UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

MAGISTRSKO DELO

OCENJEVANJE VERJETNOSTI NEPLAČILA: VPLIV ARBITRARNOSTI PRI IZDELAVI MODELA ZA OCENJEVANJE KREDITNEGA TVEGANJA FIZIČNIH OSEB

Ljubljana, november 2016

KATJA ILJAŠ

IZJAVA O AVTORSTVU

Podpisana Katja Iljaš, študentka Ekonomske fakultete Univerze v Ljubljani, avtorica predloženega dela z naslovom Ocenjevanje verjetnosti neplačila: vpliv arbitrarnosti pri izdelavi modela za ocenjevanje kreditnega tveganja fizičnih oseb, pripravljenega v sodelovanju s svetovalcem izr. prof. dr. Igorjem Lončarskim

IZJAVLJAM

1. da sem predloženo delo pripravila samostojno;

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

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

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

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

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

7. da sem pri pripravi predloženega dela ravnala v skladu z etičnimi načeli in, kjer je to potrebno, za raziskavo pridobila 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 23.11.2016 Podpis študentke:

KAZALO

UVOD **1**

1 KREDITNO TVEGANJE **3**
1.1 Komponente kreditnega tveganja 5
1.1.1 Verjetnost neplačila 7
1.1.2 Izguba v primeru neplačila 7
1.1.3 Izpostavljenost ob neplačilu 8
1.2 Baselski sporazum 8

2 MODELIRANJE KREDITNEGA TVEGANJA **10**
2.1 Proces modeliranja verjetnosti neplačila 11
2.2 Izzivi modeliranja 15
2.2.1 Problem pristranskosti izbire vzorca 16
2.2.2 Problem majhnega vzorca 18
2.2.3 Izbira razvojnega in validacijskega vzorca 20
2.2.4 Izbira modela 20
2.2.5 Izbira mere validacije 25

3 EMPIRIČNA ANALIZA **29**
3.1 Definicija neplačila 30
3.2 Baza podatkov 30
3.2.1 Spremenljivke 31
3.2.2 Priprava podatkov 32
3.3 Modeliranje verjetnosti neplačila 35
3.4 Primerjava rezultatov 44
3.5 Metoda ponovnega vzorčenja mere AUC 48

SKLEP **51**

LITERATURA IN VIRI **54**

PRILOGA

KAZALO TABEL

Tabela 1: Spreminjanje deležev posojil fizičnim osebam in posojil podjetjem v portfelju bank 4
Tabela 2: Osnovni in napredni pristop IRB 9
Tabela 3: Matrika zamenjav 25
Tabela 4: Seznam pojasnjevalnih spremenljivk 31

Tabela 5: Lastnosti vzorcev 36
Tabela 6: Primer prikaza rezultata logistične regresije 37
Tabela 7: PD pri spremembi starosti za eno enoto 38
Tabela 8: Primer matrike zamenjav pri mejni vrednosti 0,5 42
Tabela 9: Primer matrike zamenjav pri mejni vrednosti 0,5 42
Tabela 10: Matrika zamenjav pri mejni vrednosti 0,55 43
Tabela 11: Mere validacije pri različnih mejnih vrednostih 44
Tabela 12: Mera AUC – logistična regresija 44
Tabela 13: Statistike matrike zamenjav – logistična regresija 45
Tabela 14: Primerjava rezultatov odločitvenih dreves 46
Tabela 15: Primerjava mer validacije logistične regresije in odločitvenih dreves 47
Tabela 16: Primerjava mer AUC pri velikosti vzorca 4804 in različnem številu ponovitev 49
Tabela 17: Mere razpršenosti 50
Tabela 18: Primerjava mer AUC pri velikosti vzorca 1000 in različnem številu ponovitev 50
Tabela 19: Mere razpršenosti 51

KAZALO SLIK

Slika 1: Struktura nedonosnih terjatev bank po skupinah komitentov 5
Slika 2: Pričakovana izguba portfelja 6
Slika 3: Proces modeliranja verjetnosti neplačila 11
Slika 4: Delitev razvojnega in validacijskega vzorca 13
Slika 5: Problem pristranskosti izbire vzorca 16
Slika 6: Metoda ponovnega vzorčenja mere AUC 19
Slika 7: Primerjava linearne in logistične regresije 22
Slika 8: Primer prikaza odločitvenega drevesa 24
Slika 9: Krivulja ROC 28
Slika 10: Krivulja CAP 29
Slika 11: Izstopajoče vrednosti na primeru starosti 34
Slika 12: Korelacijska matrika 34
Slika 13: Prikaz odločitvenega drevesa 38
Slika 14: Prikaz odločitvenega drevesa s spremembo parametra 40
Slika 15: Primerjava krivulj ROC in mer AUC 41
Slika 16: Gostota porazdelitve ocenjene mere AUC pri različnem številu ponovitev 49
Slika 17: Gostota porazdelitve ocenjene mere AUC pri različnem številu ponovitev 51

UVOD

Kriza, ki je zajela svet po letu 2008, je pokazala na pomanjkljivosti v delovanju finančnega sektorja, zato se v zadnjih letih v gospodarstvu, predvsem v finančnih institucijah, več pozornosti usmerja v razvoj obvladovanja tveganj. V bančnem sektorju, kjer se dnevno srečujejo z izzivi ocenjevanja tveganj, je razvoj na tem področju še kako pomemben.

Kreditno tveganje je najstarejše tveganje med vsemi tveganji na finančnih trgih. Prav tako je eno izmed pomembnejših tveganj, s katerimi se banke srečujejo pri svojem poslovanju (Caouette, Altman & Narayanan, 1998), zato je pomembno, da čim bolj uspešno obvladujejo tveganja ter pri tem upoštevajo določbe Zakona o bančništvu, ki v veliki meri kot smernice upošteva kapitalski sporazum Basel II (Balant, 2005, str. 44). Z uvedbo tega sporazuma je bilo bankam omogočeno, da za oceno kreditnega tveganja uporabijo standardizirani pristop ali pristop internih bonitetnih ocen (angl. *Internal rating-based approach*, v nadaljevanju IRB). Pristop IRB je ena večjih novosti novega kapitalskega sporazuma, ki bankam omogoča, da za ocenjevanje kreditne sposobnosti komitentov uporabijo lastni sistem internih bonitetnih ocen. Cilj je zajeti tveganje posameznih naložb banke in omogočiti podrobnejše razvrščanje komitentov banke. Obstajata dva različna pristopa IRB – osnovni in napredni. Obema je skupno to, da mora banka zagotavljati svoje ocene verjetnosti neplačila (angl. *Probability of default*, v nadaljevanju PD). Za uporabo pristopa IRB mora banka izpolnjevati minimalne zahteve, v nasprotnem primeru ga ne sme uporabljati (Čargo & Štajner, 2004).

Banke se pri modeliranju parametrov kreditnega tveganja srečujejo z vprašanji, na katera ni enoznačnega odgovora. Centralna banka Republike Avstrije (v nadaljevanju OeNB) v sodelovanju s FMA (angl. *Financial Market Authority*) je leta 2004 v svoji publikaciji, izdani za namene pomoči bankam pri razvoju internih bonitetnih sistemov za ocenjevanje kreditnega tveganja, ponudila najboljše pristope na tem področju. Podaja smernice za uporabo pristopa IRB in poskuša odgovoriti na vprašanja glede izbire podatkov, ki jih lahko uporabijo pri ocenjevanju parametrov, glede najprimernejših metod modeliranja za določene segmente strank in katere postopke uporabljati za validacijo in kalibracijo modelov.

Večina literature o obvladovanju kreditnega tveganja je na temo posojil podjetjem (Kocenda & Vojtek, 2009, str. 2), medtem ko je bilo segmentu fizičnih oseb posvečeno manj pozornosti. V letih pred krizo so posojila fizičnim osebam v primerjavi s posojili podjetij predstavljala manjši delež portfelja, kar je verjetno tudi eden od razlogov, da so banke razvoj usmerjale v segment poslovanja s podjetji. Leta 2004 je delež posojil podjetjem znašal 65,36 %, fizičnim osebam pa 26,29 % celotnega portfelja terjatev bank (Banka Slovenije, 2005, str. II–12). Zadnja leta so banke zasledile rast deleža posojil fizičnim osebam, saj so v letu 2014 predstavljala 33,54 % portfelja, in padec deleža posojil podjetjem, ki so znašala 42,91

% (Banka Slovenije, 2015a, str. II-10). Zato je bankam v interesu, da posojila oz. kreditno tveganje posojil fizičnim osebam čim primerneje ovrednotijo.

Zaradi vse večjega obsega posojil fizičnim osebam se bom v svojem magistrskem delu osredotočila na ocenjevanje verjetnosti neplačila fizičnih oseb.

Pri modeliranju kreditnega tveganja oziroma ocenjevanju neplačila v vsakem koraku pridemo do razpotja oziroma arbitrarne odločitve, kjer je glede na obstoječo literaturo lahko več rešitev. Namen magistrskega dela je raziskati pomembnejša vprašanja, ki se pojavijo pri izdelavi modela za ocenjevanje kreditnega tveganja fizičnih oseb. Vprašanja in izzivi, ki jih bom raziskala v magistrskem delu, so naslednji:

- vprašanje pristranskosti izbire vzorca (angl. *Sample selection bias*),
- vprašanje majhnih vzorcev,
- vprašanje izbire velikosti razvojnega in testnega vzorca,
- vprašanje izbire modela,
- vprašanje izbire primerne mere za primerjavo različnih modelov.

Glede ne to, da v bankah razvoj modelov poteka večinoma na populaciji komitentov, katerim so odobrili posojilo, se pri modeliranju kreditnega tveganja srečamo z razvojem modelov na nenaključnih vzorcih. V kolikor pri razvoju ne upoštevamo pravilnih tehnik razvoja, lahko pride tudi do pristranskosti in nepravilnih ocen (Zmijewski, 1984, str. 54). Problem razrešimo s pomočjo procesa sklepanja o kreditni boniteti zavrnjenih komitentov (angl. *Reject inference*).

Poleg majhnega števila neplačnikov se banke lahko srečujejo tudi z manjšim številom komitentov in posledično z manjšim vzorcem za modeliranje. V tem primeru ne želimo deliti vzorca na razvojnega in validacijskega, zato lahko uporabimo metodo ponovnega vzorčenja (angl. *Bootstraping*), s pomočjo katere preverimo nepristranskost modela (Kanczes, 2012, str. 20).

Pri izdelavi samega modela je treba razvoj izvesti na razvojnem vzorcu in sámo točnost modelov preveriti na testnem oziroma validacijskem vzorcu. Tipično je 70–80 % vzorca uporabljenega za namen razvoja modela in preostalih 20–30 % za samo validacijo (Siddiqi, 2006, str. 63).

Poleg vseh omenjenih dilem je pomembno tudi vprašanje same izbire modela. V praksi je zaradi preprostosti in uporabnosti najbolj uporabljena metoda logistične regresije (Jagrič, 2011, str. 64). Druge metode so diskriminantna analiza, linearno programiranje, odločitvena drevesa in pravila, Bayesova mreža za klasifikacijo, metoda k-najbližjega soseda, metode umetne inteligence itn. V delu bom predstavila odločitvena drevesa, ki so priljubljeno orodje

za klasifikacijo in napovedovanje. Priljubljenost gre pripisati temu, da so zmogljivo orodje, katerega rezultati so jasno razumljivi (Berry & Linoff, 2004, str. 165). Zaradi razširjenosti v literaturi bom kot primerjalni model (angl. *Benchmark model*) uporabila logistično regresijo.

Na koncu bom s pomočjo različnih orodij validacije preverila, kateri nabor odločitev pri omenjenih vprašanjih da najboljše rezultate, kakor tudi vpliv različnih odločitev na kakovost oz. zanesljivost modela.

Cilj magistrskega dela je raziskati vpliv na kakovost modela glede na izbiro odločitev v posameznih korakih. Cilji so naslednji:

- ugotoviti, kako rešitve problema pristranskega vzorca in majhnosti vzorca vplivajo na kakovost modela,
- ugotoviti, kateri od izbranih modelov da boljše rezultate,
- ugotoviti, katera od mer je primernejša za primerjavo modelov za ocenjevanje kreditnega tveganja.

Pri izdelavi magistrskega dela bom za teoretični del uporabila strokovno in znanstveno literaturo ter vire (članke, knjige, publikacije različnih avtorjev in institucij).

Magistrsko delo bo sestavljeno iz treh temeljnih poglavij in več podpoglavij. V prvem delu bom predstavila problematiko, na kratko opisala kreditno tveganje in njegove komponente ter kaj je Baselski sporazum. V drugem delu se bom osredotočila na sam pregled korakov pri modeliranju in predstavitev problemov, ki pri tem nastanejo. Za vsak korak bom povzela, katere rešitve so predstavljene in uporabljene v literaturi. Zadnji, empirični del bo vseboval analizo, izdelano na osnovi podatkov. V njej bom ovrednotila vpliv vseh arbitrarnosti, ki se pojavijo pri modeliranju kreditnega tveganja. Analiza v magistrskem delu bo temeljila na podatkih, ki so prosto dosegljivi na spletni strani Kaggle. Kaggle je platforma s podatki, na podlagi katerih udeleženci preizkušajo svoje znanje o napovedovalnih modelih (angl. *Predictive-modeling*) in podatkovnem rudarjenju (angl. *Data mining*) (Give me some credit, 2015). Magistrsko delo bom zaključila s sklepnimi ugotovitvami in možnostmi za nadaljnje delo.

1 KREDITNO TVEGANJE

Banka Slovenije (2007, str. 11) v svojem priročniku Proces ocenjevanja tveganj kreditno tveganje definira kot tveganje nastanka izgube, ki je posledica dolžnikove nezmožnosti, da zaradi katerega koli razloga ne izpolni svoje finančne ali pogodbene obveznosti v celoti.
Primarna naloga banke kot finančnega posrednika je nemoten pretok finančnih sredstev v gospodarstvu. Banke delujejo kot finančni posredniki med ponudbo in povpraševanjem po kapitalu ter prevzemajo transakcijsko funkcijo (Bitenc, 2009). Zaradi vse večje konkurence

in želje po višjem dobičku banke prevzemajo vedno bolj tvegane posle. Povezavo med konkurenco in stabilnostjo v bančnem sistemu obravnavata Carletti in Hartmann (2002). Na splošno konkurenca lahko ogrozi stabilnost, vendar je za razliko od drugih sektorjev bančni dobro nadzorovan. Banke morajo pri svojem delovanju upoštevati zakonodajo, ki jo predpišejo državni organi. Banka za mednarodne poravnave je zato v želji, da bi banke v različnih državah uporabljale standardne pristope, leta 1988 objavila prvi kapitalski sporazum Basel I (Wu, 2008, str. 3). Več o baselskih sporazumih bo zapisanega v poglavju 1.2.

Ena izmed glavnih storitev, ki jih banke ponujajo svojim komitentom, je posojilo (Baesens & Van Gestel, 2009). Pomembnejši del portfelja posojil so posojila podjetjem in fizičnim osebam. Tabela 1 prikazuje spreminjanje deležev posojil fizičnim osebam in podjetjem. Razvidno je, da v zadnjih letih opazimo trend rasti deleža posojil fizičnim osebam, saj je z 26,29 % v letu 2004 delež zrastel na 33,54 % v letu 2014. Delež posojil podjetjem pa je z 65,36 % v letu 2004 v letu 2014 padel na 42,92 %.

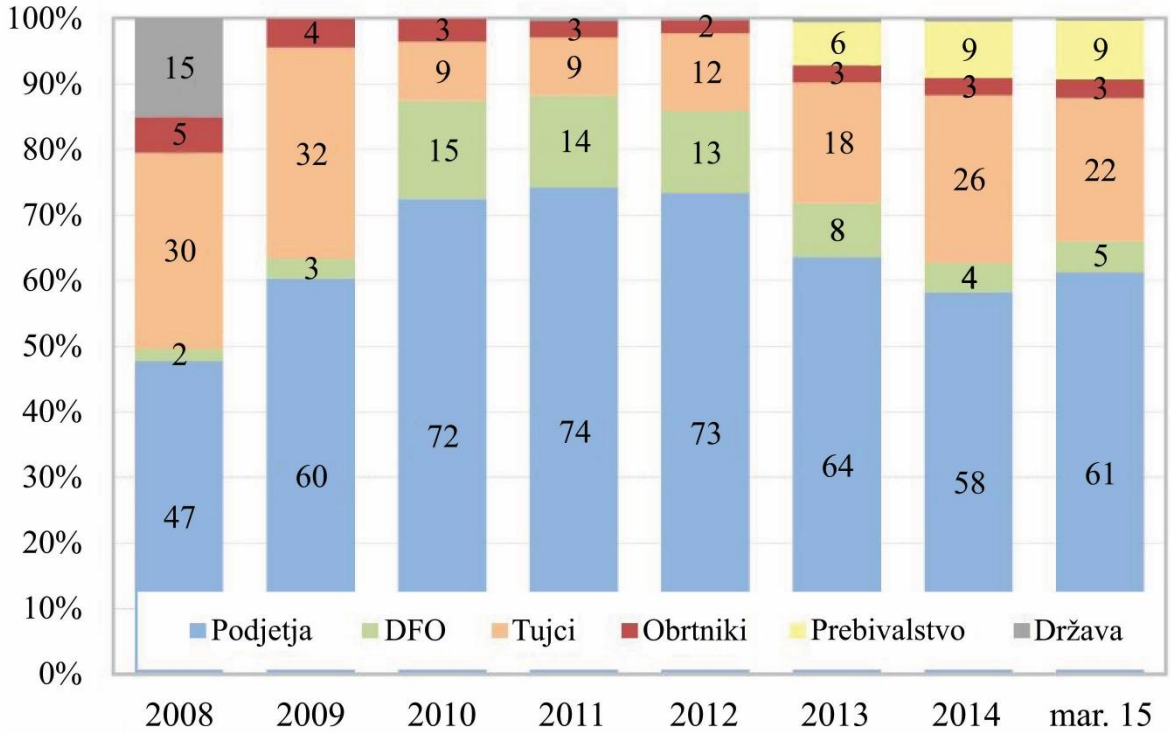
Tabela 1: Spreminjanje deležev posojil fizičnim osebam in posojil podjetjem v portfelju bank

Leto	**Delež posojil fizičnim osebam (v %)**	**Delež posojil podjetjem (v %)**
1994	21,98	69,86
2004	26,29	65,36
2010	24,53	55,05
2011	25,39	53,97
2012	25,83	51,98
2013	29,03	46,02
2014	33,54	42,92

Vir: Banka Slovenije, Bilten, 2005, str. II-12; Banka Slovenije, Bilten, 2015, str. II-10.

Menim, da je eden izmed glavnih razlogov, zakaj pozornost bank v preteklosti ni bila usmerjena v ocenjevanje kreditnega tveganja segmenta fizičnih oseb, višina deleža neplačnikov. Prebivalstvo oziroma segment fizičnih oseb sodi med najmanj tvegane skupine komitentov bank, čeprav se je delež nedonosnih terjatev (terjatve v zamudi nad 90 dni) v zadnjih letih povečal. Struktura nedonosnih terjatev bank po skupinah komitentov je razvidna iz Slike 1, kjer opazimo, da je v zadnjih letih več kot polovica zamud na strani podjetij.

Slika 1: Struktura nedonosnih terjatev bank po skupinah komitentov



Vir: Banka Slovenije, Poročilo o finančni stabilnosti, 2015, str. 52.

Ob koncu leta 2011 je delež zamud prebivalstva nad 90 dni znašal 3,5 % in konec maja 2015 dosegel 4,3 %. Delež zamud podjetij je konec marca 2015 znašal 28,1 %. Na splošno se je delež nedonosnih terjatev od druge polovice 2014 do marca 2015 znižal z 18,1 % na 11,4 %. Padec je delno posledica prenosa slabih posojil na Družbo za upravljanje terjatev bank (v nadaljevanju DUTB), delno pa je zaslediti tudi znižanje deleža zamud (Banka Slovenije, 2012; Banka Slovenije, 2015b). DUTB je bila ustanovljena marca 2013 kot pomoč pri prestrukturiranju bank sistemskega pomena, ki so se znašle v hudih solventnostnih in likvidnostnih težavah (DUTB – O nas, 2015).

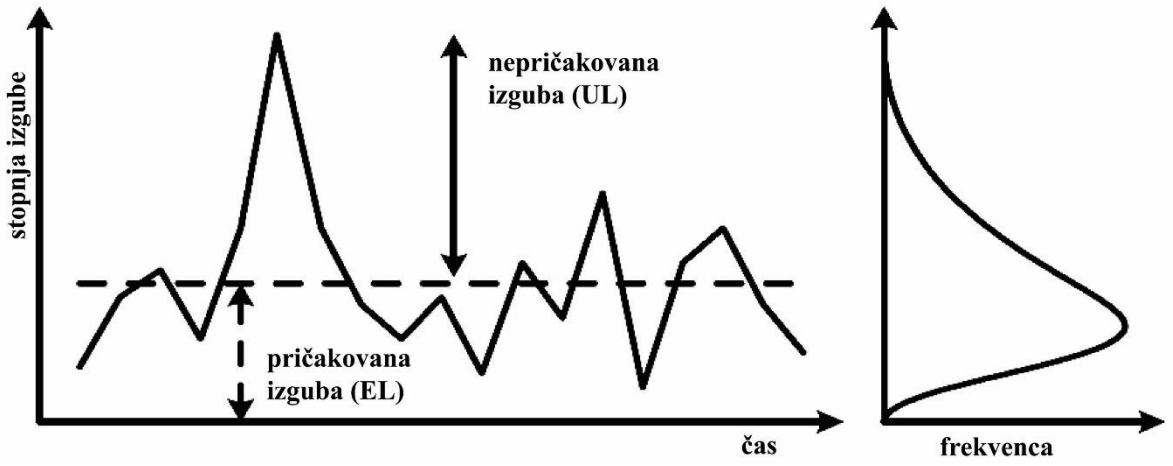
Višanje deleža posojil prebivalstva in deleža zamud je bankam dober opomnik, da je potrebno čim bolje obvladovati kreditno tveganje. V naslednjem poglavju bom opisala, katere so glavne komponente kreditnega tveganja in katere parametre mora banka oceniti, da lahko ovrednoti pričakovano izgubo.

1.1 Komponente kreditnega tveganja

Definicija kreditnega tveganja je tveganje nastanka izgube v primeru neporavnanih obveznosti dolžnika. Kako se dejanske izgube spreminjajo v času in kakšna je pripadajoča porazdelitev izgub portfelja banke, je razvidno iz Slike 2. Glede na to, da banka ne more vnaprej vedeti, kolikšno izgubo bo imela v določenem obdobju, lahko pričakovano izgubo

le oceni. Izgubo ovrednoti kot povprečno raven izgub, ki jo imenujemo pričakovana izguba (angl. *Expected loss – EL*). Na Sliki 2 je prikazana s črtkano črto. Banke vidijo pričakovano izgubo kot strošek poslovanja (angl. *Cost of doing business*), ki se obvladuje prek različnih načinov, kot je zaračunavanje provizij in premij za tveganje ob sklenitvi posojila. Vse, kar presega pričakovano izgubo in je na Sliki 2 nad črtkano črto, se imenuje nepričakovana izguba (angl. *Unexpected loss – UL*). Za razliko od pričakovane izgube banke ne vedo, kdaj se bo pojavila in kolikšna bo (BCBS, 2005).

