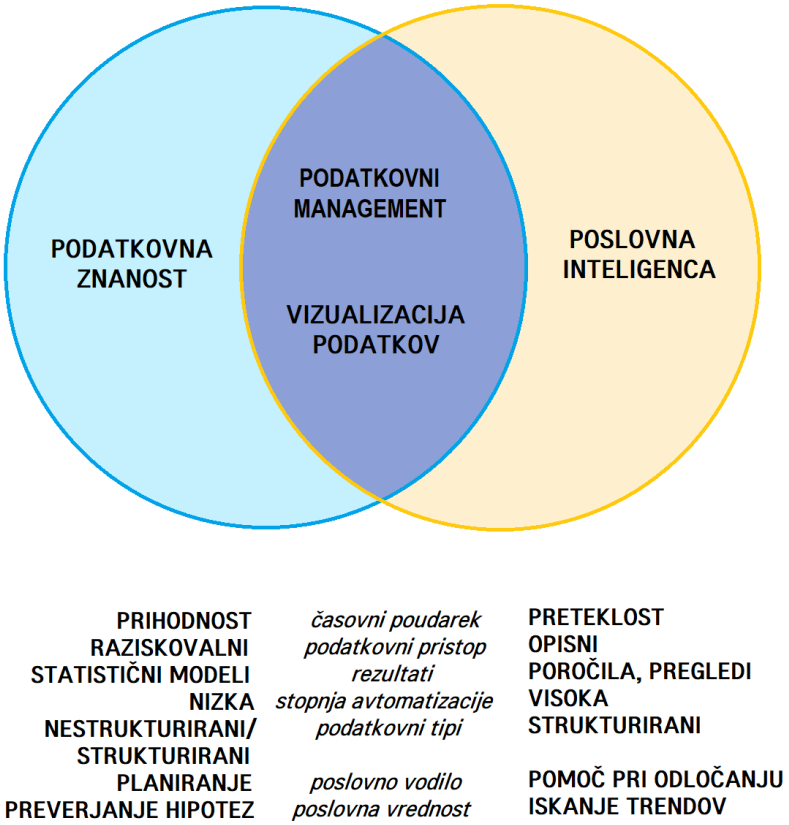
1.3.2 Razlike

Tako podatkovna znanost in poslovna inteligenca ponujata zanesljive sisteme za podporo odločanju vodstvenim delavcem, menedžerjem in vodjem oddelkov, ki so sicer strokovnjaki na svojih področju dela, vendar pa vseeno pričakujejo zanesljivo pomoč in podporo strokovnjakov s področja podatkovne analize pri sprejemanju odločitev, ki temeljijo na podatkih. Glavna razlika med podatkovno znanostjo in poslovno inteligenco se pojavi v tem, da je poslovna inteligenca namenjena bolj ravnanju s statističnimi in visoko strukturiranimi podatki, medtem ko podatkovna znanost hitreje operira z večjo količino in kompleksnostjo nestrukturiranih podatkov iz različnih virov (Jagadish, 2015). Medtem ko poslovna inteligenca lahko razume in obdeluje podatke, preoblikovane v vnaprej točno določene strukture, lahko napredne tehnologije podatkovne znanosti, med katere spadajo Big Data, internet stvari (angl. internet of things) in oblačne storitve, skupaj zbirajo, prečiščujejo, pripravljajo analizirajo ter poročajo z različnimi vrstami in oblikami podatkov, zbranimi iz različnih virov.

Na to temo je Analytics India Magazine leta 2017 objavil članek z naslovom »Business Intelligence vs Data Science« (Deoras, 2017), kjer avtorica nazorno predstavi stanje pred leti, ko bila oseba, zadolžena za delo s podatki, znana kot podatkovni analitik, osredotočena predvsem na področje poslovne inteligence. Podjetja so se sčasoma začela

oddaljevati zgolj od poročanja o preteklih rezultatih in se usmerila bolj v napovedovanje prihajajočih trendov in ponujanje ustreznih rešitev za uspeh v prihodnje. Tukaj se je izvedel velik preskok iz poslovne inteligence v podatkovno znanost. Primerjava podatkovne znanosti in poslovne inteligence je tudi nazorno prikazana na spodnji sliki.

Slika 1: Primerjava podatkovne znanosti in poslovne inteligence



Vir: lastno delo.

Danes podjetja vse bolj poudarjajo podatkovno znanost, saj ta iz leta v leto generira več podatkov, in vse kaže na to, da bo bolj uspešen tisti, ki bo to gmoto podatkov iz različnih virov znal najbolje izkoristiti. Podatkovna znanost tudi nakazuje, da bo v prihodnosti avtomatizirala večino nalog podatkovne analitike ter poslovne inteligence, medtem ko bodo poslovni uporabniki do podatkov v podatkovnih skladiščih lahko dostopali s pomočjo posebej razvitih orodij, ki bodo nudila enostavne grafične vmesnike.

Še ena pomembna razlika med podatkovno znanostjo in poslovno inteligenco je ta, da poslovna inteligenca, čeprav ima zelo pomembno vlogo pri poslovnem odločanju, še vedno pogosteje opravlja dejavnost v oddelkih informacijske tehnologije (v nadaljevanju IT), medtem ko gre podatkovna znanost še korak dlje in se bolj približa bolj poslovnim uporabnikom, ki nimajo tako močne tehnične podlage, imajo pa toliko močnejšo metodološko in poslovno podlago (Su & Chiong, 2010). Čeprav gre pri podatkovnih

znanstvenikih na nek način za napredno, bolj holistično obliko analitika poslovne inteligence, pa lahko ti med seboj še vedno dobro sodelujejo, in sicer v tej obliki, da analitiki poslovne inteligence pripravijo podatke, saj imajo dobro razumevanje in poznavanje analitičnih zahtev poslovanja in lahko tako podatkovnim znanstvenikom pomagajo pri gradnji modelov, s katerimi bodo napovedovali prihodnje trende in vzorce.

1.4 Strojno učenje

Izraz strojno učenje (angl. machine learning) je leta 1959 »skoval« ameriški pionir na področju računalniških iger in umetne inteligence, Arthur Samuel in ga definiral kot način, ki računalnikom omogoča, da se učijo, ne da bi bili za to specifično sprogramirani. Leta 1997 je Tom Mitchell Samuelove besede pretvoril še v dobro definirano matematično-relacijsko definicijo, ki pravi, da se računalnik uči iz izkušenj I z upoštevanjem določenega nabora nalog N in meritev uspešnosti U, če njegovo uspešnost nalog N, merjena z U, izboljša z izkušnjami I (Jordan & Mitchell, 2015). Strojno učenje je torej mogoče razložiti kot avtomatizacijo in izboljšanje učnega procesa računalnikov na podlagi njihovih izkušenj, ne da bi bili dejansko programirani, tj. brez kakršne koli človekove pomoči (Li, 2012).

Slika 2: Primerjava matematike, računalništva in področja dela

Vir: lastno delo.

Strojno učenje predstavlja del podatkovne znanosti. Medtem ko je strojno učenje bolj specifične narave, pa je podatkovna znanost precej širši pojem, ki poleg strojnega učenja vključuje še več različnih disciplin, med katere spadajo analiza podatkov, programski inženiring, napovedna analitika, analitika podatkov in druge (Maxwell, Warner & Fang, 2018). Proces se začne s hranjenjem kakovostnih podatkov in nato usposabljanjem naših strojev (računalnikov) z gradnjo modelov strojnega učenja z uporabo podatkov in različnih algoritmov. Izbira algoritmov je odvisna od tega, kakšne vrste podatkov imamo in kakšno nalogo poskušamo avtomatizirati.

Tudi pri strojnem učenju poznamo več različnih področij, na katera se veda lahko deli. Mi bomo uporabili Marslandovo delitev (2015), ki strojno učenje razdeli na štiri glavne tipe učenja:

- **Nadzorovano učenje (angl. supervised learning)**: Algoritem se uči iz primerljivih podatkov in s tem povezanih ciljnih odzivov, ki so lahko sestavljeni iz številskih vrednosti ali nizov lastnosti, kot so razredi ali oznake, da bi kasneje predvidel pravilen odziv na podlagi novih primerov. Ta pristop je podoben učenju ljudi pod nadzorom učitelja. Učitelj ponuja dobre primere, da si jih učenec zapomni, učenec pa nato iz teh posebnih primerov izpelje splošna pravila.

- **Nenadzorovano učenje (angl. unsupervised learning)**: Algoritem se uči iz preprostih primerov, ki niso povezani z odzivom, in prepusti algoritmu, da sam določi vzorce podatkov. Ta vrsta algoritma teži k prestrukturiranju podatkov v nekaj drugega, na primer nove lastnosti, ki lahko predstavljajo nov razred ali vrsto nepovezanih vrednosti. Koristijo nam predvsem zato, ker nam dajejo dober vpogled v pomen podatkov in hkrati dobre vhodne podatke za nadaljnje nadzorovane algoritme strojnega učenja.

  V človeškem svetu jih lahko primerjamo z metodami, ki jih ljudje uporabljamo, da ugotovimo, da so določeni predmeti ali dogodki iz istega razreda ali ne. Pri tem na primer uporabljamo opazovanje stopnje podobnosti med predmeti. Zato tudi številni spletni priporočilni sistemi temeljijo na algoritmih nenadzorovanega učenja.

- **Vzpodbujevalno učenje (angl. reinforcement learning)**: Algoritem dobi primere, ki ne vsebujejo oznak tako kot pri nenadzorovanem učenju. Dobi pa algoritem pozitivno oziroma negativno povratno informacijo glede na rešitev, ki jo ta predlaga. Taka oblika učenja se najpogosteje uporablja v odločitvenih sistemih, ko ti ne le opisujejo situacije, pač pa podajajo tudi predloge za prihodnost in imajo določene posledice. V človeškem svetu lahko to primerjamo z načinom učenja s poskušanjem in učenjem na napakah. Dejanja, ki nam prinesejo neke vrste kazen (stroški, izguba časa, bolečina itd.), na primer pripomorejo k temu, da takega dejanja ne bomo več ponovili.

  Računalnik se s pomočjo napak oziroma negativnih povratnih informacij uči, katera dejanja so dobra in katera ne ter posledično kateri postopki so najbolj uspešni.

Nazoren primer uspešnega spodbujevalnega učenja je viden pri računalnikih, ki se sami učijo igrati računalniške igrice. Najbolj znan primer je, ko je podjetje Google DeepMind izdelalo program, ki se je naučil igrati igro Atari. Računalnik namreč na koncu ni bil zgolj tehnično uspešen, pač pa je našel tudi strategijo, kako najhitreje zmagati.

- **Evolucijsko učenje (angl. evolutionary learning)**: Gre za skupino računskih modelov, ki se zgledujejo po konceptu evolucije. Algoritmi uporabljajo metodo naključnega iskanja z namenom, da bi našli rešitev (Aguilar-Ruiz et al., 2003). Pri evolucijskem računanju se ustvari začetni niz kandidatnih rešitev, ki se nato iterativno posodablja. Vsaka nova generacija se ustvari s stohastičnim odstranjevanjem manj želenih rešitev in uvajanjem majhnih naključnih sprememb.

  Ti tipi se kasneje delijo še na različne algoritme in pristope (Rose, 1998), med katerimi pa se najpogosteje uporabljajo naslednji:

- **Razvrščanje (angl. classification)**: Cilj je določiti, kateremu od nabora razredov (kategorij oziroma podpopulacij) pripada nov primer na podlagi predhodnega učenja s podatki, ki vsebujejo opažanja (ali primere), katerih pripadnost kategoriji je znana. Razvrščanje spada v kategorijo nadzorovanega učenja. Primer razvrščanja je filtriranje neželene pošte, kjer so vnosi e-poštna (ali druga) sporočila in razredi »neželena pošta« in »neželena pošta«.

- **Regresija (angl. regression)**: Algoritem, opravlja regresijsko nalogo. Regresija modelira ciljno vrednost napovedi, ki temelji na neodvisnih spremenljivkah. Večinoma se uporablja za ugotovitev razmerja med spremenljivkami in napovedno vrednostjo. Različni regresijski modeli se razlikujejo glede na vrsto razmerja med odvisnimi in neodvisnimi spremenljivkami, ki jih upoštevajo, in številom neodvisnih spremenljivk, ki se uporabljajo. Regresija – tako kot razvrščanje – spada v kategorijo nadzorovanega učenja.

- **Gručenje ali razvrščanje v skupine (angl. clustering)**: Cilj je združiti niz predmetov tako, da so predmeti v isti gruči (grozd) med seboj (v nekem smislu) bolj podobni kot tistim v drugih gručah (grozdih). Gručenje spada v kategorijo nenadzorovanega učenja. Pogosto se uporablja pri prepoznavanju vzorcev, analizi slike, iskanju informacij, bioinformatiki, stiskanju podatkov in računalniški grafiki (Arthur, 2007).

2 METODOLOŠKI PRISTOPI K PODATKOVNI ZNANOSTI

2.1 Motivacija

Ljudje, ki delujejo na področju podatkovne znanosti, vsakodnevno rešujejo težave in se ukvarjajo z vprašanji, povezanimi z analizo podatkov. Na podlagi tega gradijo modele za

napovedovanje rezultatov oziroma odkrivanje skritih vzorcev z namenom, da bi izboljšali prihodnje poslovne rezultate. V današnjem času se orodja in tehnologije, ki se uporabljajo pri analizi podatkov, zelo hitro razvijajo in s tem povečujejo sposobnosti podatkovnih znanstvenikov.

Kljub hitremu napredku orodij in tehnologij še vedno obstajajo ovire, ki zavirajo rast uspešnosti podatkovne znanosti v projektih. Pogosto še vedno ni jasno, na kateri točki naj se podjetje odloči za implementacijo le-te. Velikokrat je ta namreč primerna kasneje, kot si vodilni u podjetjih predstavljajo. Študija Beana in Davenporta (2019) je pokazala, da vpeljevanje podatkovne znanosti v podjetja predstavlja zelo velik problem, čeprav podjetja v to veliko vlagajo. Ugotovila sta, da je kar 69 % podjetij, ki so si prizadevala, da bi postala podatkovno usmerjena, ostalo neuspešnih. 53 % ni uspelo pridobiti resnih konkurenčnih prednosti iz svojih podatkov. Pogosto se zgodi, da podjetja vlagajo v podatkovno znanost namreč zgolj zato, da bi zagotovili želje svojim vlagateljem, kar pa v večini primerov ne prinaša dobrih rezultatov. Zanimiv je tudi podatek, da se čedalje manj podjetij vidi kot podatkovno usmerjenih. Leta 2017 je bilo takih 37 %, leta 2018 32,4 in leta 2019 samo 31 %. Podjetja se danes čedalje bolj zavedajo, kaj v resnici pomeni biti podatkovno usmerjen in da sama to niso.

V prvi vrsti podjetje potrebuje primerne probleme, ki bi jih s podatkovno znanostjo reševalo in predvsem potrebuje kakovostne podatke (Duggan, 2019). Pogosto namreč ljudje od podatkovnih znanstvenikov pričakujejo revolucionarna odkritja v poslovnem svetu, vendar njihovi rezultati kasneje ne prinesejo želenih izboljšav, saj podatki, ki so jim na voljo, niso dovolj kakovostni. Pri podatkovni znanosti je pomen kakovostnih podatkov, zato se je tudi razvila kratica GIGO (angl. Garbage In, Garbage Out), ker pomeni slab vhod, slab izhod.

Preden se podjetja odločijo za uvedbo podatkovne znanosti, morajo torej zagotoviti, da imajo pripravljeno podatkovno skladišče, ki je dobro dokumentirano, dobro strukturirano, prečiščeno in enostavno dostopno (Duggan, 2019). V naslednji fazi si mora podjetje zagotoviti ustrezne kadre, ki bodo zadolženi za podatkovno znanost. Če teh v podjetju nima, jih mora zaposliti. Čeprav podjetje zaposli podatkovne znanstvenike oziroma ima ljudi, ki so se pripravljeni s tem ukvarjati, pa se pogosto vseeno zgodi, da ti in podjetniki, katerih cilj je rešiti določen problem, nimajo dovolj znanja in razumevanja, kako pristopiti in uporabiti podatkovne tehnike, ki so na voljo. Običajno je težava v nerazumevanju problema in posledično tudi napačna uporaba metodologije.

Podobno kot pri tradicionalnih znanstvenikih tudi pri podatkovnih znanstvenikih ustrezna metodologija igra pomembno vlogo. Metodologija je splošna strategija, ki vodi procese in dejavnosti znotraj določene domene. Metodologija ni odvisna od rabe tehnologij ali orodij, niti ni ne predstavlja skupka različnih tehnik. Metodologija zagotavlja podatkovnim znanstvenikom ogrodje, ki se ne glede na metode, procese in hevristiko uporablja za pridobivanje rezultatov na področju podatkovne znanosti.

Podjetja se pogosto zatekajo k pristopu, v katerem razvijejo metodološki pristop k podatkovni znanosti, ki vključuje vloge, postopke in dobre prakse načina dela (Blais, 2019). Ta omogoča poslovnim, podatkovnim in drugim analitikom ter podatkovnim razvijalcem, da strokovno pravilno in učinkovito delujejo pri procesu podatkovne znanosti znotraj svojih ekip. Te skupine potem delajo na svojih ločenih projektih.

Da bi oblikovali najboljši metodološki pristop k podatkovni znanosti za zgled ostalim uporabnikom, pa podjetje potrebuje skupino izkušenih podatkovnih znanstvenikov z različnim naborom veščin (Davenport & Patil, 2012). Danes se veliko poudarja, da morajo imeti podatkovni znanstveniki zelo širok nabor znanj in veščin. To je pogosto lahko past, ki predstavlja pomankanje specializacije, kar posledično prinese težave v prihodnosti (Little Miss Data, 2018). Pomankanje poslovnih veščin lahko na primer pomeni, da namenimo ure in celo dneve za reševanje določenega problema, ki se na koncu izkaže za nepomembnega. Seveda to ne pomeni, da moramo imeti zgolj specializirana znanja. S pomankanjem širine in statističnega razumevanja lahko prihaja do subjektivnih odločanj, ki prinašajo slabe poslovne rezultate (Jagadish, 2015). Podatkovni znanstveniki morajo tako imeti pravo mero kombinacije različnih veščin ter dobrih tehničnih veščin. Ker takih ljudi podjetja pogosto nimajo in jih je na trgu težko najti, se ta zatekajo k zunanjim izvajalcem, ki podjetju zagotovijo skupino podatkovnih znanstvenikov, ki jim pomagajo postaviti za njih primeren metodološki pristop k podatkovni znanosti, na podlagi katerega podjetje v nadaljevanju razvija svoje rešitve. Podatkovni znanstveniki metodološki pristop predstavijo tako, da v praksi na primeru dokažejo delovanje koncepta (angl. proof of contept) (Blais, 2019). Ker gre za začetni projekt, na katerem bo podjetje gradilo v prihodnje, je zanj zelo pomembno, da je ta dobro razložljiv in se ga da jasno interpretirati. Podjetja so pogosto neučakana in želijo takoj doseči maksimalno učinkovitost, vendar se morajo v prvem projektu bolj osredotočiti na razumevanje tako metodološkega kot vsebinskega pristopa k problemu. Z nadaljnjimi projekti nato kasneje gradijo še povečanje učinkovitosti. Naloga podatkovnih znanstvenikov, ki gradijo metodološki pristop, je tudi, da pomagajo podjetju pri učenju, kako postati bolj podatkovno usmerjena organizacija. Pri tem jim lahko predstavljajo obstoječe primere uporabe podatkovne znanosti v drugih podjetjih in panogah.

Na to, kakšno skupino ljudi bo podjetje izbralo, pogosto vpliva že velikost podjetja. Majhna podjetja, ki so pogosto bolj prilagodljiva, imajo običajno omejena sredstva; velikokrat uporabljajo pristop, v katerem celotno ekipo podatkovnih znanstvenikov predstavlja zgolj ena oseba ali pa v skrajnem primeru dve, ki opravljata celoten proces podatkovne znanosti. En je osredotočen bolj na naloge podatkovnega inženirja, drugi pa na naloge podatkovnega znanstvenika. V srednje velikih podjetjih, kjer je sredstev nekoliko več, se tema dvema lahko pridruži še programski inženir, ki skrbi za zbiranje podatkov. Tako se lahko druga dva bolj usmerita in specializirata za analizo podatkov in gradnjo modelov (Mangini, 2019). V velikih podjetjih, kjer so finančne omejitve relativno majhne, si lahko podjetje zagotovi bolj točno določene vloge in strokovnjaka za vsak korak podatkovne znanosti posebej oziroma celo več zaposlenih za eno

posamezno fazo, kar podjetju omogoči, da izkoristi celoten potencial podatkovne znanosti.

2.2 Definiranje nalog in ciljev

Da bi ustvarili ustrezno metodologijo za nadaljnje projekte, je že na začetku potrebno definirati določene stvari, ki imajo vpliv na ustvarjeno metodologijo in posledično uspešnost podjetja pri prihodnjih projektih podatkovne znanosti. Postaviti je potrebno jasne in prilagodljive cilje, ki zadevajo poslovno strategijo podjetja (Bendheim, 1998). Dobro morajo biti določene vloge sponzorjev in vodij projekta ter njihove odgovornosti. Zagotoviti je potrebno tudi primerno tehnološko infrastrukturo, ki bo zagotavljala učinkovito analizo in pripravo podatkov ter gradnjo in postavitev modelov (Data Talent, 2019). Določiti pa je potrebno tudi vloge znotraj skupine podatkovnih znanstvenikov, ki bodo pri projektu sodelovali, določiti mesto, kjer bodo sedeli, komunicirali itd. Gre za stvar, ki na prvi pogled ni tako pomembna, vendar se v praksi pogosto izkaže, da je zagotovitev najlažje komunikacije in skupinskega dela ključno vpliva na učinkovitost projekta za vpeljevanja podatkovne znanosti.

Uspešna strategija za dosego optimalne metodologije podatkovne znanosti se začne z razumevanjem obstoječih poslovnih težav. Potrebno je dobro razumevanje vzročne zveze med tem, kar podjetja iščejo kot rezultate, in pravilnim načinom merjenja oziroma pridobivanja podatkov. Pred vsakim projektom podatkovne znanosti mora biti jasno, kako bo ta prispeval k temeljnim poslovnim ciljem in kako bo uspeh dosege ciljev opredeljen. Podjetja morajo začetni projekt podatkovne znanosti jemati kot začetek dolgotrajnega procesa, skozi katerega bo podjetje postalo podatkovno usmerjeno. Na podlagi študije Accenture (2019) je recimo kar 62 % podjetij odgovorilo, da razume odločanje na podlagi podatkov bolj učinkovito od standardnega načina, medtem ko zgolj 25 % teh podjetij svoje podatke pri odločanju dejansko uporablja. Celotna reorganizacija podjetja, da bi bilo bolj podatkovno usmerjeno, je namreč dolg in obsežen proces. Da bi proces uspešno izvedli, se ga je potrebno lotiti postopoma. Pogosto je dober način, da se najprej osredotočimo na en oddelek in v primeru uspeha podatkovno znanost razširimo tudi na druge. Ljudje so načeloma tudi bolj nagnjeni k spremembam, če predhodno vidijo, da je taka sprememba nekje že prinesla dobre rezultate.