Slika 2: Pričakovana izguba portfelja



Vir: BCBS, An Explanatory Note on the Basel II IRB Risk Weight Functions, 2005, str. 2.

Ovrednotenje izgube poteka na podlagi glavnih parametrov kreditnega tveganja. Kreditno tveganje je sestavljeno iz treh glavnih komponent, ki so izražene s pripadajočimi parametri tveganja (Jagrič, 2011, str. 43):

- tveganje neplačila (angl. *Default risk*) je tveganje, da dolžnik ne poravna pravočasno svojih obveznosti oziroma je tveganje, da nastopi dogodek neplačila. Izraženo je z verjetnostjo neplačila (v nadaljevanju PD),
- tveganje poplačila (angl. *Recovery risk*) je tveganje, povezano z dejansko izgubo, ki nastopi ob neplačilu in je zato močno odvisno od morebitnih zavarovanj in garancij. Merimo ga kot izgubo v primeru neplačila (angl. *Loss given default* , v nadaljevanju LGD),
- tveganje izpostavljenosti (angl. *Exposure risk*) je tveganje, povezano z negotovostjo višine izpostavljenosti. Tveganje izpostavljenosti je višje pri kreditnih karticah glede na posojila, saj višina izpostavljenosti ni vnaprej znana. Izražena je z oceno vrednosti izpostavljenosti oziroma z izpostavljenostjo ob neplačilu (angl. *Exposure at default,* v nadaljevanju EAD).

Za pričakovano izgubo velja:

$$EL = PD \times LGD \times EAD = PD \times LGD, \qquad (1)$$

kjer sta PD in LGD izražena v deležu in EAD v denarnih enotah.

1.1.1 Verjetnost neplačila

PD je ocena verjetnosti, da dolžnik ne bo poravnal svojih obveznosti, in je podana na podlagi kvalitativnih in kvantitativnih informacij. Po navadi je časovni horizont za oceno PD eno leto (Čargo & Štajner, 2004). Pri oceni je pomembna sama definicija, katere zamude se prepozna kot neplačilo. Je pomembna tudi višina zamude in ne le število dni zamude? V praksi je definicija dogodka neplačila pogosto vezana na bonitetne razrede. V kolikor ima dolžnik bonitetno oceno enako ali nižjo od ocene, ki določa prag za določanje neplačnikov, samodejno pridobi oznako neplačnika (Bank of Japan, 2005, str. 26).

Obstaja več različnih načinov ocenjevanja PD, od uporabe bonitetnih ocen zunanjih ponudnikov do modeliranja na podlagi internih baz podatkov. V drugem poglavju bom predstavila različne metode, ki se uporabljajo pri ocenah PD, njihove lastnosti, prednosti in slabosti.

PD nam poda samo verjetnost, da bo dolžnik postal neplačnik, dejansko izgubo pa izračunamo s pomočjo EAD in LGD (Baesens & Van Gestel, 2009, str. 25).

1.1.2 Izguba v primeru neplačila

LGD je izguba, ki se pojavi ob neplačilu in je izražena kot delež izpostavljenosti. V primeru, da celotni znesek ni poplačan, je LGD enak 100 %, v nasprotnem primeru pa 0. Praviloma je LGD lahko negativen, kar nakazuje dobiček zaradi obresti, ali pa večji od 100 %. Ta je posledica dodatnih stroškov, kot so npr. stroški sodnih postopkov (Baesens & Van Gestel, 2009).

V literaturi je LGD velikokrat izražen z deležem poplačila terjatev (angl. *Recovery rate*, v nadaljevanju RR):

$$LGD = 1 - RR, \qquad (2)$$

kjer RR pomeni, kolikšen delež terjatve je poplačan (Altman, 2006, str. 2; Baesens & Van Gestel, 2009, str. 26).

Za razliko od PD je v literaturi ocenjevanju LGD posvečene manj pozornosti. Altman temu pripisuje dva med seboj povezana razloga. Prvi razlog je usmerjanje večje pozornosti sistematičnemu delu kreditnega tveganja, na podlagi katerega je določena premija za tveganje. Drugi je predpostavka, tipična pri modeliranju kreditnega tveganja, da je RR odvisen od posameznih faktorjev, kot sta npr. zavarovanje (angl. *Collateral*) in nadrejenost posojila (angl. *Seniority*). Ta dva nista del sistematičnega tveganja in sta zato neodvisna od PD. Dokazi nakazujejo na nestanovitnost RR in vrednosti zavarovanj, ki so po navadi obratno sorazmerne s številom neplačil. Večina bank se je zato bolj posvečala PD s predpostavko, da je RR konstanten, ali pa so ga videli kot stohastično spremenljivko, neodvisno od PD (Altman, 2006).

1.1.3 Izpostavljenost ob neplačilu

EAD predstavlja dejavnik odvisnosti od zneska izpostavljenosti banke v trenutku neplačila. Praviloma je enaka nominalnemu znesku, a ni nujno znana vnaprej, saj je odvisna od same storitve. Tako je treba v primeru kreditnih kartic upoštevati še pričakovane prihodnje odlive iz naslova neizkoriščenega dela zneska limita na kreditni kartici (Čargo & Štajner, 2004; Kolar, 2014).

Poleg zgoraj omenjenih komponent je pomembna komponenta tveganja tudi zapadlost (angl. *Effective maturity*, v nadaljevanju M). M je omejeno časovno obdobje, znotraj katerega je dolžnik dolžan poravnati svoje obveznosti. Velja, da daljša kot je zapadlost, večja sta negotovost in tveganje, ki jima je banka oziroma finančna institucija izpostavljena.

Z uvedbo kapitalskega sporazuma Basel II, o katerem bo govora v naslednjem poglavju, banke za merjenje kapitalske ustreznosti lahko uporabljajo pristop, ki temelji na internih bonitetnih sistemih. V sklopu tega banka ocenjuje tudi zgoraj naštete parametre.

1.2 Baselski sporazum

Mednarodna dolžniška kriza, ki je izbruhnila leta 1980, je razkrila ranljivost bančnega sistema. Zaradi vse večje konkurenčnosti so banke prevzemale vedno bolj tvegane posle. To je vodilo do zmanjšanja kapitalske moči najpomembnejših mednarodnih bank in izkazalo se je, da propad ene banke lahko resno ogrozi ali poruši finančni sistem. Zato je Banka za mednarodne poravnave leta 1988 objavila prvi kapitalski sporazum Basel I, ki je bil namenjen spodbujanju trdnosti in stabilnosti bančnega sistema. Želeli so, da bi se v bankah po različnih državah uporabljalo standardne pristope (Wu, 2008, str. 3). Baselski standardi so priporočila, usmeritve državam pri pripravi zakonskih aktov. Tudi slovenski Zakon o bančništvu v veliki meri upošteva njihove smernice.

Pomanjkljivosti prvega sporazuma so poskusili odpraviti z uvedbo novega kapitalskega sporazuma Basel II, katerega vzpostavitev je trajala od leta 1998 do 2009. Novi sporazum še večjo pozornost usmerja v obvladovanje tveganj in poleg kreditnega obravnava tudi tržno in operativno tveganje. Basel II sestavljajo trije stebri: prvi steber predstavlja minimalne kapitalske zahteve, drugi regulatorni nadzor in tretji tržno disciplino. Za ocenjevanje kreditne sposobnosti je najpomembnejši prvi steber, katerega temeljni del in novost je uporaba tako imenovanega pristopa IRB. Pristop, ki temelji na podlagi internih bonitetnih ocen, banke lahko uporabijo pri ocenjevanju kreditne sposobnosti komitentov (Basel Commitee on Banking Supervision (v nadaljevanju BCBS), 2006).

Obstajata dva različna pristopa IRB – osnovni in napredni – s pomočjo katerih banke ocenijo zgoraj naštete parametre tveganja. Tabela 2 prikazuje razliko med pristopoma, ki je v tem, ali zagotavlja oceno parametra banka sama ali jo določi nadzorni organ. Lastnost naprednega pristopa IRB je, da banke same zagotavljajo ocene vseh naštetih parametrov.

V kolikor banka ne izpolnjuje minimalnih zahtev, ki so potrebne za uporabo pristopa IRB, ga ne sme uporabljati (Čargo & Štajner, 2004).

Tabela 2: Osnovni in napredni pristop IRB

Parameter tveganja	**Osnovni pristop IRB**	**Napredni pristop IRB**
PD	oceno zagotovi banka	oceno zagotovi banka
LGD	oceno določi nadzornik	oceno zagotovi banka
EAD	oceno določi nadzornik	oceno zagotovi banka
M	oceno zagotovi banka ali jo določi nadzornik	oceno zagotovi banka

Že pri osnovnem pristopu IRB se banke soočajo z izbiro najboljšega modela za ocenjevanje PD. V literaturi se pogosto PD ocenjuje z modeli kreditnega točkovanja (angl. *Credit scoring models*), ki so statistična orodja, s pomočjo katerih pojasnjujemo vpliv nabora neodvisnih spremenljivk na odvisno spremenljivko (Jovan & Šušterič, 2004). Podobno kot OeNB v objavi publikacije tudi Jovan (2005) primerja tri najpogostejše statistične modele, ki jih zasledimo v literaturi – logistična regresija, probit regresija in diskriminantna analiza. Uporabljene so na istih vzorcih, istih transformacijah in z istimi finančnimi kazalniki. Primerjava pokaže, da je kakovost razločevanja uporabljenih metod odvisna od transformacije kazalnikov. Diskriminantna analiza je manj občutljiva na sámo transformacijo, vendar zahteva normalno porazdelitev neodvisnih spremenljivk.

Prednosti, ki jih prinaša pristop IRB, so (BCBS, 2001):

- uporaba pristopa za interno upravljanje tveganj in za namene regulatornih zahtev,

- nižje regulatorne kapitalske zahteve, ki so občutljivejše na parametre kreditnega tveganja in izgube,
- večja preglednost kreditnega portfelja banke in boljše razlikovanje med razredi tveganj,
- upoštevanje dodatnih faktorjev tveganja, ki jih zunanji ponudniki bonitetnih ocen slabo ali sploh ne poznajo.

Leta 2010 je baselski komite objavil novi kapitalski sporazum Basel III, katerega implementacija bo trajala vse do marca 2019. Novi sporazum naj bi še okrepil kapitalske zahteve s pomočjo večje likvidnosti bank in z zmanjšanjem bančnega finančnega vzvoda.

V naslednjem poglavju se bom osredotočila na sam proces modeliranja PD in vsa pomembnejša vprašanja, ki se pri tem pojavijo.

2 MODELIRANJE KREDITNEGA TVEGANJA

Drugi odstavek 1. člena Sklepa o izračunu kapitalske zahteve za kreditno tveganje po standardiziranem pristopu za banke in hranilnice določa, da mora banka za izračun kapitalske zahteve za kreditno tveganje uporabljati standardizirani pristop, razen če ima dovoljenje Banke Slovenije za uporabo pristopa IRB. Standardizirani pristop je pristop tehtanja terjatev z različnimi utežmi glede na njihovo tveganost, ki določa višino uteži glede na kategorijo izpostavljenosti. Za izpostavljenosti bančništva na drobno, kamor spada tudi segment prebivalstvo, se dodeli utež tveganja 75 %. To pomeni, da so vsa posojila z enakimi značilnostmi, dana različnim osebam, enako obravnavana, in banka za tveganja, povezana s temi posojili, zagotavlja enak znesek kapitala. S pristopom IRB pa je kapitalska zahteva za kreditno tveganje nižja pri komitentu z višjo bonitetno oceno in višja pri tistem z nižjo oceno. Pristop bankam omogoča boljše obvladovanje tveganj, saj je portfelj banke odzivnejši na dejanska tveganja, ki jih banka prevzema (BCBS, 2001). Ocenjevanje PD je temeljni del pristopa IRB in ga mora banka oceniti ne glede na izbrani pristop (BCBS, 2001, str. 3).

V povezavi z ocenjevanjem PD v literaturi največkrat zasledimo modele kreditnega točkovanja, ki se jih uporablja za ločevanje med slabimi in dobrimi posojilojemalci (Wu, 2008; Baesens & Van Gestel, 2009; Thomas, Edelman & Crook, 2002; Siddiqi, 2006). Na podlagi vhodnih podatkov model komitentu pripiše točke, na podlagi katerih se banka odloči, ali bo posojilo odobrila ali ne. Načeloma v praksi višje točke pomenijo manjšo verjetnost, da komitent posojila ne bo odplačal. Banka mora določiti mejo (angl. *Cut-off*), nad katero so posojila komitentov samodejno odobrena, in mejo, pod katero so samodejno zavrnjena. Tista, ki padejo v vmesni interval, se pošlje v nadaljnjo obravnavo. Pri določanju meje mora banka zasledovati svoje cilje in najti čim boljše razmerje med stopnjo odobritve in dobičkom. Večje kot bo število odobrenih posojil potencialnim neplačnikom, nižji bo dobiček. Višja stopnja zavrnjenih pa lahko vpliva na tržni delež (Siddiqi, 2006). Ker se neodobrena posojila potencialno dobrim komitentom odrazijo v obliki oportunitetnih

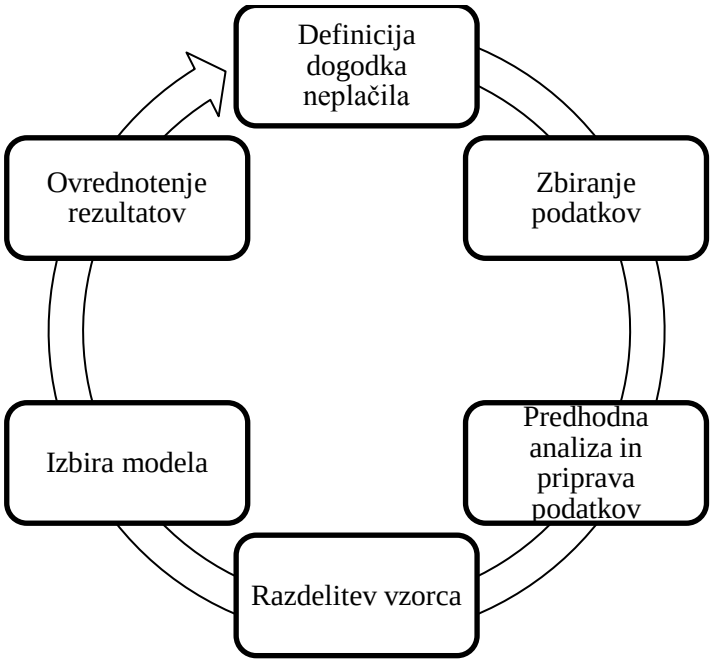
stroškov, v primeru neplačila pa pride do neposrednih stroškov, je za banke pomembnejše učinkovito prepoznavanje slabih komitentov (Liu, 2002, str. 17). Zato mora banka izbrati model, ki bo uspešno zasledoval zastavljene cilje.

V splošnem se banke ukvarjajo z dvema različnima odločitvama – odobritev novega posojila in povečanje zadolženosti obstoječim komitentom (npr. povečanje limita na kreditni kartici). Medtem ko se za odobritev novih posojil uporabljajo metode kreditnega točkovanja, se v primeru povečanja zadolženosti uporablja metode vedenjskega točkovanja (angl. *Behavioral scoring*). Te imajo več predpostavk glede vedenja komitenta in z njihovo pomočjo lahko zaznajo potencialne neplačnike (Thomas et al., 2002). V sklopu magistrskega delu se bom osredotočila na ocenjevanje PD za nove komitente banke.

2.1 Proces modeliranja verjetnosti neplačila

V samem procesu modeliranja PD morajo biti v različnih korakih sprejete odločitve, ki lahko vplivajo na kakovost modela. Katere podatke uporabiti, kateri model je najprimernejši, katero tehniko uporabiti za validacijo, so primeri vprašanj pri izdelavi modela PD. Cilj je izbrati njihovo najoptimalnejšo kombinacijo. Glavni koraki procesa modeliranja PD tako za fizične osebe kot podjetja so prikazani na Sliki 3.

Slika 3: Proces modeliranja verjetnosti neplačila



Povzeto in prirejeno po R. Anderson, The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation, 2007, str. 74.

Modeliranje verjetnosti neplačila je v osnovi binarni model izbire, saj je odvisna spremenljivka binarna in zavzame vrednosti 0 ali 1. 1 po navadi pomeni, da je prišlo do dogodka neplačila in je posojilojemalec postal neplačnik; 0 pa pomeni, da do dogodka neplačila ni prišlo in ima posojilojemalec status plačnika:

$$y = \begin{cases} 1, \text{ neplačnik} \\ 0, \text{ plačnik} \end{cases}$$

V kolikor bi pri izdelavi modela upoštevali še tako imenovane vmesne komitente[1], imamo opravka z modelom z diskretno odvisno spremenljivko. Ti v praksi niso pogosti, saj večina avtorjev uporablja binarni model (Siddiqi, 2006; Anderson, 2007).

Prvi korak modeliranja je definicija dogodka neplačila, saj je od tega odvisno, kateri komitenti bodo pridobili status neplačnika. Tipična definicija neplačila, ki je tudi skladna po baselski definiciji, je zamuda nad 90 dni (Jagrič, 2011; Kocenda & Vojtek, 2009; Siddiqi, 2006). Več o sami definiciji bo zapisano v nadaljevanju.

Naslednja dva koraka, ki sledita in sta najobsežnejša, sta zbiranje podatkov in predhodna analiza ter priprava podatkov. Glede na to, za kateri segment se ocenjuje PD, mora banka določiti, katere podatke in vire podatkov bo uporabila. Na splošno se podatki ločijo na tri skupine: kvantitativni, kvalitativni in zunanji podatki (OeNB, 2004, str. 62). Za modeliranje PD fizičnih oseb se po navadi uporabljajo:

- kvantitativni podatki, kot so višina prihodka, razpoložljiv prihodek, zadolženost itn.,
- kvalitativni podatki, kot so na primer sociodemografski podatki, podatki o stanovanjski situaciji, namen posojila itn.,
- zunanji podatki o preteklem poslovanju z drugimi bankami.

Prvi dve skupini podatkov banke pridobijo iz lastnih virov, medtem ko zunanje podatke pridobijo od zunanjih ponudnikov. V Sloveniji od leta 2008 deluje informacijski sistem, imenovan SISBON. Ustanovljen je bil pod okriljem Združenja bank Slovenije, katerega upravljanje je leta 2016 prevzela Banka Slovenije (SISBON, b.l.). SISBON je informacijski sistem, katerega namen je:

- zagotovitev odgovornega kreditiranja,
- preprečevanje prezadolženosti posameznikov,
- izvajanje nadzora nad upravljanjem kreditnega tveganja v bankah.

[1] Kot vmesni so poimenovani tisti komitenti, ki imajo zamudo med 30 in 90 dni. V tem primeru so plačniki vsi tisti, ki zamujajo do 30 dni.

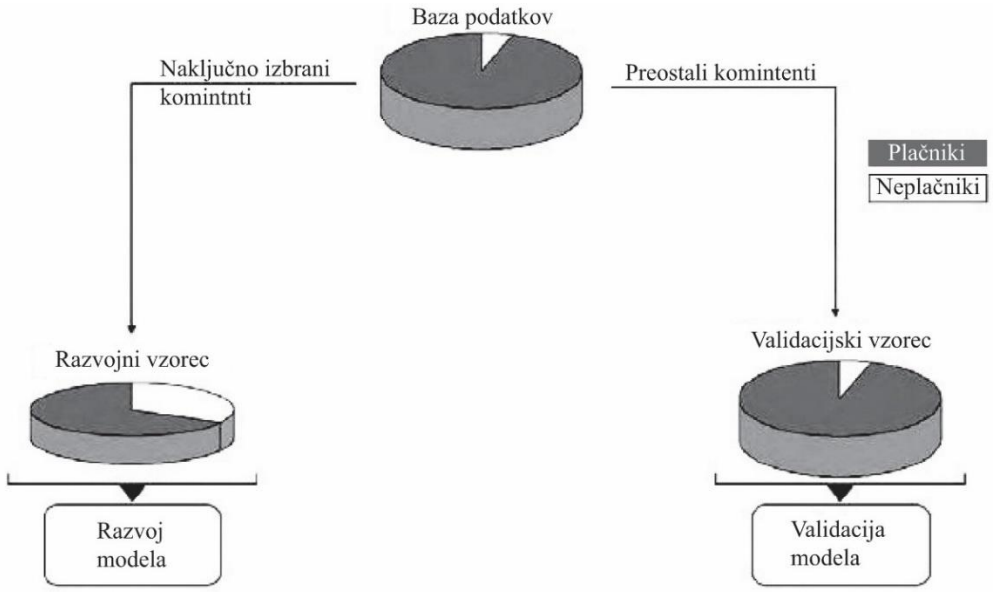
Podatki o komitentih (ime, datum rojstva, naslov prebivališča, podatki o stečaju ...) in podatki o poslih (vrsta posla, datum sklenitve, znesek ob odobritvi, znesek anuitete, datum zapadlosti, odplačilna doba …) se v sistemu hranijo 4 leta po prenehanju obveznosti. Banke so uporabnice in upravljavke zbirk osebnih podatkov svojih komitentov. Hkrati so odgovorne za pravilnost in ažurnost podatkov, ki jih na dnevni ravni posredujejo v SISBON. V magistrski nalogi bosta uporabljeni le prvi dve vrsti podatkov, saj poleg dejstva, da izbrana baza podatkov ni posnetek slovenskega bančnega sistema, velja tudi, da so podatki v SISBON-u zaupne narave in niso razkriti širši javnosti.

Ko so podatki zbrani, jih je treba analizirati in pripraviti na nadaljnjo obravnavo, treba je preveriti tudi njihovo pravilnost in smiselnost, pa tudi kakovost – manjkajoče vrednosti, napake in osamelce (OeNB, 2004).

Po analizi in pripravi podatkov, na podlagi katerih bom izvedla modeliranje, je naslednji korak delitev le-teh na razvojni in validacijski vzorec, kar predstavlja Slika 4. Vzorca morata biti izbrana naključno in podatki, ki jih uporabimo za razvoj, ne smejo biti vključeni v validacijskem vzorcu – vzorca morata biti disjunktna (OeNB, 2004; Siddiqi, 2006). Da to dosežemo, najprej naključno izberemo komitente za razvojni vzorec, preostale pa vključimo v validacijskega.

V kolikor baza podatkov ne vsebuje dovolj velikega števila opazovanj in želimo model razviti na celotnem vzorcu, si lahko pomagamo z metodo ponovnega vzorčenja. Na ta način podatkov ni treba deliti in model lahko razvijemo na celotnem vzorcu.

Slika 4: Delitev razvojnega in validacijskega vzorca



Vir: OeNB & FMA, Guidelines on credit risk management: Rating models and validation, 2004, str. 73.

Zadnja dva koraka procesa modeliranja sta izbira modela in ovrednotenje rezultatov. V splošnem ne obstaja model, za katerega bi lahko trdili, da je najprimernejši za ocenjevanje PD (Bolton, 2010). Namen modelov za ocenjevanje PD je čim boljše razlikovanje med plačniki in neplačniki, ki ga dosežemo z izbiro najustreznejše metode glede na dane okoliščine. Različne metode bodo različno uspešne glede na kakovost podatkov, vrednost odvisne spremenljivke, velikost vzorca, segment, za katerega ocenjujemo model, in podobno. Zato je vedno priporočljivo preizkusiti več metod in izbrati najustreznejšo. OeNB in FMA (2004) modele za ocenjevanje PD deli na tri oblike:

- hevristični modeli (angl. *Heuristic models*),
- statistični modeli (angl. *Statistical models*),
- vzorčni modeli (angl. *Casual models*).

Poleg zgoraj naštetih se uporabljajo kombinacije le-teh, ki se imenujejo hibridni modeli (angl. *Hybrid models*). Hibridni model je kombinacija hevrističnega in enega od preostalih dveh. V literaturi zasledimo različna poimenovanja in delitve modelov. Baesens in Van Gestel (2009) delita modele na ekspertne, modele, ki temeljijo na empiričnih podatkih, in finančne modele. Spet drugi delijo pristope na tradicionalne in sodobne; med tradicionalne pristope štejejo ekspertne sisteme, kreditno točkovanje in sisteme bonitetnih ocen (Allen, DeLong & Saunders, 2004).

Hevristični modeli temeljijo na ideji učenja iz preteklih izkušenj, saj velja, da se na podlagi preteklih izkušenj poda oceno kreditne sposobnosti, ki jo banka pričakuje v prihodnosti. V literaturi so poimenovani tudi kot ekspertni modeli (angl. *Expert models*), saj parametre določajo finančni strokovnjaki (angl. *Financial experts*) in niso določeni statistično. Uporabljeni so v primeru nezadostne količine podatkov, saj poleg vseh znanih dejavnikov upoštevajo še mnenje svetovalca. Možna subjektivna ocena je tudi potencialna slabost hevrističnih modelov (Baesens & Van Gestel, 2009). Zaradi narave vhodnih podatkov, kot so vrednost kapitala in dolga podjetja ter volatilnost sredstev, vzorčni modeli niso primerni za uporabo na segmentu prebivalstva (OeNB & FMA, 2004).

V skupino statističnih modelov se uvrščajo tehnike, kot so diskriminantna analiza, regresijski modeli, linearno programiranje, odločitvena drevesa in pravila, Bayesova mreža za klasifikacijo, metoda k-najbližjega soseda in metode umetne inteligence. Za razliko od hevrističnih modelov statistični uporabljajo statistične metode na bazi podatkov, ki mora vsebovati dovolj veliko število komitentov (OeNB & FMA, 2004). Durand (1941) je prvi, ki je na vzorcu posojil podjetjem, za razlikovanje med dobrimi in slabimi posojili, uporabil metodo diskriminantne analize. Diskriminantna analiza je multivariantna statistična tehnika, katere cilj je poiskati kombinacijo spremenljivk, ki najbolj razlikujejo dve skupini. Dosega slabše klasifikacijske sposobnosti kot druge metode (Desai, Crook & Overstreet, 1996; Lee, Chiu, Lu & Chen, 2002). Izmed v literaturi omenjenih metod naj bi metoda nevronskih mrež,

ki sodi med metode umetne inteligence, imela boljše klasifikacijske sposobnosti kot druge metode (Desai et al., 1996; Lee et al., 2002; West, 2000). Vendar pa imajo metode nevronskih mrež tudi naslednje slabosti; v primerjavi z logistično regresijo in odločitvenimi drevesi pri metodi nevronskih mrež ni preprosto določiti, katere spremenljivke najbolj vplivajo na PD. Poleg tega je uporaba in interpretacija nevronskih mrež zapletenejša kot interpretacija in razumevanje rezultatov drugih naštetih metod. Za sam razvoj modela s pomočjo metode nevronskih mrež je zaradi računske zahtevnosti treba imeti veliko znanja ter zmogljivejšo računalniško tehnologijo. Slabost nevronskih mrež je morda tudi preveliko prilagajanje modela razvojnemu vzorcu (Tu, 1996). Zaradi omenjenih slabosti bom v nalogi primerjala logistično regresijo in odločitvena drevesa. Zaradi razširjenosti v literaturi in dobrih rezultatov bom logistično regresijo uporabila kot primerjalni model. Več o uporabljenih metodah bom zapisala v nadaljevanju.

Katera izmed preizkušenih metod je najustreznejša, preverimo s pomočjo različnih mer za presojo kakovosti modela. V literaturi najdemo različne vrste mer, ki se uporabljajo v praksi. Med najbolj poznanimi so: krivulja ROC (angl. *Receiver operating characteristic*), mera AUC (angl. *Area under curve*), matrika zamenjav (angl. *Confusion matrix*), krivulja CAP (angl. *Cumulative accuracy profile*) in Ginijev koeficient. Krivulja ROC je grafična mera, ki predstavlja razmerje med kumulativno frekvenco neplačnikov in kumulativno frekvenco plačnikov. Matrika zamenjav je osnova za izračun statistik napačnih klasifikacij (angl. *Misclassification statistics*), s pomočjo katerih primerjamo število pravilno in napačno razvrščenih. Želimo imeti čim manj napačno in čim več pravilno razvrščenih plačnikov in neplačnikov. Kot že ime pove, je AUC mera površine pod krivuljo ROC (OeNB & FMA, 2004; Siddiqi, 2006; Engelmann, Hayden & Tasche, 2003). Sam izračun in prikaz omenjenih mer bom predstavila v nadaljevanju.

Del procesa modeliranja je tudi redna spremljava prek postopkov rednih validacij. Tako banke preverjajo kakovost modelov in lahko razvijajo vedno boljše modele. Sčasoma se baze podatkov povečajo, kar lahko povzroči spremembe v lastnostih populacije (Siddiqi, 2006). Zato je sam proces modeliranja kreditnega tveganja ponavljajoč proces, ki ga morajo banke neprestano izboljševati in prilagajati trenutnim razmeram v ekonomiji in svojim komitentom.

Kateri izzivi in vprašanja ter v katerem koraku modeliranja PD se pojavijo, bom opisala v naslednjem poglavju.

2.2 Izzivi modeliranja

Pri procesu modeliranja se v določenem koraku srečujemo z različnimi izzivi. Za spopadanje z njimi lahko izberemo različne metode, ki lahko vplivajo na kakovost modelov. Cilj magistrskega dela je raziskati vpliv odločitev v posameznih korakih na kakovost modela.

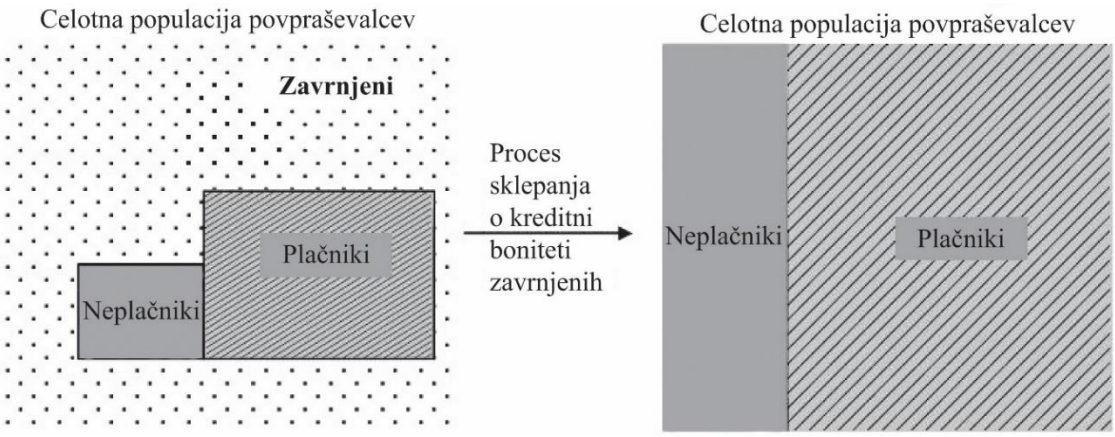
Prva dva izziva se pojavita že na začetku, pri samem zbiranju podatkov. Banka se pri pripravi podatkov sooči z dvema problemoma: prvi je problem pristranskosti izbire vzorca, ki je posledica nedostopnosti podatkov o zavrnjenih komitentih; drugi problem se pojavi v primeru, ko baza podatkov za razvoj in validacijo modela na dveh disjunktnih vzorcih ni dovolj velika. Drugi izzivi so del samega razvoja modela.

2.2.1 Problem pristranskosti izbire vzorca

Preden banka dejansko začne proces modeliranja verjetnosti neplačila, se spopade s prvim izzivom. Večina bank ima v svojih evidencah zgolj podatke o komitentih, katerim je bilo posojilo odobreno, oziroma ima podatke o zavrnjenih komitentih, vendar ne pozna njihovega statusa (plačnik oziroma neplačnik). Zato so banke prisiljene razvijati modele na populaciji, v kateri ni podatkov o vseh komitentih, katerim posojila niso odobrili. Razvoj za namene uporabe modela na celotni populacijo komitentov, tako zavrnjenih kot sprejetih, temelji na populaciji sprejetih komitentov in je lahko pristranski (Banasik, Crook, & Thoma, 2003; Siddiqi, 2006). Značilnosti zavrnjenih strank banki niso poznane in se lahko razlikujejo od značilnosti sprejetih komitentov (Feelders, 2000). Zato gre v primeru manjkajočih podatkov zavrnjenjih lahko za pristransko izbiro že v prvem koraku, saj smo iz populacije vseh izbrali le tiste, ki so bili sprejeti. Z izbiro razvojnega vzorca tako pridobimo pristranski vzorec, ki lahko povzroči pristranskost modela.

Slika 5 prikazuje, kako poteka proces sklepanja o kreditni boniteti zavrnjenih. S pomočjo različnih tehnik sklepamo o bonitetni oceni zavrnjenih in tako dobimo množico, kjer ni nobenega komitenta, ki bi ga zavrnili. Tako lahko apliciramo model na celotno množico komitentov, ki zaprosijo za posojilo.

Slika 5: Problem pristranskosti izbire vzorca



Vir: N. Siddiqi, Credit risk scorecards: developing and implementing intelligent credit scoring, 2006, str. 99.

Thomas (2000) problem manjkajočih podatkov o statusu zavrnjenih vidi na dva načina: če bi vsem zavrnjenim pripisali status neplačnika, bi model drugim posojilojemalcem s podobnimi značilnostmi vedno pripisal slabo oceno. V tem primeru bi jih banka vedno zavrnila. Po drugi strani pa so bili verjetno zavrnjeni zaradi utemeljenih razlogov in zato lahko trdimo, da je zelo verjetno, da je stopnja neplačila v skupini zavrnjenih dejansko višja od skupine sprejetih. Problem pristranskosti izbire vzorca in proces sklepanja o bonitetni oceni zavrnjenih je v literaturi pogosto predmet analize (Banasik et al., 2003; Crook & Banasik, 2004; Greene, 1998). Heckman (1979) je med prvimi avtorji, ki so se ukvarjali z vprašanjem pristranske izbire vzorca. Problem se pojavi zaradi manjkajočih podatkov in ga predstavi kot problem tehtanja, ki je posledica neprimernih ali ničelnih uteži, ki jih imajo potencialni posojilojemalci. Hand and Hanley (1993) sta ugotovila, da če le ne predpostavimo določenih povezav med porazdelitvami plačnikov in neplačnikov, ki držijo za obe skupini, tako zavrnjenih kot sprejetih, problema ni mogoče ustrezno odpraviti. Če bi želeli pridobiti realno sliko značilnosti populacije, bi banke morale za kratek čas sprejeti vse, ki zaprosijo za posojilo. Vendar pa za razliko od bank, ki bi zaradi prevelikih izgub ta način lahko uporabile le za redke storitve, podjetja, kot so na primer ponudniki telekomunikacijskih storitev, lahko sprejmejo vse komitente, ki zaprosijo za storitev. S tem pridobijo bazo podatkov, na podlagi katere si lahko ustvarijo primeren vzorec za razvoj modela (Thomas, 2000, str. 2).

Metodi sklepanja o boniteti zavrnjenih, ki sta največkrat opisani v literaturi, sta ekstrapolacija (angl. *Extrapolation*) in povečanje (angl. *Augmentation*), imenovano tudi metoda ponovnega tehtanja (Hsia, 1978; Verstraeten & Van den Poel, 2005; Banasik et al., 2003; Joanes, 1993). Nekateri avtorji uporabijo metode, kot so multinominalni logistični model (Reichert, Cho & Wagner, 1983), algoritem EM (Feelders, 2000) in bivariatni probit model (Crook & Banasik, 2004). Pomembne zaključke v svojem delu izpostavita Crook in Banasik (2004):

- tudi pri visokem deležu zavrnjenih je izboljšanje z metodo skromno,
- v primeru visokega deleža zavrnjenih so se pri tehniki ponovnega tehtanja modeli še poslabšali, v splošnem pa niso dosegli boljše klasifikacije,
- metoda ponovnega tehtanja brez dodatnih koristi celo preprečuje uporabo spoznanj o porazdelitvi dobrih in slabih v populaciji,
- metoda ekstrapolacije ne koristi in ne škoduje – koeficienti se ne spremenijo bistveno in tudi sama spoznanja o porazdelitvi dobrih in slabih ne prinašajo nobene koristi.

Podobno kot drugi je Banasik et al. (2003) odkril le majhno razliko v klasifikaciji ob vključitvi zavrnjenih v model. Izbrali so si različne mejne vrednosti, ki so določale, ali bo komitent zavrnjen. Vendar poleg dejstva, da je izboljšanje mogoče, niso dokazali pretiranega izboljšanja v kakovosti.

Metoda ponovnega tehtanja temelji na ideji, da je obnašanje zavrnjenih z določeno bonitetno oceno enako obnašanju sprejetih z isto oceno. Model je razvit na vzorcu vseh in je primeren le v primeru, da za obe skupini poznamo vse spremenljivke, uporabljene v modelu. Ko ocenimo model, razdelimo posojilojemalce v razrede, v katerih poznamo število plačnikov in neplačnikov. Nato v vsakem razredu zavrnjenim pripišemo enako porazdelitev dobrih in slabih. Ponovno ocenimo model le na vzorcu odobrenih, pri čemer popravimo uteži glede na število zavrnjenih v pripadajočem razredu osnovnega modela. Glede na to, da je cilj tehnike popravek uteži, ki so lahko neprimerne ali celo ničelne, seveda uteži lahko spremeni le pri tistih, ki so prisotni v vzorcu – kar pa ne odpravi problema ničelnih uteži (Crook & Banasik, 2004, str. 3). Uteži, ki jih model pripiše sprejetim, so tehtane obratno sorazmerno z verjetnostjo sprejetja posojila in vzamejo v zakup vpliv vseh zavrnjenih s podobnimi karakteristikami (Verstraeten & Van den Poel, 2005).

Ekstrapolacija je po drugi strani metoda, ki klasificira množico zavrnjenih na modelu, ki je ocenjen samo na množici sprejetih. S tem se pripiše porazdelitev dobrih in slabih s pomočjo osnovnega modela na množico zavrnjenih. Model, ki ga ponovno ocenimo na množici vseh, predstavlja končni model za ocenjevanje verjetnosti neplačila (Crook & Banasik, 2004; Verstraeten & Van den Poel, 2005).

2.2.2 Problem majhnega vzorca

V fazi zbiranja podatkov banka lahko ugotovi, da njihova baza podatkov ne vsebuje zadostnega število komitentov, na podlagi katerega bi lahko razdelili podatke na razvojni in validacijski vzorec. Siddiqi (2006) pravi, da je dovolj veliko število za razvoj modela okoli 2000 neplačnikov in 2000 plačnikov, ki jih naključno izberemo iz baze.

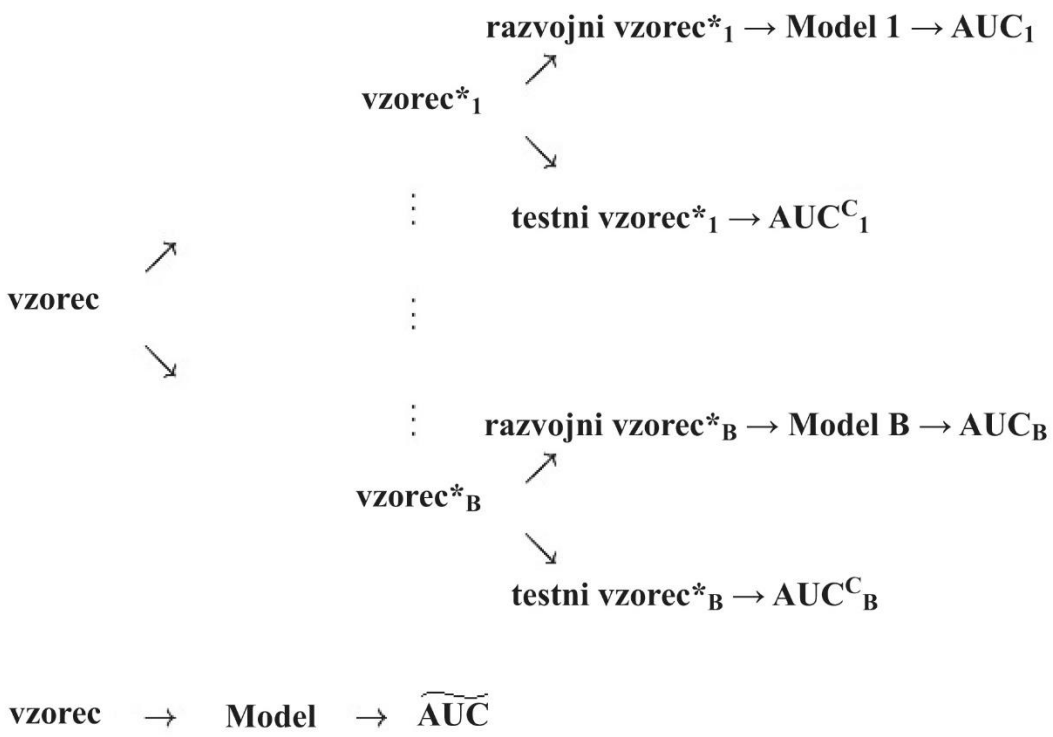
V primeru majhnega števila komitentov v bazi praktično ni mogoče testiranje zunaj vzorca (angl. *Out-of-sample validation*), saj s tem zmanjšamo zanesljivost modela (OeNB & FMA, 2004, str. 73). Da banke lahko razvijejo model na podlagi vseh razpoložljivih podatkov, lahko uporabijo tako imenovano metodo ponovnega vzorčenja, s pomočjo katere preverimo nepristranskost modela. Poleg te metode Thomas et al. (2002) opiše postopek križnega preverjanja (angl. *Cross validation*). Medtem ko so pri postopku križnega preverjanja podatki razdeljeni na dva vzorca, pri metodi ponovnega vzorčenja delitve podatkov ni.

Pri postopku križnega preverjanja ločimo dve metodi izbire vzorca, to je metodo rotacije (angl. *Rotation method*) in metodo, kjer testiranje poteka na eni opazovani enoti (angl. L*eave-one-out method*). Pri obeh metodah večkrat ponovimo testiranje uspešnosti modela in vzamemo povprečje dobljenih rezultatov. Pri prvi metodi je vzorec razdeljen na m disjunktnih podskupin, vsaka pa predstavlja validacijski vzorec za model, ki je zgrajen na preostalih. Povedano drugače – modeliranje poteka na m-1 vzorcih, validacija pa na preostalem vzorcu. Mera uspešnosti modela je povprečje napak (angl. *Error terms*). Če

uspešnost modela merimo na primer z mero AUC, potem vzamemo povprečje vseh *m* mer AUC, ki smo jih izračunali. Pri drugi metodi vsako opazovano enoto uporabimo za validacijski vzorec modela, zgrajenega na preostalih podatkih. Postopek se ponovi za vse opazovane enote in je zato lahko zamuden. Rezultati metode rotacije niso tako robustni kot v primeru metode testiranja na eni opazovani enoti (Thomas et al., 2002, str. 111).

Metoda ponovnega vzorčenja s ponavljanjem, z razliko od že omenjenih metod, za razvoj modela uporabi razvojne vzorce, ki so enake velikosti, kot je velikost baze podatkov. S pomočjo vzorčenja s ponavljanjem pridobimo *m* razvojnih vzorcev, medtem ko testiranje lahko poteka na originalnem vzorcu oziroma bazi podatkov. V razvojem vzorcu se zaradi vzorčenja s ponavljanjem opazovana enota lahko pojavi večkrat (Thomas et al., 2002, str. 111). Kanczes (2012) je v svojem delu opisal način, kako pridobiti oceno mere AUC s pomočjo metode ponovnega vzorčenja. Bazo podatkov s pomočjo vzorčenja s ponavljanjem razdelimo na *B* razvojnih in validacijskih vzorcev. Slika 6 ponazarja možni način metode ponovnega vzorčenja. Za vsak vzorec se za razvojni in testni vzorec izračuna meri AUC, s pomočjo katerih pridobimo oceno AUC, ki bi jo pridobili na neodvisnem validacijskem vzorcu. Ta je pričakovano nižja kot pa ocena mere AUC, pridobljena na razvojnem vzorcu, kar je posledica prevelikega prilagajanja modela razvojnemu vzorcu (angl. *Overfitting*). Omenjeno metodo bom uporabila v svoji magistrski nalogi.

Slika 6: Metoda ponovnega vzorčenja mere AUC



Vir: L. Kanczes, Bootstrapping and its Applications to Credit Scoring, 2012, str. 22.

Naj bo N velikost vzorca in B število ponovitev. Naj bo b = 1, 2 … B. Mera AUC, pridobljena na razvojnem vzorcu, naj bo AUC_b, na testnem pa AUC_b^c. Predpostavka, ki velja, je, da je povprečje napak $AUC_b - AUC_b^c$ dobra ocena med $\widehat{AUC}$ in $\widetilde{AUC}$. Prva predstavlja mero AUC, pridobljeno na modelu, kjer testiranje poteka na istem vzorcu, in druga mero AUC, ko je testiranje opravljeno zunaj vzorca.