V okviru celotnega projekta se je potrebno zavedati, da uspešnost projekta ni odvisna zgolj od skupine podatkovnih znanstvenikov. V ozadju ima pomembno vlogo tudi vodja projekta, ki je v malih podjetjih običajno vodja oddelka, v velikih pa vodja oddelka za podatke CDO. Ena glavnih nalog te vloge je spodbujati kulturo spodbujanja podatkovne znanosti in ljudem pripisovati odgovornost za rezultate. Ta oseba bo odgovorna tudi za določitev smernic v zvezi z upravljanjem in etiko (vprašanja zasebnosti, varnost podatkov). Pomembno vlogo ima tudi sponzor projekta. Njegova naloga je, da širi in spodbuja sprejemanje strategije podatkov (Data Talent, 2019). Dopolnjuje vodilno vlogo v smislu, da z zunanje perspektive ocenjuje prednosti, ki jih odločanje na podlagi

podatkov prinaša. Običajno je nekdo z veliko izkušnjami v podjetju, lahko tudi ista oseba kot vodja oddelka za podatke.

Pri opredelitvi odgovornosti je tudi pomembno, da se vključi strategija, kako bodo analitične dejavnosti postavljene znotraj podjetja. Projekt je namreč lahko centraliziran ali pa razporejen po različnih oddelkih in poslovnih enotah in celo zunanjih izvajalcih, kar lahko bistveno vpliva na njegovo dinamiko in na zadovoljstvo tako podatkovnih znanstvenikov kot naročnikov projekta. Tako kot na številnih drugih področjih tudi tukaj ni ene unikatne najboljše metode, saj je pristop odvisen lastnosti, ki so od podjetja do podjetja različne.

2.3 Tehnološko ogrodje in infrastruktura

Tehnološko ogrodje in infrastruktura je pogosto stvar, ki jo naročniki projektov zanemarjajo, čeprav igra veliko vlogo. Modeli podatkovne znanosti so običajno zelo iterativni, kar pomeni, da morajo imeti podatkovni znanstveniki na voljo prožno tehnološko ogrodje, ki je odporno tudi proti številnim napakam. Neustrezno tehnološko ogrodje lahko delo podatkovnih znanstvenikov drastično upočasni, kar pomeni slabo izkoriščenost časa in znanja ter posledično denarja za podjetje. Za podjetja, ki obdelujejo zelo velike količine podatkov, je potrebna zelo stroga organizacija, saj bo veliko podatkovnih projektov neposredno preizkušeno v proizvodnem okolju. Stroški vzdrževanja razvojnega okolja, ko se poveča število podatkov, postanejo namreč previsoki. Okolja morajo tako podpirati sočasne procese, pri čemer uporabniška izkušnja ne sme čutiti posledic.

Da bi lahko tehnološko ogrodje dobro delovalo in zagotavljalo vse želene funkcije, pa mora imeti podjetje pripravljeno tudi ustrezno infrastrukturo, ki bo to okolje podpirala. Že na začetku je potrebno imeti v mislih, s kakšno količino podatkov se bo delalo v proizvodnem okolju in temu primerno prilagoditi infrastrukturo. Pogosto se zgodi napaka, ko podjetja ne zagotovijo dovolj zmogljive infrastrukture in to ugotovijo šele, ko v proizvodnem okolju že tečejo analize podatkov, saj na razvojnem okolju količine podatkov niso bile tako velike. To lahko privede do velike poslovne in posledično finančne zgube.

2.4 Organizacijski pristop k projektu ter življenjski cikel podatkovne znanosti

Temeljni del vsake podatkovne strategije je izbor organizacijskega pristopa k projektu. Dolžina projekta je običajno omejena, zato podatkovni znanstveniki nimajo neomejenega časa za pripravo metodoloških napotkov (Molenberghs, Kenward, Fitzmaurice & Tsiatis). Ta čas predstavlja pogosto od 6 do 12 tednov. Da bi razpoložljive tedne karseda dobro izkoristili, morajo podatkovni znanstveniki izbrati način dela, s katerim bodo poskrbeli za organizacijo projekta. Najpogosteje uporabljeni so naslednji pristopi (Saltz et al., 2017):

- **Agilni scrum**: delo se razdeli v časovno omejene dele, imenovane šprinti. Na začetku šprinta se definira, kaj naj bi šprint vključeval; na koncu šprinta naj bi imeli vsaj delno uporaben izdelek. Prednosti agilnega scruma so, da omogoča hitro in efektivno izvedbo projekta. Skozi dnevne sestanke se vidi, kdo je naredil kaj; v fazi predstavitve rezultatov šprinta ekipa dobiva povratne informacije naročnika projekta in na podlagi tega prilagodi delo v naslednjem šprintu. Slabost agilnega scruma je ta, da zahteva več kooperativnosti sodelujočih in predstavlja velik izziv v večjih projektnih skupinah.

- **CRISP**: pristop je podoben metodi slapa. Delo se razdeli po fazah podatkovne znanosti od priprave podatkov do modeliranja in postavitve, pri čemer se postavi mejnike, do katerih je potrebno doseči vnaprej zastavljene cilje. Prednosti CRISP-a so te, da je dobro časovno zasnovan, enostaven za razumevanje in omogoča enostavno testiranje. Slabost CRISP-a je, da ne omogoča velike fleksibilnosti, saj je delujoča rešitev vidna naročniku šele na koncu projekta, zato njegova povratna informacija ne igra tako velike vloge kot pri agilnem scrumu.

- **Agilni Kanban**: delo se razdeli na manjše aktivnosti, ki jih nato postavimo na Kanban tablo. Tabla je razdeljena na več sekcij, med katerimi lahko te aktivnosti prehajajo (npr. za narediti, v delu, narejeno ipd.). Agilni kanban ni zasnovan na časovnici, pač pa na dogodkih oziroma aktivnostih. Prednost tega je, da lahko aktivnost kadarkoli poljubno dodamo, ne da bi pri tem podrli obstoječ projektni načrt. Težava agilnega kanbana se lahko pojavi, če tabla aktivnosti postane preveč zapletena ali če ta ni dovolj pogosto posodobljena. Študija Saltza, Shamshurina in Crowstona (2017) je sicer pokazala, da je ta pristop najbolj učinkovit.

- **Brez definirane metodologije**: podatkovni znanstveniki se projekta lotijo brez posebnih metodoloških napotkov. Ta način članom projektne skupine sicer omogoča največ svobode, vendar je pri takem načinu organizacije verjetnost, da projekt ne bo uspešen, največja.

Ko podatkovni znanstveniki izberejo organizacijski pristop, sledi izvajanje življenjskega cikla podatkovne znanosti. Ta vključuje faze oziroma korake podatkovne znanosti, potrebne za uspešno izpeljavo projekta. Faze se v vseh projektih podatkovne znanosti enake ne glede na izbor organizacijskega pristopa; njihov vrstni red oziroma organizacije te-teh se lahko spreminja glede na organizacijski pristop in projektni načrt. Rollins (2015) je faze življenjskega cikla podatkovne znanosti razdelil na 10 korakov:

1. poslovno razumevanje (angl. business understanding),

2. opredelitev analitičnega pristopa (angl. analytic approach),

3. podatkovne zahteve (angl. data requirenments),

4. zbiranje podatkov (angl. data collection),

5. razumevanje podatkov (angl. data understanding),

6. priprava podatkov (angl. data preparation),

7. modeliranje (angl. modeling),

8. vrednotenje (angl. evaluation),

9. postavitev (angl. deployment),

10. povratna informacija (angl. feedback).

2.4.1 Poslovno razumevanje

Vsaka zahteva stranke se začne z določeno težavo; naloga podatkovnih znanstvenikov je, da jo najprej razume in pristopi k tej težavi s statističnimi in strojnimi tehnikami učenja. Vsak projekt se torej začne s poslovnim razumevanjem. V tej fazi imajo najbolj kritično vlogo poslovni sponzorji, ki potrebujejo analitično rešitev z opredelitvijo problema, ciljev projekta in zahtev rešitve z vidika poslovanja. Ta prva faza je temelj za uspešno reševanje poslovnega problema. Da bi zagotovili uspešnost projekta, bi morali sponzorji sodelovati v celotnem projektu, da bi zagotovili strokovno znanje o domeni, pregledali vmesne ugotovitve in zagotovili, da bo delo ostalo na pravi poti, da bi ustvarili načrtovano rešitev.

2.4.2 Opredelitev analitičnega pristopa

Naslednji korak je opredelitev analitičnega pristopa, kjer lahko podatkovni znanstveniki s podatki, ko je jasno opredeljen poslovni problem, opredelijo analitični pristop za rešitev problema. Ta korak vključuje izražanje težav v okviru statističnih in strojnih tehnologij učenja. Zato je ta korak ključnega pomena, saj pomaga določiti, katere vrste vzorcev bodo potrebne za najučinkovitejšo obravnavo problema. Če je težava določiti verjetnost nečesa, je smiselno uporabiti napovedni model; če je vprašanje prikazati razmerja med različnimi entitetami, bo morda potreben opisni pristop in če naš problem zahteva štetje, je statistična analiza najboljši način za njegovo rešitev. Za vsako vrsto pristopa lahko uporabimo različne algoritme.

2.4.3 Podatkovne zahteve

Ko se določi način, kako rešiti svojo težavo, je potrebno odkriti ustrezne podatke za podatkovni model. Podatkovne zahteve so faza, v kateri določimo potrebno vsebino podatkov, oblike, formate in vire za začetno zbiranje podatkov in jih uporabimo v algoritmu izbranega pristopa. Ta faza ima zelo pomembno vlogo, saj se na podlagi podatkovnih zahtev v naslednji fazi zbira podatke, kar predstavlja dolgotrajen proces, ki je v primeru napačnih podatkovnih zahtev lahko povsem nekoristen.

2.4.4 Zbiranje podatkov

V začetni fazi zbiranja podatkov podatkovni znanstveniki s podatki prepoznajo in zberejo razpoložljive vire podatkov – strukturirane, nestrukturirane in polstrukturirane – ustrezne za problematično področje. Običajno se morajo odločiti, ali bodo vlagali dodatne naložbe za pridobitev manj dostopnih podatkovnih elementov. Priporočljivo je, da se odločitev o naložbi odloži, dokler se ne ve več o podatkih in modelu. Če se pri zbiranju podatkov pojavijo vrzeli, je smiselno, da podatkovni znanstvenik ustrezno spremeni zahteve po podatkih in se loti ponovnega zbiranja novih in/ali večje količine podatkov. Čeprav sta vzorčenje in uporaba podmnožic podatkov še vedno pomembna, današnje visoko zmogljive platforme in analitične funkcije v bazi podatkov omogočajo, da podatkovni znanstveniki uporabljajo veliko večje zbirke podatkov, ki vsebujejo veliko ali celo vse razpoložljive podatke. Z vključitvijo večjega števila podatkov se odpirajo možnosti, da lahko napovedni modeli na bolj enostaven način predstavijo redke dogodke, kot so na primer pojavnost bolezni ali odpoved sistema.

2.4.5 Razumevanje podatkov

Ko je faza zbiranja podatkov končana, podatkovni znanstveniki uporabljajo opisno statistiko in tehnike vizualizacije, da bi bolje razumeli pridobljene podatke. Podatkovni znanstveniki raziskujejo nabor podatkov, na podlagi česar pridobijo razumevanje njihove vsebine in ugotovijo, ali so potrebni dodatni podatki za zapolnitev vrzeli; obenem tudi preverijo kakovost pridobljenih podatkov. Preveriti morajo vrsto posameznih podatkov in izvedeti več o atributih in njihovem pomenu.

2.4.6 Priprava podatkov

Ta faza vključuje vse aktivnosti za sestavljanje nabora podatkov, ki se bo uporabljal v naslednji fazi modeliranja. Dejavnosti faze priprave podatkov vključujejo čiščenje podatkov (obravnavanje manjkajočih ali neveljavnih vrednosti, odstranjevanje dvojnikov, pravilno oblikovanje), združevanje podatkov iz več virov (datoteke, tabele, platforme) in pretvarjanje podatkov v bolj uporabne spremenljivke. V postopku, imenovanem inženiring lastnosti (angl. feature engineering), lahko znanstveniki s kombinacijo poznavanja domen in obstoječih strukturiranih spremenljivk s podatki ustvarijo dodatne pojasnjevalne spremenljivke, ki jih imenujejo tudi napovedovalci (angl. predictors) ali lastnosti. Kadar so na voljo besedilni podatki, na primer dnevniki klicnih centrov za stranke ali zdravniške opombe v nestrukturirani ali polstrukturirani obliki, je analitika besedila koristna pri pridobivanju novih strukturiranih spremenljivk, ki obogatijo nabor napovedovalcev in izboljšajo natančnost modela.

Priprava podatkov je običajno najbolj zamuden korak v projektu na področju podatkovne znanosti. V večini primerov predstavlja 70 % vsega časa na projektu, a ta številka je lahko tudi precej višja oziroma nižja, k čemur lahko pripomorejo dobro

prečiščeni in strukturirani podatki iz prejšnjih faz (Balatsko, 2019).

Številni koraki priprave podatkov so sicer pogosti pri različnih tipih težav. Postopek priprave podatkov je v takih primerih mogoče pospešiti z uvedbo avtomatizacije. Z današnjimi visokozmogljivimi, množično vzporednimi sistemi in analitičnimi funkcionalnostmi, lahko podatkovni znanstveniki tako lažje in hitreje pripravijo podatke z zelo velikimi množicami podatkov.

2.4.7 Modeliranje

Korak modeliranja se začne, ko imamo pripravljeno prvo različico nabora podatkov. Modeliranje se osredotoča na razvoj opisnih in napovednih (prediktivnih) modelov v skladu s predhodno definiranim analitičnim pristopom. Opisno modeliranje je matematični postopek, ki opisuje dogodke v resničnem svetu in razmerja med dejavniki, ki so zanje odgovorni, na primer opisni model lahko preuči stvari, kot so: če je oseba to storila, potem ji bo to gotovo všeč. Prediktivno modeliranje je postopek, ki za pridobivanje podatkov in verjetnost napoveduje rezultate, na primer napovedovalni model se lahko uporabi za določitev, ali je e-pošta neželena pošta. Z napovednimi modeli podatkovni znanstveniki uporabljajo nabor učnih podatkov (angl. training set), ki jih predstavljajo zgodovinski podatki z znanim rezultatom spremenljivke zanimanja. Postopek modeliranja je običajno iterativen, saj organizacije pridobijo vmesne informacije in vpoglede v podatke, kar vodi do izboljšav pri pripravi podatkov in specifikaciji modela. Za določeno tehniko lahko podatkovni znanstveniki preizkusijo več algoritmov z različnimi parametri, da bi našli najboljši model za razpoložljive spremenljivke.

2.4.8 Vrednotenje

Med razvojem modela in pred fazo postavitve podatkovni znanstvenik oceni kakovost modela in se prepriča o njegovi pravilnosti delovanja in o uspešnosti reševanja poslovnega problema. Podatkovni znanstvenik razlaga kakovost modela in njegovo učinkovitost pri reševanju problema z uporabo izračunavanja različnih diagnostičnih meritev in drugih končnih vrednosti, kot so tabele in grafi. Za napovedni model podatkovni znanstveniki uporabljajo testni nabor podatkov (angl. test set), ki je neodvisen od učnega nabora podatkov, vendar pa vseeno sledi isti porazdelitvi verjetnosti in ima znan rezultat iskane spremenljivke. Testni nabor podatkov se uporablja za oceno kakovosti modela; na podlagi ugotovitev se lahko model še izpopolni. Taki metodi pravimo prečno preverjanje (angl. cross-validation). Za končno oceno se včasih uporabi tudi potrditveni nabor podatkov (angl. validation set). Taki metodi s tremi tipi podatkovnih naborov pravimo metoda zadrževanja (angl. hold-out method).

V nekaterih primerih lahko podatkovni znanstveniki modelu dodelijo tudi dodatne

preizkuse statistične značilnosti (angl. statistical significance test) kot dodatni dokaz njegove kakovosti. Ta dodatni dokaz je lahko koristen za utemeljitev izvajanja modela ali ukrepanja, kadar je vložek velik – na primer drag dodatni zdravstveni protokol ali kritičen sistem letalskega leta.

2.4.9 Postavitev

Ko se razvije zadovoljiv model in tega odobrijo poslovni sponzorji, se model uporabi v proizvodnem okolju ali primerljivem testnem okolju. Običajno je postavljen na omejen način, dokler se njegova učinkovitost v celoti ne oceni. Zahtevnost postavitve je lahko zelo različna. Lahko gre za rutinsko operacijo, ki se v celoti opravi v nekaj urah, lahko pa so stvari bolj zapletene in zahteva ogromno dodatne konfiguracije in zapletenih namestitvenih procesov. Uvedba modela v operativni poslovni proces običajno vključuje dodatne skupine, spretnosti in tehnologije znotraj podjetja. Prodajna skupina lahko na primer uvede model nagnjenosti k odzivu (angl. response model) na oglaševalsko akcijo, ki ga je ustvaril razvojni tim in ga upravlja marketinška skupina.

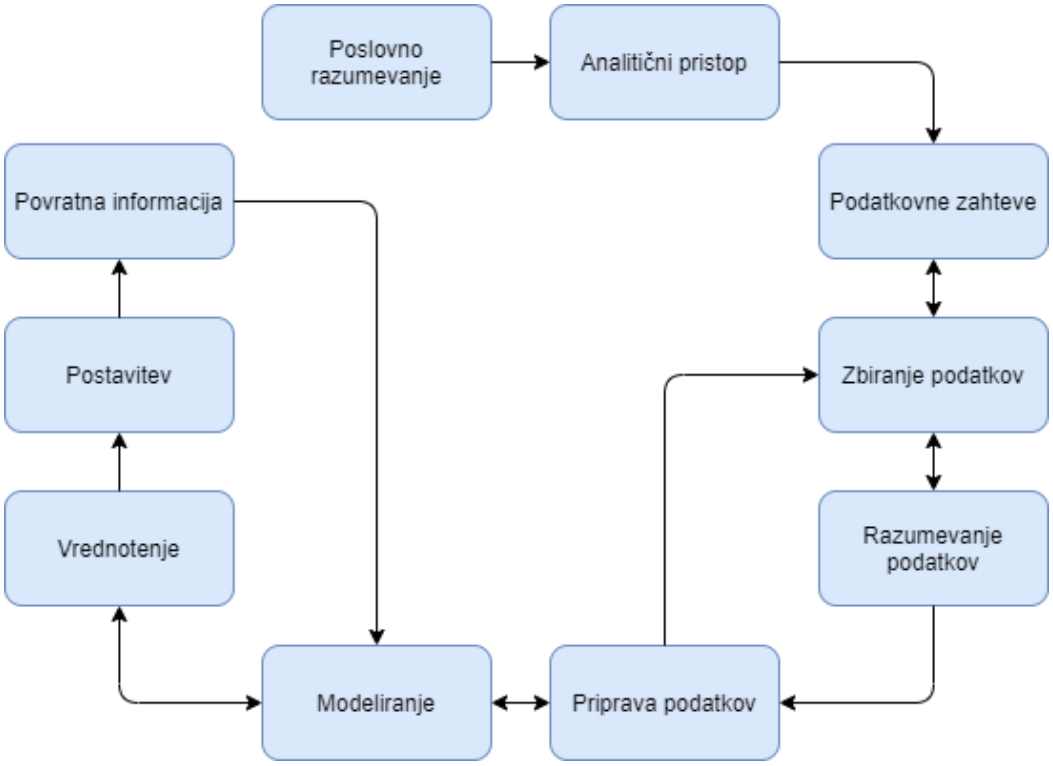
2.4.10 Povratna informacija

Z zbiranjem rezultatov iz implementiranega modela organizacija dobi povratne informacije o uspešnosti modela in njegovem vplivu na okolje, v katerem je bil nameščen. Na primer, povratne informacije bi lahko bile v obliki stopenj odzivanja na promocijsko kampanjo, usmerjeno v skupino strank, ki jih model prepozna kot odzivne z velikim potencialom. Analiza teh povratnih informacij omogoča, da podatkovni znanstveniki izpopolnijo model, da izboljšajo njegovo natančnost in uporabnost. Lahko avtomatizirajo nekatere ali vse korake za zbiranje povratnih informacij in oceno modela, izboljšave in prerazporeditve, da pospešijo postopek osveževanja modela za boljše rezultate.

2.4.11 Zagotavljanje nadaljnje vrednosti za podjetje

Tok metodologije ponazarja iterativno naravo postopka reševanja problemov. Ko podatkovni znanstveniki dobijo več informacij o podatkih in modeliranju, se pogosto vrnejo na prejšnjo stopnjo, da bi prilagodili. Modeli niso ustvarjeni enkrat, nameščeni in puščeni na mestu, kot je; namesto tega se s povratnimi informacijami, izpopolnjevanjem in prerazporeditvijo modeli nenehno izboljšujejo in prilagajajo razvijajočim se pogojem in potrebam, ki se pojavijo šele po začetku projekta. Tako lahko model in delo, ki stoji za njim, data organizaciji nenehno vrednost, dokler je rešitev potrebna.

Slika 3: Koraki projekta podatkovne znanosti



Vir: lastno delo

2.5 Uspešnost

Podjetja se podatkovne znanosti lotevajo, ker želijo biti bolj uspešna. Pri podatkovni znanosti je pogosto tako, da zaradi njenega pogostega omenjanja naročniki velikokrat pričakujejo preveč drastične ugotovitve in spremembe, ki pa jih na koncu ne dobijo in so razočarani, čeprav jim lahko projekt poveča poslovno uspešnost. Da ne bi prihajalo do nenadnih razočaranj, je na koncu zelo pomembno, da podatkovni znanstveniki skrbijo za dober management pričakovanj že od začetka vpeljevanja podatkovne znanosti.

Pilotni projekt je manjši predhodni projekt, ki ga izvedemo, da bi se lažje odločili, kako najbolje metodološko pristopiti in izvesti obsežnejše projekte (Zailinawati, Mazza & Schattner, 2006). S pilotnim projektom lahko ugotovimo, katere metode so najboljše za izvajanje nadaljnjih projektov, in ocenimo, koliko časa in sredstev bo v nadaljnjih projektih potrebno (Crossman, 2020). Pri pilotskem projektu je pomembno, da se uspešnost projekta gleda dolgoročno, saj je namen pilotnega projekta za uporabnike ta, da se naučijo uporabe podatkovne znanost in ne toliko sami takojšnji rezultati in finančne metrike. Pilotski projekt podatkovne znanosti je uspešen takrat, ko izpolnjuje naslednja merila uspešnosti (Kasunic, 2004):

- **prikaz poslovne vrednosti**: po končanem pilotskem projektu mora biti jasno videno, kakšno poslovno vrednost prinaša uvedba podatkovne znanosti. Poslovna vrednost je glavno vodilo pri sprejemanju odločitev, zato je za pilotni projekt zelo pomembno, da je

nazorno prikazana;

- **identifikacija dejavnikov dodane vrednosti**: po končanem pilotnem projektu mora biti znano, na katere dejavnike se moramo v nadaljnjih projektih posebej osredotočati in prinašajo dodano vrednost za podjetje. Podjetje mora vedeti, na kaj mora biti posebej pozorno, ko se bodo zaposleni brez nadzora izkušene ekipe lotevali nadaljnjih projektov podatkovne znanosti;

- **ozaveščanje deležnikov o poslovnih in tehničnih zmožnostih**: predstavlja tako zavedanje deležnikov o možnostih, ki jih podatkovna znanost prinaša kot tudi prenos tako metodološkega kot tehničnega znanja. Da bi bil pilotni projekt uspešen, morajo podatkovni znanstveniki poskrbeti, da deležniki spoznajo zmožnosti ogrodja podatkovne znanosti, saj lahko na podlagi teh identificirajo številne dodatne poslovne priložnosti;

- **implementacija delujočega koncepta metodološkega pristopa**: vključuje delujočo rešitev ter postopek, kako priti do take rešitve. Pomembno je, da je rešitev enostavna za uporabo, dobro deluje na obstoječi infrastrukturi in se navezuje zgolj na podatke, ki so podjetju na voljo. Taka rešitev podjetju v prihodnje omogoča, da bo lahko gradilo modele, ki jim bodo v prihodnje prinašali konkurenčno prednost.

Podjetja pogosto zamenjujejo uspešnost projekta z uspešnostjo modela. Če je model uspešen pri svoji napovedih, še ne pomeni, da je uspešen tudi projekt kot celota, saj morda s tem podjetju ni prinesel velike dodane vrednosti (Hotz, 2019). Model lahko na primer napoveduje neko lastnost zelo natančno, vendar pa je poznavanje te lastnosti za podjetje neuporabno oziroma je ne zna v nadaljevanju dobro izkoristiti. Pogosto se namreč naredi model, ki nam sicer do neke mere pomaga, ne more pa zaradi pomankanja vseh podatkov napovedovati točno tiste lastnosti, ki bi jo podjetje želelo. Lahko pa sicer napove lastnost, ki je s to povezana. Na drugi strani se lahko pojavi tudi obratno. Podjetja zaradi nižje uspešnosti modela avtomatično sklepajo, da je neuspešen tudi projekt podatkovne znanosti, kar tudi ne drži vedno. Da bi uspešno končali projekt podatkovne znanosti, morajo biti upoštevani trije ključni dejavniki (Simpson, 2019):

- **Motivacija in zagnanost sponzorja projekta**: sponzor projekta mora biti na voljo projektni skupini, jim pomagati pri zagotavljanju sredstev in skrbeti za pozitivno naravnanost k podatkovni znanosti pri uporabnikih.

- **Sprejemanje spremembe s strani deležnikov**: za uspešnost projekta je zelo pomembno, da so vsi deležniki pripravljeni sprejeti spremembe, ki jih bo projekt prinesel. Pogosto se namreč dogaja, da ljudje zavračajo novosti, kar pa v tem primeru predstavlja veliko oviro pri doseganju ciljev. Tukaj pomembno vlogo poleg sponzorja projekta igrajo tudi vodje deležnikov, ki morajo sponzorju projekta pomagati pri pozitivni naravnanosti k spremembam.

- **Razpoložljivo znanje pri vodenju projekta in izdelavi rešitve**: podjetje mora imeti na voljo strokovnjake, ki imajo tako organizacijsko, vsebinsko in tehnično znanje, ki se

zadeva projekta. Če teh nima znotraj podjetja, lahko najame zunanje izvajalce. Brez teh znanj projekt ne more biti uspešno izveden.

3 RAZISKOVALNO DELO

V tem delu magistrskega dela bom predstavil projekt uvajanja podatkovne znanosti v slovenskem podjetju, kjer sem sodeloval kot član ekipe podatkovnih znanstvenikov, čigar naloga je bila, da naredimo koncept metodološkega pristopa podatkovne znanosti za naslednje projekte. Pri tem se bom navezoval na teoretično podlago, ki sem jo predstavil v prejšnjem poglavju.

Podjetje, s katerim bomo delali, je ena največjih bančnih in finančnih skupin v Sloveniji. V sklopu podjetja je več podjetij s sedežem v Sloveniji; več hčerinskih podjetij imajo razpršenih tudi po jugovzhodni Evropi. Tako, kot je danes trend pri večini drugih bank, ima tudi naša banka cilj, da postane bolj prilagojena stranki (angl. customer oriented). Zaveda se, da ima pri tem pomembno vlogo izkoriščanje podatkovnih platform in tehnologij, ki bi ji omogočilo, da svojo ponudbo bolj prilagodi ciljnim skupinam.

3.1 Metodologija dela na projektu

Podjetje nas je najelo kot strokovnjake na področju podatkovne znanosti, da bi mu pomagali pri njegovih začetkih na področju podatkovne znanosti in razvili primeren metodološki pristop k podatkovni znanosti. Pred približno pol leta so za potrebe svoje rasti na področju podatkovne znanosti kupili licence za IBM Watson Studio, ki naj bi predstavljal ogrodje njihovih rešitev. IBM Watson Studio je programska rešitev s police (angl. off-the-shell) za podatkovne znanstvenike in inženirje. Ponuja nabor orodij za podatkovno znanost, kot so prenosniki RStudio, Spark, Jupyter in Zeppelin, ki so integrirani z lastniškimi tehnologijami IBM. S tehničnega vidika njihovi podatkovni znanstveniki in analitiki izkušenj z Watson Studiem nimajo,a so se v zadnjih 6 mesecih od nakupa licenc intenzivneje izobraževali po izbiri na področju programskih jezikov R ali Python, zato želijo, da bi imeli tudi v prihodnje možnost dela z obema jezikoma. V preteklosti so z drugimi partnerji tudi že razvili en napovedni model, ki pa jim ni prinesel želenih rezultatov.

Odločili smo se, da bomo uporabili metodologijo razvoja metodološkega pristopa k podatkovni znanosti, opisanega v teoretičnem delu magistrske naloge in ga predstavili s pilotnim projektom, v katerem bomo naredili napovedni model, ki bo reševal enega od njihovih problemov ter tako dokazali delovanje zastavljenega koncepta. Za izvedbo celotnega procesa uvajanja procesov podatkovne znanosti smo dobili šest tednov. Odločili smo se, da si bomo delo znotraj teh šestih tednov organizirali po kombinaciji metod agilnega scruma in agilnega kanbana. Agilni scrum smo izbrali zato, ker bomo tako najlažje sproti obveščali naročnike o našem delu in rezultatih ter v primeru njihovih dodatnih zahtev iterativno prilagodili proces oziroma ciljni izdelek. Za kombinacijo

metode z agilnim kanbanom smo se odločili, saj ta v teoriji prinaša najboljše rezultate; hkrati nam bo omogočil bolj strukturirano vodenje aktivnosti.

Šest tednov smo razdelili na tri dvotedenske enote, imenovane šprinti. Da bi lahko dobro definirali vse cilje in potrebe stranke ter naredili dober projektni plan, smo predhodno organizirali dvodnevno oblikovno delavnico (angl. design thinking workshop). Gre za pristop k reševanju problemov, katerega cilj je upoštevati človeške težave, raziskati možnosti z različnimi ljudmi, preizkusiti nove rešitve in v proces vnesti inovacije. S pomočjo oblikovne delavnice smo skupaj s sponzorji projekta, vodji oddelkov in skupino strankinih podatkovnih znanstvenikov in analitikov izvedli številne različne vaje in delavnice, s katerimi smo dobili boljšo celostno sliko in razumevanje o problemu. Skupaj s stranko smo identificirali probleme in naredili okvirni plan za šesttedenski projekt.

3.2 Oblikovna delavnica

Oblikovne delavnice smo se lotili z določenimi vnaprej pripravljenimi cilji, na podlagi katerih smo tudi izvedli delavnice, s pomočjo katerih bi prišli do želenih rezultatov. Osredotočili smo se na to, kdo naj bi bil na koncu uporabnik ustvarjene rešitve, s katerimi težavami (angl. pain points) se trenutno sooča ta uporabnik, in predvsem, kakšni naj bi bili poslovni izidi od kakršnih koli inovacij. Ti odgovori bi nam pomagali doseči skupno razumevanje primera uporabe in identifikacija osebnosti (Francis, Sumathi & Shiva, 2019). Oceniti želimo tudi trenutno kakovost podatkov, ki jih bomo uporabili pri izvajanju naše rešitve in preverjanju hipotez, povezanih s problematiko. Namen delavnice je osredotočen predvsem na razumevanje trenutne situacije in identifikacijo problemov in želja uporabnikov in v tej fazi še ne toliko na reševanje njihovih problemov.

Poleg sponzorjev projekta so tu še podatkovni znanstveniki ter podatkovni in poslovni analitiki, vključno z zaposlenimi v hčerinskih podjetjih v jugovzhodni Evropi, saj je v tej fazi bistvenega pomena, da je prisotnih čim več različnih profilov ljudi, ki bi nam omogočili kvalitetno in obsežno povratno informacijo, od katere bo kasneje odvisna uspešnost šesttedenskega projekta.

3.2.1 Uvodna delavnica

Oblikovno delavnico smo začeli s predstavitvijo namena delavnice in okvirnih načrtov za šesttedenski projekt; nato smo začeli s krajšimi manjšimi delavnicami, kjer so uporabniki na podlagi naših navodil risali določene objekte, nepovezane s projektom. Namen tega je bil sprostiti udeležene ter ponazoriti, da ljudje kljub enakim napotkom in navodilom zadeve vizualiziramo različno in narišemo drugačne skice. Podobno je tudi v poslovnem svetu, kjer različni uporabniki isto rešitev vidijo z različnih vidikov in se marsikatera prednost za nekoga drugemu zdi slabost.

3.2.2 Upi in strahovi

Prva večja delavnica se je osredotočala na upe in strahove udeležencev delavnice. Cilj delavnice je bil, da se z udeleženimi uskladimo in bolje spoznamo. Želimo dobiti ideje, kaj bo ključno pri uspešnosti prihajajočega projekta ter kaj so nevarnosti, na katere bomo morali biti posebej pozorni. Udeležencem smo razdelili samolepilne listke, na katere so zapisali svoje upe in strahove za naš projekt ter jih prilepili na tablo upov oziroma tablo strahov. Listke smo nato organizirali v smiselne skupine, na podlagi katerih smo se nato dodatno pogovorili z udeleženimi in bodo v našem projektu imeli večji poudarek.

Ugotovljene skupine so naslednje:

Upi:

- metodologija podatkovne znanosti (upajo, da se bodo naučili pravilnega pristopa k podatkovni znanosti in bili sposobni sami ugotoviti nadaljnje rešitve od začetka do konca (angl. end-to-end solution)),
- Watson Studio (upajo, da bodo bolje spoznali okolje Watson Studio in bili sposobni sami postaviti prihodnje modele),
- ocenjevanje poslovne vrednost (upajo, da bodo v prihodnje znali dobro oceniti poslovno vrednost razvitih modelov),
- splošno (v to kategorijo spadajo različni upi, kot so dobro sodelovanje, dober začetek ipd.).

Strahovi:

- podatki (strah jih je, da bo kvaliteta zbranih podatkov slaba oziroma bodo imeli težave z dostopom),
- tehnične omejitve (strah jih je, da jim platforma ne bo dovoljevala nekaterih želenih funkcionalnosti),
- zmogljivost modela (strah jih je, da model ne bo imel nobene dodane vrednosti),
- postavitev (strah jih je, da bo postavitev prezapletena za vzdrževanje),
- poslovni učinek (strah jih je, da bo fokus projekta preveč usmerjen samo v tehnične zadeve),

    projektno vodenje (strah jih je, da bodo cilji projekta preohlapno postavljeni in da bo na koncu zmanjkalo časa za primere dobrih praks).

Slika 4: Rezultati delavnice Upi in strahovi

Vir: lastno delo

3.2.3 Razumevanje situacije

Ko smo spoznali pričakovanja in skrbi udeležencev, je prišel na vrsto drugi pomembnejši sklop, ki ga predstavlja razumevanje situacije. Aktivnost smo razdelili na dve delavnici. Prva se imenuje empatijska preglednica (angl. emphaty map), cilj katere je identificirati pomembna področja, ki bi jih lahko izboljšali. V njej tudi identificiramo glavnega uporabnika naše rešitve in analiziramo njegove pomembne odločitve in težave, ki jih ima pri uporabi. Druga se imenuje kot-je scenarij (angl. as-is scenario), ki nam pomaga pri razumevanju trenutne situacije in težav, s katerimi se soočajo. Spoznati želimo uporabnikovo trenutno uporabniško izkušnjo ter procese, s katerimi se ta sooča vsak dan.

V prvi delavnici smo na tablo narisali namišljeno osebo Popy (zgolj primer imena za lažje navajanje v prihodnje), ki je po poklicu podatkovni znanstvenik v njihovem podjetju in s katero so se udeleženci morali poistovetiti. Popy torej predstavlja njihovega povprečnega podatkovnega znanstvenika. Tablo smo razdelili na štiri dele: stvari, ki jih Popy reče, si jih misli, jih počne in jih čuti. Udeleženci so podobno kot v prejšnji delavnici za vsako od rubrik na tabli na samolepilni listek napisali, kar oni mislijo, da je resnični odgovor za Popya in ga prilepili na tablo. Na koncu smo ponovno liste uredili po smiselnih skupinah.

Na podlagi delavnice smo ugotovili, da imajo strankini podatkovni znanstveniki mnenje, da Popy pogosto sprašuje, kaj je pravzaprav sploh potrebno, da naredi, in kaj zares želijo od njega, ter pravi, da si želi boljše pogoje in predvsem več časa. Popy misli, da ga poslovna stran ne razume, da model nima nobene dodane vrednosti, ter se sprašuje, ali bodo sploh uporabili njegov model. Popy sicer pridobiva vhodne podatke, jih razume in prečiščuje, gradi model in razume poslovni problem. Čuti pa, da njegovi modeli ne izpolnjujejo pričakovanj, da je sam nenehno pod stresom ter da ima težavo z nenehnim omejevanjem s strani sistema. S Popyem smo dobili precej tipičen primer uporabnika, ki pa ga bomo pri gradnji modela imeli v mislih kot nekoga, za kogar gradimo naš model in bo imel koristi od našega projekta.

Slika 5: Rezultati delavnice Empatijska preglednica

Vir: lastno delo

V drugi delavnici v sklopu razumevanja situacije smo se osredotočili na kot-je situacijo. V prvi fazi delavnice, smo želeli poiskati težavne točke (angl. pain-points), s katerimi se njihovi podatkovni znanstveniki vsakodnevno srečujejo. Ponovno smo uporabili koncept lepljenje samolepilnih listkov na tablo in grupiranje teh po vsebini iz prejšnjih delavnic. Ugotovili smo, da imajo štiri glavne probleme – zbiranje in agregiranje podatkov je zelo časovno potratno; težko je ustrezno definirati ključni kazalnik uspešnosti (angl. key performance indicator, v nadaljevanju KPI); ni jasno, kako oceniti stroške implementacije rešitve ter ročno postaviti modele.

V drugi fazi delavnice smo se bolj kot na njihove težave osredotočili na njihove vsakodnevne naloge ter kaj pri njih počnejo in razmišljajo. Tudi tokrat smo uporabili princip lepljenja in grupiranja samolepljivih listkov. Identificirali so 6 glavnih nalog: definiranje problema, zbiranje in procesiranje podatkov, modeliranje, analiza stroškov, testno napovedovanje kampanj ter definiranje vrednosti KPI. Drugi del delavnice aktivnosti ni prinesel veliko ugotovitev, saj so se večinoma v rubriki »kaj počne« odzvali z dejanskim delom, medtem ko so pri počutju zgolj opisovali razpoloženja (vesel, navdušen, žalosten ipd.).

Slika 6: Rezultati delavnice kot-je

Vir: lastno delo

3.2.4 Podatki

Ko smo imeli zbrane informacije o tem, kaj podatkovni znanstveniki počnejo, s kakšnimi težavami se soočajo ter kako se ob tem počutijo, smo se želeli usmeriti še v model, ki ga bomo razvili v našem šesttedenskem projektu. Udeleženi so predlagali, da razvijemo model nagnjenosti k potrošniškem posojilu, katerega rešitev bo posplošljiva tudi na druge modele v prihodnosti in bo imel vlogo neke vrste pilota za njihov razvoj podatkovne znanosti. V času projekta bosta z nami prisotna tudi dva njihova podatkovna znanstvenika oziroma analitika, ki sta predvsem močna na vsebinskem področju, saj se ukvarjata pretežno z trženjem; hkrati bosta z nami pridobivala znanja, ki jih bosta skupaj z nami kasneje poskušala prenesti na ostale strankine podatkovne znanstvenike.

Po pogovoru s prisotnimi smo ugotovili, da bomo pri modeliranju imeli na voljo demografske podatke, bonitetno oceno, finančni status (prihodki/plača), sredstva in

obveznosti, kanale in produkte ter kartične transakcije. Kakovosti podatkov v tem koraku še ne poznamo, a nam je stranka obljubila, da bo podatke karseda dobro strukturirala pred začetkom projekta, kar nam bo močno olajšalo delo.

V sklopu podatkov smo z udeleženci želeli oblikovati nekaj hipotez, ki jih bomo v začetku projekta skupaj z razumevanjem podatkov želeli potrditi oziroma ovreči. Na podlagi njihovih predlogov smo prišli do treh osrednjih hipotez:

- stranke, ki so v zadnjem času spremenile njihovo potrošno obnašanje ali so kupile drugo posojilo ali imajo relativno nizek prihodek ali si lastijo manjše podjetje, so bolj nagnjene k nakupu potrošniškega posojila;
- obstajajo sezonski trendi nakupov potrošniških posojil;
- obstaja pozitivna korelacija med fluktuacijo in nagnjenostjo k nakupu potrošniškega posojila.

3.2.5 Povzetek naslednjih korakov

V tej fazi smo z udeleženimi naredili povzetek in oblikovali okviren načrt za naš projekt. Ker je Watson Studio kot delovno okolje za podatkovne znanstvenike pri naši stranki še v eksperimentalni fazi, posebej postavitveni del, smo se odločili oblikovati postavitveni cevovod (angl. deployment pipeline) za uvedbo modela nagnjenosti k nakupu potrošniškega posojila, ki bo posplošljiv tudi za druge napovedne modele. V času projekta bomo rešitev razvijali skupaj z dvema strankinima podatkovnima znanstvenika, ki trenutno delata v oddelku za trženje; ta model se močno navezuje na njiju in njune potrebe.

Cilji, ki jih moramo doseči v našem šesttedenskem projektu, so torej naslednji:

- analizirati podatke z namenom, da bi pridobili dobro razumevanje o motivacijah potrošnikov, da vzamejo potrošniško posojilo. Obenem je potrebno preveriti tudi prej definirane poslovne hipoteze;
- zagotoviti merila za ocenjevanje modela nagnjenosti k potrošniškemu posojilu;
- zagotoviti model strojnega učenja in celovito rešitev;
- dati predloge za nadaljnje korake po našem projektu;
- prenos znanja pristopa tako k podatkovni znanosti, kot k delu z Watson Studiem,

3.3 Oblikovanje šprintov

Na podlagi ugotovitev v oblikovni delavnici smo razvili seznam ciljev, ki jih bomo želeli

uresničiti v našem šesttedenskem projektu. Da bi se zadeve lotili čim bolj organizirano, smo se po metodi agilnega scruma odločili, da celoten projekt razdelimo v tri dvotedenske šprinte, znotraj katerih bodo izvedene smiselne celote v povezavi s končno rešitvijo.

Posamezen šprint se nanaša na določeno časovno obdobje, v katerem se določena naloga ali dejavnost zaključi in nato pregleda. Dnevne sestanke se sklicuje, da se projektna skupina lahko pogovori o napredku projekta in sprejema odločitve ter se spopada z izzivi. Ko se šprint konča, je naloga projektne skupine, da sponzorju oz. lastniku projekta predstavi svoje dokončano delo in sponzor projekta na podlagi vnaprej določenih meril ugotovi, ali so bila pričakovanja izpolnjena ali ne.

3.3.1 Šprint 1

Prvi šprint se nanaša na raziskovanje podatkov (angl. data exploration) ter gradnji prve verzije modela. Znotraj prvega šprinta bomo raziskali in poskušali razumeti tako nabor podatkov, ki se uporablja v trenutnem modelu nagnjenosti k nakupu potrošniškega posojila kot podatke o kartičnih transakcijah. Raziskati bo tudi potrebno veljavnost v oblikovni delavnici definiranih hipotez ter na podlagi ugotovitev pripraviti podatkovni nabor, ki ga bomo uporabili v našem modelu. Ko bomo imeli pripravljene podatke, bomo naredili prvo verzijo našega modela; hkrati bo potrebno pripraviti arhitekturno rešitev celovite rešitve.

3.3.2 Šprint 2

V drugem šprintu bo poudarek na izboljšanju začetne verzije modela ter implementaciji celovite rešitve (angl. end-to-end solution). Model se bo izboljševal s testiranjem modela na drugih, podobnih podatkih (angl. fine tuning), nastavljanjem hiperparametrov (angl. hyperparameters) ter nadaljnjo izbiro algoritma ter inženiringom lastnosti (angl. feature engineering), kjer neobdelane podatke pretvorimo v obliko, ki je bolj primerna za modeliranje. V drugem šprintu bo potrebno tudi določiti način merjenja uspešnosti modela in na koncu še postaviti celoten cevovod naše rešitve. V tej fazi bi morali imeti delujočo celovito rešitev, ki bo pripravljena za uporabo.

3.3.3 Šprint 3

Tretji, zaključni šprint bo namenjen predvsem končnemu povzetku, kjer bomo dokončali vse od prej odprte naloge ter delili dobro prakso tudi s podatkovnimi znanstveniki, ki v teh šestih tednih z nami niso sodelovali. V zadnjem šprintu bomo tako dokončali naš model nagnjenosti k nakupu potrošniškega posojila in napisali dokumentacijo tako za vsebinski kot tehnični del projekta. Organizirali bomo tudi srečanja s širšo publiko, kjer bomo predstavili naša priporočila za nadaljnje delo na področju podatkovne znanosti ter predstavili delo v našem projektu in obenem podali primere dobrih praks tako za projekte

podatkovne znanosti na splošno kot tudi bolj tehnično o možnostih Watson Studia. Na koncu bomo na prošnjo stranke predstavili tudi nekaj primerov uporabe v drugih podjetjih.

3.3 Priprava na projekt

Pred začetkom projekta, preden se lotimo tehničnega in vsebinskega dela, je potrebno definirati vse cilje in pogoje, ki morajo biti predhodno izpolnjeni ter narediti projektni načrt za šesttedensko izvedbo le-tega. Poslovni cilj je stranki pomagati na poti, da postane bolj usmerjena k strankam (angl. customer-oriented) in izboljša uspešnost poslovanja z uporabo umetne inteligence. Za dosego teh ciljev se bomo v tem projektu združili z njimi in jim pomagali razviti ponovljivo metodologijo pristopa k podatkovni znanosti s primerom napovednega modela za nakup potrošniškega posojila, ki bo banki v prihodnosti pomagal, da bo lahko ponujala bolj stranki prilagojene programe.

Tehnični cilj projekta je stranki pokazati zmožnosti Watson Studio platforme, ki podpirajo:

- raziskovanje (angl. explore) podatkov za boljše razumevanje potrošnikovih namenov ter preverjanje poslovnih hipotez,
- ocenjevanje uspešnosti napovednega modela za nakup potrošniškega posojila,
- implementacijo celotne rešitve strojnega učenja (angl. machine learning).

Med tehnične cilje spada tudi predstavitev naslednjih korakov po končanem projektu ter prenos znanj, ki se navezujejo na metodologijo podatkovne znanosti ter Watson Studia.

Za dosego teh ciljev smo sestavili projektni načrt, ki bo, kot smo že zapisali, vključeval tri agilne šprinte, ki bodo trajali vsak po dva tedna:

Pred začetkom šprinta:

- uvodni sestanek oziroma oblikovna delavnica (smo jo že izvedli);
- namestitev in priprava platforma (že opravljeno);
- dodelitev pravic naši ekipi s strani naročnikovih administratorjev;
- naročnik pripravi nabor podatkov za projekt, vključno z naborom podatkov, ki se uporablja v trenutnem modelu, in agregiranimi podatki o transakcijah s kartico.