Verjetnost, da vzorec $*_b$ ne vsebuje elementa prvotne baze podatkov, je enaka $\left(1 - \frac{1}{N}\right)^N$. Ko se N približuje neskončnosti, verjetnost konvergira k $\frac{1}{e}$, kar pomeni, da je verjetnost, da vzorec $*_b$ vsebuje element, enaka $1 - \frac{1}{e}$. Tako lahko pridobimo oceno mere $\widehat{AUC}$ s pomočjo cenilke:

$$\widehat{AUC} \approx \sum_{b=1}^{B} \left(\left(1 - \frac{1}{e}\right) AUC_b^c + \left(\frac{1}{e}\right) AUC_b \right) \qquad (3)$$

2.2.3 Izbira razvojnega in validacijskega vzorca

Analizi in pripravi podatkov, ko imamo zagotovljeno veliko število opazovanih enot, sledi delitev na razvojni in validacijski oziroma testni vzorec. Prvi vzorec je namenjen ocenjevanju modela, ki ga preverimo na drugem. Kot že zapisano: tipično je 70–80 % podatkov namenjeno razvoju, preostalih 20–30 % pa za testiranje zunaj vzorca. Pomembno je, da so opazovane enote naključno vključene v vzorec in vedno v le enega od vzorcev. Poleg same delitve vzorca na razvojnega in validacijskega lahko prilagajamo tudi število neplačnikov in plačnikov, vključenih v posamezen vzorec (Siddiqi, 2006). V delu bom raziskala vpliv spreminjanja razmerja delitve celotne populacije na razvojni in validacijski vzorec ter spreminjanja števila plačnikov ter neplačnikov v posameznem vzorcu na kakovost modela. Velja, da večji kot bo razvojni vzorec, bolje bo model razvrstil komitente, večji testni vzorec pa nam poda natančnejšo oceno stopnje napake (Liu, 2002).

Testiranje zunaj vzorca, ki ga lahko zagotovimo le ob zadostnem številu komitentov v bazi, je pomembno zaradi možnosti pojava prevelikega prilagajanja modela (Siddiqi, 2006, str. 127). Katere mere lahko uporabimo za testiranje kakovosti modela, bom opisala v nadaljevanju.

2.2.4 Izbira modela

Ključni korak pri modeliranju PD je izbira modela. Cilj je izbrati model, ki bo najbolje razlikoval med plačniki in neplačniki. V praksi je zaradi svoje preprostosti največkrat uporabljena logistična regresija, ki pa je v literaturi velikokrat uporabljena kot primerjalni

model (Jagrič, 2011; Lee et al., 2002; Desai et al., 1996; West, 2000). Logistična regresija spada med skupino regresijskih analiz in je statistična metoda, s pomočjo katere analiziramo odnos med odvisno spremenljivo in eno ali več neodvisnimi spremenljivkami. Najbolj poznana med vsemi regresijskimi analizami je linearna regresija, kjer je povezava med odvisno in neodvisnimi oziroma pojasnjevalnimi spremenljivkami linearna. V primeru logistične regresije predpostavljamo linearno povezavo med logaritmom obeta (angl. *Odds*), odvisne spremenljivke in pojasnjevalnih spremenljivk. Pri problemu ocenjevanja PD je logistična regresija primernejša, saj že po sami zasnovi problema pride do kršitev klasičnih predpostavk linearne regresije. Prva predpostavka linearne regresije, ki je kršena, je zvezna porazdelitev odvisne spremenljivke. V primeru ocenjevanja PD je odvisna spremenljivka binarna, saj nas zanima status komitentov, ki zavzema le dve različne vrednosti. Druge predpostavke so še (Jagrič, 2011, str. 17):

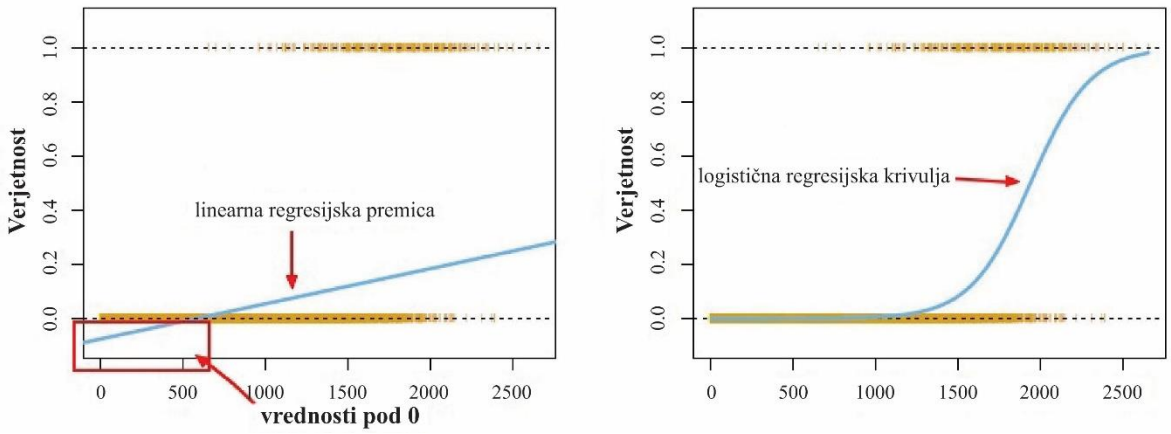
- normalna porazdelitev naključne spremenljivke, kjer je njena povprečna vrednost enaka nič,
- homoskedastičnost – naključna spremenljivka ima konstantno pogojno varianco,
- odsotnost avtokorelacije – ničelna korelacija med vrednostmi odvisne spremenljivke,
- ortagonalnost – odvisna spremenljivka in neodvisne spremenljivke niso korelirane.

Vse naštete predpostavke pri logistični regresiji niso potrebne. Čeprav ni predpostavke o normalni porazdelitvi odvisne spremenljivke, pa se predpostavlja, da je porazdelitev iz skupine eksponentnih porazdelitev – normalna, Poissonova, binomska in gama. Poleg zgoraj omenjenih linearne povezave in binarne odvisne spremenljivke veljajo pri logistični regresiji še naslednje predpostavke:

- ni multikolinearnosti – nobena neodvisna spremenljivka ni linearna kombinacija drugih,
- osamelci in manjkajoči podatki so nezaželeni – zaželeno je, da se pred modeliranjem odstrani osamelce in ustrezno popravi manjkajoče vrednosti v podatkih,
- veliki vzorci,
- naključni vzorec – neodvisnost opazovanih enot.

Slika 7 je grafični prikaz, zakaj je linearna regresija neprimerna za uporabo v primeru binarne odvisne spremenljivke. Regresijska premica se namreč slabo prilega vrednostim v primerjavi z logistično regresijsko krivuljo.

Slika 7: Primerjava linearne in logistične regresije



Vir: Statstutor,. Pearson's correlation , b.l..

Enostavni linearni regresijski model lahko zapišemo v obliki enačbe:

$$E(Y|X = x) = \beta_0 + \beta_1 x, \tag{4}$$

kjer Y predstavlja odvisno in X neodvisne spremenljivke, β_0 in β_1 pa sta regresijska koeficienta, ki ju želimo oceniti. V primeru binarnega modela želimo omejiti $E(Y|X = x)$ med meji 0 in 1. Zgornjo enačbo označimo kot:

$$E(Y|X = x) = \pi(x) = p, \tag{5}$$

kjer je p delež neplačnikov v vzorcu:

$$p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \tag{6}$$

in velja, da predstavlja nelinearno povezavo med odvisno spremenljivko in neodvisnimi spremenljivkami. Če enačbo nekoliko preuredimo, dobimo tako imenovane obete:

$$o = \frac{p}{1 - p} = e^{\beta_0 + \beta_1 x}. \tag{7}$$

S pomočjo naravne logaritemske funkcije enačbo (7) transformiramo in tako pridobimo linearno povezavo med neodvisnimi spremenljivkami in logaritemsko transformacijo odvisne spremenljivke:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x. \qquad (8)$$

Enačba (8) je tako imenovana logit transformacija in predstavlja populacijski regresijski model, iz katerega lahko izpeljemo obete, kot so prikazani v enačbi (7). Transformacija ima številne želene značilnosti linearnega regresijskega modela, saj je linearna v koeficientih, lahko je zvezna in zavzema neskončne vrednosti (Hosmer Jr. & Lemeshow, 2004). Koeficienti regresije so ocenjeni z metodo največjega verjetja, kjer je funkcija verjetja v splošnem enaka

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{y_i}, \qquad (9)$$

pri čemer je β vektor koeficientov, ki jih ocenjujemo, $\pi(x_i)^{y_i}(1-\pi(x_i))^{y_i}$ prispevek k funkciji verjetja vsakega para (x_i,y_i). Zaradi preprostejšega izračuna enačbo (9) logaritmiramo in cenilke pridobimo s pomočjo naslednje enačbe:

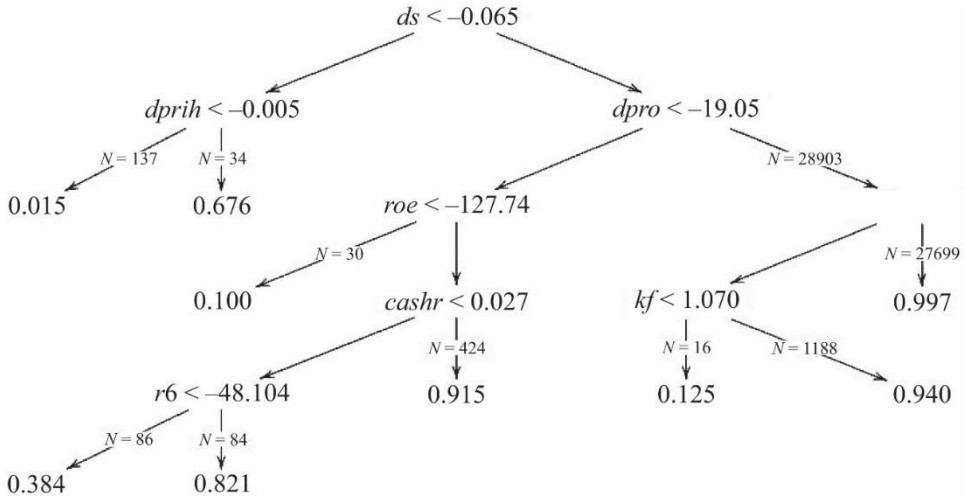
$$\max_{\beta} L(\beta) : L(\beta) = ln(l(\beta)) = \sum_{i=1}^{n}\{y_i ln[\pi(x_i)] + (1-y_i) ln[1-\pi(x_i)]\}. \qquad (10)$$

Poleg logistične regresije bom v magistrski nalogi uporabila metodo odločitvenih dreves. Logistična regresija je preprosta za uporabo, saj je računsko in interpretacijsko nezahtevna. Uporaba logistične regresije ni primerna v primeru manjkajočih ali koreliranih podatkov (Bolton, 2010, str. 29). Za razliko od logistične regresije, ki predpostavlja, da je porazdelitev odvisne spremenljivke iz skupine eksponentnih, pri odločitvenih drevesih ni predpostavke v zvezi s porazdelitvijo odvisne spremenljivke. Zato odločitvena drevesa veljajo za neparametrično metodo (Lewis, 2000). Poleg tega za razliko od logistične regresije dopuščajo multikolinearnost spremenljivk. Podobno kot logistična regresija je tudi metoda odločitvenih dreves preprosta za uporabo in interpretacijo. Deluje tudi v primerih, ko so v podatkih prisotne manjkajoče vrednosti ali so numerični ter kategorični podatki nesimetrično porazdeljeni. Njene glavne pomanjkljivosti so računska zahtevnost v primeru večjih baz podatkov ter nestabilnost metode v primeru majhnih sprememb vhodnih podatkov (Bolton, 2010).

Odločitveno drevo (angl. *Decision tree*) je statistična metoda klasifikacije s strukturo drevesa. S pomočjo odločitvenih pravil iz korena, ki predstavlja celotno bazo podatkov, proti zunanjim vozliščem razdeli bazo na manjše skupine, ki so bolj homogene. Drevesa so sestavljena iz vozlišč, med katere sodi tudi začetno vozlišče, imenovano koren, vej in listov. Vozlišča predstavljajo neodvisne spremenljivke, njihove vrednosti ustrezajo vejam in listi razredom, v našem primeru neplačnikom in plačnikom (Berry & Linoff, 2004). Vozlišča vsebujejo preizkus neodvisnih spremenljivk, kjer se baza razdeli glede na vrednost atributa.

Zato je pomembno, kako dobro posamezni atribut ločuje plačnike od neplačnikov. V ta namen uporabimo meri, kot sta informacijski prispevek (angl. *information gain*) ali Ginijev koeficient. Primer prikaza odločitvenega drevesa je na Sliki 8. Iz njega je razvidno, kateri so tisti kazalniki, ki ločujejo med plačniki in neplačniki. Skupno ima drevo 8 vozlišč, kar pomeni, da se baza podatkov glede na definirane vrednosti v vozliščih razdeli na manjše skupine podatkov, katerim model priredi ocenjeno PD. Tako je v primeru, ko je kazalnik ds manjši od –0,065 in kazalnik dprih manjši od –0,005, PD pa enaka 0,015. Iz Slike je razvidno, da je število takih komitentov enako 137.

Slika 8: Primer prikaza odločitvenega drevesa



Vir: A. Brezigar-Masten & I. Masten, CART-based selection of bankruptcy predictors for the logit model, 2012, str. 12.

V praksi so zaželena manj kompleksna odločitvena drevesa, saj je njihova interpretacija lažja. Kako kompleksno je drevo, se določi na podlagi skupnega števila vozlišč, dolžine in števila atributov. Kompleksnost pomembno vpliva na natančnost modela, ki jo lahko uravnavamo prek kriterija zaustavitve (angl. *Stopping criteria*) in rezanja (angl. *Pruning*) (Dahan, Cohen, Rokach & Maimon., 2014). Pri kriteriju zaustavitve, kamor sodi tudi rezanje vnaprej (angl. *Prepruning*), se s postavitvijo mejne vrednosti prepreči preveliko prilagajanje razvojnemu vzorcu. Pri naknadnem rezanju (angl. *Postpruning*) dodatno odstranimo meje, potem ko je celotno drevo že zgrajeno (Osei-Bryson, 2007). V praksi se je, čeprav je zamudnejše, izkazalo naknadno rezanje (Han, Pei & Kamber, 2011, str. 346).

Obstajajo različne metode, kot so CART (angl. *Classification and regression trees*) in algoritma C4.5 in ID3. CART in C4.5 sta sestavljena iz faze rasti in rezanja drevesa (Rokach & Maimon, 2005), ID3 in C4.5 temeljita na entropiji, kjer je kriterij za razdelitev informacijski prispevek spremenljivke, CART pa temelji na Ginijevem koeficientu (Hssina, Merbouha, Ezzikouri & Erritali, 2014). Kljub priljubljenosti v literaturi je veliko avtorjev zasledilo slabšo klasifikacijsko sposobnost odločitvenih dreves v primerjavi z drugimi

metodami, kot so diskriminantna analiza, logistična regresija in nevronske mreže (Finlay, 2011; Huang, Chen & Wang, 2007; Ong et al., 2005; Xiao, Zhao, & Fei, 2006). Brezigar-Masten in Masten (2012) sta sicer na vzorcu podjetij pokazala, da je kombinacija logistične metode in metode CART dosegla boljše rezultate kot vsaka metoda posebej.

2.2.5 Izbira mere validacije

Zadnji korak procesa modeliranja PD je validacija modela. Kot pove že ime validacijskega vzorca, je ta namenjen preverjanju kakovosti izbranega modela. Katere mere so najpogosteje uporabljene, sem omenila že v prejšnjih poglavjih, sam izračun mer pa bo predstavljen v nadaljevanju. V praksi ne obstaja metoda validacije, ki bi bila splošno sprejeta (Medema, Koning & Lensink, 2009), zato se banka sama odloči, katero bo porabila. Baselski sporazum določa, da morajo imeti banke robustne sisteme validacije natančnosti in konsistentnosti bonitetnih sistemov. Medema et al. (2009) proces validacije deli na teoretično in statistično validacijo ter validacijo podatkov. Prva preveri pravilnost predpostavk uporabljenega modela, medtem ko druga meri njegovo natančnost. Cilj tretje je reprezentativnost podatkov glede na ciljno populacijo. V nalogi se bom v večji meri ukvarjala s statistično validacijo, v sklopu katere bom preverila moč razločevanja (angl. *Discrimination*) in kalibracijo (angl. *Calibration*) modela. Moč razločevanja je sposobnost modela, da zadovoljivo ločuje med neplačniki in plačniki, kalibracija pa je zmožnost nepristranske ocene statusa komitenta (Medema et al., 2009). Mere razločevanja so npr. krivulje ROC, AUC, CAP in Ginijev koeficient.

Matrika zamenjav predstavlja tabelarični prikaz razmerja med pravilno in napačno razvrščenimi plačniki in neplačniki; prikazana je v Tabeli 3. TP (angl. *True positive*) in TN (angl. *True negative*) predstavljata pravilno razvrščene plačnike in neplačnike, FP (angl. *False positive*) in FN (angl. *False negative*) pa napačno razvrščene neplačnike, ki jih model razvrsti kot plačnike, in plačnike, ki jih razvrsti kot neplačnike.

Tabela 3: Matrika zamenjav

		Napovedana vrednost	
		Plačniki	Neplačniki
Dejanska vrednost	Plačniki	TP	FN
	Neplačniki	FP	TN

Vir: N. Siddiqi, Credit risk scorecards: developing and implementing intelligent credit scoring, 2006, str. 120.

Statistike oziroma kazalniki, ki jih definiramo s pomočjo matrike zamenjav, so:

- občutljivost ali preklic (angl. *Sensitivity or recall or True positive rate*): meri delež pravilno razvrščenih plačnikov:

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{11}$$

- specifičnost (angl. S*pecificity or True negative rate*): meri delež pravilno razvrščenih neplačnikov:

$$TNR = SPC = \frac{TN}{N} = \frac{TN}{FP + TN} \tag{12}$$

- vrednost pozitivno napovedanih (angl. *Precision or Positive predictive value*): predstavlja delež vseh pravilno razvrščenih plačnikov med vsemi komitenti, ki jih model prepozna kot plačnike:

$$PPV = \frac{TP}{TP + FP} \tag{13}$$

- vrednost negativno napovedanih (angl. *Negative predictive value*): predstavlja delež vseh pravilno razvrščenih neplačnikov med vsemi komitenti, ki jih model prepozna kot neplačnike:

$$NPV = \frac{TN}{TN + FN} \tag{14}$$

- stopnja lažno pozitivnih vrednosti (angl. *False positive rate*): predstavlja delež napačno razvrščenih neplačnikov med vsemi neplačniki:

$$FPR = \frac{FP}{N} \tag{15}$$

- stopnja napačnega odkritja (angl. *False discovery rate*): predstavlja delež napačno razvrščenih neplačnikov med vsemi komitenti, ki jih je model prepoznal kot plačnike:

$$FDR = \frac{FP}{TP + FP} = 1 - PPV \tag{16}$$

- stopnja lažno negativnih vrednosti (angl. *False negative rate*): predstavlja delež napačno razvrščenih plačnikov med vsemi plačniki:

$$FNR = \frac{FN}{P} \tag{17}$$

- stopnja natančnosti (angl. *Accuracy*): predstavlja delež pravilno razvrščenih komitentov glede na vse komitente:

$$ACC = \frac{TP + TN}{P + N} \tag{18}$$

- stopnja napake (angl. *Error rate*): predstavlja delež napačno razvrščenih komitentov glede na vse komitente:

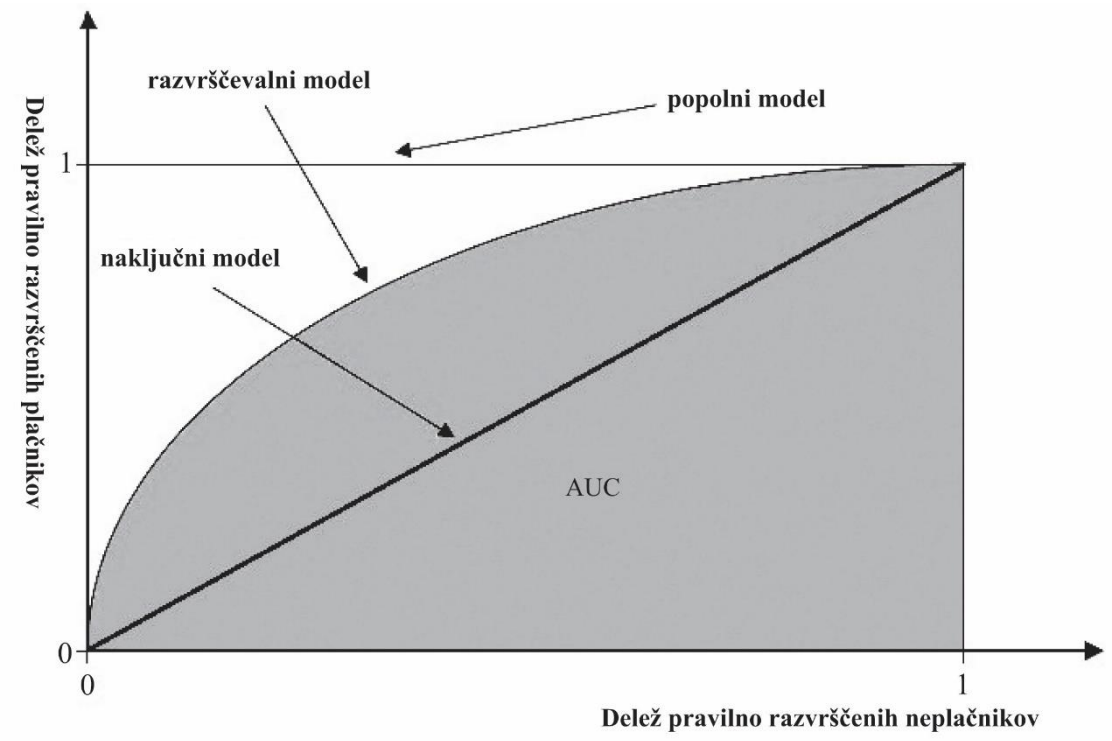
$$ERR = \frac{FP + FN}{P + N} \tag{19}$$

V literaturi so najbolj priljubljene statistike za primerjavo napovedne moči modelov stopnja natančnosti, občutljivost in specifičnost (Bolton, 2010; Jagrič, 2011; Siddiqi, 2006).

V splošnem želimo, da ima model čim višjo stopnjo natančnosti oziroma čim nižjo stopnjo napake. Glede na to, da ima banka večjo izgubo od plačnika, kot pa je njegov dobiček, je v splošnem zato pomembnejša specifičnost kot pa občutljivost. Vpliv na zgornje mere ima tudi izbira mejne vrednosti, ki določa, ali bo komitentu pripisan status plačnika ali neplačnika. S spreminjanjem mejne vrednosti lahko vplivamo na spreminjanje statistik in s tem v model uvedemo subjektivnost (Tsai, Lin & Cheng., 2009). Mejna vrednost, ki je velikokrat definirana na stopnji 0,5, vpliva na kalibracijo modela (Medema et al., 2009), saj z njenim spreminjanjem vplivamo na pravilno razvrstitev komitentov.

Krivulja ROC je grafični prikaz razmerja med deležem pravilno razvrščenih plačnikov in deležem neplačnikov, ki jih je model ocenil kot neplačnike. Idealni model bi pravilno razvrstil vse neplačnike v razred neplačnikov, kar bi pomenilo, da bi krivulja ROC potekala od točke (0,0) prek (0,1) do (1,1). V primeru modela, ki ne bi ločeval med neplačniki in plačniki in bi napovedoval zgolj naključno, bi krivulja ROC potekala po diagonali od točke (0,0) do (1,1). Primer krivulje ROC je prikazan na Sliki 9.