Šprint 1:

- raziskava in razumevanje nabora podatkov, ki se uporablja v trenutnem modelu nagnjenosti za potrošniška posojila ter agregirane podatke o transakcijah s kartico;

- preverjanje hipotez o poslovnih predpostavkah oz. podatkih, ki so bile zastavljene v oblikovni delavnici;

- priprava nabora podatkov, ki ga bomo uporabili v našem napovednem modelu;

- osnutek napovednega modela za nakup potrošniškega kredita;

- oblikovanje arhitekture celovite rešitve.

Šprint 2:

- dokončanje napovednega modela s strojnim učenjem, vključno z nastavitvijo hiperparametrov, nadaljnjo izbiro modela in inženiringom funkcij;

- določitev načina merjenja uspešnosti modela;

- implementacija celovite rešitve, vključno s prenosom podatkov iz podatkovnega skladišča, funkcijskim inženiringom, modeliranjem, ocenjevanjem uspešnosti modela, napovedovanjem (angl. scoring), zapisovanjem izhodnih rezultatov nazaj v podatkovno skladišče in postavitvijo (angl. deployment) rešitve v Watson Studio. Napovedovanje se lahko izvaja ročno ali avtomatično.

Šprint 3:

- končati model strojnega učenja;

- pripraviti dokumentacijo celotnega projekta;

- svetovanje o nadaljnjih korakih, vključno z nadaljnjim izboljšanjem uspešnosti modela in vrednotenjem poslovne vrednosti za model napovedovanja;

organizacija delavnic za izmenjavo izkušenj s tem primerom uporabe, splošno najboljšo prakso in predstavitev možnosti Watson Studia širšemu občinstvu.

4 PROJEKT

Glavni del projekta se je začel z uvodnim sestankom ekipe podatkovnih znanstvenikov, kjer se je naloge za šprint 1, definirane v uvodni delavnici in projektnemu planu, strukturiralo v več manjših nalog. Nato se je te vloge razdelilo med podatkovne znanstvenike, kjer je vsak dobil svoj pripadajoči del, o katerem je nato poročal na vsakodnevnih krajših sestankih. V grobem lahko rečemo, da se je delo razdelilo v analizo in razumevanje podatkov, pripravo podatkov in modeliranje ter analizo arhitekture celostne rešitve. Za organizacijo nalog smo uporabili orodje Trello, kjer smo imeli 5 kategorij za celoten projekt: pred začetkom, šprint 1, šprint 2, šprint 3 in delo v ozadju (backlog), kamor so spadale aktivnosti, ki se niso navezovale specifično na noben šprint, a jih je bilo potrebno razrešiti v teku celotnega projekta.

4.1 Šprint 1

4.1.1 Analiza in razumevanje podatkov

Namen tega dela je bil, da se spoznamo z danimi podatki in jih razumemo. Naročnik nam je pripravil dva glavna nabora podatkov – podatke o potrošnikih, ki so stranke naše banke, in podatke o njihovih transakcijah. Poleg tega smo dobili tudi nabor podatkov o bančnih produktih, ki jih potrošniki uporabljajo, skupaj z njihovimi stanji, kanale, kjer so podatki o tem, kateri potrošniki uporabljajo mobilna in internetna plačila ter v kakšni meri ter podatkovni slovar, v katerem so bila razložena polja, ki so v naborih podatkov. Potrošnikovi podatki so vključevali 96 lastnosti za zadnjih 24 mesecev, med katere spadajo demografski podatki (starost, spol, segment) ter mesečna finančna stanja. Transakcijski podatki so vključevali mesečna poročila kartičnih transakcij za vse potrošnike v zadnjih 24 mesecih. Na voljo smo imeli 91 različnih lastnosti, ki vključujejo razlage potrošnikih transakcij vključno s številom transakcij, njihovimi vrednostmi, imeni trgovin in kategorijami nakupov, ki nam povedo, za kakšno vrsto nakupa gre (hrana, pohištvo, dvig na bankomatu itd.). Težava drugega nabora podatkov pa je bila ta, da je od 91 lastnosti zgolj 10 bilo takih, ki so imeli manj kot 40 % manjkajočih vrednosti. Manjkajoče vrednosti nam povedo zgolj to, da se večina potrošnikov še vedno primarno zanaša na plačevanje z gotovino in v večini primerov ne uporablja plačevanja s karticami. Skupno smo dobili 8 milijonov zapisov, razdeljenih v 4 datoteke, kjer je vsaka od njih imela približno 100 različnih lastnosti.

Da bi se lažje lotili razumevanja problema, smo najprej opravili analizo zunanje literature (Barrueta-Meza, Castillo-Villarreal & Armas-Aguirre, 2018), da ugotovili, kateri so glavni razlogi, da potrošniki kupujejo potrošniška posojila, da bi te razloge potem lahko preverili z našimi podatki. Ugotovili smo, da je glavnih pet razlogov naslednjih:

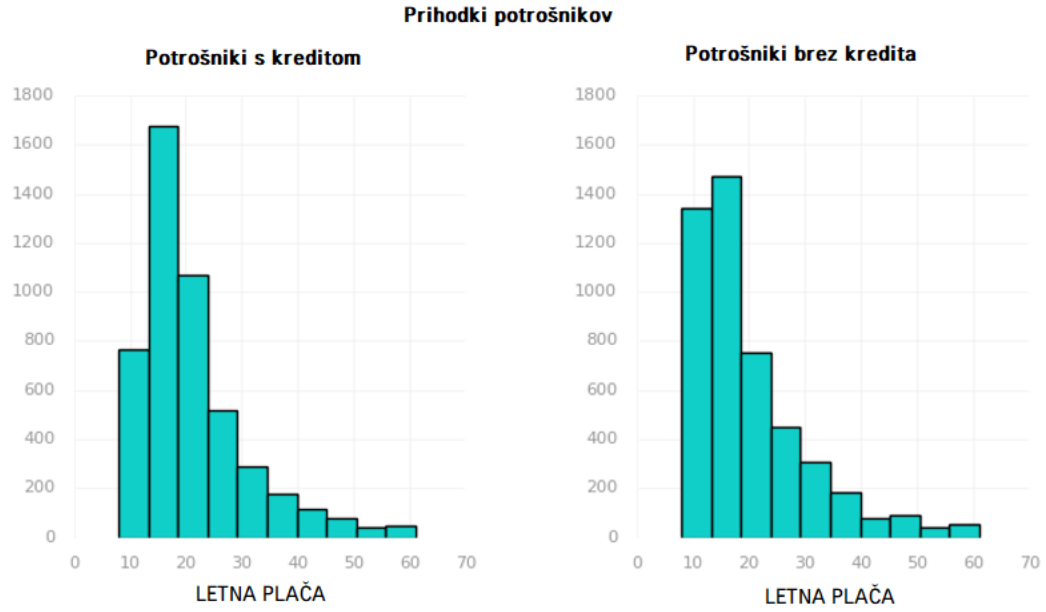
- konsolidacijski dolg,
- nujni stroški, kamor spadajo stroški pogrebov in medicinski računi,
- osebni dogodki: poroke, počitnice, nakup avta,
- prenova doma,
- začetek ali širjenje podjetja.

Edini preverljiv razlog za nas je začetek oziroma širjenje podjetja. Na podlagi naših podatkov namreč lahko ugotovimo, v čigavem primeru gre za samostojnega podjetnika in na ta način vidimo, ali so ti bolj nagnjeni k nakupu potrošniškega posojila kot ostali.

V naslednji fazi smo se lotili preverjanja predhodno postavljenih hipotez in predpostavk. Z uporabo pythona smo vizualizirali lastnosti nabora podatkov, da bi tako lažje uvideli, ali hipoteze držijo ali ne. Kot kreditojemalce smo jemali zgolj potrošnike, ki so kupili

svoje prvo posojilo oziroma ga pred tem niso imeli. Tako smo izločili tiste, ki posojila že imajo, saj nas ta skupina ne zanima, ker posojilo pri nas že imajo. Naš cilj pa bo pridobiti nove kreditojemalce. Ker pa je takih, ki so kupili posojilo, precej manj kot tistih, ki ga niso, smo za lažjo primerjavo in vizualizacijo iz tistih brez posojila naredili vzorec, ki je po velikosti ekvivalenten naboru potrošnikov, ki so posojilo kupili. Prva hipoteza je bila, da imajo potrošniki, ki prvič kupujejo potrošniško posojilo, v povprečju višje prihodke.

Slika 7: Primerjava prihodkov potrošnikov s kreditom in brez njega

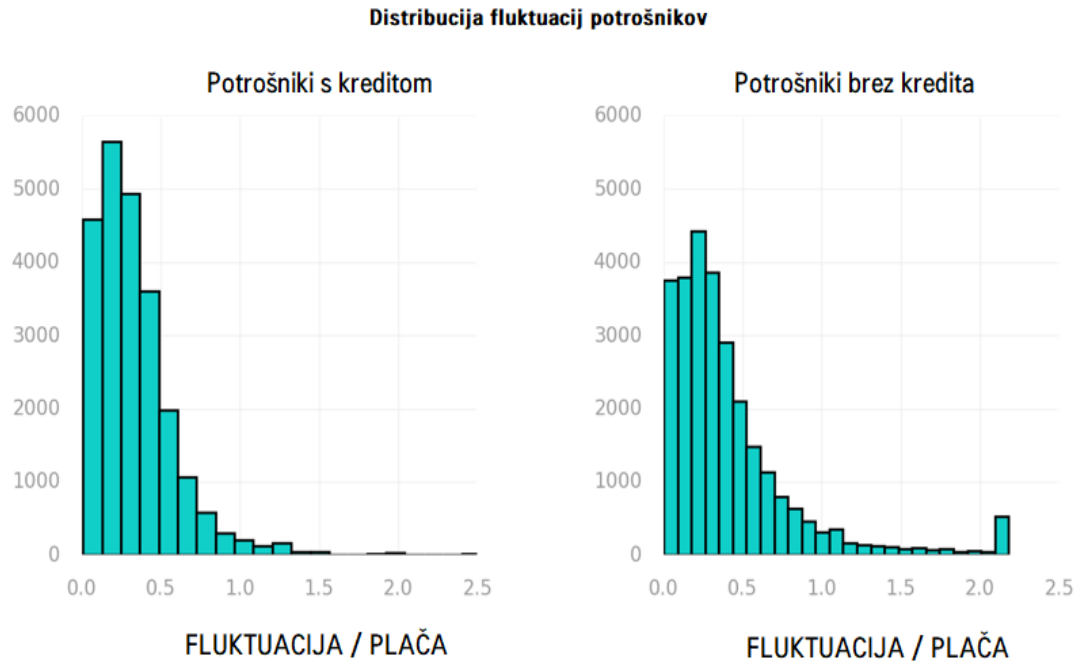


Vir: lastno delo

Grafa prikazujeta distribucijo glede na mesečno plačo za potrošnike, ki so kupili posojilo (levo) in za tiste, ki ga niso, čeprav so bili kreditno sposobni (desno). Na podlagi rezultatov ne moremo trditi, da gre za bistvene razlike med obema grafoma. Na začetku krivulje sicer vidimo rahlo večje povprečje pri potrošnikih s posojilom, vendar gre za stolpca, ki imata relativno podobno vrednost plače in razlika med njima ni bistvenega pomena. Gledano z vidika celotne krivulje pomembnih razlik ni.

Naslednja hipoteza je bila, da je za potrošnike, ki se jim prihodek na mesečni ravni manj spreminja, več možnosti, da bodo kupili potrošniški kredit. Spreminjanje prihodka oziroma fluktuacija pa je pojem, ki si ga lahko različno razlagamo. Na podlagi razlik med plačami v naboru podatkov smo prišli do absolutnih fluktuacij prihodka. Če te seštejemo na nivoju 12 mesecev, dobimo seštevek vseh sprememb v enem letu, ki pa nam za v kontekstu primerjave z ostalimi potrošniki, ne pove veliko, saj ima vsak posameznik različne prihodke. Mesečna fluktuacija 200 evrov pri nekomu s 700 evri mesečne plače na primer pomeni precej večjo realno vrednost kot pri nekomu, ki ima 5000 evrov mesečne plače. Da bi prišli do bolj smiselnih številk za primerjavo, smo povprečno absolutno fluktuacijo delili s povprečnim zaslužkom in tako dobili fluktuacijsko vrednost za primerjavo med potrošniki.

Slika 8: Primerjava distribucij fluktuacij potrošnikov s kreditom in brez njega



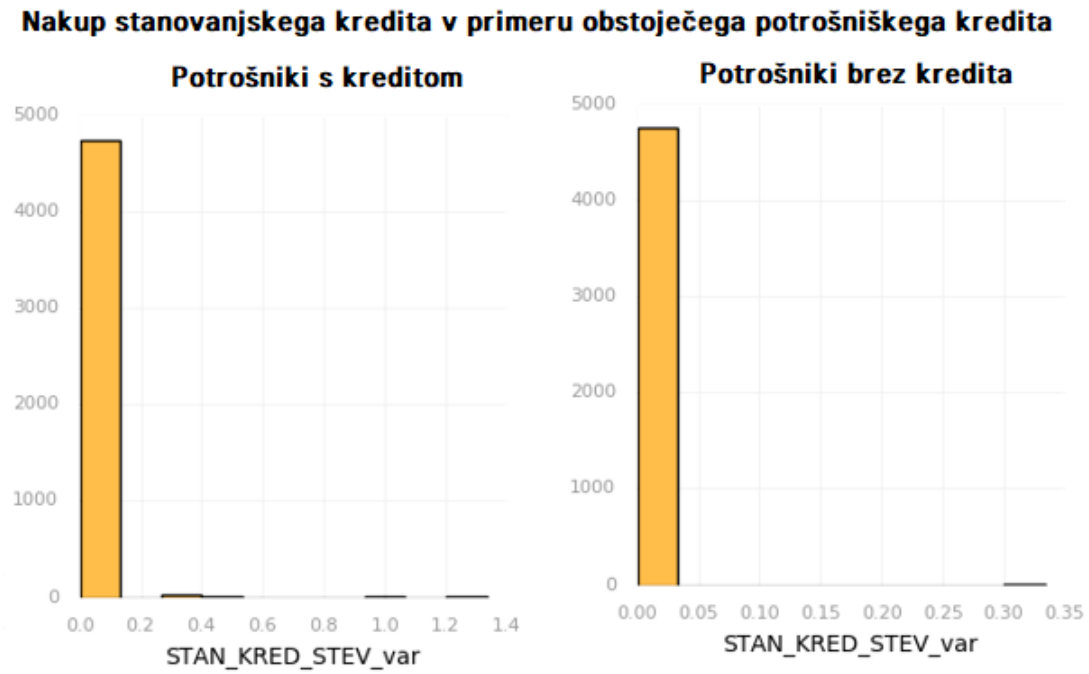
Vir: lastno delo

Grafa na sliki 8 prikazujeta distribucijo potrošnikov, ki so kupili kredit (levo), in tistih, ki ga niso (desno), glede na definirano fluktuacijo v zgornjem odstavku. Tudi v tem primeru nismo našli izrazitih sprememb. Desni graf potrošnikov brez posojila je sicer nekoliko bolj top, kar je posledica večjega števila stolpcev. To smo storili zaradi lepše ponazoritve rezultatov. Zanimiv je sicer skok števila potrošnikov na desnem grafu potrošnikov brez kredita okoli vrednosti razmerja 2. Tukaj imajo tako vrednost vsi potrošniki, ki celoletno plačo dobijo v enem kosu, kar pomeni, da imajo v letu dve veliki spremembi prihodka (na mesec, ko dobijo plačo, in na mesec po njem) ter en velik priliv, kar pri razmerju privede do vrednosti 2:1.

Naslednja hipoteza se je nanašala na povezavo med potrošniškim kreditom in stanovanjskim kreditom. V tem primeru smo se odločili narediti dve primerjavi. Prva je bila, kako potrošniški kredit vpliva na nakup stanovanjskega kredita, druga pa, kako stanovanjski kredit vpliva na potrošniškega.

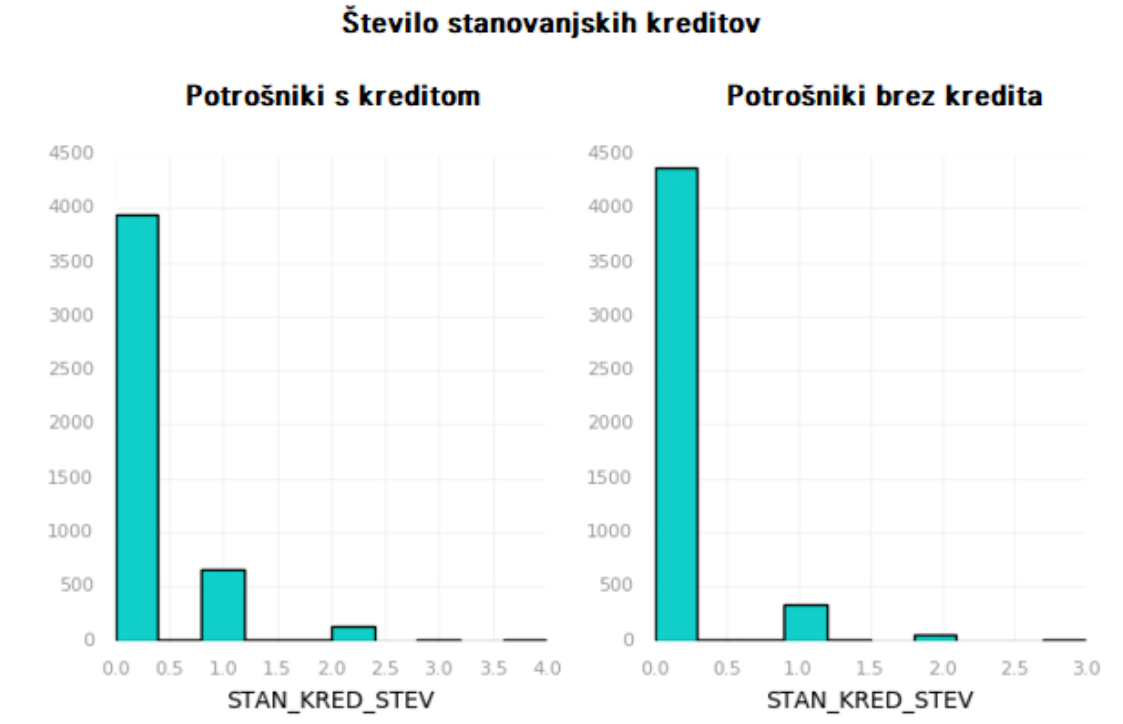
Na podlagi distribucije števila stanovanjskih kreditov potrošnikov, ki že imajo oziroma nimajo potrošniškega kredita (slika 9), smo ugotovili, da imetje potrošniškega kredita ne vpliva na nakup stanovanjskega kredita, saj sta grafa skoraj identična. Glede na to, da je v obeh primerih vrednost 0 izrazito prevladujoča, bi bilo morda smiselno analizo postaviti nekoliko drugače, vendar zaradi izredno majhnih številk temu primeru nismo posvečali nadaljnje pozornosti.

Slika 9: Primerjava nakupov stanovanjskega kredita za potrošnike s kreditom in brez njega



Vir: lastno delo

Slika 10: Primerjava števila stanovanjskih kreditov za potrošnike s kreditom in brez njega



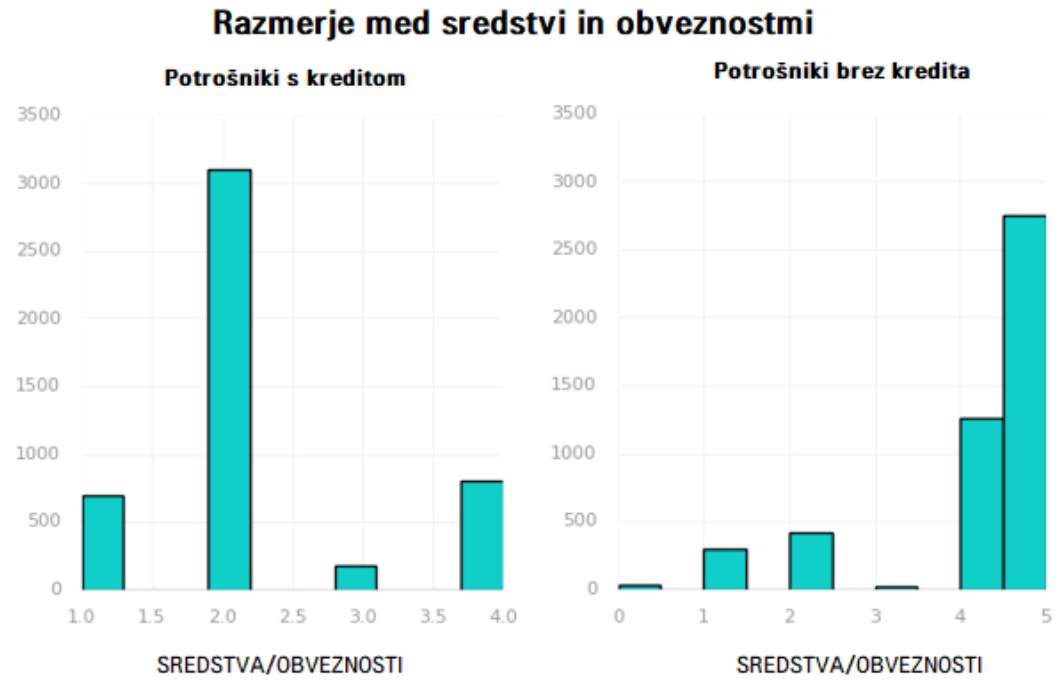
Vir: lastno delo

Pri primerjavi grafov distribucije potrošniških kreditov potrošnikov, ki že imajo oziroma nimajo stanovanjskega kredita, smo odkrili manjši vpliv. Čeprav številke na prvi pogled

niso drastične, se nazorno vidi, da je pri potrošnikih, ki že imajo stanovanjski kredit, verjetnost, da bodo vzeli še potrošniški kredit, dvakrat večja kot pri tistih, ki stanovanjskega kredita nimajo.

Naša naslednja predpostavka je bila, da so manj konservativni potrošniki bolj nagnjeni k jemanju kreditov. Predpostavka je precej široka, zato smo tudi to razdelili na več manjših predpostavk. Primerjali smo razlike med razmerji sredstva/obveznosti (asset liability ration), številom kreditnih kartic, številom varčevalnih računov in zneskom limita za prekoračitev med potrošniki s potrošniškim kreditom in tistimi brez njega.