Slika 9: Krivulja ROC



Vir: S. E. Satchell & W. Xia, Analytic models of the ROC curve: Applications to credit rating model validation, 2007, str. 7.

V kolikor bi želeli primerjati več kot dva modela, bi graf krivulje ROC lahko postal nepregleden. Zato uvedemo mero AUC, ki meri površino pod krivuljo. V primeru popolnega modela je mera AUC enaka 1, kjer pa model naključno napoveduje neplačnike, pa 0,5. Pri primerjavi modelov je boljši tisti, ki ima višjo vrednost AUC (OeNB & FMA, 2004).

Medtem ko krivulja ROC predstavlja mero zmožnosti diskriminiranja modela med plačniki in neplačniki, AUC predstavlja verjetnost, da je ocenjena PD za neplačnika višja kot za plačnika (Jovan & Šušterič, 2004, str. 109). Kot trdita Jovan in Šušterič, velja v praksi pravilo pri evalvaciji mere AUC, in sicer:

- če je mera AUC enaka 0,5, model ne razlikuje med plačniki in neplačniki,
- če je mera AUC med 0,7 in 0,8, potem je razločevanje modela sprejemljivo,
- če pa je mera AUC višja od 0,8, je razločevanje modela odlično.

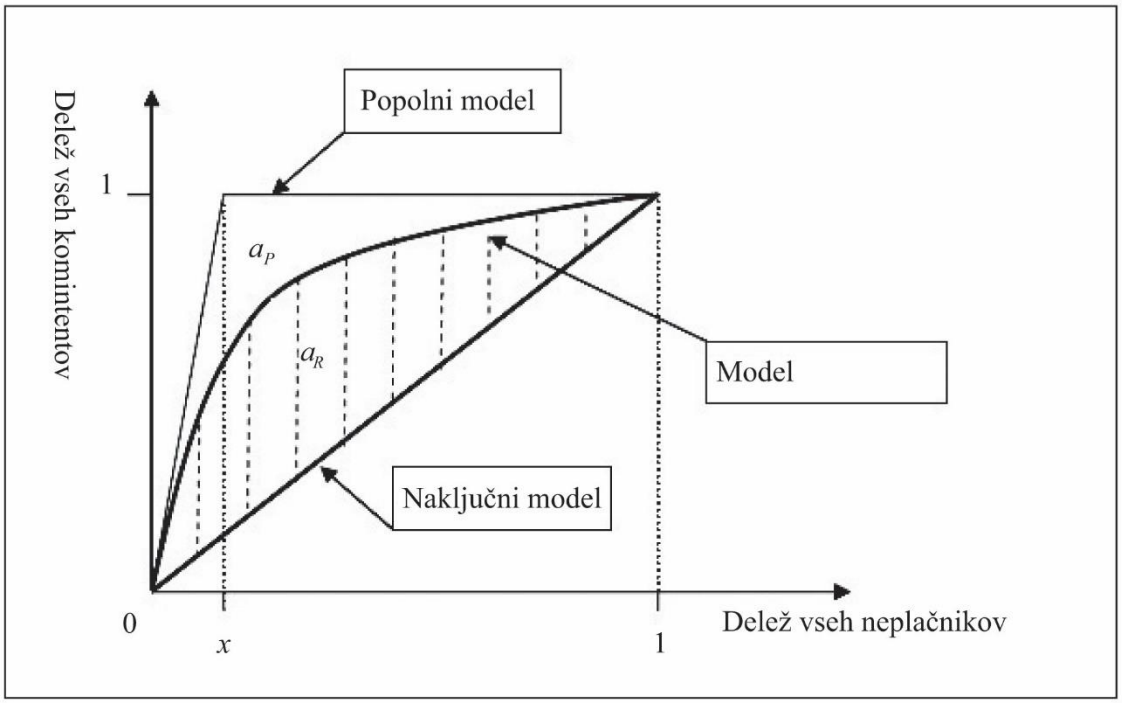
Če namesto kumulativne frekvence plačnikov uporabimo kumulativno frekvenco vseh plačnikov in neplačnikov skupaj, dobimo krivuljo CAP, poznano tudi kot Ginijeva ali Lorenzova krivulja (prikazana na Sliki 10). Podobno kot je mera AUC izpeljana iz krivulje ROC, je Ginijev koeficient definiran na osnovi krivulje CAP. V literaturi Ginijev koeficient poimenujejo tudi razmerje natančnosti (angl. *Accuracy ratio*) (Engelmann et al., 2003).

Predstavlja razmerje površine med diagonalo in krivuljo CAP (na Sliki 10 označena z a_R) glede na površino med diagonalno in krivulo CAP popoldnega modela (na Sliki 10 označena z a_P). Velja:

$$Ginijev\ koeficient = 2*AUC - 1. \tag{20}$$

Podobno kot pri AUC tudi pri Ginijevem koeficientu velja, da višja vrednost koeficienta pomeni višjo napovedno moč modela.

Slika 10: Krivulja CAP



Vir: S. E. Satchell & W. Xia, Analytic models of the ROC curve: Applications to credit rating model validation, 2007, str. 6.

Poleg krivulje ROC je priljubljena validacijska tehnika tudi krivulja CAP in pripadajoči Ginijev koeficient (Engelmann et al., 2003, str. 1). BCBS (2005) je poleg krivulje CAP kot najpomembnejšo mero navedel krivuljo ROC, na temo katere obstaja veliko literature (Satchell & Xia, 2007; Medema et al., 2009; Vuk & Curk, 2006; Engelmann et al., 2003).

3 EMPIRIČNA ANALIZA

Modeliranje v sklopu magistrskega dela temelji na prosto dostopnih podatkih spletne strani Kaggle, ki so bili del tekmovanja o napovedovanju verjetnosti neplačila. Cilj je bila izdelava modela z najvišjo vrednostjo AUC. Najvišja mera AUC, ki je bila dosežena, je znašala

0,869558. Podatka o tem, katera od metod za ocenjevanje verjetnosti neplačila je bila uporabljena v tem primeru, nisem pridobila. Izdelava empirične analize je izvedena v prosto dostopnem programskem paketu za statistične analize in manipulacijo podatkov, imenovanem R (Matloff, 2011).

3.1 Definicija neplačila

Prvi korak modeliranja je sama definicija neplačnika. Sklep o izračunu kapitalske zahteve za kreditno tveganje po pristopu na podlagi notranjih bonitetnih sistemov za banke in hranilnice v 28. členu splošno opredeljuje dogodek neplačila kot:

- neplačilo posameznega dolžnika, ki nastopi, ko se zgodi eden ali oba od naslednjih dogodkov:
  - banka meni, da obstaja majhna verjetnost, da bo dolžnik poravnal svoje kreditne obveznosti do banke, njene nadrejene družbe ali katere koli njej podrejene družbe v celoti, ne da bi bilo za poplačilo treba uporabiti ukrepe, kot je unovčenje zavarovanja (če obstaja),
  - dolžnik več kakor 90 dni zamuja s plačilom katere koli pomembne kreditne obveznosti do banke, njene nadrejene družbe ali katere koli njej podrejene družbe,
- dnevi zamude začnejo teči, ko dolžnik svojih obveznosti ne poravna v pogodbeno dogovorjenih rokih oziroma prekorači odobrena okvirna posojila in je zadevni znesek v zamudi pomemben,
- dnevi zamude za kreditne kartice začnejo teči od dne, ko bi dolžnik moral poravnati obveznosti.

Definicija je skladna s priporočili oziroma opredelitvijo definicije neplačila po baselskem kapitalskemu sporazumu (BCBS, 2006, str. 100).

V bazi podatkov iz opisa spremenljivk opazimo, da je definicija neplačila skladna z zgornjo opredelitvijo. Odvisna spremenljivka oziroma status komitenta je spremenljivka, ki določi, ali je komitent v obdobju dveh let zamujal s plačilom več kot 90 dni in si s tem pridobil status neplačnika. Vendar pa ne moremo z gotovostjo trditi, ali so pri definiciji neplačnika upoštevani le materialno pomembni zneski, kar je opredeljeno v drugem odstavku člena.

3.2 Baza podatkov

Poleg statusa plačnika oziroma neplačnika vsebuje uporabljena baza podatkov še deset drugih spremenljivk. Skupno število opazovanih enot je 150.000 komitentov, med katerimi je 139.974 plačnikov in 10.026 neplačnikov. Obdobje, v katerem opazujemo obnašanje komitenta, je dolgo 24 mesecev. Po navadi zajema 12 mesecev, vendar je lahko razširjeno tudi na daljše obdobje (OeNB & FMA, 2004, str. 67). Komitent pridobi status neplačnika,

če s plačilom v opazovanem obdobju zamuja 90 dni ali več. Podatki niso posnetek slovenskega bančnega sistema, prav tako pa ni znano, iz katerega obdobja so. Vse, kar je znanega, je dejstvo, da so podatki o posojilojemalcih v bančnem sektorju.

3.2.1 Spremenljivke

Kot že omenjeno, baza podatkov vsebuje podatek o statusu komitenta in 10 številskih pojasnjevalnih spremenljivk. Tip spremenljivk in njihova imena so predstavljeni v Tabeli 4.

Tabela 4: Seznam pojasnjevalnih spremenljivk

Oznaka spremenljivke	Ime spremenljivke	Tip spremenljivke	Število manjkajočih vrednosti
X_1	izkoriščenost posojil	delež/odstotek	0
X_2	starost	celoštevilska	1
X_3	zamuda med 30 in 59 dni	celoštevilska	269
X_4	delež dolga	delež/odstotek	0
X_5	mesečni prihodek	realna	29.731
X_6	število kreditnih kartic in limitov	celoštevilska	0
X_7	zamuda nad 90 dni	celoštevilska	269
X_8	število posojil	celoštevilska	0
X_9	zamuda med 60 in 89 dni	celoštevilska	269
X_{10}	število družinskih članov	celoštevilska	3.924

Poleg odvisne spremenljivke, ki določa status komitenta, so prisotne še naslednje pojasnjevalne spremenljivke:

- X_1: izkoriščenost posojil predstavlja skupno stanje na kreditnih karticah in limitov brez zavarovanih posojil, kot so nepremičninska in avtomobilska, deljeno s skupno vrednostjo vseh obstoječih posojil;
- X_2: starost komitenta v letih;
- X_3: kolikokrat je komitent v preteklem obdobju zamujal s plačilom več kot 29 in manj kot 60 dni;
- X_4: mesečni izdatki komitenta, kot so obroki posojil, preživnina in življenjski stroški, deljeno z mesečnim prihodkom komitenta;
- X_5: mesečni prihodek komitenta;
- X_6: število kreditnih kartic in limitov;
- X_7: kolikokrat je komitent v preteklem obdobju zamujal s plačilom nad 89 dni;
- X_8: število hipotekarnih in nepremičninskih posojil;
- X_9: kolikokrat je komitent v preteklem obdobju zamujal s plačilom več kot 59 dni in manj kot 90 dni;

- X_{10}: število vzdrževanih družinskih članov: če je vrednost spremenljivke enaka 1, potem je v skupnem gospodinjstvu poleg komitenta še ena oseba.

Med podanimi spremenljivkami sta le dve iz skupine sociodemografskih podatkov, X_2 – starost in X_{10} – število vzdrževanih članov. Vse ostale spremenljivke opisujejo obnašanje komitenta v preteklosti, njegovo finančno stanje in podatke o številu posojil ter kreditnih linij. To so tako imenovane vedenjske spremenljivke in opisujejo vedenje komitenta. Opisujejo njegovo zadolženost, mesečni prihodek, zamude v preteklosti in podobno. Med sociodemografske podatke je poleg starosti komitenta in števila vzdrževanih članov, ki so del baze podatkov, primerno vključiti še delovno dobo, vrsto zaposlitve, izobrazbo, status bivanja in stan (Siddiqi, 2006; Jagrič, 2011), ki pa niso del baze podatkov, uporabljene v magistrski nalogi. V literaturi je veliko pozornosti posvečene lastnostim komitentov, ki vplivajo na status neplačnika. Kot primer vzemimo starost, kjer je skladno s teorijo življenjskega cikla pričakovano, da je zadolževanje mladih, ko se pričakuje rastoče prihodke, višje kot pri starejši populaciji. Z leti so ljudje vedno manj naklonjeni zadolževanju in tekoči prihodek v večji meri zadovolji njihove potrebe. Višja zadolženost je prisotna tudi v primeru višje izobraženih posameznikov in v primeru velikih družin (Vandone, 2009, str. 18). Namen magistrske naloge je raziskati kakovost modela glede na izbiro odločitve v posameznih korakih modeliranja kreditnega tveganja, zato pozornost ne bo usmerjena na lastnosti posameznika, ki najbolj vplivajo na pridobitev statusa neplačnika.

3.2.2 Priprava podatkov

Prvi korak, ki ga je treba izvesti, je pregled podatkov in njihova kakovost. Pri tem sem preverila manjkajoče vrednosti, napake in osamelce. V Tabeli 4 so zapisani podatki o številu manjkajočih vrednosti. Med njimi sta spremenljivki X_5 – mesečni prihodek in X_{10}. Vsem 3924 komitentom, katerim manjka podatek o številu vzdrževanih članov, hkrati manjka tudi podatek o mesečnem prihodku. 269 komitentom manjkajo podatki o številu dni zamude plačil – spremenljivke X_3, X_7 in X_9. Kot napako v podatkih vzamemo starost komitenta, ki je enaka 0, in tiste, katerih starost je več od 100.

Pri obdelavi manjkajočih podatkov, napak in osamelcev poznamo več metod. Hand, Mannila in Smyth (2001) navajajo:

- odstranitev komitentov z manjkajočimi vrednostmi,
- popravki vrednosti,
- nadomestitev z izbrano vrednostjo.

Odstranitev komitentov z manjkajočimi podatki je mogoča v primeru, ko s tem ne zmanjšamo razpoložljivosti in velikosti baze podatkov ter je še vedno primerna za nadaljnjo raziskavo. Pri metodi popravkov vrednosti je pomembno, da manjkajočo vrednost ali napako

bodisi lahko nadomestimo s pravo vrednostjo na podlagi sklepanja bodisi jo naknadno pridobimo iz svojih virov. Pri kvantitativnih spremenljivkah lahko obstoječo vrednost nadomestimo z izbrano vrednostjo, ki je lahko na primer aritmetična sredina, modus itn. Katero metodo uporabiti pri ravnanju z manjkajočimi podatki, je odvisno od kakovosti poznavanja podatkov. Zaradi zadostne količine podatkov in nesmiselne zamenjave manjkajočih podatkov z ničelnimi vrednostmi komitente z manjkajočimi vrednostmi odstranim iz baze podatkov. Na ta način pridobim bazo podatkov, ki vsebuje 8272 neplačnikov in 111.842 plačnikov.

Osamelci ali anomalije v podatkih so lahko napake ali izstopajoče vrednosti, ki jih je treba primerno obravnavati, saj močno vplivajo na samo modeliranje (Hand et al., 2001). Ghosh in Vogt (2012) navajata tri metode obravnavanja osamelcev:

- izstopajočih vrednosti ne spreminjamo in jih obravnavamo kot vse druge vrednosti,
- metoda preslikovanja robnih vrednosti (angl. *Winsorize*), s pomočjo katere približamo izstopajočo vrednost drugim,
- metoda obrezovanja (angl. *Trimming*) oziroma izločitev izstopajočih vrednosti iz vzorca.

Zadnji dve metodi sta standardna načina za obravnavanje osamelcev. Najpogosteje uporabljena je metoda preslikovanja robnih vrednosti, kjer vse izstopajoče vrednosti zamenjamo z določeno vrednostjo. Primer postopka je, da najvišjih 5 % vrednosti zamenjamo z vrednostjo 95. percentila in najnižjih 5 % z vrednostjo 5. percentila. Pri metodi obrezovanja se vse izstopajoče vrednosti odstranijo iz podatkov (Ghosh & Vogt, 2012).

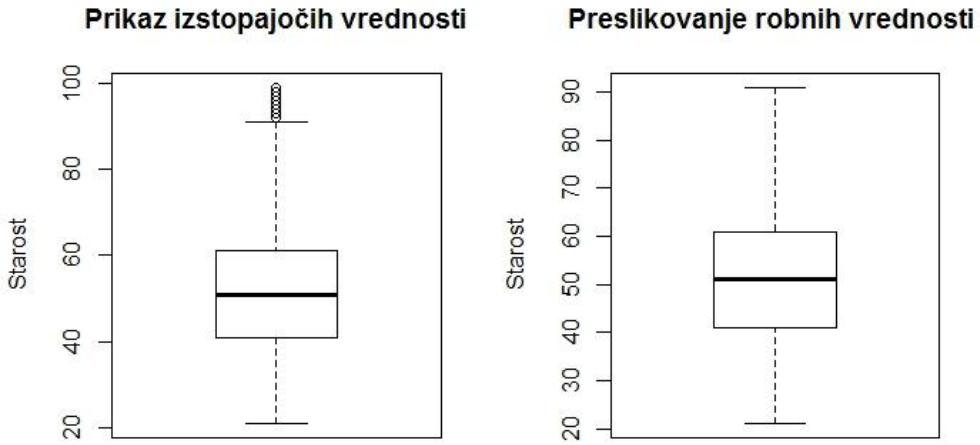
Pri prepoznavanju izstopajočih vrednosti ali osamelcev lahko uporabimo tudi grafični prikaz. Na Sliki 11 sta prikazana kvantilna diagrama (angl. *Boxplot*) pred in po obdelavi izstopajočih vrednosti. Pred obdelavo podatkov iz kvantilnega diagrama spremenljivke X_2 razberemo naslednje podatke:

- mediana ali srednja vrednost je enaka 51 let,
- prvi in tretji kvantil sta enaka 41 in 61 let,
- minimalna in maksimalna vrednost sta 21 in 99 let, minimalna (spodnja meja) in maksimalna (zgornja meja) vrednost, brez upoštevanih izstopajočih vrednosti, sta enaki 21 in 91 let,
- vse, kar je nad minimalno in maksimalno vrednostjo, brez upoštevanih izstopajočih vrednosti, so osamelci ali izstopajoče vrednosti.

V prvem koraku obdelave podatkov so bili izločeni vsi komitenti, katerih starost je bila enaka 0 ali več kot 100 let. V teh primerih je po vsej verjetnosti napaka v podatkih. Ostale izstopajoče vrednosti so bile obravnavane kot izstopajoče vrednosti in so bile odpravljene s

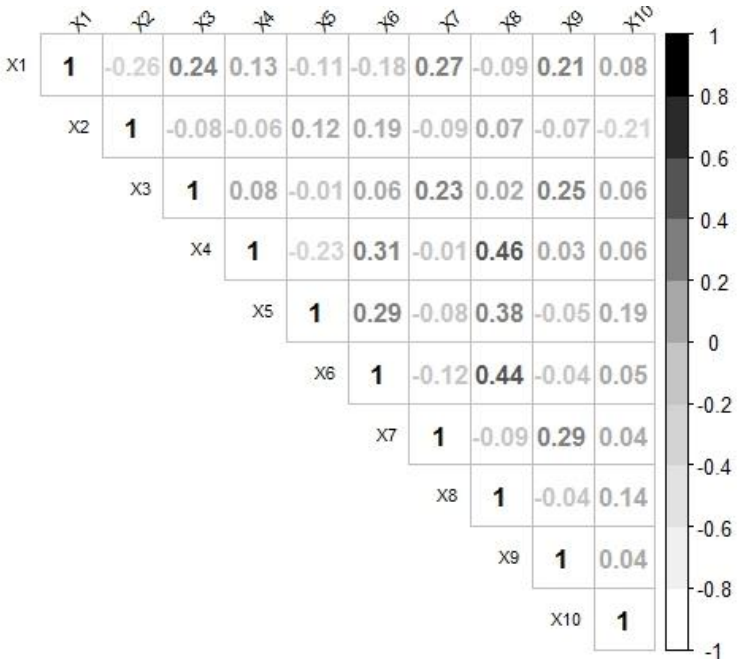
pomočjo metode preslikovanja robnih vrednosti. S tem so bile vse starosti komitentov nad zgornjo mejo zamenjane z 91 leti. Podobno so bile obravnavane vse preostale spremenljivke.

Slika 11: Izstopajoče vrednosti na primeru starosti



Pri pripravi podatkov je treba preveriti korelacijo med spremenljivkami, saj visoka korelacija med njimi lahko vpliva na nestabilnost modela. Bolj sta spremenljivki korelirani, bolj se korelacijski koeficient bliža 1 ali –1. Na splošno velja, da je korelacija med spremenljivkama visoka, kjer je korelacijski koeficient višji od 0,6 ali –0,6 (Statstutor, b.l.). Na Sliki 12 je predstavljena korelacijska matrika, iz katere je razvidno, da nobeni dve spremenljivki ne korelirata previsoko. V obravnavanih podatkih je najvišji korelacijski koeficient 0,46 in ni potrebe po izključitvi katere koli od spremenljivk.

Slika 12: Korelacijska matrika



Po pripravi podatkov je baza primerna za nadaljnjo obdelavo oz. modeliranje.

3.3 Modeliranje verjetnosti neplačila

V nadaljevanju bom opisala, kako je potekal postopek modeliranja verjetnosti neplačila in korake, po katerih sem raziskala določena vprašanja. V tem delu so raziskani vsi problemi, razen problema majhnega vzorca, ki je opisan v poglavju 3.5. Opisani bodo postopek in rezultati na primeru enega od vzorčenj, medtem ko bo primerjava podatkov vseh vzorčenj in metod sledila v naslednjem poglavju. Po predhodni obdelavi podatkov sledi raziskava prvega problema, opisanega v poglavju 2.2 – problem pristranskosti vzorca. Po pregledu podatkov je bilo ugotovljeno, da baza ne vsebuje informacij o zavrnjenih kandidatih, zato je edina možna metoda, ki sem jo lahko uporabila, neupoštevanje zavrnjenih komitentov pri izdelavi modela. Model razvijem na podlagi sprejetih komitentov in bi bil lahko nato uporabljen na vseh posojilojemalcih. V kolikor bi model, glede na predhodno definirano mejno vrednost verjetnosti neplačila, posojilojemalca prepoznal kot neplačnika, mu banka ne odobri posojila (Siddiqi, 2006, str. 103).