Slika 11: Primerjava razmerij med sredstvi in obveznostmi za potrošnike s kreditom in brez njega



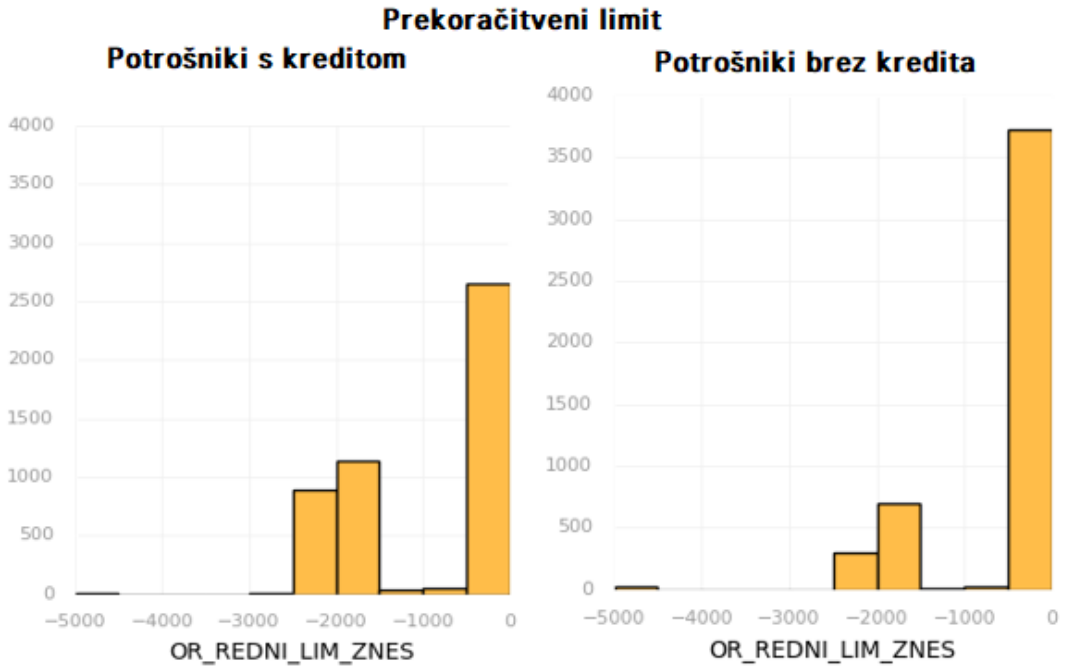
Vir: lastno delo

Graf distribucij razmerij med sredstvi in obveznostmi prikazuje, da imajo potrošniki s potrošniškim kreditom v povprečju precej nižje razmerje (najpogostejša vrednost je 2) med sredstvi in obveznostmi kot tisti brez potrošniškega kredita (najpogostejša vrednost je 5). To potrjuje našo delno predpostavko, da bolj konservativni potrošniki manj pogosto kupujejo kredite.

Graf na sliki 12 prikazuje distribucije limita prekoračitve zneska na bančnem računu za potrošnike s potrošniškim kreditom in tiste brez njega. Tako kot v zgornjem primeru lahko tudi tukaj potrdimo svojo predpostavko, saj imajo tisti s potrošniškim kreditom v povprečju precej višji limit kot tisti brez potrošniškega kredita. Našo predpostavko sta potrdila tudi podatka o tem, da imajo kreditojemalci v povprečju več kreditnih kartic in manj varčevalnih računov. Na podlagi teh štirih ugotovitev lahko res rečemo, da so manj konservativni ljudje bolj nagnjeni k nakupu potrošniškega kredita.

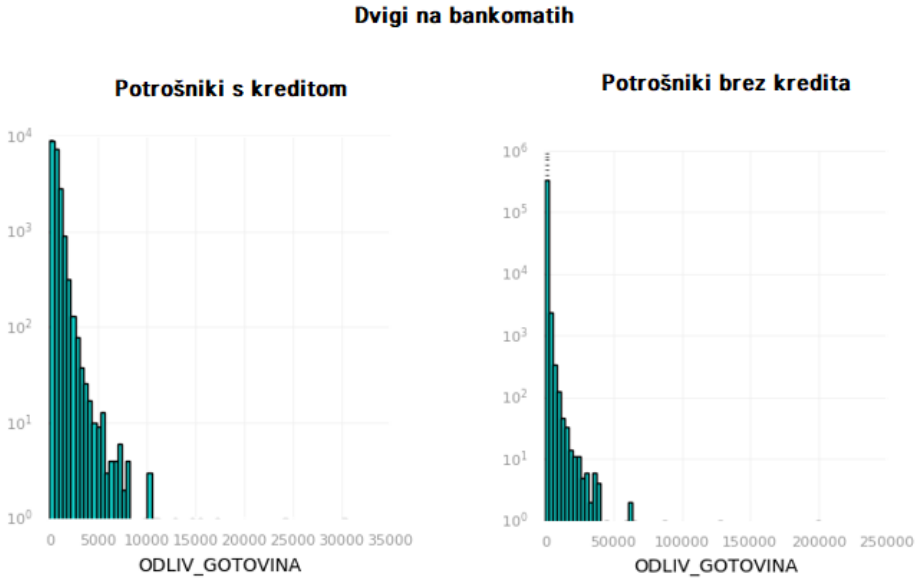
Naša četrta predpostavka je bila, da potrošniki, ki imajo potrošniški kredit, manj plačujejo z gotovino in več s kreditno kartico. Na spodnjem grafu distribucij gotovinskih odlivov tega sicer nismo opazili, zato predpostavke ne moremo potrditi.

Slika 12: Primerjava prekoračitvenega limita za potrošnike s kreditom in brez njega



Vir: lastno delo

Slika 13: Primerjava dvigov na bankomatu za potrošnike s kreditom in brez njega

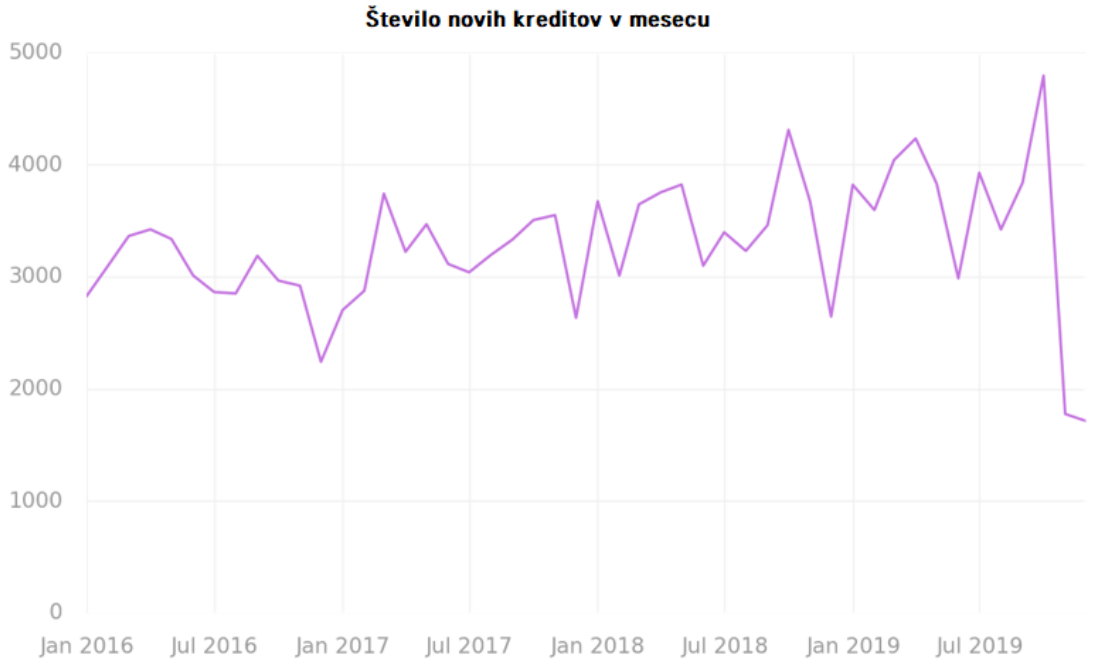


Vir: lastno delo

Zadnja hipoteza, ki smo jo želeli preveriti, je bila, da pri potrošniških kreditih obstajajo

določeni sezonski trendi. Ker iz zadnjih 24 mesecev nismo uspeli dobiti koristnih informacij, smo stranko zaprosili za večji nabor podatkov, ki je vključeval zadnjih 48 mesecev.

Slika 2: Pregled števila novih kreditov v mesecu v obdobju od januarja 2016 do novembra 2019



Vir: lastno delo

Zgornja slika prikazuje število novih posojil od januarja 2016 do decembra 2019. Načeloma velikih fluktuacij števila kreditov ni opaziti. Gre pa omeniti rahle padce v decembru in poletnih mesecih, kar je pravzaprav pričakovano, saj se ljudje finančno že prej pripravijo na te mesece (darila, počitnice). Še ena zanimivost je večji skok in vrhunec jemanja kreditov v oktobru 2019. Ta je posledica sprememb v zakonodaji, saj so novembra začela veljati strožja pravila, zato so potrošniki, ki so razmišljali o nakupu kredita, tega vzeli že prej. To potrjuje tudi velik upad nakupa kreditov od novembra dalje, ko so številke dosegle najnižje vrednosti v opazovanih zadnjih štirih letih.

Po koncu hipotez smo želeli še preveriti predpostavko iz literature o glavnih vzrokih jemanja kreditov, kot smo zapisali zgoraj, da so lastniki manjših podjetij bolj nagnjeni k jemanju potrošniških kreditov. Na podlagi podatkov, ki nam jih je pripravil naročnik, smo naredili analizo in ugotovili, da lastništvo manjšega podjetja sicer vpliva na nakup potrošniškega kredita, je pa ta vpliv relativno majhen.

4.1.2 Priprava podatkov in modeliranje

V tej fazi smo že pridobili razumevanje o podatkih, ki jih imamo; dobili smo določene vpoglede v trende in načine obnašanja in v naslednjem koraku smo se lotili priprave podatkov za naš model. Najprej je bilo potrebno definirati, kdo bo naše ciljno občinstvo,

na koga se bo kasneje navezovala oglaševalna kampanja, ki bo ustvarjena na podlagi rezultatov napovednega modela. Odločili smo se, da iščemo posameznike, ki se bodo v naslednjih dveh mesecih odločili za nakup kredita. Potencialen kreditojemalec v tem trenutku še ne sme imeti drugega potrošniškega kredita, saj banka ne želi spodbujati jemanja novih, boljših kreditov, s katerimi bi enostavneje odplačali starega. Poslovna pravila za naše ciljno občinstvo lahko torej definiramo na naslednji način: iščemo potrošnika, ki ta trenutek še nima potrošniškega kredita, je pa na podlagi naših trenutnih podatkov primeren (angl. eligible). Pogoji, da je potrošnik primeren za kredit, so naslednji: potrošnik mora biti živ, starejši od 18 let, državljan Slovenije, njegove skupne dolžnosti ne smejo presegati 150.000 evrov, čisti prihodek po odštetih mesečnih plačilih za kredit mora presegati 822 evrov in imeti mora kreditno oceno vsaj razreda C. Cilj našega napovednega modela bo torej poiskati takega potrošnika; nato se bo v odnosu do tega potrošnika izvedlo bolj intenzivno oglaševanje. Pri pogovoru o tej strategiji smo sicer izpostavili dve slabosti. Prva je ta, da naj bi izbrani potrošniki v vsakem primeru kupili kredit in gre zgolj za nekoristen strošek oglaševanja, druga pa ta, da določene potrošnike pretirano oglaševanje odvrne od nakupa. Težava v prvem primeru so sicer lahko res odvečni stroški, vendar lahko ob pomankanju oglaševanja potrošnik sicer res kupi kredit, a to stori pri konkurenčni banki, česar pa si ne želimo. Drugo težavo do neke mere lahko odpravimo tako, da oglašujemo zgolj potrošnikom, ki so v oglaševanje ob podpisu pogodbe z banko privolili.

Podatki, ki smo jih dobili od stranke, so bili v CSV obliki in so bili načeloma sprejemljive kakovosti z izjemo transakcijskih podatkov, kjer je bila večina podatkov manjkajočih. Med drugim so manjkali tudi trije meseci, ki smo jih zapolnili tako, da smo za manjkajoče vrednosti povprečili vrednost iz prejšnjega in vrednost iz naslednjega meseca. Čeprav bomo kasneje podatke pobirali z R skripto neposredno iz podatkovnega skladišča, smo se v tej fazi odločili uporabiti kar datoteko, ki nam jo je posredovala stranka. Vseeno pa je bilo potrebno podatke še nekoliko urediti. V prvi fazi smo na podlagi poslovnih pravil, ki nam jih je zagotovila banka, iz podatkov izločili vse potrošnike, ki niso kreditno sposobni. Glede na to, da je razmerje med kreditojemalci in ostalimi izrazito neuravnoteženo, smo razdelili potrošnike na kreditojemalce in ostale ter nato naredili vzorec tistih, ki kredita niso vzeli, da bi pridobili enako število obeh. Podatki, ki smo jih dobili, so sicer vključevali informacije za 24 mesecev od decembra 2017 do novembra 2019, vendar zaradi sprememb v zakonodaji, kot smo zapisali zgoraj, podatkov po oktobru 2019 nismo mogli uporabiti. Odločili smo se, da bomo model zgradili na podlagi 12 mesecev; njegovo uspešnost bomo preverili na naslednjih dveh mesecih. Za učenje smo tako torej namenili podatke med junijem 2018 in majem, medtem ko bo učenje potekalo na podatkih za junij in julij 2019.

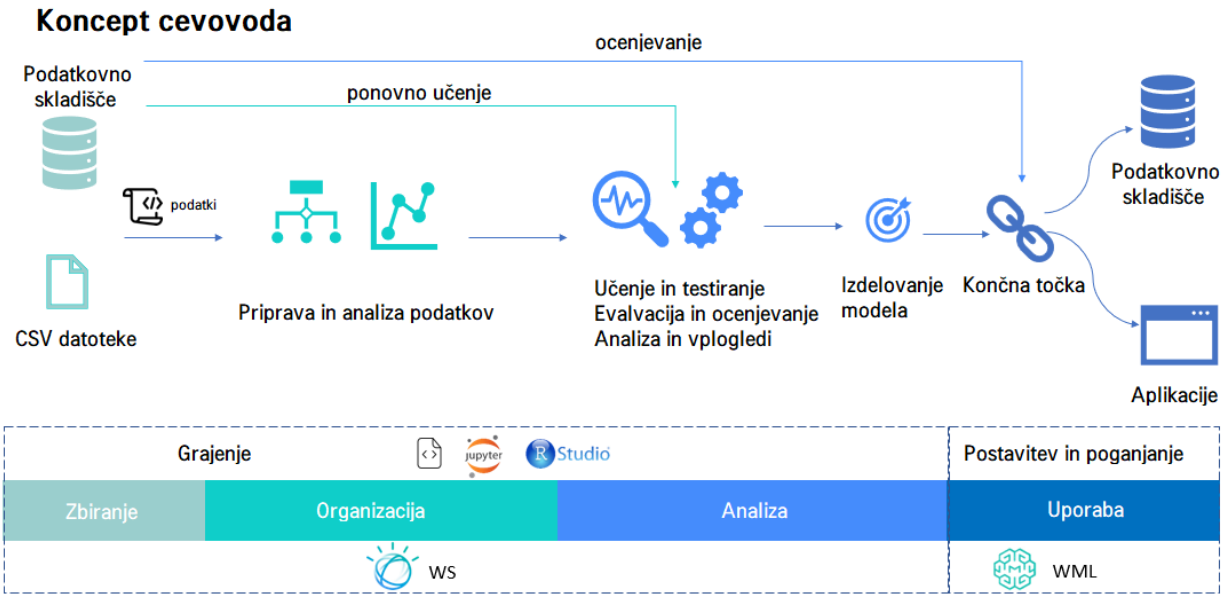
Ko imamo pripravljene podatke, lahko pričnemo z modeliranjem. Naš začetni model smo zastavili tako, da smo mu podali vse lastnosti, ki jih imamo v prečiščenem naboru podatkov z namenom, da bi kasneje ugotovili, katere lastnosti nam prinašajo izboljšanje rezultatov in katere ne. V prvem modelu smo uporabili algoritem naključnega gozda (Livingston, 2005) in ga sprogramirali v programskem jeziku R. Tako smo prišli do faze,

kjer imamo začeten napovedni model, medtem ko bomo v drugem šprintu izvedli še izbiro ustreznih lastnosti, poskusili nove lastnosti glede na poslovno razumevanje, naredili model bolj stabilen in robusten, razvili dobro razumevanje modela in njegovih rezultatov ter po potrebi preizkusili kakšen drug algoritem oziroma modelarski pristop.

4.1.3 Arhitektura cevovoda celostne rešitve

Znotraj prvega šprinta je bila naša naloga tudi razviti koncept cevovoda celostne rešitve, ki ga bomo na koncu uporabili. Znotraj Watson Studia imamo dve komponente – Watson Studio in Watson Machine Learning. Prvi je namenjen zajemu in analizi podatkov ter gradnji modela, drugi pa postavitvi (angl. deployment) in avtomatizaciji postopka. Ideja celotne rešitve je, da podatke zajamemo z uporabo python oziroma R skript bodisi iz podatkovnega skladišča, bodisi iz datotek CSV ter njihovo analizo naredimo v jupyter zvezkih v poljubnem jeziku (R, python). Ko podatke razumemo in jih počistimo in uredimo, te podamo modelu, ki ga prav tako razvijemo znotraj Watson Studia. Postopek čiščenja podatkov lahko tudi izpustimo in jih neposredno posredujemo modelu. Ko je model naučen, ga postavimo (angl. deploy) v okolje Watson Machine Learninga, kjer lahko kasneje avtomatsko teče; podatke zapisuje bodisi v podatkovno skladišče, bodisi v aplikacijo, ki te podatke potrebuje. Tam se tudi izvaja testiranje uspešnosti.

Slika 15: Koncept cevovoda celostne rešitve



Vir: lastno delo

4.1.4 Povzetek

Na koncu prvega šprinta smo širši publiki podatkovnih znanstvenikov, zaposlenih v banki, ter njihovim vodjem predstavili naš prispevek in delo, ki smo ga opravili v prvih dveh tednih. Ključnih pet stvari, ki smo jih dosegli v tem času, so:

- razumevanje trenutne situacije;
- preverili smo poslovne hipoteze;
- pripravili smo nabor podatkov za potrebe našega napovednega modela;
- razvili smo koncept cevovoda celovite rešitve;
- razvili smo prvo verzijo napovednega modela.

4.2 Šprint 2

V prvem šprintu je bil glavni namen raziskovanje in razumevanje situacije in podatkov; v drugem šprintu je bil poudarek predvsem na razvoju in implementaciji rešitve. Na začetku drugega šprinta smo imeli ponovno daljši sestanek, kjer smo naredili podrobnejši plan za naslednja dva tedna. Skupaj z zaposlenimi v naročnikovem podjetju smo tudi razvijali in diskutirali o idejah za izboljšanje modela. Posamično smo naredili raziskave in analize potencialnih rešitev in jih nato na skupnem sestanku predstavili še ostalim, na podlagi česar smo nato oblikovali smer, v katero bomo šli. Odločili smo se, da bomo uporabnike najprej segmentirali po skupinah in šele nato nad njimi pognali napovedni model in tako izboljšali učinkovitost našega modela, saj se različne skupine na različne dogodke različno odzivajo, kar pa se pri povprečenju vsega lahko izgubi. Delo smo tako tokrat razdelili na segmentacijo potrošnikov, izboljšanje modela ter implementacijo celostne rešitve.

4.2.1 Segmentacija potrošnikov

Ideja segmentacije potrošnikov je, da se primerja profile in podatke za vse kreditno sposobne potrošnike in se izračuna podobnosti med njimi. Na podlagi teh podobnosti se nato naredi skupine, v katerih so podobni potrošniki s podobnimi navadami in lastnostmi. To nam omogoči bolj učinkovito prilagajanje parametrov in boljše razumevanje modela. Nudi nam tudi možnost, da lahko bolj prilagojeno oglašujemo in lažje ugotavljamo, kdo so potencialni kreditojemalci, saj se bo model vsakemu tipu skupine bolj natančno prilegal. Prav tako bomo lahko odkrili priložnosti za nove kampanje glede na trende posamezne skupine (npr. spodbujanje ljudi, ki veliko zapravljajo, da uporabljajo kreditne kartice ipd.). Slabost tega je sicer ta, da potrebujemo toliko različnih modelov, kot imamo skupin. Vseeno pa v primeru dobrih rezultatov prednosti odtehtajo nekaj več dela pri vzdrževanju.

Z uporabo K-meadians algoritma za gručenje smo nabor vseh 295.931 potrošnikov razdelili v štiri različne skupine potrošnikov. V grobem bi lahko rekli, da smo dobili skupino kreditojemalcev, skupino varčevalcev, skupino, ki plačuje pretežno z gotovino, in skupino upokojencev.

Skupina kreditojemalcev vključuje 32.862 potrošnikov, ki so v 70 % aktivni moški z visokim dohodkom. Njihova povprečna starost je 47 let. Glede na celotno populacijo imajo zelo visok dohodek; vsak mesec izvedejo veliko transakcij in so finančno stabilni. Imajo tudi visoke obveznosti (angl. liabilities), visok limit prekoračitve stanja na bančnem računu ter veliko uporabljajo kreditno kartico. Stanje na njihovem računu je nadpovprečno, čeprav ne toliko, kot so nadpovprečne prej naštete lastnosti. Razmerje med sredstvi in obveznostmi imajo nizko; enako velja za varčevalne račune. Predstavljajo idealen profil kreditojemalca, kot smo ga odkrili v prvem šprintu (ni konservativen, uporablja kreditne kartice, ima visok limit prekoračitve, ima nizko razmerje med sredstvi in obveznostmi). 1 % potrošnikov v tej skupini je v zadnjih 24 mesecih vzelo potrošniški kredit, kar je z naskokom največji odstotek med vsemi štirimi skupinami.

Skupina varčevalcev vključuje 60.723 potrošnikov, ki so v 68 % aktivne ženske, ki imajo podobno kot prva skupina relativno visok dohodek. Povprečna starost druge skupine je 52 let. Podobno kot skupina kreditojemalcev imajo tudi varčevalci poleg visokega prihodka tudi veliko mesečnih transakcij in dobro finančno stabilnost. Se pa od kreditojemalcev najbolj razlikujejo v tem, da imajo zelo nizke skupne stroške, malo uporabljajo kreditne kartice in imajo nastavljene nizke limite prekoračitve stanja na bančnem računu. Njihovo stanje na računu je zelo visoko; nadpovprečno je tudi njihovo razmerje med sredstvi in obveznostmi ter število varčevalnih računov. Od vseh štirih skupin varčevalci najmanj pogosto vzamejo potrošniški kredit. Med njimi je samo 0,2 % takih.

Skupina upokojencev je daleč največja in s 112.289 potrošniki predstavlja več kot tretjino vseh potrošnikov banke. Čeprav smo skupino poimenovali upokojenci, to sicer ne pomeni, da so vsi v pokoju; gre za najstarejšo skupino z lastnostmi upokojencev. Povprečna starost je 63 let; gre tudi za najmanj aktivno skupino. Od vseh skupin imajo v povprečju najnižje prihodke in relativno slabo finančno stabilnost. Ne uporabljajo kreditnih kartic. Hkrati tudi zelo redko dvigujejo denar, ki ga imajo na računu. Imajo zelo nizko stanje bančnem računu. 0,3 % jih ima potrošniški kredit.

Skupina ljudi, ki plačujejo z gotovino, je druga največja in vključuje 90.057 potrošnikov. Kar 95 % je aktivnih moških; povprečna starost je 53 let. V povprečju imajo podpovprečen prihodek, čeprav je ta še vedno precej višji od prihodka upokojencev. Podobno kot upokojenci tudi oni ne uporabljajo kreditnih kartic in imajo slabo finančno stabilnost. Stanje na njihovem računu je boljše kot pri upokojencih, čeprav je še vedno podpovprečno. Imajo največ dvigov denarja na bankomatih izmed vseh štirih skupin. V skupini potrošnikov, ki plačuje z gotovino, jih ima 0,4 % potrošniški kredit.