Naslednji korak modeliranja je določitev razvojnega in validacijskega vzorca. Glede na literaturo je večina podatkov, to je med 70 in 80 %, namenjenih razvoju modela, medtem ko so preostali uporabljeni za njegovo validacijo. V sklopu magistrskega dela sem preverila, kako različni načini delitve baze vplivajo na kakovost modela. Ne glede na načine delitev, ki so opisani v nadaljevanju, so bili komitenti naključno vključeni v vzorec in so bili vedno le del enega od vzorcev – razvojnega ali validacijskega. Pri delitvi vzorca sem uporabila dve različni metodi. Prva je bila odstotkovna delitev celotne baze na razvojni in validacijski vzorec; pri drugi jih je bilo za razvojni vzorec iz množice vseh neplačnikov naključno izbrano določeno število komitentov. Nato sem jih iz množice plačnikov naključno izbrala toliko, da sem dobila različna razmerja med obema skupinama. Vse preostale komitente sem vključila v validacijski vzorec. Načini vzorčenja so na kratko opisani v nadaljevanju:

- 1. način: iz celotne baze podatkov sem naključno izbrala 50 % komitentov in jih uvrstila v razvojni vzorec. Preostale komitente sem vključila v validacijski vzorec,
- 2. način: iz celotne baze podatkov sem naključno izbrala 70 % komitentov in jih uvrstila v razvojni vzorec. Preostalih 30 % je bilo vključenih v validacijski vzorec,
- 3. način: podobno kot pri prvih dveh vzorčenjih je bilo 80 % komitentov naključno uvrščenih v razvojni vzorec, preostalih 20 % predstavlja validacijski vzorec,
- 4. način: podobno kot pri prvih treh vzorčenjih je bilo 40 % komitentov naključno uvrščenih v razvojni vzorec, preostalih 60 % predstavlja validacijski vzorec,
- 5. način: celotno bazo podatkov sem razdelila na plačnike in neplačnike. Iz množice plačnikov je bilo naključno izbranih 5000 komitentov. Tako sem pridobila razvojni vzorec, v katerem je bilo razmerje med neplačniki in plačniki 1 : 1. Vse preostale neplačnike in plačnike sem vključila v validacijski vzorec,

- 6. način: ponovno sem iz množice neplačnikov naključno izbrala 5000 komitentov in dodala 15.000 komitentov iz množice plačnikov. S tem sem pridobila razvojni vzorec, v katerem je bilo razmerje 1 : 3. Preostali komitenti so bili vključeni v validacijski vzorec,
- 7. način: ponovno sem iz množice neplačnikov naključno izbrala 5000 komitentov in iz množice plačnikov 45.000 komitentov. Pridobljeni razvojni vzorec je imel razmerje 1 : 9. Preostali komitenti so bili vključeni v validacijski vzorec.

V Tabeli 5 so zapisane velikosti vseh vzorcev, koliko plačnikov ter neplačnikov vsebuje vsak razvojni vzorec in skupno število preostalih komitentov, vključenih v validacijski vzorec.

Tabela 5: Lastnosti vzorcev

Načini vzorčenja	Razvojni vzorec			Velikost validacijskega vzorca
	Velikost	Št. neplačnikov	Št. plačnikov	
1	60.057	4.117	55.940	60.057
2	84.079	5.791	78.288	36.035
3	96.091	6.645	89.446	24.023
4	48.045	3.360	44.685	72.069
5	10.000	5.000	5.000	110.114
6	20.000	5.000	15.000	100.114
7	50.000	5.000	45.000	70.114

V nadaljevanju bom na primeru vzorčenja, kjer je prva polovica naključno izbranih podatkov, vključenih v razvojni vzorec, in druga polovica v validacijskega, predstavila rezultate modelov ter opisala postopek. Sledila bosta primerjava rezultatov in vpliv na kakovost modela glede na izbiro odločitve v posameznih korakih.

Delitvi podatkov na razvojni in validacijski vzorec sledi izdelava modela s pomočjo logistične regresije in odločitvenih dreves. V Tabeli 6 je primer prikaza rezultatov logistične regresije za prvi način vzorčenja. Regresijski model vključuje 10 pojasnjevalnih spremenljivk, ki so predstavljene v Tabeli 4. Drugi stolpec v tabeli predstavlja regresijske koeficiente, kjer velja, da pozitivni koeficient povečuje PD – višja vrednost vpliva na večje povečanje. Nasprotno pa spremenljivke z negativnim koeficientom zmanjšujejo PD. Negativni koeficient ima le ena spremenljivka, njegova vrednost je blizu 0. Vse preostale spremenljivke pozitivno vplivajo na PD. Tako na primer višje število vzdrževanih članov (X10) zvišuje PD. Višja starost v nasprotnem primeru zmanjšuje PD.

Vsi koeficienti so statistično značilni pri stopnji značilnosti, manjši od 0,01, kar je mogoče razbrati iz zadnjega stolpca. Kljub temu da spremenljivka zaradi višine koeficienta lahko vpliva na spremembo verjetnosti neplačila, ni nujno, da jo je smiselno obdržati pri analizi.

V kolikor ni statistično značilna, jo je bolj smiselno odstraniti iz regresijskega modela. Dejstvo, da ni statistično značilna, nakazuje, da ni povezana z odvisno spremenljivko (Bolton, 2010, str. 85).

Tabela 6: Primer prikaza rezultata logistične regresije

Spremenljivka	Prikaz rezultata logistične regresije			
	Koeficient	Std. napaka	Z-statistic	Pr(>|z|)
C	–3,7195	0,0917	–40,5635	0,0000
X_1	1,8885	0,0550	34,3669	0,0000
X_2	–0,0151	0,0015	–9,9654	0,0000
X_3	0,4462	0,0195	22,8820	0,0000
X_4	0,2159	0,0811	2,6619	0,0078
X_5	0,0000	0,0000	–6,3078	0,0000
X_6	0,0438	0,0049	8,9098	0,0000
X_7	1,5434	0,0484	31,8787	0,0000
X_8	0,1142	0,0249	4,5936	0,0000
X_9	1,0143	0,0509	19,9098	0,0000
X_{10}	0,0469	0,0156	3,0049	0,0027

Ena možnih interpretacij višine regresijskega koeficienta je, da regresijski koeficient izraža spremembo v ocenjeni logit vrednosti ob spremembi vrednosti pojasnjevalne spremenljivke za eno enoto ob nespremenjenih vrednostih drugih pojasnjevalnih spremenljivk (Jagrič, 2011, str. 150). Če vzamemo primer spremenljivke X_7 – kolikokrat je komitent v opazovanem obdobju zamudil 90 dni ali več – je koeficient enak 1,5434. Povečanje spremenljivke za eno enoto v povprečju logaritem obeta poveča za 1,5434 enote, medtem ko ostale spremenljivke pustimo nespremenjene. Najvišja vrednost koeficienta in s tem pozitivni vpliv na PD sta ugotovljena pri spremenljivki X_1, ki predstavlja izkoriščenost kreditnih kartic in limitov.

Možna interpretacija je tudi s pomočjo obetov. Če enačbo (4) zapišemo splošno, velja:

$$o = \frac{p}{1-p} = e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_i X_i} = e^{\beta_0} * e^{\beta_1 X_1} * \ldots * e^{\beta_i X_i}. \qquad (21)$$

V primeru, ko vse spremenljivke ostanejo nespremenjene in se na primer spremenljivka X_{10} spremeni za eno enoto, bo to povečalo obet za $e^{0,0469}$. Če bi se število vzdrževanih članov povišalo za eno osebo, je obet v prid nastanka dogodka neplačila 1,048017-kratnik tistega, ki je veljal pred spremembo (Jagrič, 2011).

Poleg omenjene interpretacije koeficientov je pomembno tudi, za koliko se pri spremembi ene od spremenljivk spremeni ocenjena vrednost neplačila. Iz baze podatkov pridobim podatke za komitenta in izračunam PD s prvotnimi podatki in s podatki, kjer je le vrednost

spremenljivka X_2, povečana za eno enoto. Podatki so predstavljeni v Tabeli 7. Ob pogoju, da se starost komitenta poviša za eno leto, medtem ko druge spremenljivke ostanejo nespremenjene, je verjetnost, da bo komitent postal neplačnik, manjša. Kot vidimo, se slednje sklada z negativnim koeficientom v Tabeli 6, kar pomeni, da je PD manjša v primeru, ko starost povečamo za eno leto.

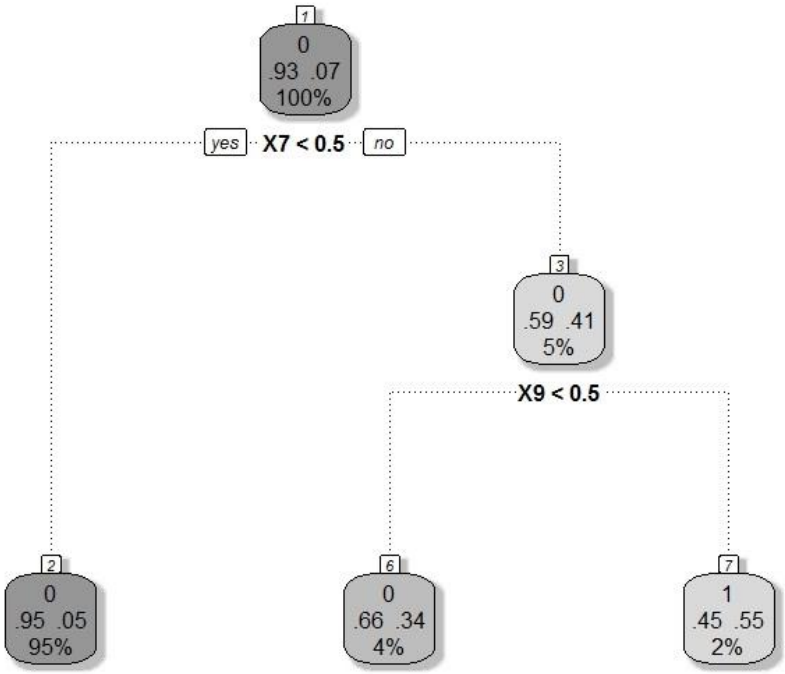
Tabela 7: PD pri spremembi starosti za eno enoto

	Logit vrednost	Obet	PD
Prvotni podatki	–2,306923	0,09956711	0,09055119
Spremenjeni podatki	–2,322048	0,09807254	0,08931335

Na Sliki 13 je prikazano razvrščevalno odločitveno drevo, ki vsebuje pravila za razvrščanje plačnikov in neplačnikov. Interpretacija pravil za razvrščanje je naslednja:

- če je vrednost kazalnika oziroma spremenljivke X_7 manjša od 0,5, je komitent prepoznan kot plačnik,
- če je vrednost spremenljivke X_7 večja od 0,5 in je spremenljivka X_9 manjša od 0,5, je komitent prepoznan kot plačnik,
- če je vrednost spremenljivke X_7 večja od 0,5 in je spremenljivka X_9 večja od 0,5, je komitent prepoznan kot neplačnik.

Slika 13: Prikaz odločitvenega drevesa

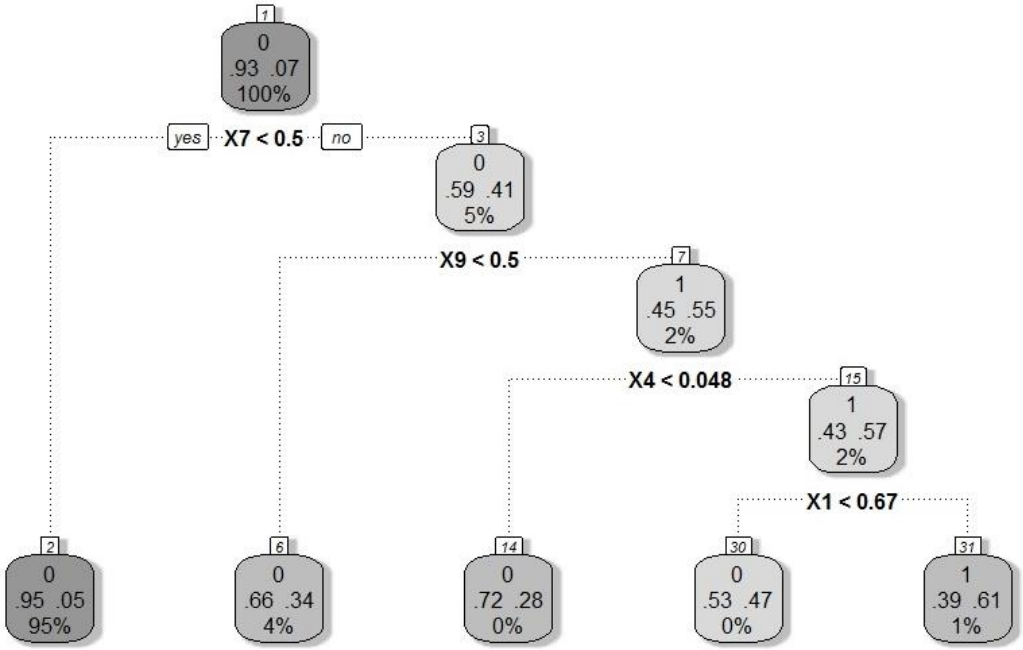


V kolikor je komitent v zadnjih dveh letih zamujal s plačilom več kot 60 dni, je prepoznan kot neplačnik. Pri logistični regresiji je napovedana PD izračunana glede na vrednosti spremenljivk in pridobljenih koeficientov in tako lahko različna za vsakega komitenta. Za razliko od napovedi verjetnosti neplačila pri logistični regresiji je pri napovedi verjetnosti neplačila z odločitvenim drevesom moč opaziti, da so prisotne le tri različne verjetnosti neplačila – 0,0493, 0,3419 in 0,5498. Velja, da je število različnih vrednosti odvisno od števila končnih listov v drevesu. Glede na to, da je mogoče s pomočjo različnih parametrov spreminjati število vozlišč v drevesu, s tem lahko vplivamo tudi na končno število listov. Na Sliki 14 je predstavljeno odločitveno drevo, kateremu smo spremenili tako imenovani parameter kompleksnosti (angl. *Complexity parameter*) in s tem pridobili drevo z večjim številom vozlišč in listov.

Sama interpretacija rezultatov odločitvenih dreves je s pomočjo grafičnega prikaza preprosta. Na Sliki 13 vidimo, da je skupno število vozlišč enako 2 in listov 3. Prekinjene črte predstavljajo veje drevesa. Prva delitev baze v začetnem vozlišču je glede na vrednosti spremenljivke X_7, ki bazo podatkov razdeli na dva dela. Naslednja delitev dela baze je določena z vrednostjo spremenljivke X_9. Več kot je vozlišč, večje je število delitev baze, kar vidimo na Sliki 14. Kolikšen je delež baze v posameznem vozlišču, je razvidno iz odstotka, zapisanega v grafu. Začetno vozlišče ima celotno bazo, kar predstavlja 100 %. Prva delitev glede na vrednost pojasnjevalne spremenljivke 95 % komitentom pripiše oceno verjetnosti neplačila 0,0493; preostalih 5 % komitentov se glede na vrednost spremenljivke X_9 ponovno razdeli v dve skupini: prva ima ocenjeno PD 0,3419, druga 0,5498. Vrednost ocenjene verjetnosti neplačila preberemo iz številk, zapisanih nad deležem baze. V primeru prvega lista, kjer imamo 95 % celotne baze, je ocenjena vrednost neplačila razvidna iz vrednosti, ki je zapisana nad deležem baze, ki jo vsebuje posamezni list. V vsakem vozlišču je razvidno, katera je mejna vrednost spremenljivke, ki deli bazo podatkov. V primeru začetnega vozlišča je mejna vrednost spremenljivke enaka 0,5.

Če primerjamo Sliki 13 in 14, opazimo, da smo s spreminjanjem parametrov število vozlišč povečali z 2 na 4 in število listov s 3 na 5. Razlika je tudi v številu različnih napovedanih verjetnosti neplačila. V primeru večjega števila vozlišč in listov pridobimo tudi večje število različnih ocen verjetnosti. Kako to vpliva na kakovost modela, sem preverila s pomočjo različnih mer.

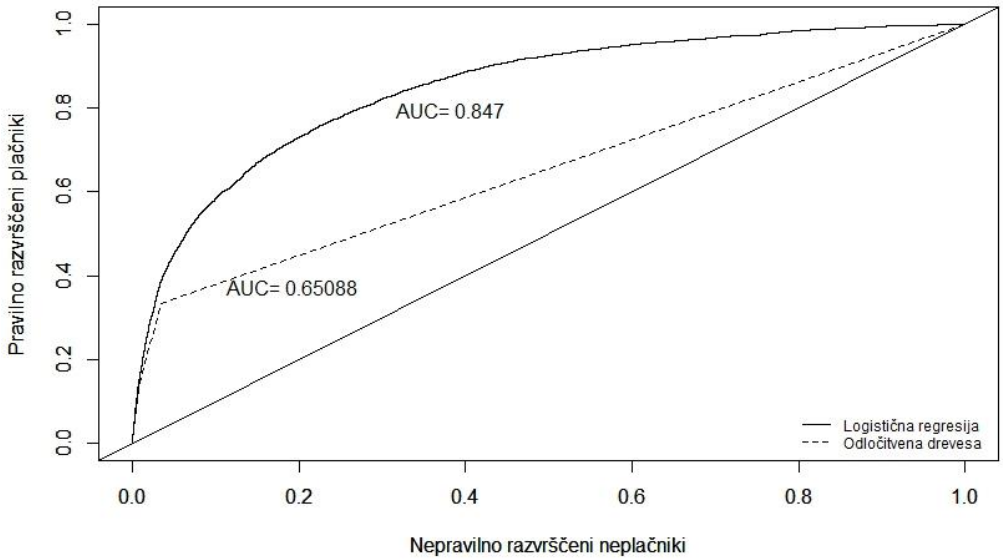
Slika 14: Prikaz odločitvenega drevesa s spremembo parametra



Naslednji korak je primerjava mer validacije pridobljenih rezultatov. Prvi meri validacije sta predstavljeni na Sliki 15. Za oba modela sta prikazani krivulji ROC in zapisani pripadajoči meri AUC. Za primerjavo sem uporabila prvotno odločitveno drevo s tremi vozlišči, saj je razlika v meri AUC minimalna. Mera AUC prvotnega drevesa je enaka 0,65088, medtem ko je mera AUC za spremenjeno drevo 0,65092. Prav tako se krivulji ROC skoraj prekrivata in s pomočjo grafičnega prikaza ne opazimo razlike.

Kot je razvidno iz grafičnega prikaza na Sliki 15, se metoda logistične regresije izkaže za veliko boljšo kot metoda odločitvenega drevesa, saj dosega znatno razlikovanje med plačniki in neplačniki. Mera AUC logistične regresije je 0,847 in po izsledkih iz literature predstavlja zelo dobro ločevanje med komitenti. Po drugi strani pa je mera AUC 0,65088 nizka in slabo ločuje med plačniki in neplačniki.

Slika 15: Primerjava krivulj ROC in mer AUC



Za krivuljo CAP velja, da je ekvivalent krivulji ROC in Ginijevega koeficienta mera, ki se izpelje iz mere AUC (Engelmann et al., 2003), zato sem se pri primerjavi modelov omejila zgolj na krivuljo ROC in pripadajočo mero AUC. V kolikor primerjamo več krivulj ROC, je grafični prikaz težko berljiv in primerjava ni preprosta kot v primeru mere AUC. Zaradi obeh navedenih razlogov bo primerjava končnih rezultatov temeljila na meri AUC, kjer višja vrednost pomeni boljši model.

Poleg mere AUC sem končne rezultate primerjala s pomočjo mer, ki so izpeljane iz matrike zamenjav. Rezultat modela je za vsakega komitenta v obliki ocenjene vrednosti neplačila, kateri je treba glede na postavljeno vrednost pripisati status plačnika ali neplačnika. Zato izdelava matrike zahteva določitev mejne vrednosti, ki ločuje med plačniki in neplačniki. Prikazane so matrike zamenjav pri treh različnih mejnih vrednostih. Tabele vsebujejo števila pravilno in napačno razvrščenih komitentov ter nekatere mere, kot so natančnost (v nadaljevanju ACC), stopnja napake (v nadaljevanju ERR), občutljivost (v nadaljevanju TPR), specifičnost (v nadaljevanju TNR), vrednost pozitivno napovedanih (v nadaljevanju PPV), vrednost negativno napovedanih (v nadaljevanju NPV), stopnja lažno pozitivnih (v nadaljevanju FPR), stopnja napačnega odkritja (v nadaljevanju FDR) in stopnja lažno negativnih (v nadaljevanju FNR). V primeru logistične regresije se število pravilno in napačno razvrščenih komitentov razlikuje v vseh matrikah. V Tabeli 8, 9 in 10 so predstavljene matrike zamenjav pri različnih mejnih vrednostih. V tabelah je zapisano število pravilno in napačno napovedanih plačnikov in neplačnikov glede na dejanske vrednosti.

Tabela 8: Primer matrike zamenjav pri mejni vrednosti 0,5

Logistična regresija		Napovedana vrednost	
		Plačniki	Neplačniki
Dejanska vrednost	Plačniki	55.017	885
	Neplačniki	3.159	996
Odločitvena drevesa		Napovedana vrednost	
		Plačniki	Neplačniki
Dejanska vrednost	Plačniki	55.418	484
	Neplačniki	3.555	600
Odločitvena drevesa s spremembo parametra		Napovedana vrednost	
		Plačniki	Neplačniki
Dejanska vrednost	Plačniki	55.460	442
	Neplačniki	3.584	571

Tabela 9: Primer matrike zamenjav pri mejni vrednosti 0,5

Logistična regresija		Napovedana vrednost	
		Plačniki	Neplačniki
Dejanska vrednost	Plačniki	55.331	571
	Neplačniki	3.433	722
Odločitvena drevesa		Napovedana vrednost	
		Plačniki	Neplačniki
Dejanska vrednost	Plačniki	55.418	484
	Neplačniki	3.555	600
Odločitvena drevesa s spremembo parametra		Napovedana vrednost	
		Plačniki	Neplačniki
Dejanska vrednost	Plačniki	55.595	307
	Neplačniki	3.699	456

Opazimo, da se matrika zamenjav za odločitvena drevesa pri zvišanju mejne vrednosti z 0,4 na 0,5 ne spremeni. Omenjeno je posledica števila različnih ocenjenih vrednosti neplačila. Za razliko od logistične regresije, kjer je število različnih vrednosti neplačila 60.020, je v primeru odločitvenega drevesa enako 3. Pri spremembi mejne vrednosti tako ni spremembe v pripisanem statusu komitenta, saj so verjetnosti neplačila enake 0,0493, 0,3419 in 0,5498. Podobno opazimo pri primerjavi matrike zamenjav odločitvenega drevesa s spremembo parametra, kjer se pri zvišanju mejne vrednosti z 0,5 na 0,55 matrika ne spremeni.

V Tabeli 10 je razvidno, da je pri mejni vrednosti 0,55 vsem komitentom pripisan status plačnika, saj je najvišja ocenjena PD enaka 0,5498, iz česar sledi, da model pri mejni vrednosti 0,55 nobenega komitenta ne prepozna kot neplačnika.