4.2.2 Modeliranje

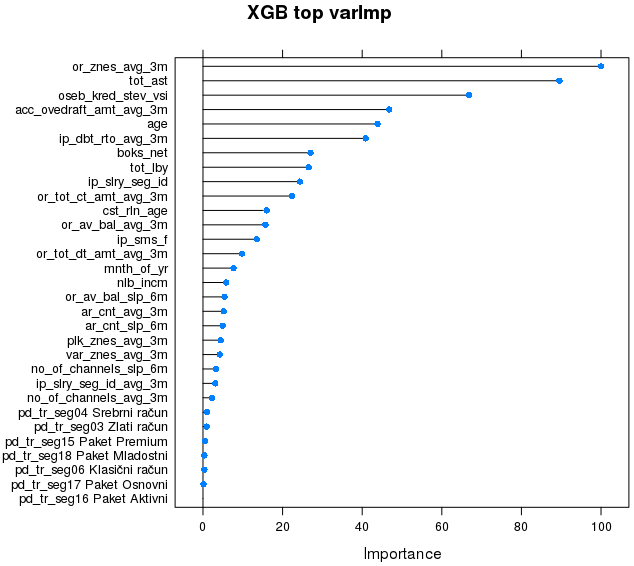
V prvem šprintu smo definirali našo ciljno občinstvo, uporabili poslovna pravila, definirali časovni razpon naših podatkov, na katerih se bo model učil in testiral, pripravili

podatke ter izdelali prvi model z metodo naključnega gozda. V drugem šprintu se bomo osredotočili na inženiring lastnosti, izbor lastnosti, primerjavo modelov ter testiranje modela. Najprej smo se lotili izbora lastnosti. Odstranili smo lastnosti, ki so imele pretežno eno vrednost, niso imele poslovne dodane vrednosti ali pa so bile zelo povezane z drugimi lastnostmi in ni bilo smiselno, da bi obdržali obe. Na podlagi ugotovitev iz analize podatkov in tudi ugotovitev iz prve iteracije modela smo dodali kar nekaj novih lastnosti. Dodali smo četrtletne in polletne agregacije naslednjih lastnosti (nekatere so že v osnovi izpeljane iz drugih lastnosti):

- razmerje med sredstvi in obveznostmi,
- prihodki,
- stanje na računu,
- mesečni znesek transakcij s kreditnimi in debetnimi karticami,
- limit prekoračitve stanja na računu,
- stanje na kreditni kartici,
- stanje na varčevalnem računu,
- število izdelkov, ki jih pri banki potrošnik uporablja,
- število različnih lokacij, kjer potrošnik plačuje, dviguje denar.

V naslednji fazi je bilo potrebno izbrati lastnosti, ki jih bomo uporabili v našem modelu. Naredili smo korelacijsko analizo, na podlagi katere smo lahko razvrstili lastnosti po pomembnosti.

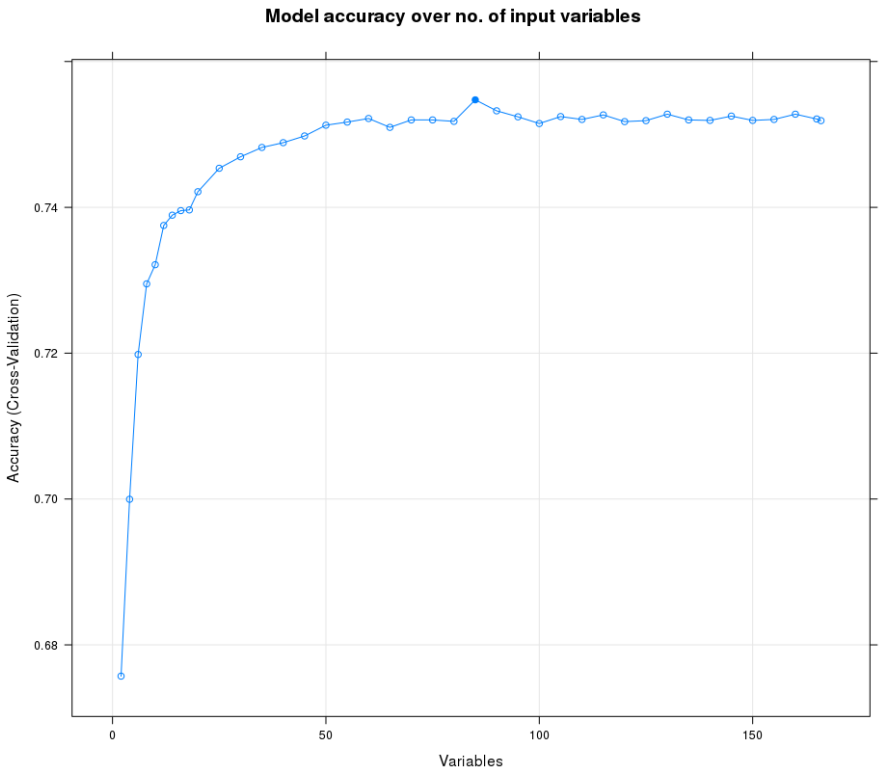
Slika 16: Lastnosti, razvrščene po pomembnosti



Vir: lastno delo

Z uporabo rekurzivne tehnike odstranjevanja lastnosti, ki se prilega modelu in odstranjuje najšibkejšo lastnost oziroma lastnosti, smo nato uspeli število lastnosti zmanjšati iz 166 na zgolj 24, ne da bi pri tem pomembno izgubili natančnost modela.

Slika 17: Graf natančnosti modela v odvisnosti od števila uporabljenih lastnosti

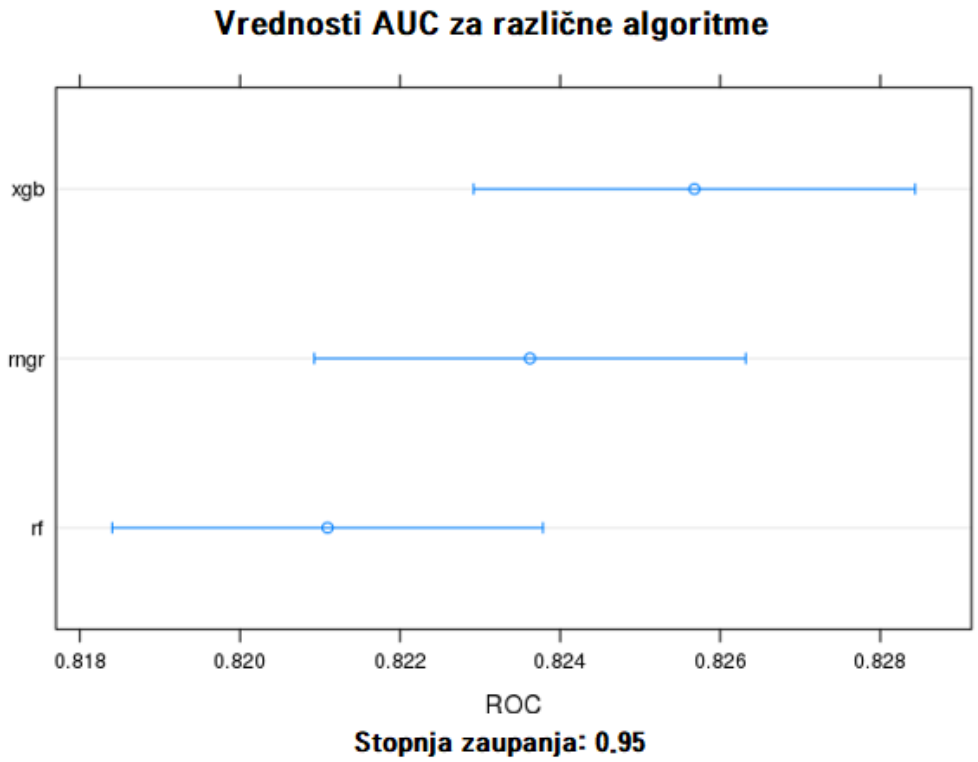


Vir: lastno delo

Pri izbiranju lastnosti in izbiri algoritma gre za iterativen primer, saj se obe dejavnosti v veliki meri prekrivata in prilagajata druga drugi. Pri gradnji našega modela smo uporabili tri različne algoritme, ki smo jih kasneje primerjali med seboj. Prvi algoritem je bil naključen gozd, ki smo ga uporabili že kot naš izhodiščni algoritem. Gre za algoritem nadzorovanega učenja, ki naključno ustvari in združi več odločitvenih dreves v en »gozd«. Naslednji algoritem je bil ranger, ki pravzaprav predstavlja hitro implementacijo naključnega gozda za visoko dimenzijske podatke. Tretji algoritem, ki smo ga uporabili, pa je bil eXtreme Gradient Boosting (v nadaljevanju XGBoost). Gre za algoritem strojnega učenja, ki, podobno kot prejšnja dva algoritma, temelji na odločitvenih drevesih in uporablja ogrodje za pospeševanje gradientov (angl. gradient boosting framework). Za vse tri algoritme smo opravili primerjavo AUC krivulj (slika 18). Najboljše rezultate smo dosegli z uporabo algoritma XGBoost, čeprav so bili rezultati zelo blizu.

Ko smo pripravili podatke in izbrali algoritem, je bil model pripravljen na testiranje. Izvedli smo analizo občutljivosti in specifičnosti, kjer se na podlagi v naprej določenega praga (angl. threshold) za napoved v modelu razglasi, ali gre za pozitivno ali negativno vrednost.

Slika 18: Primerjava vrednosti AUC za različne algoritme



Vir: lastno delo

Slika 19: Tabela pravilno in napačno napovedanih vrednosti glede na postavljeno mejo napovedovanja

Meja	Pravilno negativni	Pravilno pozitivni	Napačno negativni	Napačno pozitivni	Napovedani pozitivni
0,00	0	1.249	0	294.682	295.931
0,10	73.032	1.239	10	221.650	222.889
0,20	118.883	1.215	34	175.799	177.014
...	...	...	...	...	...
0,80	278.894	421	828	15.788	16.209
0,90	293.499	59	1.190	1.183	1.242
1,00	294.682	0	1.249	0	0

Vir: lastno delo

Na sliki zgoraj vidimo, kako se spreminja razmerje med pravimi pozitivnimi, napačnimi pozitivnimi, pravimi negativnimi in napačnimi negativnimi napovedmi. Na podlagi tega lahko izberemo, koliko potrošnikov (in posledično kakšen prag) bomo izbrali in jim oglaševali bankin potrošniški kredit. V tem koraku se mora banka odločiti, kakšne prioritete ima pri izbiri števila potrošnikov, ki se jim bo oglaševalo, saj mora izbirati med

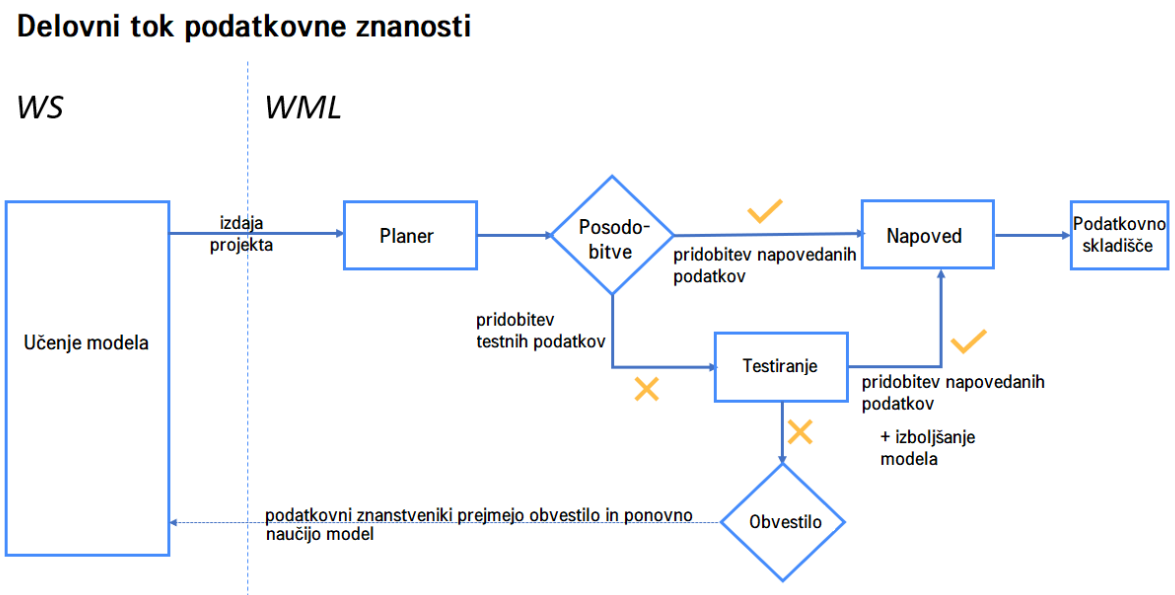
stroški napačnega oglaševanja (napačne pozitivne napovedi) in stroški izpuščene priložnosti za oglaševanje (napačni negativni).

4.2.3 Cevovod celostne rešitve

Po končanem modeliranju pride na vrsto implementacija cevovoda celostne rešitve, ki predstavlja najbolj tehnični del projekta. Cevovod smo zasnovali za dve različni osebi, podatkovnega znanstvenika in upravitelja postavitve (angl. deployment manager). Da bi obema uporabnikoma zagotovili karseda dobro uporabniško izkušnjo, je rešitev zasnovana tako, da uporablja dve različni okolji. Glavna naloga podatkovnega znanstvenika je razviti model strojnega učenja z analizo nabora podatkov, preizkušanjem različnih algoritmov in izbiro ustrezne meritve vrednotenja. Te aktivnosti naj bi potekale v razvojnem okolju Watson Studio (WS). Upravitelj postavitve ima medtem nalogo, da načrtuje, da bo cevovod strojnega učenja redno potekal oziroma se bo izvajal na zahtevo Watson Machine Learning (WML) ter uporabljal končno točko modela oziroma zunanje aplikacije, ki niso znotraj področja Watson Studia.

Osnovna ideja je, da imamo postopek karseda avtomatiziran, zato smo ustvarili orkestralno skripto, ki v ustreznem vrstnem redu kliče ostale skripte. S tem smo dosegli, da se vsak mesec izvaja načrtovano delo, ki naloži podatke iz podatkovnega skladišča, oceni uspešnost modela, podatkovne znanstvenike obvesti o uspešnosti, naredi napoved v primeru dobre ocene in zapiše rezultat nazaj v podatkovno skladišče. Celoten avtomatiziran proces se lahko izvede tudi ročno, če želimo kakšen korak izvesti drugače oziroma želimo rezultate ob drugem času, kot je bil planiran s planerjem.

Slika 20: Delovni tok celostne rešitve podatkovne znanosti



Vir: lastno delo

V primeru cevovoda strojnega učenja imamo tako dva možna scenarija. V prvem imamo

relativno star model, ki ga je potrebno ponovno učiti, v drugem pa imamo nov model, ki je pripravljen za napovedovanje. V prvem scenariju torej ob zajetju podatkov iz podatkovnega skladišča ugotovimo, da so se ti spremenili, zato je potrebno naš model oceniti še na novih podatkih. Če so rezultati uspešnosti še vedno zadovoljivi, model nespremenjen uporabimo še za napoved in podatke zapišemo nazaj v podatkovno skladišče. Če pa rezultati ocene uspešnosti niso dovolj dobri, se o tem preko elektronske pošte obvesti ustrezne osebe in podatkovni znanstveniki morajo pognati ponovno učenje modela. Če bi prišlo zaradi drastičnih sprememb do povsem drugačnih podatkov, bi v skrajnem primeru morali podatkovni znanstveniki popraviti tudi algoritem. Drugi scenarij je precej krajši. Ob zajemu podatkov iz podatkovnega skladišča ugotovimo, da se ti niso spremenili od zadnjega preverjanja uspešnosti napovedi, zato model ponovno uporabimo za napoved.

4.2.4 Povzetek

Tako kot na koncu prvega šprinta smo tudi tokrat za širše občinstvo podatkovnih znanstvenikov in analitikov banke pripravili predstavitev našega prispevka v zadnjih dveh tednih. Uspelo nam je realizirati vse tri začrtane cilje:

- segmentacija potrošnikov,
- izboljšava modela,
- implementacija avtomatiziranega cevovoda strojnega učenja.

Na predstavitvi smo za občinstvo poleg predstavitve rezultatov pripravili tudi krajši demo, kjer smo pokazali delovanje cevovoda strojnega učenja. Najprej smo pognali cevovod, kjer je bil model učen že dolgo tega, zato je pri preverjanju uspešnosti napovedovanja ta vsem prisotnim na predstavitvi poslal elektronsko pošto z obvestilom o težavi. Nato smo pred njimi izvedli ponovno učenje modela in cevovod ponovno pognali. Ta je pozitivno opravil test uspešnosti in opravil napoved, ki jo je zapisal v podatkovno skladišče. Udeleženci predstavitve so bili z enostavnostjo in hitrostjo naše rešitve zelo zadovoljni, saj so tako pridobili funkcionalnosti, ki je prej niso imeli. Prav tako so bili zadovoljni tudi z rezultati modela, čeprav bi bilo mogoče v prihodnosti z dodatno analizo tega še izboljšati.

Na koncu predstavitve smo izvedli še krajšo delavnico, ki se je navezovala na tretji šprint. S prisotnimi smo namreč diskutirali o želenih delavnicah, ki jih bomo izvedli v okviru prenosa znanja v tretjem šprintu.

4.3 Šprint 3

Tretji šprint je namenjen zaključku projekta. S stranko smo se dogovorili, da bomo šprint razdelili na dve celoti. Prva se bo nanašala na prenos znanja in bolj tehnične predstavitve

cevovoda strojnega učenja ustreznim uporabnikom, druga pa na celosten pristop podatkovne znanosti, kjer bomo s stranko izvedli delavnice, kjer bomo stranki podrobneje razložili delovanje našega napovednega modela ter s stranko delili tudi več primerov uporabe v drugih podjetjih. V tej fazi je potrebno tudi napisati vso dokumentacijo tako za vsebinski kot tehnični del projekta.

4.3.1 Prenos znanja

Za prenos znanja smo organizirali tri štiriurne izobraževalne delavnice. Na prvo smo povabili tako podatkovne znanstvenike, ki bodo analizirali podatke in razvijali model, kot sistemske skrbnike, ki bodo imeli v banki vlogo postavitvenega upravnika. Še enkrat smo jim podrobneje predstavili cevovod strojnega učenja ter jim predstavili, kako se ga upravlja s tehničnega vidika ter jim podali podrobno dokumentacijo za oba scenarija. Naslednji dve delavnici pa sta bili ločeni. Ena je bila za podatkovne znanstvenike, druga pa za sistemske administratorje. Znotraj teh delavnic smo še enkrat podrobno pregledali njihove vloge in naloge, ki jih bodo morali opravljati. Na teh delavnicah smo tudi vsem udeležencem zagotovili prenosnik z dostopom do njihovega pripadajočega okolja, tako da so stvari iz predstavitve lahko preizkusili tudi sami, pri čemer smo jim pomagali, če so se jim pojavile težave.

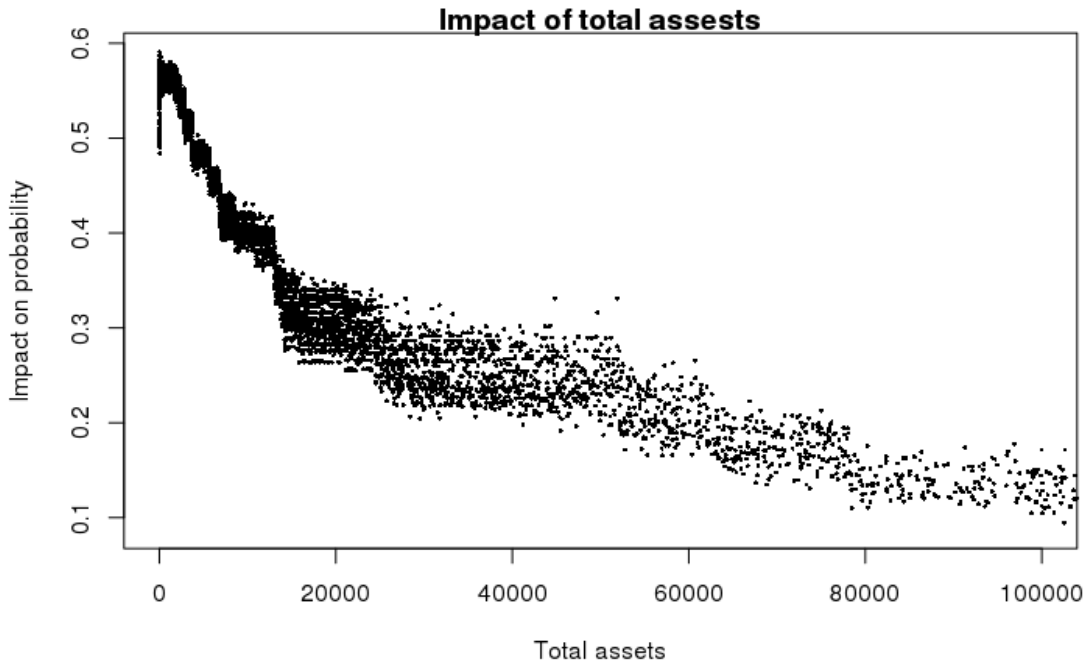
4.3.2 Pristop k podatkovni znanosti

Za predstavitev pravega pristopa k podatkovni znanosti smo organizirali dve celodnevne delavnice. Na prvi smo predstavili metodologijo podatkovne znanosti na splošno, kjer smo dali velik poudarek predstavitvi pomembnosti faz – razumevanje poslovnega problema, razumevanje podatkov, priprava podatkov, modeliranje, preverjanje rezultatov (angl. evaluation), postavitev in pospeševanje (angl. deployment & acceleration).

V naslednji fazi delavnice smo udeležencem predstavili razumevanje modela. Da bi dosegli boljše razumevanje, smo aktivnost razdelili na tri dele – primerjava vseh lastnosti, predstavitev vpliva ene lastnosti na skupen rezultat ter razlaga napovedi modela na podlagi izbranega potrošnika. Pri primerjavi lastnosti smo jim predstavili rezultate za korelacijsko analizo, ki smo jo naredili v drugem šprintu pri izboru lastnosti, kjer smo lastnosti razvrstili po pomembnosti. Tako so lahko videli, katere lastnosti imajo v modelu največjo težo in bistveno vplivajo na odločitev napovedi modela. Pri predstavitvi vpliva ene lastnosti na skupen rezultat smo vizualizirali razmerja med vrednostjo najpomembnejših lastnosti in verjetnostjo, da bo potrošnik vzel posojilo. Tako smo na primer pokazali, kako verjetnost, da bo potrošnik vzel kredit, pada z naraščanjem stanja na bančnem računu v zadnjih treh mesecih, medtem ko verjetnost narašča s številom preteklih potrošnikovih potrošniških posojil. Zanimiv primer, ki smo ga pokazali, je bil tudi vpliv vseh sredstev potrošnika na verjetnost, da bo ta kupil potrošniški kredit. Na začetku ta namreč kratek čas narašča, a se potem krivulja obrne in začne verjetnost padati

z višanjem potrošnikovih skupnih sredstev.

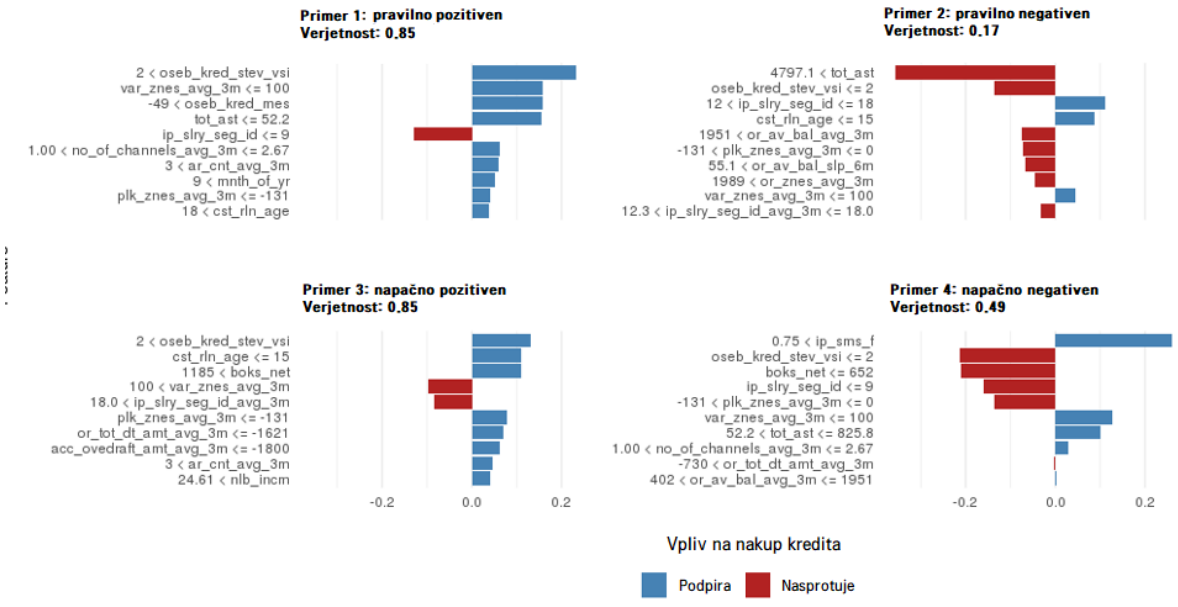
Slika 21: Graf verjetnosti za nakup potrošniškega kredita v odvisnosti od vseh sredstev



Vir: lastno delo

Zadnja aktivnost v tej delavnici je bila razlaga odločitev modela na posameznem potrošniku. Za udeležence smo pripravili štiri primere – pravilno pozitivnega, pravilno negativnega, napačno pozitivnega in napačno negativnega potrošnika, na podlagi katerih smo razložili odločitev modela.

Slika 22: Primerjava štirih različnih napovedi (pravilno pozitivna, pravilno negativna, napačno pozitivna, napačno negativna)



Vir: lastno delo

Na sliki zgoraj so primeri štirih potrošnikov, za katere je model izdelal napoved. Na primeru zgoraj levo vidimo, da večina njegovih vrednosti lastnosti podpira verjetnost, da bo vzel potrošniški kredit, zato je ta relativno visoka s 85 %. Model je napovedal, da bo potrošnik vzel kredit in bil pri napovedi uspešen. V primeru zgoraj desno vidimo, da večina potrošnikovih vrednosti lastnosti nakazuje, da potrošnik kredita ne bo vzel, zato je verjetnost nizka pri 17 %. Model v tem primeru napove, da potrošnik kredita ne bo vzel in je ponovno uspešen. V primeru levo spodaj so vrednosti potrošnikovih lastnosti ponovno precej bolj nagnjene k nakupu kredita in ima verjetnost 84 %. Model ponovno napove, da bo potrošnik kredit vzel, a se tukaj zmoti. Čeprav se je model tukaj zmotil, je napaka razumljiva, saj ima potrošnik zelo podoben profil prvemu, ki je kredit vzel, a je iz nam neznanih razlogov sprejel drugačno odločitev. V primeru desno spodaj je odločitev težja, saj ima kljub močnim argumentom proti jemanju kredita potrošnik kar nekaj lastnosti, ki kažejo na možen nakup le-tega. Na podlagi verjetnosti 49 % model napove, da potrošnik potrošniškega kredita ne bo vzel. Ponovno se zmoti, vendar je napaka iz podobnih razlogov kot v prejšnjem primeru razumljiva, saj je imel model utemeljene razloge za negativno odločitev o nakupu potrošniškega kredita.

Na drugi delavnici smo udeležencem predstavili štiri primere uporabe – pametno postavljanje cen v letalski družbi (angl. smart pricing for airlines), napovedovanje strank, ki bodo zapustile telekomunikacijsko podjetje, prilagajanje ponudbe potrošnikom v argentinski banki in označevanje klicev, v katerih so se potrošniki pritoževali v ameriški banki. Vsebina teh primerov uporabe ne spada v obseg tega magistrskega dela.

4.3.3 Povzetek

Tako kot po prvih dveh šprintih sta tudi tretjemu šprintu sledila povzetek in predstavitev opravljenega dela. Tokrat se sicer nismo osredotočili zgolj na predstavitev tretjega šprinta, ki je bil osredotočen predvsem na pisanje dokumentacije in prenos znanja, pač pa smo pogledali razvoj celotnega projekta v vse treh šprintih in podali oceno. V prvem šprintu nam je uspelo razložiti podatke in dokazati obstoje oziroma neobstoje povezav med poslovnimi hipotezami. Pripravili smo podatke, naredili prvi osnutek napovednega modela in začrtali arhitekturo celovite rešitve, ki smo jo na koncu implementirali. V drugem šprintu smo naredili segmentacijo potrošnikov, izboljšali model in izmerili njegovo učinkovitost pri napovedovanju. V sklopu drugega šprinta smo tudi uspeli implementirati celoten cevovod celovite rešitve. V tretjem šprintu smo zaključili z delom na modelu in znanje o uporabi predali zaposlenim v banki (tako podatkovne znanstvenike kot sistemske administratorje). Naredili smo jim tudi ustrezno tehnično dokumentacijo ter jim podali predloge za nadaljevanje njihove poti na področju podatkovne znanosti. Poleg prenosa znanja smo jim predstavili tudi projekte s strojnim učenjem pri drugih strankah, ki bi jih lahko uporabili v prihodnosti. Za projekt v celoti bi lahko rekli, da smo izpolnili vse predhodno definirane cilje.

Po koncu naše predstavitve smo povabili pred ostale še podatkovna znanstvenika, ki sta

zaposlena pri stranki, da sta z ostalimi delila svojo izkušnjo. Pohvalila sta način dela ter način komuniciranja ter sodelovanja skozi celoten projekt ter izrazila pozitivno tehnično izkušnjo z Watson Studiom. Glede na videno so bili pozitivno presenečeni tudi ostali prisotni na tej predstavitvi, ki v času pisanja te seminarske naloge že aktivno uporabljajo pripravljeno platformo in predstavljeno metodologijo. Strinjali so se, da je bil to dober začetni projekt, ki jim bo dobra iztočnica za implementacijo in vzdrževanje ostalih modelov, ki jim bodo pomagali postati karseda k stranki usmerjena banka (angl. customer oriented).

4.3 Tehnični izzivi pri projektu

V tem podpoglavju bom opisal tehnične težave, s katerimi smo se soočili v sklopu celotnega projekta, vendar te vsebinsko niso spadale v noben šprint. Gre za večinoma tehnične zadeve, ki smo jih dali v rubriko delo v ozadju (angl. backolog).

Že na začetku projekta smo imeli težave z opremo. Zaradi strankinih varnostnih zadev se s svojimi prenosniki nismo mogli povezati s sistemom, zato nam je bilo potrebno zagotoviti dostope in prostor v podjetju, kar je trajalo nekaj ur. Prav tako sta dva naša podatkovna znanstvenika imela v prvem tednu težave z uporabo njihovih tipkovnic, saj sta prihajala iz tujine in nista bila vajena slovenske opreme. Od drugega tedna naprej sta imela svojo opremo.

Naslednja težava se nam je pojavila pri nameščanju tako python kot R knjižnic. Banka ima zaradi varnostnih razlogov namreč onemogočen dostop do zunanjega interneta na strežnikih, kjer je nameščen Watson Studio. Začasno smo rešili zadevo tako, da smo v python skriptah za nameščanje dodali ustrezen parameter, ki obide proxy; v RStudiu smo lokalno naložili paket, ki smo ga predhodno naložili z interneta in nato iz njega naredili namestitev. Rešitev stranki zaenkrat odgovarja, vendar smo predlagali, da v prihodnosti vzpostavijo lokalen repozitorij knjižnic, ki jih bodo potem lahko poljubno nameščali.

Na začetku projekta so se začele pojavljati težave z uporabo tehnologije za verzioniranje - git. Shranjevanje popravkov in nalaganje le-teh s strežnika je včasih vzelo ogromno časa oziroma se včasih sploh ni izvedlo do konca. Ko analizi konfiguracije okolja Watson Studio smo ugotovili, da je za to kriva napačna konfiguracija, saj je z vsako operacijo mogoče naložiti in prenesti vsebino celotnega repozitorija in ne zgolj sprememb. Ker je bil repozitorij zelo velik (že na začetku skoraj 10 GB) in povezava zgolj povprečna, je bil čas nalaganja in prenašanja zelo dolg. Po spremembi konfiguracije do težav ni več prihajalo.

V obdobju modeliranja so se začele pojavljati performančne težave. Okolje je postajalo neodzivno; operacije so potrebovale nadpovprečno veliko časa, da so se izvedle ipd. Pri diagnostiki težav smo stopili v stik tudi z razvojno ekipo produkta, skupaj s katero smo na koncu ugotovili, da je težava v prezasedenosti virov, saj je istočasno teklo zelo veliko sej. Ugotovili smo, da so odprte seje stare tudi več kot pol leta, zato smo večino sej po

posvetu z odgovornimi na banki zaprli. Da se to v prihodnje ne bi ponovilo, so odgovorni v banki vse uporabnike okolja obvestili, da morajo seje po končani uporabi zapreti, saj okolje za to ne skrbi samo. Za ta korak smo jim poslali tudi navodila.

Na težavo smo naleteli tudi pri postavitvi rešitve (angl. deployment), ko smo želeli postaviti naš XGBoost model v Watson Machine Learning. Ugotovili smo, da verzija, ki jo ima stranka nameščeno, ne podpira tega algoritma, ustvarjenega s caret knjižnico v R-u oziroma scikitlearn v pythonu. Težavi smo se izognili tako, da smo v Watson Machine Learning implementirali zgolj plansko skripto. Slednja nato kliče XGBoost model v Watson Studio Local. Da bi se izognili lokalni kopiji modela, bi lahko naredili API in model klicali prek njega.

4.4 Vpeljevanje podatkovne znanosti v podjetje

Čeprav smo potek projekta in aktivnosti vnaprej dobro definirali in strukturirali, smo se v teku projekta vseeno soočali z izzivi, ki jih pred pričetkom projekta nismo pričakovali. V tem poglavju se bom osredotočil na te izzive in ugotovitve, ki smo jih v času projekta pridobili.

Uvodna oblikovna delavnica se je izkazala za zelo uspešen pristop k spoznavanju uporabnikov, na katere se bo navezovala naša rešitev, njihovih počutij, aktivnosti, težav, s katerimi se soočajo, in njihovimi željami ter pričakovanji za naš projekt. Poznavanje uporabnikov in njihovih želja je zelo pomemben dejavnik uspešnosti takega projekta, saj mora biti tako projekt kot končna rešitev prilagojena njihovim željam in potrebam, kar kasneje vpliva tudi na uporabniško izkušnjo in zadovoljstvo stranke. Na podlagi poznavanja uporabnikov lahko tudi lažje obvladujemo njihova pričakovanja v zvezi s projektom, kar je pri pilotnih projektih podatkovne znanosti še toliko večjega pomena. Pogosto imajo namreč ljudje, ki ne poznajo tako dobro področja podatkovne znanosti, prevelika pričakovanja, kar na koncu pomeni razočaranja. Na drugi strani pa tudi uporabniki, ki so v preteklosti že doživeli razočaranje na področju podatkovne znanosti in zaradi pomankanja zaupanja v projekt, ne pristopijo k zadevi tako resno, kot bi sicer. Tukaj je bila naša naloga, da smo take posameznike dodatno motivirali in jim predstavili potencialne prednosti, ki jih bo projekt prinesel. Bistvenega pomena za uspešnost tega projekta je bilo tudi sodelovanje naročnika projekta in strankinega vodje oddelka za podatke, ki se je trudil, da je širil dobro klimo in spodbujal pozitivno naravnanost k pilotnemu projektu.

Na uvodni oblikovni delavnici so nam zaposleni pri stranki povedali, da so pred nami že najeli skupino podatkovnih znanstvenikov pri konkurenčnem podjetju, a jim proces vpeljave podatkovne znanosti ni uspel. Čeprav so se tako kot mi tudi oni lotili vpeljave podatkovne znanosti s pomočjo pilotnega projekta, pa z njihovo storitvijo naročnik ni bil zadovoljen. Po strankinem mnenju je bilo za rabo napovednih modelov potrebno preveč ročnega dela; rezultati modela niso prinesli drastičnih sprememb poslovne uspešnosti. Hkrati zaposleni niso dobili dovolj tehničnega znanja, da bi v prihodnosti gradili oziroma

izboljšali obstoječi model. Stranko smo na tej točki vprašali, ali bi želeli, da nadaljujemo tam, kjer so s prejšnjim izvajalcem ostali glede na to, da je bilo kar nekaj dela že opravljenega. S tem se niso strinjali, saj so želeli sodelovati pri postopku grajenja pilotskega modela od začetka do konca, da bi s tem pridobili čim več znanja. S tem smo se strinjali tudi mi, zato smo se odločili, da začnemo celoten proces na začetku. Obstoječ model nagnjenosti k nakupu potrošniškega posojila lahko uporabimo na koncu za primerjavo z našim. Primerjava sicer ni povsem pravična, saj je obstoječ model star že približno eno leto in podatki niso več tako relevantni. A to je vseeno na nek način pokazatelj, ali smo izbrali dober pristop k našemu modelu ali ne. Na koncu se je odločitev za grajenje od začetka izkazala za uspešno, saj smo uspeli zgraditi bolj učinkovit model. Poleg tega je bila tudi stranka s pridobljenim znanjem bistveno bolj zadovoljna. Bistvenega pomena za dosego uspešnosti na takem projektu je bilo tudi to, da smo se bolj kot na učinkovitost modela osredotočili na njegovo interpretativnost. V pilotskem projektu gre namreč v veliki meri za učenje in razumevanje bodočih uporabnikov in razvijalcev, zato je pomembno, da je model razumljiv in je njegove napovedi mogoče logično obrazložiti. Teoretično bi lahko naredili model z natančnejšimi napovedmi, a če pri tem strankini zaposleni ne bi pridobili razumevanja dela, bi bila precej velika verjetnost, da bi imeli večje težave pri gradnji prihodnjih modelov, kot jih bodo imeli sedaj.

Najpomembnejša aktivnost, ki jo podatkovni znanstveniki pogosto zanemarjajo pri projektih strojnega učenja, je razmislek o celostni rešitvi. Ta lahko vzame veliko časa, vendar je slabo zastavljena rešitev najpogosteje glavni razlog za neuspel projekt. Čeprav so se s tem strinjali vsi udeleženci oblikovne delavnice, so hkrati priznali, da zaradi kratkih časovnih omejitev tudi sami pogosto delajo enake napake. Ideja strankine ekipe je bila, da bi razvit model poganjali na vsake 3 mesece; ta bi vračal uporabnike, ki naj bi čez dva meseca vzeli kredit. Banka tako dobi dovolj časa, da lahko k zadevi ustrezno pristopi. Glede na to, da je imela stranka že prej razvit, a sicer ne najbolj učinkovit model, smo imeli tudi dobro izhodiščno merilo, kako uspešen je naš model.

Preden smo se lotili prvega šprinta, smo stranko prosili, da njihovi vodilni pri izvajanju projekta stopi nazaj in prepusti celotno vodenje projekta nam. Zato smo se tudi odločili, da na koncu vsakega od šprintov naredimo predstavitev naših rezultatov, ki ji je sledila diskusija, kjer so tako vodilni kot posamezniki, ki jih projekt zadeva, izražali svoja mnenja in ideje, na podlagi katerih je nato lahko prišlo do sprememb v projektu in projektnem načrtu.

Znotraj prvega šprinta nam je podatkovni znanstvenik, zaposlen pri stranki, pripravil nabor podatkov, nad katerimi smo nato izvedli analizo in jih uporabili pri učenju modela. Ker so pred enim letom pri projektu podatkovne znanosti že sodelovali z drugim podjetjem, so znanje in informacije o relevantnih podatkih že imeli ter nam pripravili relativno dobro strukturiran nabor podatkov, ki je temeljil na naboru podatkov, ki so jih uporabili pred enim letom. Razlika je bila sicer ta, da so bili naši podatki bolj aktualni. Sčasoma smo sicer ugotavljali, da nekaterih podatkov ne potrebujemo, a bi bilo nekatere

potrebno še dodati, zato je proces zbiranja nabora podatkov potekal iterativno. Proces bi imel še veliko več iteracij, če stranka ne bi imela predhodno definiranega nabora in bi celoten postopek morali izvesti od začetka.

Razumevanje poslovnega ozadja in podatkov je na področju podatkovne znanosti ključnega pomena. Da bi podatke bolje razumeli, smo se pri analizi podatkov močno opirali na vizualizacijo. Tako smo si lažje predstavljali razporeditve vrednosti spreminjanja podatkov glede na trende ipd. Vizualizacija nam je pomagala tudi pri predstavitvi rezultatov naročnikom in kasnejšim uporabnikom, za katere bo razumevanje vsebine podatkov v prihodnjih projektih igralo ključno vlogo.

Pri razumevanju podatkov smo naleteli tudi na izziv, saj nekaterih podatkov iz nabora nismo mogli dobiti. Na podlagi pogovorov s strankinimi vsebinskimi strokovnjaki smo namreč ugotovili, da potrošniki najpogosteje vzamejo posojilo zaradi konsolidacijskega dolga, nujnih stroškov, osebnih dogodkov, prenove doma ali začetka oziroma širjenja podjetja. Večine od teh razlogov pa z našim naborom podatkov nismo mogli preveriti, saj stranka nima podatkov o takih dogodkih. Na podlagi davčne številke nam je sicer uspelo priti do informacij o tem, kdo si lasti manjše podjetje, kar je na nek način izpolnilo kriterije za en pogoj, ostalih pa nismo mogli dobiti. Ker gre za ključne informacije, smo stranki svetovali, naj v prihodnje poskuša odkriti način, kako bi lahko pridobili vsaj kakšno od teh lastnosti, saj bi lahko to bistveno izboljšalo natančnost modela, če bi se izkazalo, da gre za res tako velik vpliv, kot so nam povedali strankini vsebinski strokovnjaki.

V času projekta se je s strani našega poslovnega partnerja, ki je pri stranki namestil razvojno okolje, pojavil dvom o načinu dela, saj smo vsi podatkovni znanstveniki na projektu dobili administratorske pravice v okolju Watson Studio. To po njihovem mnenju namreč ne izpolnjuje varnostnih zahtev in začetnega dogovora, da imajo administratorske pravice zgolj sistemski skrbniki. Po posvetu s stranko so se odločili, da v času projekta vsi podatkovni znanstveniki obdržimo administratorske pravice; po koncu projekta se bodo dogovorili, komu se bo kaj omogočilo za potrebe vzdrževanja rešitve.

S prvim šprintom so bili tako vodilni kot kasnejši uporabniki zadovoljni. Slednji so bili sicer presenečeni nad nizko stopnjo potrjenih hipotez. Tukaj je bilo pomembno, da smo že na začetku poskušali obvladovati njihova pričakovanja, da razočaranje ni bilo preveliko. Izkazalo se je tudi, da je za projekte podatkovne znanosti pomembno, da delamo karseda malo predpostavk in se opiramo zgolj na preverjena dejstva. Na diskusiji po prvem šprintu so strankini zaposleni izrazili tudi zanimanje za način komuniciranja med projektom; tehničnih ter vsebinskih pripomb niso imeli, kar je bilo glede na to, da je končna rešitev šele v delu, tudi pričakovati. Vseeno pa smo opazili, da je zaposlenim način dela in komunikacije na projektu zelo pomemben, kar je gotovo ena od lastnosti, ki jih je dobro upoštevati pri oblikovanju metodoloških pristopov ne le za podatkovno znanost, pač pa tudi projekte na drugih področjih.

V drugem šprintu smo ugotovitve v prvem morali uporabiti pri gradnji modela in cevovoda celotne rešitve. Glavni izziv, ki se nam je pojavljal, je bil dobra izbira pravega razmerja med enostavnostjo in generalizacijo rešitve. Glede na to, da so prav zaradi zapletenosti procesa pri prejšnjem poskusu uvedbe podatkovne znanosti obupali, smo na eni strani želeli narediti karseda enostaven proces za uporabnika s čim več avtomatizacije, medtem ko smo morali na drugi strani upoštevati robustnost rešitve in rešitev, ki bo primerna za različne vrste modelov in bo nadzorniku vseeno dopuščala dovolj veliko mero nadzora. Na koncu drugega šprinta smo že imeli delujočo rešitev, zato je bilo na predstavitvi šprinta poleg predstavljene funkcionalnosti pomembno tudi to, da smo se s prisotnimi pogovorili o njihovih željah in potrebah v tretjem, zadnjem šprintu. Namen zadnjega šprinta je namreč prenos znanja še na ostale uporabnike, predstavitev nadaljnjih možnosti in predvsem razumevanje cevovoda strojnega učenja ter koncepta podatkovne znanosti v celoti. Tako s prisotnimi vodilnimi kot s kasnejšimi uporabniki smo se dogovorili za delavnice v tretjem šprintu, ki bodo vključevale tako vsebinski kot tudi tehnični del podatkovne znanosti, kjer bomo predstavili zmožnosti Watson Studio platforme, na kateri bodo v prihodnje gradili svoje projekte podatkovne znanosti.

Strankina podatkovna znanstvenika sta imela že v teku prvih dveh šprintov številna tehnična vprašanja, kot so, kako naložiti podatke iz podatkovnega skladišča neposredno v Watson Studio, kako lahko skrijejo podatke za prijavo v podatkovno skladišče, kje lahko vidijo odprte seje, shranjujejo svoje konfiguracije ipd. Na ta vprašanja smo jim v teku projekta odgovarjali in jih učili pravilne uporabe. Pogosto je šlo tudi za situacije, ko odgovora nismo poznali niti sami, zato smo morali pogledati v dokumentacijo orodij.

5 DISKUSIJA

Vpeljevanje podatkovne znanosti v podjetje ni enostaven proces (angl. straight forward). Za uspešno izpeljavo pilotnega projekta morajo podjetja oziroma projektne ekipe že na začetku zasledovati merila uspešnosti, na podlagi katerih se na koncu oceni uspešnost le-tega. Da bi zagotovile prikaz poslovne vrednosti, morajo projektne ekipe pilotni projekt zasnovati tako, da podjetju nazorno prikažejo ugotovitve v času projekta ter predstavijo, kaj lahko z njimi naredijo. V našem projektu je bilo to vidno predvsem pri vizualnih predstavitvah, ko smo ključnim deležnikom podjetja na predstavitvah s pomočjo grafov in vizualnih izrisov potrdili oziroma zavrgli vnaprej postavljene hipoteze, na podlagi katerih se bo podjetje v prihodnje lahko precej bolj smiselno lotilo oglaševanja potrošniških kreditov. Pri oglaševanju jim bo pomagala tudi segmentacija strank, ki je prej niso imeli oziroma niso vedeli, kako bi jo najbolj smiselno naredili. Za zagotovitev identifikacije dejavnikov dodane vrednosti morajo projektne skupine v okviru celotnega projekta spremljati in voditi dogajanje ter opozarjati na pomen posameznih dejavnikov. V našem projektu se je to videlo predvsem pri poudarjanju pomembnosti analize in vizualizacije podatkov, na podlagi katerih smo na koncu tudi dobili veliko informacij in dodane vrednosti. To je bilo občasno zahtevno, saj so strankini zaposleni želeli

nadaljevati z ostalimi fazami in tej niso pripisovali velike pozornosti.