Tabela 10: Matrika zamenjav pri mejni vrednosti 0,55

Logistična regresija		Napovedana vrednost	
		Plačniki	Neplačniki
Dejanska vrednost	Plačniki	55.481	421
	Neplačniki	3.566	589
Odločitvena drevesa		Napovedana vrednost	
		Plačniki	Neplačniki
Dejanska vrednost	Plačniki	55.902	0
	Neplačniki	4.155	0
Odločitvena drevesa s spremembo parametra		Napovedana vrednost	
		Plačniki	Neplačniki
Dejanska vrednost	Plačniki	55.595	307
	Neplačniki	3.699	456

V Tabeli 11 so prikazane mere validacije glede na mejno vrednost in uporabljeni model. Za kazalnike ACC, TPR, TNR, PPV in NPV želimo, da so čim bližje 1, preostali pa čim bližje 0. Ker je dobiček od poplačanega posojila nižji, kot je izguba od neplačanega, je TPR pomembnejši kazalnik kot TNR. Poleg tega, da banka želi čim večji dobiček in hkrati zasleduje tudi cilj čim večjega tržnega deleža, mora biti pozorna tudi na višino kazalnika FNR in FDR. Pri visokih TRP in TNR je treba imeti čim manjši FNR in FDR. Čim manjša sta FNR in FDR, manj je zavrnjenih potencialnih plačnikov in dejanskih neplačnikov.

Najvišjo natančnost doseže logistična regresija pri mejni vrednosti 0,55, hkrati pa odločitvena drevesa pri isti vrednosti najslabšo. Pri isti vrednosti logistična regresija doseže najnižjo stopnjo napake, drevesa pa najvišjo, saj vsem komitentom pripiše status neplačnikov. Logistična regresija pri mejni vrednosti 0,55 doseže najvišji vrednosti kazalnikov TPR, TNR, FPR in FNR glede na ostale mejne vrednosti. Iz tega lahko sklepamo, da kljub najvišji natančnosti in najnižji stopnji napake ni nujno, da je meja 0,55 res najboljša izbira za logistično regresijo.

Tabela 11: Mere validacije pri različnih mejnih vrednostih

Mejna vrednost	Model	Mere validacije								
		ACC	ERR	TPR	TNR	PPV	NPV	FPR	FDR	FNR
0,4	LOG	0,9327	0,0673	0,9842	0,2397	0,9457	0,5295	0,7603	0,0543	0,0158
	DT1	0,9327	0,0673	0,9913	0,1444	0,9397	0,5535	0,8556	0,0603	0,0087
	DT2	0,9330	0,0673	0,9921	0,1374	0,9393	0,5637	0,8626	0,0607	0,0079
0,5	LOG	0,9333	0,0667	0,9898	0,1738	0,9416	0,5584	0,8262	0,0584	0,0102
	DT1	0,9327	0,0673	0,9913	0,1444	0,9397	0,5535	0,8556	0,0603	0,0087
	DT2	0,9333	0,0667	0,9945	0,1097	0,9376	0,5976	0,8903	0,0624	0,0055
0,55	LOG	0,9336	0,0664	0,9925	0,1418	0,9396	0,5832	0,8582	0,0624	0,0750
	DT1	0,9308	0,0692	1,0000	0,0000	0,9308	N/A	1,0000	0,0692	0,0000
	DT2	0,9333	0,0667	0,9945	0,1097	0,9376	0,5976	0,8903	0,0624	0,0055

Glede na primerjavo kazalnikov bi bila najprimernejša mejna vrednost 0,5. Pri povišanju vrednosti z 0,4 na 0,5 se pri logistični regresiji in odločitvenih drevesih s spremenjenim parametrom pomembnejši kazalniki spremenijo do te mere, da je njihova kombinacija sprejemljiva. Kot želimo, se ACC in TPR povišata, ERR in FNR pa znižata. Za kazalnik TNR, ki se zniža, in FPR, ki je višji, bi želeli ravno obratno. Vendar pa je njuna sprememba manjša od spremembe vseh kazalnikov, ki kažejo na boljše razlikovanje med komitenti.

3.4 Primerjava rezultatov

Cilj magistrskega dela je bilo raziskati, kako izbira odločitev v posameznih korakih vpliva na kakovost modela. Postopek, ki je opisan v prejšnjem poglavju, sem ponovila za vse različne načine vzorčenja, ki so predstavljeni v Tabeli 6. Tabeli 12 in 13 predstavljata rezultate v primeru logistične regresije: prva vsebuje mere AUC, druga matriko zamenjav pri dveh različnih mejnih vrednostih.

Tabela 12: Mera AUC – logistična regresija

Vzorec	1	2	3	4	5	6	7
AUC	0,8470	0,8413	0,8350	0,8428	0,8446	0,8481	0,8450

Najvišja vrednost AUC je v primeru 6. vzorčenja, kjer je bilo razmerje med neplačniki in plačniki enako 1 : 3. Ne glede na vzorčenje opazimo, da logistična regresija dosega odlično razlikovanje med plačniki in neplačniki, saj vsi dosegajo mero AUC, ki je višja od 0,8.

Tabela 13 razkriva, da so na splošno modeli logistične regresije, razviti na odstotkovno razdeljenih podatkih, dosegli visoko stopnjo natančnosti in nizko stopnjo napake. Prav tako visoke vrednosti kazalnika TPR in nizke vrednosti FNR omogočajo dobro prepoznavanje in razvrščanje plačnikov. Nekoliko slabše je prepoznavanje neplačnikov, saj kazalnik TNR zavzema nizke vrednosti in FPR visoke. Na splošno je najslabšo natančnost in najvišjo stopnjo napake moč zaslediti pri modelu, razvitem na vzorcu, kjer je razmerje med plačniki

in neplačniki 1 : 1. V nasprotju z drugimi načini vzorčenja TNR in FPR dosegata najvišjo in najnižjo vrednost. Iz tega sklepamo, da model sicer dobro razvrsti neplačnike, vendar pa nizka vrednost TPR kaže na slabše prepoznavanje plačnikov. Kot že omenjeno, banka v osnovi zasleduje dva cilja – čim večji dobiček in tržni delež. Zato je treba izbrati kombinacijo kazalnikov, da bo uporabljeni model glede na izbrani cilj kar se da dobro napovedoval status komitenta. Ker je strošek neplačanega posojila višji kot dobiček poplačanega, je treba poleg visoke natančnosti upoštevati tudi dobro prepoznavanje neplačnikov. Zato v primeru logistične regresije pri mejni vrednosti 0,5 kombinacija kazalnikov za vzorčenje, kjer je delež neplačnikov in plačnikov enak 1 : 3, kaže na najvišjo napovedno moč modela pri dokaj visokem deležu prepoznavanja neplačnikov. Natančnost je v primerjavi z drugimi načini vzorčenja manjša in stopnja napake večja, vendar pa sta TNR in FPR med najvišjimi v primerjavi z drugimi vzorčenji. Omenjeno vzorčenje doseže tudi najvišjo mero AUC.

Tabela 13: Statistike matrike zamenjav – logistična regresija

Vzorec	Mejna vrednost	Mere validacije					
		ACC	ERR	TPR	TNR	FPR	FNR
1	0,4	0,9327	0,0673	0,9842	0,2397	0,7603	0,0158
	0,5	0,9333	0,0667	0,9898	0,1738	0,8262	0,0102
2	0,4	0,9323	0,0677	0,9838	0,2370	0,7630	0,0162
	0,5	0,9338	0,0662	0,9897	0,1769	0,8231	0,0103
3	0,4	0,9313	0,0687	0,9821	0,2317	0,7683	0,0179
	0,5	0,9340	0,0660	0,9885	0,1832	0,8168	0,0115
4	0,4	0,9317	0,0683	0,9827	0,2343	0,7657	0,0173
	0,5	0,9332	0,0668	0,9885	0,1777	0,8223	0,0115
5	0,4	0,7088	0,2912	0,7055	0,8172	0,1828	0,2945
	0,5	0,7927	0,2073	0,7948	0,7265	0,2735	0,2052
6	0,4	0,9013	0,0987	0,9127	0,5617	0,4383	0,0873
	0,5	0,9248	0,0752	0,9396	0,4869	0,5131	0,0604
7	0,4	0,9427	0,0573	0,9736	0,3123	0,6877	0,0264
	0,5	0,9482	0,0518	0,9825	0,2466	0,7534	0,0175

Če primerjamo kazalnike ACC, TPR in TNR, ki so v literaturi največkrat uporabljeni, ne obstaja vzorčenje, ki bi v celoti nadvladalo preostala vzorčenja. Najvišjo natančnost pri mejni stopnji 0,5 doseže 7. vzorec, najvišji TPR pri mejni vrednosti 0,5 1. vzorec, najvišji TNR pri mejni vrednosti 0,4 pa 5. vzorec.

Pri izdelavi se je pokazalo, da je odločitveno drevo z večjim številom vozlišč dosegalo nekoliko boljše prepoznavanje plačnikov in slabše prepoznavanje neplačnikov kot tisto z manjšim številom vozlišč. Rezultati so prikazani v Tabeli 14. Izračun je potekal na mejnih vrednostih, ki so bile uporabljene v prejšnjem poglavju. Ker so rezultati med seboj

primerljivi, so v tabeli prikazane mere, izpeljane iz matrike zamenjav pri mejni vrednosti 0,5. Model DT1 predstavlja odločitveno drevo z manjšim številom vozlišč.

Če primerjamo mero AUC, je pri vseh vzorčenjih mera drevesa z večjim številom vozlišč (v nadaljevanju DT2) enaka ali višja od mere manjšega drevesa. V večini vzorcev je natančnost modela DT2 boljša in stopnja napake manjša. TPR in FNR v večini kažeta na natančnejše napovedovanje plačnikov modela DT2 v primerjavi z modelom DT1. Model DT2 v primerjavi z modelom DT1 doseže slabše napovedovanje neplačnikov, saj ima nižje vrednosti TNR in višje vrednosti FPR.

Tabela 14: Primerjava rezultatov odločitvenih dreves

Vzorec	Model	Mere validacije						
		AUC	ACC	ERR	TPR	TNR	FPR	FNR
1	DT1	0,6509	0,9327	0,0673	0,9913	0,1444	0,8556	0,0087
	DT2	0,6509	0,9333	0,0667	0,9945	0,1097	0,8903	0,0055
2	DT1	0,6478	0,9329	0,0671	0,9914	0,1415	0,8585	0,0086
	DT2	0,6482	0,9335	0,0665	0,9931	0,1278	0,8722	0,0069
3	DT1	0,6456	0,9340	0,0660	0,9913	0,1444	0,8556	0,0087
	DT2	0,6461	0,9342	0,0658	0,9921	0,1377	0,8623	0,0079
4	DT1	0,6439	0,9330	0,0670	0,9915	0,1336	0,8664	0,0085
	DT2	0,6444	0,9324	0,0676	0,9905	0,1386	0,8614	0,0095
5	DT1	0,7577	0,6563	0,3437	0,6507	0,8380	0,1620	0,3493
	DT2	0,8146	0,7401	0,2599	0,7391	0,7726	0,2274	0,2609
6	DT1	0,7853	0,9015	0,0985	0,9139	0,5330	0,4670	0,0861
	DT2	0,7897	0,9110	0,0890	0,9239	0,5290	0,4710	0,0761
7	DT1	0,6489	0,9470	0,0530	0,9815	0,2430	0,7570	0,0185
	DT2	0,6489	0,9482	0,0518	0,9843	0,2112	0,7888	0,0157

Primerjava mere AUC ter pomembnejših kazalnikov za vse tri modele je prikazana v Tabeli 15. V splošnem opazimo, da primerjava mere AUC nakazuje boljšo napovedno moč logistične regresije. Prikazane so mere validacije pri mejni vrednosti 0,5. Najvišjo natančnost in najnižjo napako imata LOG in DR2 pri 7. načinu vzorčenja. Dosegla sta pravilno klasifikacijo pri 94,82 % komitentov in napačno pri 5,18 % komitentov. Najvišji delež pravilno razvrščenih plačnikov (99,45 %) doseže DT2 pri 1. vzorčenju, najvišji delež pravilno razvrščenih neplačnikov (83,80 %) in najnižjo stopnjo lažno pozitivnih vrednosti (16,2 %) DT1 pri 5. vzorčenju ter najnižjo stopnjo lažno negativnih vrednosti (1,03 %) DT2 pri 1. vzorčenju. Sledi, da nobena izmed metod ne predstavlja optimalne izbire pri danem načinu vzorčenja, zato je treba izbrati metodo, katere kombinacija kazalnikov najbolje odraža želene cilje banke. Pri slabšem napovedovanju plačnikov se izkaže, da je napovedovanje neplačnikov boljše. Dobro napovedno moč in razvrstitev neplačnikov ter plačnikov doseže logistična regresija pri 6. načinu vzorčenja, ki glede na ostala vzorčenja in

modele pri visokih vrednostih ACC in TPR ter nizkih vrednostih ERR in FNR doseže dokaj visok TNR in nizek FPR.

Tabela 15: Primerjava mer validacije logistične regresije in odločitvenih dreves

Vzorec	Model	Mere validacije						
		AUC	ACC	ERR	TPR	TNR	FPR	FNR
1	LOG	0,8470	0,9333	0,0667	0,9898	0,1738	0,8262	0,0102
	DT1	0,6509	0,9327	0,0673	0,9913	0,1444	0,8556	0,0087
	DT2	0,6509	0,9333	0,0667	0,9945	0,1097	0,8903	0,0055
2	LOG	0,8413	0,9338	0,0662	0,9897	0,1769	0,8231	0,0103
	DT1	0,6478	0,9329	0,0671	0,9914	0,1415	0,8585	0,0086
	DT2	0,6482	0,9335	0,0665	0,9931	0,1278	0,8722	0,0069
3	LOG	0,8350	0,9340	0,0660	0,9885	0,1832	0,8168	0,0115
	DT1	0,6456	0,9340	0,0660	0,9913	0,1444	0,8556	0,0087
	DT2	0,6461	0,9342	0,0658	0,9921	0,1377	0,8623	0,0079
4	LOG	0,8428	0,9332	0,0668	0,9885	0,1777	0,8223	0,0115
	DT1	0,7577	0,9330	0,0670	0,9915	0,1336	0,8664	0,0085
	DT2	0,8146	0,9324	0,0676	0,9905	0,1386	0,8614	0,0095
5	LOG	0,8446	0,7927	0,2073	0,7948	0,7265	0,2735	0,2052
	DT1	0,6439	0,6563	0,3437	0,6507	0,8380	0,1620	0,3493
	DT2	0,6444	0,7401	0,2599	0,7391	0,7726	0,2274	0,2609
6	LOG	0,8481	0,9248	0,0752	0,9396	0,4869	0,5131	0,0604
	DT1	0,7853	0,9015	0,0985	0,9139	0,5330	0,4670	0,0861
	DT2	0,7897	0,9110	0,0890	0,9239	0,5290	0,4710	0,0761
7	LOG	0,8450	0,9482	0,0518	0,9825	0,2466	0,7534	0,0175
	DT1	0,6489	0,9470	0,0530	0,9815	0,2430	0,7570	0,0185
	DT2	0,6489	0,9482	0,0518	0,9843	0,2112	0,7888	0,0157

Na splošno vzorčenja, kjer je baza podatkov odstotkovno razdeljena na razvojni in validacijski vzorec, dosegajo višje vrednosti ACC, TPR in FPR ter nižje vrednosti ERR, TNR in FNR kot vzorčenja, pri katerih razvojni vzorec vsebuje vnaprej določeno razmerje med neplačniki in plačniki, zato lahko zaključimo, da dosegajo boljšo natančnost in manjšo napako, bolje napovedujejo neplačnike, a slabše plačnike. Da bi pridobili optimalni klasifikacijski model, bi morali poznati stroške, ki nastanejo v primeru nepravilne napovedi plačnikov in neplačnikov. Optimalni model je tak, ki minimizira pričakovano prihodnjo izgubo banke (Liu, 2002, str. 19).

Primerjava mer validacij je v primeru mere AUC pokazala prednost uporabe logistične regresije nad odločitvenimi drevesi za ocenjevanje kreditnega tveganja fizičnih oseb. To se

sklada z rezultati iz literature, kjer so avtorji pokazali superiornost logistične regresije nad odločitvenimi drevesi (Kocenda & Vojtek, 2009).

Cilj magistrskega dela je bilo raziskati vpliv arbitrarnosti, ki se pojavijo pri izdelavi modela za ocenjevanje kreditnega tveganja fizičnih oseb. Na podlagi pridobljenih rezultatov ugotovim, da odločitve v posameznem koraku vplivajo na samo kakovost modela. Glede na to, da je sama kakovost modela odvisna od izbire mere validacije, je odločitev izbire mere povezana z vsemi predhodnimi koraki. Šele glede na izbrano mero namreč lahko določimo, ali je odločitev posameznega koraka vplivala na kakovost modela. Pri delitvi baze podatkov na razvojni in validacijski vzorec je moč opaziti spremembe v merah validacije. Spremembe mere AUC in kazalnikov matrike zamenjav nakazujejo na dejstvo, da način delitve vpliva na kakovost modela.

Kateri način vzorčenja dosega najboljše rezultate, je odvisno od izbire mere validacije. Če bi za primerjavo izbrali mero AUC, so razlike med vzorčenjem najmanjše in najbolj konstantne. Najvišjo vrednost doseže pri 6. vzorčenju, kjer sem naključno izbrala 5.000 neplačnikov ter naključnih 15.000 plačnikov. Z uporabo kazalnikov ACC in ERR se 7. vzorčenje izkaže za najprimernejšega, 5. vzorčenje pa za najmanj primernega. Tako kot vzorčenje izbira modela vpliva na njegovo kakovost, ki je močno odvisna od izbrane mere validacije. Pri izbiri mere AUC logistična regresija veliko bolje razločuje med plačniki in neplačniki glede na odločitvena drevesa, ki slabo razlikujejo med izbranima skupinama. Tudi primerjava drugih kazalnikov nakazuje, da izbira modela vpliva na kakovost. Zadnji korak modeliranja je že omenjena izbira mere validacije, ki ima izmed vseh korakov modeliranja najmočnejši vpliv, saj prav z njo določimo kakovost modela. Mera AUC podaja enoznačni odgovor na vprašanje, kateri model bolje ločuje med plačniki in neplačniki – višja kot je mera AUC, boljši je model. Kazalniki, ki so izpeljani iz matrike zamenjav, pa morajo biti obravnavani sinhrono in ne posamično, saj je treba za določitev najboljšega modela izbrati primerno kombinacijo glede na cilje banke.

3.5 Metoda ponovnega vzorčenja mere AUC

Zadnji izziv, ki ga je treba raziskati, je vprašanje majhnih vzorcev oziroma kako lahko banke postopajo v primeru, kadar njihove baze podatkov za razvoj modela niso dovolj obsežne. V primeru nezadostne količine podatkov delitev na razvojni in validacijski vzorec ni priporočljiva – razvoj in testiranje potekata na istem vzorcu. Podobno kot je Kanczes (2012) v svojem delu uporabil metodo ponovnega vzorčenja na meri AUC, sem v sklopu magistrskega dela poskusila prikazati način, kako banka lahko postopa v primeru premajhnega števila komitentov. Metodo ponovnega vzorčenja sem uporabila v dveh primerih: v prvem primeru sem naključno izbrala komitente, ki so skupno predstavljali 4 % celotne populacije, in pridobila vzorec velikosti 4.804; v drugem primeru sem iz množice plačnikov in neplačnikov naključno izbrala 500 komitentov, ki so skupaj tvorili vzorec

velikosti 1.000. V obeh primerih sem postopek metode ponovnega vzorčenja ponovila 100-krat, 1.000-krat, 3.000-krat in 10.000-krat in s pomočjo enačbe (3) izračunala oceno mere AUC. Kolikokrat sem ponovila vzorčenje, sem označila s parametrom B.

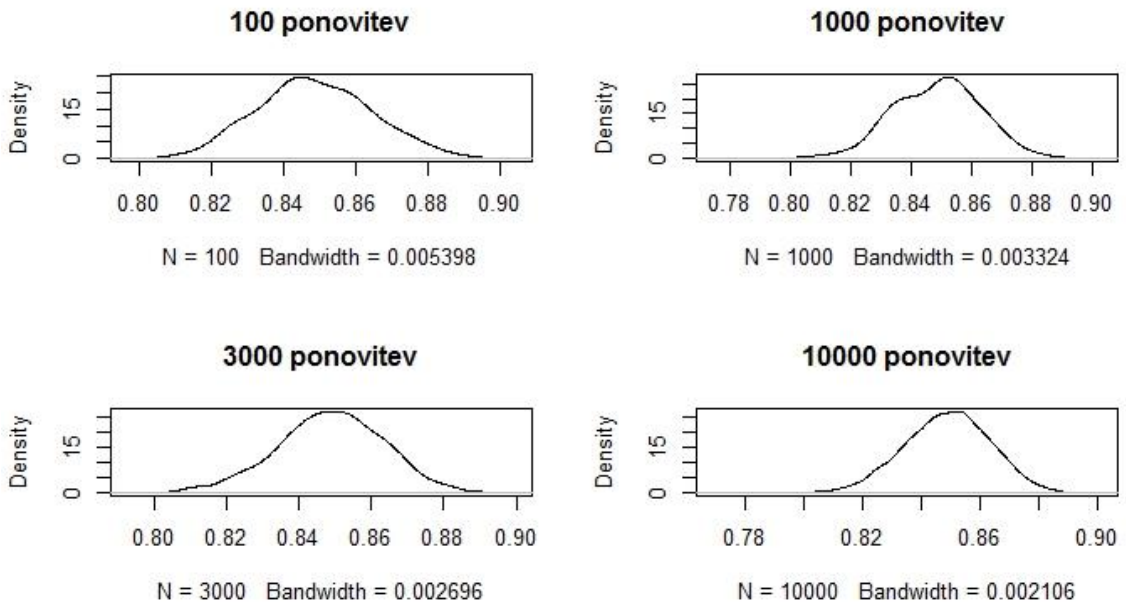
Iz celotne baze podatkov sem naključno izbrala 4804 enot oziroma 4 % celotne populacije. V vsaki ponovitvi sem izračuna meri AUC_b in AUC_b^c, s pomočjo katerih sem pridobila oceno mere AUC. Pričakovano je ta nižja od mere AUC, pridobljene v primeru testiranja na istem vzorcu – $\widehat{AUC}$, kar je predstavljeno v Tabeli 16. Celotni postopek sem ponovila za vse 4 vrednosti parametra B. Kot je razvidno, so vse izračunane mere nižje v primerjavi z AUC na prvotnem vzorcu. Več kot je ponovitev, bolj se ocena mere AUC približuje meri AUC prvotnega vzorca.

Tabela 16: Primerjava mer AUC pri velikosti vzorca 4804 in različnem številu ponovitev

	AUC
Prvotni vzorec	0,8503166
Metoda ponovnega vzorčenja (100)	0,8486008
Metoda ponovnega vzorčenja (1.000)	0,8486613
Metoda ponovnega vzorčenja (3.000)	0,8487707
Metoda ponovnega vzorčenja (10.000)	0,8489695

Na Sliki 16 so predstavljene gostote porazdelitve ocenjene mere AUC. Z večanjem števila ponovitev se porazdelitev približuje normalni porazdelitvi.

Slika 16: Gostota porazdelitve ocenjene mere AUC pri različnem številu ponovitev



Kakšne so mere razpršenosti, je predstavljeno v Tabeli 17. Prikazani so pričakovana vrednost (angl. *Expected value*), varianca (angl. *Variance*), standardni odklon (angl. *Standard deviation*) in kvantili (angl. *Quantiles*).