Kot smo zapisali že v teoretičnem delu, je poleg poslovne vrednosti in njenih dejavnikov za uspešno izpeljavo pilotnega projekta podatkovne znanosti tudi pomembno, da se podjetje zaveda svojih tehničinih zmožnosti in se zaveda, kaj vse mu tehnologija omogoča. Da bi stranko seznanili z vsemi možnostmi, ki jih podatkovna znanost ponuja, smo organizirali predstavitev podobnih primerov podatkovne znanosti še v drugih podjetjih. Pomembno in ključno za uspešnost v nadaljnjih projektih podatkovne znanosti se mi zdi, da podjetje razume, kaj jim podatkovna znanost vse omogoča in kje vse jo je mogoče uporabiti. Gre namreč za vedo, ki se je dobro razvila šele v zadnjih 10 letih in podjetja velikokrat niti ne vedo, kako bi jo uporabila. Pomembno je tudi, da podjetja razumejo, s kakšnimi problemi se lahko pri takih problemih soočajo ter kako odreagirati nanje. Zadevati se morajo, da se uvedba podatkovne znanosti ne zgodi čez noč; celoten proces ni enostaven (angl. straight forward). Pri ozaveščanju je pomembno tudi, da deležnikom predamo njim potrebno znanje, ki ga bodo uporabljali v nadaljnjih projektih podatkovne znanosti. Mi smo se odločili, da bomo znanje predali s pomočjo stvarnega učenja (angl. hands-on) na delavnicah. Tako so lahko deležniki s pomočjo testnih primerov z našo pomočjo spoznavali in se naučili tehnik in postopkov, ki jih bodo kasneje uporabljali. Ta način prenosa znanja se mi zdi dober in učinkovit, saj lahko uporabnik zadevo, ki mu jo poveš, tudi sam preizkusi in si zato lažje zapomni oziroma lahko v primeru nejasnosti takoj dobi pomoč. Zdi se mi, da to predstavlja veliko prednost v primerjavi z delavnicami, kjer lahko udeleženci snov zgolj poslušajo ali pa o njej berejo v dokumentaciji.

Za uspešnem pilotni projekt je pomembna tudi implementacija delujočega koncepta metodološkega pristopa, definirana v teoretičnem delu naloge. Da bi podjetja razvila dober koncept metodološkega pristopa, morajo pri razvoju upoštevati mnenja kasnejših uporabnikov, biti pozorna na enostavnost uporabe ter poskrbeti, da rešitve delujejo dovolj dobro na infrastrukturi, ki je na voljo, in uporabljajo podatke, ki bodo na voljo tudi v prihodnje. V našem primeru smo se osredotočili na uporabniku prijazno rešitev, saj to poveča verjetnost, da bo slednjo ta uporabljal. Na oblikovni delavnici smo od deležnikov izvedeli, da se jim zdi zelo pomembno, da je rešitev enostavna in ne predstavlja veliko ročnega dela pri poganjanju modela. Da bi dosegli enostavnost rešitve, smo se odločili, da našo rešitev karseda avtomatiziramo, ne da bi pri tem izgubili funkcionalnost. Za dosego cilja smo uporabili tudi skripte, ki so ob določenih časih glede na izpolnjene pogoje klicale ustrezne ukaze. Vse delo se prav tako lahko še vedno izvede ročno, vendar je avtomatizacija prinesla zelo pozitivno povratno informacijo s strani stranke. Avtomatizacija je proces, ki ga priporočam tudi drugim podjetjem, ki se bodo v prihodnosti lotevala takšnih projektov, saj tako ne prihranimo zgolj časa, pač pa tudi močno izboljšamo uporabniško izkušnjo.

Eden od dejavnikov uspešnosti koncepta rešitve je tudi ta, da ta dobro deluje na obstoječi infrastrukturi. Da bi to dosegli, smo že na začetku podrobno načrtovali rešitev, pri kateri smo celotno funkcionalnost lahko naredili s komponentami znotraj strankine

infrastrukture, namenjene temu projektu oziroma projektom podatkovne znanosti. S tem smo se izognili kasnejšim večjim zapletom, kjer bi lahko ugotovili, da naše okolje nečesa ne podpira, kar bi pomenilo, da mora stranka kupiti dodatno strojno oziroma programsko opremo. Kljub načrtovani rešitvi smo sicer v določeni točki ugotovili, da strankina verzija ne podpira želenega modula. Težavo smo (podobno kot avtomatizacijo) rešili z uporabi skripte. V tej fazi se je izkazalo, kako pomembno je dobro predhodno načrtovanje rešitve. V našem primeru se je sicer dobro izšlo, vendar bi v primeru, da s skripto ne bi mogli rešiti zadeve, stvar lahko precej zapletla projekt. Za podjetja je pomembno, da pri projektih podatkovne znanosti delajo karseda malo predpostavk in vsako stvar predhodno preverijo in se s tem izognejo morebitnim poznejšim zapletom.

Pri rešitvi smo morali uporabiti podatke, ki so stranki ves čas dostopni in lahko redno pridobivajo posodobljene vrednosti. Da bi to dosegli, smo uporabili neposredne podatke iz njihovega obstoječega podatkovnega skladišča, kjer stranka vsak dan prejema posodobljene aktualne podatke. Rešitev je bila za nas sicer najlažja, vendar pa obstaja na tem področju še kar nekaj prostora za izboljšave. Glede na to, da nam posamični podatki velikokrat ne povejo veliko, bi lahko naredili še dodatne izpeljanke podatkov, ki bi lahko še dodatno izboljšale model. Še večji napredek bi lahko dosegli, če bi našli način, kako pridobiti še podatke, ki jih podjetje trenutno nima in jih ne more izpeljati niti iz obstoječih podatkov. Mislim, da je za vsa podjetja, ki se lotevajo podatkovne znanosti, zelo pomembno, da poskrbijo za dobre in kvalitetne podatke, saj so ti ključnega pomena za dobre rezultate tako napovednih modelov kot kasnejše poslovne uspešnosti.

Tabela 1: Merila uspešnosti in koraki, ki smo jih izvedli, da bi jih dosegli

AKTIVNOSTI V PROJEKTU	MERILA USPEŠNOSTI
Vizualna predstavitev skupin uporabnikov in ugotovitev, povezanih s predhodno postavljenimi hipotezam	**Prikaz poslovne vrednosti**
Vodenje projektnih faz in opozarjanje na pomembnejše dejavnike dodane vrednosti	**Identifikacija dejavnikov dodane vrednosti**
Organizacija delavnic s stvarnim učenjem (angl. hands-on)	**Prikaz tehničnih zmožnosti**
Predstavitev primerov uporabe podatkovne znanosti v drugih podjetjih	**Prikaz poslovnih možnosti podatkovne znanosti**

Avtomatizacija rešitve	**Enostavna uporaba rešitve**
Podrobno načrtovanje rešitve; uporaba skript	**Delovanje na obstoječi infrastrukturi**
Uporaba podatkovnega skladišča	**Uporaba razpoložljivih podatkov**

Vir: lastno delo

Za uspešno izpeljavo projekta je poleg zasledovanja meril uspešnosti – tako kot smo zapisali že v teoretičnem delu naloge – pomembno tudi upoštevanje ključnih dejavnikov uspešnosti. Če želijo uspešno izpeljati pilotni projekt podatkovne znanosti, morajo podjetja imeti zagnanega in motiviranega sponzorja podjetja, ki skrbi za pozitivno naravnanost k podjetju. V našem projektu smo imeli srečo, saj je bil sponzor projekta zelo zainteresiran za področje podatkovne znanosti in je dal projektu tudi veliko prioriteto. Zagotovil je okolje in pogoje dela, da smo lahko nemoteno razvijali rešitev. Prisoten je bil pri na vseh srečanjih s širšimi deležniki, kjer je podajal svoja pričakovanja in povratne informacije o dosedanjem delu, na podlagi katerih smo nato naprej razvijali rešitev. Če bi se sponzor projekta po začetku projekta umaknil in ne bi bil več prisoten, bi se projekt dogajal precej počasneje. Hkrati bi bila večja verjetnost, da bi se projekt začel dogajati v napačno smer. Skrbel je tudi za pozitivno naravnanost k podatkovni znanosti celotnega podjetja, kar je močno pripomoglo tudi pri drugem kritičnem dejavniku – odprtost deležnikov za spremembe. Čeprav je sponzor s svojim zgledom in spodbujanjem podatkovne znanosti deležnike delno že pripravil k nagnjenju k spremembam, pa so bili tej v fazi, ko smo začeli s projektom, še vedno nekoliko negotovi. Pomemben dejavnik, ki ga morajo podjetja zagotoviti, je tudi sprejemanje sprememb s strani deležnikov. Da podjetja to dosežejo, morajo v pilotnem projektu prisluhniti deležnikom, se z njimi pogovoriti in na podlagi ugotovitev razvijati rešitev. Tako lahko podjetje lažje pridobi njihovo zaupanje, ki v nadaljevanju igra pomembno vlogo pri uspešnosti vpeljevanja podatkovne znanosti v podjetje. V našem primeru smo za zagotavljanje zaupanja deležnikov na uvodni oblikovni delavnici deležnikom predstavili način dela v pilotskem projektu in z njimi preko krajših delavnic oblikovali njim prilagojeno rešitev. Deležniki so bili na začetku negotovi, saj si niso dobro predstavljali, kakšna bo na koncu rešitev in so predvsem zaradi negativnih izkušenj v preteklosti mislili, da jim rešitev ne bo prinesla pozitivnih poslovnih rezultatov, pač pa bo predstavljala zgolj novo breme. Po končani delavnici so deležniki začeli verjeti v uspeh projekta, saj se je predlagana rešitev prilagajala njihovim željam in potrebam; hkrati smo sproti na delavnici reševali dvome in potencialne težave, na katere so pomislili. Tretji ključni dejavnik je razpoložljivo znanje vodenja podjetja in razvoja rešitve. Podjetje, ki šele začenja z vpeljavo podatkovne znanosti mora v projektni skupini zagotoviti vsaj enega ali dva strokovnjaka na podlagi podatkovne znanosti, ki imata že izkušnje s

podobnimi pilotskimi projekti, saj se je brez izkušenj precej težje izogniti in rešiti iz težav, ki se pri takem projektu lahko pojavijo. V našem projektu sta bila v naši ekipi podatkovnih znanstvenikov dva člana, z večletnimi izkušnjami na področju pilotnih projektov podatkovne znanosti, kar nam je pomagalo pri načrtovanju dejavnosti. Tako smo se tudi lažje izognili morebitnim pogostim napakam in težavam.

Pilotskega projekta smo se lotili z namenom pomoči vpeljave podatkovne znanosti v podjetje. Glede na to, da je bil cilj projekta postaviti metodološki pristop za vse naslednje projekte, se bo prava uspešnost pilotnega projekta pokazala šele čez nekaj časa, ko se bo podjetje začelo ukvarjati še z drugimi modeli in ko se bo pokazalo, kakšne poslovne prednosti jim bo sploh prinesel obstoječ model. Lahko pa z gotovostjo trdim, da sem z izpeljavo pilotnega projekta zadovoljen, saj ne le da smo uspeli uresničiti vse tehnične cilje, pač pa smo tudi uspeli na strankine zaposlene prenesti znanje, ki ga prej niso imeli, kar so nam potrdili tudi sami. Za uspešno izpeljavo projekta je bilo ključno, da smo si v začetku vzeli čas tako za spoznavanje uporabnikov ter prepoznavanje potreb in želja kot za pripravo in analizo podatkov ter načrtovanje celotnega cevovoda rešitve. Če bi katerega od teh segmentov slabo izvedli, bi kasneje lahko prišlo do težav, ki bi nam preprečili uspešno dokončanje projekta v naprej zastavljenem času.

SKLEP

Iz leta v leto narašča količina podatkov, ki jih imajo podjetja na voljo; hkrati narašča tudi pomembnost podatkovne znanosti, ki danes že predstavlja področje, ki pogosto vpliva na to, ali bo podjetje poslovalo uspešno ali ne. Zaradi porasta pomena podatkovne znanosti se danes čedalje več podjetij odloča, da se s tem želijo ukvarjati tudi v sama. Vpeljevanje podatkovne znanosti v podjetje pa ni enostaven proces in podjetja velikokrat ne vedo, kako bi k zadevi pristopila za dosego najboljši rezultatov. Pogosto se namreč zgodi, da zaradi napačnega metodološkega pristopa vpeljevanje podatkovne znanosti postane neuspešno in podjetja to področje opustijo.

Namen magistrskega dela je bil na podlagi realnega primera oblikovati predlog učinkovitega pristopa k vpeljevanju podatkovne znanosti v podjetje, ki je na tem področju novo in si želi postati bolj podatkovno usmerjeno, kar nam je tudi uspelo. S pomočjo analize projekta vpeljave podatkovne znanosti v podjetje na primeru napovednega modela smo identificirali potencialne težave in možne rešitve ter dejavnike, ki pozitivno vplivajo na uspešnost vpeljave podatkovne znanosti v podjetje. Magistrsko delo smo razdelili na dva dela. V prvem smo predstavili teoretično podlago za podatkovno znanost in pilotne projekte na njenem področju. Podatke in informacije v tem delu smo večinoma pridobili s pregledom literature; delno pa ti slonijo tudi na lastnih izkušnjah. V drugem delu magistrske naloge sem se osredotočil na praktični del in analizo primera vpeljevanja podatkovne znanosti v slovensko podjetje, na podlagi katerega sem kasneje oblikoval tudi predloge in nasvete za podjetja, ki s podatkovno znanostjo šele začenjajo. V analizi sem se opiral na teoretično podlago, predstavljeno v prvem delu magistrskega dela in kritično ovrednotil ter ocenil, katerih stvari

smo se v projektu držali in katerih ne. Pri delu nisem imel veliko omejitev, saj sem kot član ekipe pri vpeljevanju pilotnega projekta sodeloval pri celotnem procesu in imel vpogled v vse dokumente, povezane z izvedbo projekta.

Kot je omenjeno že zgoraj, je bil osrednji cilj magistrskega dela analiza projekta vpeljave podatkovne znanosti na primeru napovednega modela, kar nam je tudi uspelo doseči. Lahko bi rekli, da magistrsko delo predstavlja zaključeno celoto in dobro začetno gradivo za vsa podjetja, ki začenjajo svojo pot na področju podatkovne znanosti in bi si želela postati bolj podatkovno usmerjena. Analiza uporabe teoretične podlage na praktičnemu primeru uvajanja podatkovne znanosti v podjetje daje največjo dodano vrednost prav podjetjem, ki so na področju podatkovne znanosti nova oziroma se na tem področju še niso najbolje znašla in potrebujejo pomoč pri postavitvi pravih metodoloških temeljev. Naš prispevek je koristen predvsem za managerje in vodje oddelkov, ki skrbijo za vpeljavo podatkovne znanosti, saj smo se osredotočili bolj na metodološki pristop in managerski vidik in ne toliko na tehnično vsebino. Podjetja lahko magistrsko nalogo uporabijo kot dobro teoretično osnovo pri njihovih začetkih podatkovne znanosti; obenem jim naloga nudi tudi dober šolski primer vpeljevanja podatkovne znanosti v podjetje, na podlagi katerega lahko ta gradijo svoje pilotske projekte podatkovne znanosti. Predstavljen projekt namreč ne vključuje zgolj projektnega načrta, pač pa tudi predstavi ovire in izzive, ki so se v projektu pojavljali in reševali.

Čeprav magistrsko delo predstavlja zaključeno celoto, bi lahko to še dodatno razširili. V nadaljevanju bi lahko dodali še opažanja in rezultate, ki bi jih podjetje pridobilo v daljšem časovnem obdobju. Tako bi ugotovili, kako uspešen je bil v resnici naš pilotski projekt in kakšne rezultate je prinesel. Na podlagi tega bi lahko podjetjem, ki začenjajo na področju podatkovne znanosti, podali še več predlogov za učinkovitejši pristop k projektu. Magistrsko delo bi v nadaljevanju lahko razširil tudi tako, da bi izvedli analize različnih primerov vpeljevanja in nato primerjali rezultate, ugotovitve in različne načine pristopov k projektu. Podobno kot pri dodajanju dolgoročnih rezultatov bi nam tudi ta primerjava omogočila, da bi prišli do dodatnih spoznanj, na podlagi katerih bi lahko še boljše svetovali podjetjem, ki se lotevajo podatkovne znanosti.

LITERATURA IN VIRI

1. Aguilar-Ruiz, J. S., Riquelme, J. C. & Toro, M. (2003). Evolutionary learning of hierarchical decision rules. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *33*(2), (str. 324–331).

2. Arthur, D. & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, *07-09-January-2007*, (str. 1027–1035). New Orleans: Society for Industrial and Applied Mathematics

3. Balatsko, M. (2019). All you want to know about preprocessing: Data preparation. *Medium*. Pridobljeno 25. marca 2020 iz https://towardsdatascience.com/all-you-want-to-know-about-preprocessing-data-preparation-b6c2866071d4

4. Barrueta-Meza, R., Castillo-Villarreal, J. P. & Armas-Aguirre, J. (2018). Predictive Model to Determine Customer Desertion in Peruvian Banking Entities. *2018 Congreso Internacional de Innovacion y Tendencias En Ingenieria, CONIITI 2018 - Proceedings*, (str. 1–5). Bogota: Universidad Catolica de Colombia

5. Bendheim, C. L., Waddock, S. A. & Graves, S. B. (1998). Determining best practice in corporate-stakeholder relations using data envelopment analysis: An industry-level study. In *Business and Society* (Vol. 37, Issue 3).

6. Blais, O. (2019). The Hitchhiker's Guide to a Successful Data Science Practice. *Moov AI*. Pridobljeno 25. marca 2020 iz https://medium.com/@ODSC/the-hitchhikers-guide-to-a-successful-data-science-practice-7e65bb979f6a

7. Crossman, Ashley. (2020, February 11). Pilot Study in Research. Pridobljeno 25. marca 2020 iz https://www.thoughtco.com/pilot-study-3026449

8. Data Talent. (2019). Data Strategy – A Roadmap for Successful Implementation of Data Science. *Data Talent*. Pridobljeno 4. junija 2020 iz https://www.datatalent.io/blog-post?ID=3

9. Davenport, T. H. & Bean, R. (2019). Companies Are Failing in Their Efforts to Become Data-Driven. *Harvard Business Review*, (str. 1–5).

10. Davenport, T. H. & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, *90*(10), 5.

11. Dean, J. & Ghemawat, S. (2010). Map Reduce: A flexible data processing tool. *Communications of the ACM*, *53*(1), 72–77. https://doi.org/10.1145/1629175.1629198

12. Deoras, S. (2017). Business Intelligence vs Data Science. *Analytics India Magazine*. Pridobljeno 29. junija 2020 iz https://analyticsindiamag.com/business-intelligence-different-data-science/

13. Dhar, V. (2012). Data Science and Prediction Vasant Dhar Professor, Stern School of Business Director, Center for Digital Economy Research. *Communications of the ACM*, *May*, 64–73.

14. Duggan, J. (2019). *When to Implement Data Science Into Your Business*. Cognetik. Pridobljeno 4. junija 2020 iz https://www.cognetik.com/blog/when-to-implement-data-science-into-your-business/

15. Felipe, R. (2019). Small companies deserve Data Science. *Medium*. Pridobljeno 4.

junija 2020 iz https://towardsdatascience.com/small-company-big-data-science-48a3ccaf400

16. Foster, P., T. F. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking (1st ed.). *O'Reilly Media, Inc*.

17. Francis, X., Sumathi, V. P. & Shiva, J. (2019). An exploratory data analysis for loan prediction based on nature of the clients. *International Journal of Recent Technology and Engineering*, *7*(4), 176–179.

18. Gutierrez, D. (2018). Data Science Teams on the Rise, Give Unicorns a Break. *ODSC Conferences*. Pridobljeno 4. junija 2020 iz https://opendatascience.com/give-unicorns-a-break-its-time-for-data-science-teams/

19. H. Li, Z. Z. (2012). Bussiness-driven automatic IT change management based on machine learning. *IEEE Network Operations and Management Symposium*, 1374–1377. Budimpešta: IEEE Communications Society

20. Hotz, N. (2019). 9 Ways to Measure Data Science Project Performance. *Data Science Project Management*. Pridobljeno 4. junija 2020 http://www.datascience-pm.com/9-ways-to-measure-data-science-project-performance/

21. Jagadish, H. V. (2015). Big Data and Science: Myths and Reality. *Big Data Research*, *2*(2), 49–52.

22. Jordan, M. I. & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), (str. 255–260).

23. Kasunic, M. (2004). Conducting Effective Pilot Studies. *Defense*, 1–39.

24. Larson, D. (2019). A Review and Future Direction of Business Analytics Project Delivery. (str. 95–114).

25. Livingston, F. (2005). Implementation of Breiman's Random Forest Machine Learning Algorithm. *Machine Learning Journal Paper*, (str. 1–13).

26. Little Miss Data (2018). How to solve a business problem using data. *Little Miss Data.* Pridobljeno 4. junija 2020 https://www.littlemissdata.com/blog/businessproblem

27. Mangini, F. (2019). Implementing a Data Science Process in Your Company. *ThinkingOnData*. Pridobljeno 4. junija 2020 https://www.thinkingondata.com/implementing-data-science-process-in-your-company/

28. Marsland, S. (2014). Machine learning: An Algorithmic Perspective, Secon Edition (2nd. Ed.). Boca Raton: Chapman & Hall/CRC

29. Maxwell, A. E., Warner, T. A. & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, *39*(9), (str. 2784–2817).

30. Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., Verbeke, G. (2015). *Handbook of Missing Data Methodology*. New York: Chapman and Hall/CRC

31. Patel, K., Bancroft, N., Drucker, S. M., Fogarty, J., Ko, A. J. & Landay, J. A. (2010). Gestalt: Integrated support for implementation and analysis in machine learning. *UIST 2010 - 23rd ACM Symposium on User Interface Software and Technology*, (str. 37–46). New York: SIGCHI & SIGGRAPH

32. Rollins, J. (2015). Foundational Methodology for Data Science. (str. 1–4).

33. Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, *86*(11), (str. 2210–2239).

34. Saltz, J., Shamshurin, I. & Crowston, K. (2017). Comparing Data Science Project Management Methodologies via a Controlled Experiment. *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)*, (str. 1013–1022). Waikoloa Village: Leibniz Center for Informatics

35. Silipo, R. (2019). Practicing Data Science – Asking for Directions in a Data Science Project. *Medium*. Pridobljeno 4. junija 2020 https://towardsdatascience.com/practicing-data-science-5487e9f88aad

36. Simpson, P. (2019). Your Data Science Project Will Fail Unless It Meets These 3 Donditions. Pridobljeno 4. junija 2020 https://towardsdatascience.com/3-conditions-for-data-science-project-success-e31d3a798ec2

37. Su, S. I. & Chiong, R. (2010). Business intelligence. *Encyclopedia of Knowledge Management*, *1*(January), 72–80. https://doi.org/10.4018/978-1-59904-931-1.ch008

38. Ucros, M. (2018). The 10 Areas of Expertise in Data Science, and Why You should Choose One. *Medium*. Pridobljeno 4. junija 2020 https://medium.com/@melodyucros/interested-in-data-science-heres-a-list-of-of-10-specializations-to-choose-from-cd342c53b673

39. Vicario, G. & Coleman, S. (2019). A review of data science in business and industry and a future view. *Applied Stochastic Models in Business and Industry*, *November 2018*, (str. 6–18).

40. Waller, A., Fawcett, S. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logics*, *34*(2), (str. 77–84).

41. Zailinawati, H., Mazza, D. & Schattner, P. (2006). Doing A Pilot Study. *Malays Fam*

Physican, (2018), (str. 70-73).