Tabela 17: Mere razpršenosti

B	E(X)	Var(X)	Sd(X)	Kvantili				
				0 %	25 %	50 %	75 %	100 %
100	0,8486	0,000234	0,0153	0,8127	0,8385	0,8473	0,8587	0,8880
1000	0,8487	0,000216	0,0147	0,7844	0,8383	0,8495	0,8586	0,8933
3000	0,8488	0,000221	0,0149	0,8004	0,8391	0,8491	0,8590	0,8918
10.000	0,8490	0,000218	0,0148	0,7755	0,8391	0,8495	0,8591	0,8955

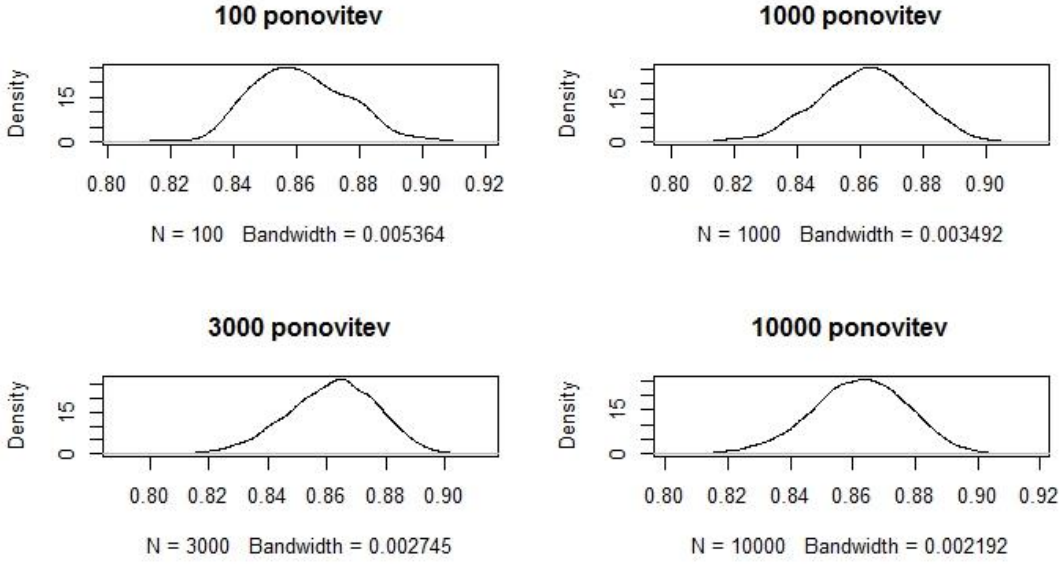
Podobno ponovim za primer, kjer sem naključno izbrala 500 plačnikov in 500 neplačnikov. Pridobljeni vzorec velikosti 1.000 ima razmerje med plačniki in neplačniki enako 1 : 1. Ocene mer AUC so predstavljene v Tabeli 18. Z večanjem števila ponovitev se ocena mere AUC približuje meri AUC prvotnega vzorca.

Tabela 18: Primerjava mer AUC pri velikosti vzorca 1000 in različnem številu ponovitev

	AUC
Prvotni vzorec	0,863292
Metoda ponovnega vzorčenja (100)	0,861214
Metoda ponovnega vzorčenja (1000)	0,861913
Metoda ponovnega vzorčenja (3000)	0,861744
Metoda ponovnega vzorčenja (10.000)	0,862062

Podobno kot v prejšnjem primeru so na Sliki 17 predstavljene gostote porazdelitve. Z večanjem števila ponovitev se porazdelitev približuje normalni porazdelitvi.

Slika 17: Gostota porazdelitve ocenjene mere AUC pri različnem številu ponovitev



Mere razpršenosti so predstavljene v Tabeli 19.

Tabela 19: Mere razpršenosti

B	E(X)	Var(X)	Sd(X)	Kvantili				
				0 %	25 %	50 %	75 %	100 %
100	0,8612	0,000224	0,0149	0,8193	0,8507	0,8599	0,8719	0,9035
1.000	0,8619	0,000239	0,0154	0,8096	0,8516	0,8626	0,8724	0,9047
3.000	0,8617	0,000229	0,0151	0,7972	0,8518	0,8627	0,8728	0,9048
10.000	0,8621	0,000236	0,0154	0,8075	0,8521	0,8626	0,8727	0,9115

Mere AUC, pridobljene s ponovnim vzorčenjem, dosegajo visoke vrednosti. Podoben postopek bi lahko ponovili še za preostale mere.

SKLEP

Bankam je v interesu, da pri odobravanju posojil čim bolje ocenijo potencialni status komitenta. Višjo kot ima napovedno moč metoda, s pomočjo katere razvrščajo potencialne posojilojemalce, manjše so izgube zaradi neodplačanih posojil. V magistrskem delu sem poskusila pokazati, kako različne izbire v posameznih korakih modeliranja vplivajo na kakovost modela.

V prvem delu magistrskega dela sem predstavila kreditno tveganje in njegove komponente ter vsebino Baselskega sporazuma, ki se navezuje na ocenjevanje kreditnega tveganja in PD. Vpliv njegovih določb zasledimo tudi v slovenskih zakonih, ki jih morajo banke upoštevati pri svojem delovanju. Zaradi teh je bankam v interesu pridobiti čim boljšo oceno kreditnega tveganja, ki je del bančnega poslovanja.

Drugi del magistrskega dela opisuje problematiko modeliranja kreditnega tveganja, s katero se srečujejo banke. S pomočjo literature sem poskusila opisati različne metode, ki so uporabljene v praksi. Pri vprašanju pristranskosti izbire vzorca je v literaturi večina metod razkrila majhne razlike v klasifikaciji, kadar so v model vključili zavrnjene komitente. V primeru premajhnega vzorca so avtorji uporabljali metodi križnega preverjanja in ponovnega vzorčenja. V svojem delu sem pridobila oceno mere AUC s pomočjo metode ponovnega vzorčenja. V literaturi avtorji uporabljajo različne načine delitve baze podatkov, in sicer na razvojni in validacijski vzorec. Tipična delitev je vključitev 70–80 % komitentov v razvojni vzorec in preostale v validacijskega. Pri izbiri metode za modeliranje PD se v literaturi največkrat pojavlja logistična regresija, medtem ko najboljše rezultate avtorji dosežejo s pomočjo nevronskih mrež. V primeru validacije modelov avtorji uporabljajo krivulji ROC in CAP in obe pripadajoči meri, AUC in Ginijev koeficient.

V zadnjem, empiričnem delu magistrske naloge sem ugotovitve iz literature poskusila uporabiti za raziskavo vseh izzivov, ki se pojavijo pri procesu modeliranja kreditnega tveganja za fizične osebe. Zaradi nedosegljivosti podatkov o zavrnjenih komitentih ni bilo moč raziskati različnih metod, ki so jih avtorji uporabili v procesu sklepanja o bonitetni oceni zavrnjenih. Zaradi slednjega je raziskovanje potekalo na bazi podatkov sprejetih komitentov.

Vprašanje izbire velikosti razvojnega in testnega vzorca sem poskusila raziskati z različnimi načini vzorčenja. Zaradi možne pristranskosti vzorca sem vse vzorce izbirala naključno. Uporabila sem dva glavna različna načina delitve baze: pri prvem načinu sem celotno bazo odstotkovno razdelila na razvojni in validacijski vzorec. Za namene razvojnih vzorcev sem iz baze podatkov naključno izbrala 40 %, 50 %, 70 % in 80 % komitentov, preostale pa vključila v validacijski vzorec, tj. 60 %, 50 %, 30 % ter 20 % komitentov. Pri drugem načinu sem iz baze naključno izbrala neplačnike ter plačnike in tako pridobila razvojne vzorce z različnimi razmerji med neplačniki in plačniki. Pridobljena razmerja so enaka 1 : 1, 1 : 3 in 1 : 9. Vse preostale komitente sem vključila v validacijski vzorec. V splošnem so vzorci, pridobljeni s prvim načinom, dosegli boljšo napovedno moč. Prav tako so boljše napovedovali potencialne neplačnike, vendar slabše plačnike. Vzorci z različnimi razmerji med plačniki in neplačniki so sicer dosegli nekoliko slabšo napovedno moč, a so boljše napovedali potencialne plačnike.

Vprašanja izbire modela sem se lotila s primerjavo dveh metod, ki sta pogosto uporabljeni v literaturi. V večini primerov avtorji kot primerjalni model uporabijo logistično regresijo. Zaradi preprostosti izdelave in interpretacije avtorji velikokrat uporabljajo tudi metode odločitvenih dreves. V splošnem se je metoda logistične regresije izkazala za boljšo odločitev. Čeprav razlikovanje v napovedni moči na podlagi statistik, izpeljanih iz matrike zamenjav, ni podalo enotnega odgovora na vprašanje kakovosti modela, kaže razlika v meri AUC v prid logistični regresiji.

Ugotovila sem, da je vprašanje izbire primerne mere za primerjavo kakovosti modelov tesno povezano s preostalimi vprašanji, ki smo jih raziskali v magistrski nalogi. Ker kakovost modela ugotavljamo prek mer, je zato njihova izbira vplivala na primerjavo različnih metod v korakih procesa kreditnega tveganja. Primerjava AUC je boljšo napovedno moč za vse načine vzorčenja pripisala logistični regresiji, ki je dosegla odlično razlikovanje med plačniki in neplačniki. Pri primerjavi statistik, izpeljanih iz matrike zamenjav, se pojavi tudi vprašanje izbire mejne vrednosti. Pri spreminjanju vrednosti se glede na ocenjeno PD spreminja status komitenta, zato vpliva na vrednosti v matriki zamenjav in nadalje na statistike. Pri mejni vrednosti 0,5, ki je tudi po navadi uporabljena v literaturi, so v splošnem kombinacije kazalnikov kazale na najboljšo napovedno moč. Za končno odločitev nadvlade metode je potrebna dodatna analiza stroškov.

V primeru manjšega števila podatkov o komitentih lahko uporabimo metodo ponovnega vzorčenja in s tem pridobimo oceno mere AUC. Ugotovila sem, da je pridobljena vrednost mere nižja kot vrednost mere na prvotnem vzorcu. Sama izdelava metode ponovnega vzorčenja in njeno delovanje sta odvisni od zmogljivosti računalniške tehnologije. Sam proces ponovnega vzorčenja je v primeru večjega števila ponovitev lahko zamuden, saj je preračun trajal tudi več kot uro.

Banke se morajo pri izdelavi zavedati, da sta potrebna stalno spremljanje delovanja modela in njegova dodelava, saj se z leti spreminjajo ne le podatki o komitentih, ampak tudi značilnosti populacije in ekonomskega okolja.

Izdelava magistrske naloge je pokazala, da obstaja veliko možnosti za nadaljnje raziskave na področju ocenjevanja kreditnega tveganja. V kolikor imajo banke možnost, analizo lahko razširijo z dodatnimi modeli, kot so nevronske mreže, naključni gozdovi ali kombinacije različnih metod. Prav tako bi lahko dodatno raziskali, kolikšen je vpliv rezanja in kriterija zaustavitve pri odločitvenih drevesih. Podobno bi lahko pri logistični regresiji uporabili različne metode, kot je na primer tako imenovana postopna metoda, pri kateri postopoma dodajamo spremenljivke. V primeru dostopnih podatkov o zavrnjenih komitentih je mogoče raziskati še različne metode za sklepanja o bonitetni oceni zavrnjenih. Za nadaljnjo delo bi v primeru metode ponovnega vzorčenja model lahko razširili na več različnih mer validacije ter raziskali razlike med njimi.

LITERATURA IN VIRI

1. Allen, L., DeLong, G., & Saunders, A. (2004). Issues in the credit risk modeling of retail markets. *Journal of Banking & Finance, 28*(4), 727–752.
2. Altman, E. (2006). *Default recovery rates and LGD in credit risk modeling and practice: an updated review of the literature and empirical evidence*. New York University: Stern School of Business.
3. Anderson, R. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*. New York: Oxford University Press.
4. Baesens, B., & Van Gestel, T. (2009). *Credit Risk Management: Basic Concepts*. New York: Oxford University Press.
5. Balant, A. (2005). *Mednarodni bančni standardi - Basel II* (diplomsko delo). Maribor: Ekonomsko-poslovna fakulteta.
6. Banasik, J., Crook, J., & Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society, 54*(8), 822–832.
7. Bank of Japan. (2005) *Advancing Credit Risk Management through Internal Rating Systems.* Najdeno 26. septembra 2015 na spletni strani https://www.boj.or.jp/en/research/brp/ron.../fsk0509a.pdf
8. Banka Slovenije. (2007). *Proces ocenjevanja tveganj* (interno gradivo). Najdeno 18. avgusta 2015 na spletni strani https://www.bsi.si/library/includes/datoteka.asp?DatotekaId=243
9. Banka Slovenije. (2005). *Bilten*. Najdeno 18. avgusta 2015 na spletni strani http://www.bsi.si/iskalniki/bilteni.asp?MapaId=229#
10. Banka Slovenije. (2012). *Poročilo o finančni stabilnosti*. Najdeno 21. oktobra 2015 na spletni strani https://www.bsi.si/publikacije-in-raziskave.asp?MapaId=285
11. Banka Slovenije. (2015a). *Bilten*. Najdeno 18. avgusta 2015 na spletni strani http://www.bsi.si/iskalniki/bilteni.asp?MapaId=229#
12. Banka Slovenije. (2015b). *Poročilo o finančni stabilnosti*. Najdeno 21. oktobra 2015 na spletni strani https://www.bsi.si/publikacije-in-raziskave.asp?MapaId=285
13. Basel Commitee on Banking Supervision. (2001). *Consultative Document: The Internal Rating-Based Approach*. Basel: Bank for International Settlements.
14. Basel Commitee on Banking Supervision. (2005). *An Explanatory Note on the Basel II IRB Risk Weight Functions*. Basel: Bank for International Settlements.
15. Basel Commitee on Banking Supervision. (2006). *International convergence of capital measurement and capital standards: a revised framework*. Basel: Bank for International Settlements.
16. Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. Canada: John Wiley & Sons.
17. Bitenc, M. (2009). *Temelji bančnega poslovanja*. Ljubljana: Ministrstvo za šolstvo in šport Republike Slovenije.

18. Bolton, C. (2010). *Logistic regression and its application in credit scoring* (magistrsko delo). Pretoria: University of Pretoria.
19. Brezigar-Masten, A., & Masten, I. (2012). CART-based selection of bankruptcy predictors for the logit model. *Expert systems with applications, 39*(11), 10153–10159.
20. Caouette, J. B., Altman, E. I., & Narayanan, P. (1998). *Managing credit risk: the next great financial challenge*. Canada: John Wiley & Sons.
21. Carletti, E., & Hartmann, P. (2002). Competition and Stability: what's Special about Banking? *European Central Bank Working Paper Series No. 146,* 8-33. Najdeno 19 avgust 2016 na spletni strani http://www.pedz.uni-mannheim.de/daten/edz-ki/ezb/02/w-paper/ecbwp146.pdf
22. Crook, J., & Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance, 28*(4), 857–874.
23. Čargo, A., & Štajner, M. (2004). Minimalne zahteve za uvedbo Pristopa IRB. *Zbornik, 9,* 89–105.
24. Dahan, H., Cohen, S., Rokach, L., & Maimon, O. (2014). Proactive data mining using decision trees. V *Proactive Data Mining with Decision Trees* (str. 21-33). New York: Springer.
25. Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research, 95*(1), 24–37.
26. Durand, D. (1941). *Risk elements in consumer instalment financing*. National Bureau of Economy Research Books.
27. *Družba za upravljanje terjatev bank – O nas*. Najdeno 26. oktobra 2015 na spletni strani http://www.dutb.eu/si/o-nas
28. Engelmann, B., Hayden, E., & Tasche, D. (2003). *Measuring the discriminative power of rating systems*. Frankfurt: Deutsche Bundesbank.
29. Feelders, A. (2000). Credit scoring and reject inference with mixture models. *International Journal of Intelligent Systems in Accounting, Finance & Management, 9*(1), 1–8.
30. Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research, 210*(2), 368–378.
31. Ghosh, D., & Vogt, A. (2012). *Outliers: An evaluation of methodologies*. Najdeno 10. februarja 2016 na spletni strani https://www.amstat.org/sections/srms/proceedings/y2012/files/304068_72402.pdf
32. *Give me some credit*. Najdeno 26. oktober 2015 na spletni strani https://www.kaggle.com/c/GiveMeSomeCredit/data
33. Greene, W. (1998). Sample selection in credit-scoring models. *Japan and the world economy, 10*(3), 299–316.
34. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques.* Waltham, Massachusetts: Elsevier.

35. Hand, D. J., & Henley, W. E. (1993). Can reject inference ever work? *IMA Journal of Management Mathematics, 5*(1), 45-55.
36. Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, Massachusetts: the MIT press.
37. Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–161.
38. Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression*. Canada: John Wiley & Sons.
39. Hsia, D. C. (1978). Credit scoring and the equal credit opportunity act. *The Hastings Law Journal 30*, 371–448 .
40. Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications, 33*(4), 847–856.
41. Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications*, 13–19.
42. Jagrič, V. (2011). *Ocenjevanje verjetnosti neplačila za kredite prebivalstvu: nelinearen pristop s samoorganizirajočimi se mrežami* (doktorska disertacija). Maribor: Ekonomsko-poslovna fakulteta.
43. Joanes, D. N. (1993). Reject inference applied to logistic regression for credit scoring. *IMA Journal of Management Mathematics, 5*(1), 35–43.
44. Jovan, M. (2005). *Od česa je odvisno razločevanje statističnih modelov?* Ljubljana: Banka Slovenije.
45. Jovan, M., & Šušterič, M. (2004). Statistično ocenjevanje verjetnosti neplačila za slovenska podjetja. *Banka Slovenije.* Najdeno 18 oktober 2015 na spletni strani https://www.bsi.si/library/includes/datoteka.asp?DatotekaId=543
46. Kanczes, L. (2012). *Bootstrapping and its Applications to Credit Scoring* (magistrsko delo). London: Imperial College London.
47. Kocenda, E., & Vojtek, M. (2009). *Default predictors and credit scoring models for retail banking.* CESifo Working Paper, No. 2862. München: CESifo Group Munich, 2006.
48. Kolar, J. (2014). *IRB bonitetni model za slovensko prebivalstvo* (magistrsko delo). Maribor: Fakulteta za naravoslovje in matematiko.
49. Lee, T.-S., Chiu, C.-C., Lu, C.-J., & Chen, I.-F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with applications, 23*(3), 245–254.
50. Lewis, R. J. (2000). An introduction to classification and regression tree (CART) analysis. *2000 Annual Meeting of the Society for Academic Emergency Medicine* (str. 1–14). San Francisco, California: Society for Academic Emergency Medicine.
51. Liu, Y. (2002). *The evaluation of classification models for credit scoring*. Göttingen, Germany: Institut for Economical Informatics.

52. Matloff, N. (2011). *The Art of R Programming*. California: No Starch Press Series
53. Medema, L., Koning, R. H., & Lensink, R. (2009). A practical approach to validating a PD model. *Journal of Banking & Finance, 33*(4), 701–708.
54. Oesterreichische Nationalbank & Financial Market Authority. (2004). *Rating models and validation*. Vienna: Oesterreichische Nationalbank.
55. Ong, C.-S., Huang, J.-J., & Tzeng, G.-H. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications, 29*(1), 41–47.
56. Osei-Bryson, K.-M. (2007). Post-pruning in decision tree induction using multiple performance measures. *Computers & operations research, 34*(11), 3331–3345.
57. Reichert, A. K., Cho, C.-C., & Wagner, G. M. (1983). An examination of the conceptual issues involved in developing credit-scoring models. *Journal of Business & Economic Statistics, 1*(2), 101–114.
58. Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics – Part C (Applications and Reviews), 35*(4), 476–487.
59. Satchell, S. E., & Xia, W. (2007). *Analytic models of the ROC curve: Applications to credit rating model validation*. Sydney: Quantitative Finance Research Centre.
60. Statstutor. (b.l.). *Pearson's correlation*. Najdeno 3. avgusta 2016 na spletni strani http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf
61. Siddiqi, N. (2006). *Credit risk scorecards: developing and implementing intelligent credit scoring*. Canada: John Wiley & Sons.
62. SISBON (b.l.). *Uvodna stran*. Najdeno 15. oktober 2015 na spletni strani http://sisbon.si/sl
63. Sklep o izračunu kapitalske zahteve za kreditno tveganje po standardiziranem pristopu za banke in hranilnice. *Uradni list* RS št. 135/06, 104/07, 85/10, 97/10, 62/11, 100/11, 22/12, 100/12, 60/13 in 25/15 – ZBan-2.
64. Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting, 16*(2), 149–172.
65. Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Philadelphia: Society for Industrial and applied Mathematics.
66. Tsai, M.-C., Lin, S.-P., Cheng, C.-C., & Lin, Y.-P. (2009). The consumer loan default predicting model - An application of DEA-DA and neural network. *Expert Systems with applications, 36*(9), 11682–11690.
67. Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology, 49*(11), 1225–1231.
68. Vandone, D. (2009). *Consumer credit in Europe: risks and opportunities of a dynamic industry*. Berlin, Heidelberg: Springer Science & Business Media.
69. Verstraeten, G., & Van den Poel, D. (2005). The impact of sample bias on consumer credit scoring performance and profitability. *Journal of the operational research society, 56*(8), 981–992.

70. Vuk, M., & Curk, T. (2006). ROC curve, lift chart and calibration plot. *Metodoloski zvezki*, *3*(1), 89–108.
71. West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(11), 1131–1152.
72. Wu, X. (2008). *Credit Scoring Model Validation* (magistrsko delo). Amsterdam: University of Amsterdam.
73. Xiao, W., Zhao, Q., & Fei, Q. (2006). A comparative study of data mining methods in consumer loans credit scoring management. *Journal of Systems Science and Systems Engineering*, *15*(4), 419–435.
74. Zakon o bančništvu. *Uradni list* RS, št. 99/10, 52/11 popr., 9/11 ZPlaSS-B, 35/11, 59/11, 85/11, 48/12, 105/12, 56/13, 63/13 ZS-K, 96/13 in 25/15 ZBan-2.
75. Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research*, 33, 59–82.

PRILOGA

PRILOGA: Seznam kratic

ACC	Accuracy
AUC	Area under curve
BCBS	Basel Committee on Banking Supervision
CAP	Cumulative accuracy profile
CART	Classification and regression trees
DUTB	Družba za upravljanje terjatev bank
EAD	Exposure at default
EL	Expected loss
ERR	Error rate
FDR	False discovery rate
FNR	False negative rate
FPR	False positive rate
IRB	Internal rating-based approach
LGD	Loss given default
NPV	Negative predictive value
OENB	Oesterreichische Nationalbank
PD	Probability of default
PPV	Positive prediction value
ROC	Receiver operating characteristic
RR	Recovery rate
TNR	True negative rate
TPR	True positive rate
UL	Unexpected loss