UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

MAGISTRSKO DELO

KREDITNO TOČKOVANJE KOMITENTOV V LIZING PODJETJIH

Ljubljana, februar 2017 BORUT VALENTINČIČ

IZJAVA O AVTORSTVU

Podpisani Borut Valentinčič, študent Ekonomske fakultete Univerze v Ljubljani, avtor predloženega dela z naslovom Kreditno točkovanje komitentov v lizing podjetjih, pripravljenega v sodelovanju s svetovalcem prof. dr. Markom Košakom

IZJAVLJAM

1. da sem predloženo delo pripravil samostojno;

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

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

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

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

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

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

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

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

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

V Ljubljani, dne 21.02.2017 Podpis študenta: _____

KAZALO

KAZALO TABEL

KAZALO SLIK

UVOD

Podobno kot pri vseh finančnih institucijah, je tudi pri lizing podjetjih pri poslovanju prisotno kreditno tveganje, zato tudi v tej panogi obstaja potreba po čim boljšem ravnanju ob sklepanju novih lizing poslov. S tem se zmanjšuje tveganje in posledično minimalizirajo možne izgube iz naslova slabih posojil. Na splošno to pomeni, da se lizing posel odobri zgolj tistemu povpraševalcu, ki je ekonomsko sposoben vračila ter izkazuje tudi pripravljenost, da bo prevzete obveznosti voljan poravnavati (Jagrič, 2012). Ob tem je potrebno opozoriti tudi na morebitne nasprotne interese, saj kredit odobrava kreditni agent (problem agenta in principala), ki ima lahko interes odobriti čim večje število lizing poslov, ne glede na morebitne izgube (Jagrič, 2012). Ta problem je še posebej verjeten pri napačnem načinu nagrajevanja. Dodaten problem, ki je lahko prisoten v fazi odobravanja kreditnega agenta, je možnost pojava diskriminacije zaradi njegove subjektivne presoje, čemur pa se lahko v veliki meri izognemo le s t. i. kreditnim točkovanjem (angl. *Credit scoring*). V tem primeru se analiza strankine kreditne sposobnosti izvaja bolj objektivno (Koh, 2015), predvsem pa s tem izpolnimo zahtevo, da za vse potencialne stranke veljajo enaki kriteriji. To seveda velja v primeru, da model točkovanja ne vsebuje kriterijev, kot so rasa, spolna usmerjenost, verska pripadnost in podobno. V državah ZDA so leta 1974 v ta namen celo sprejeli zakon o enakih možnostih (angl. *Equal Credit Opportunity Act*), ki prepoveduje vsakršno diskriminacijo, razen statistične (Koh, 2015).

V zadnjem času, ko smo priča ekstremnemu razvoju informacijske tehnologije in imamo na voljo neprimerno večje količine podatkov kot v preteklosti, v bančnem sistemu prihaja do nenehnih teženj po čim večji stopnji avtomatizacije delovnih procesov na vseh področjih. Cilj je zmanjševanje časa, potrebnega za izvedbo delovnih procesov, kar vodi v čim večje zmanjševanje stroškov poslovanja (Leonard, 1995). Še posebej to velja za transakcijsko bančništvo, financiranja fizičnih oseb in manjših ali srednje velikih podjetij (angl. *Small or Medium Enterprise*), kot so lizing, faktoring, hipotekarna posojila in posojila na podlagi poslovnih izkazov.

Eden glavnih ciljev finančnih podjetij je vzpostavitev lastnega točkovanja kreditne zanesljivosti oziroma kreditnega točkovanja, ki bi na podlagi nekaj vhodnih podatkov stranke z veliko verjetnostjo napovedal njeno kreditno sposobnost v prihodnosti. S tem bi si bistveno olajšali proces odločanja o morebitnem financiranju, lažje bi določali obrestne mere, vezane na rizičnost posla, in izboljšali opredeljevanje potrebnih zavarovanj za sklenitev posla (Prakash, 1995).

V lizing panogi v Sloveniji se v zadnjem času sicer vzpostavljajo točkovni sistemi, vendar so zaenkrat redki in zaradi krajše prakse tudi niso tako zanesljivi. Zaradi specifičnosti slovenskega trga in lizing poslovanja tudi lokalizacija točkovnih modelov drugih držav, kjer se takšni sistemi uspešno uporabljajo že dlje časa, ni možna. Druga opcija je še lokalizacija

modelov bank, ki so lastniško povezane in kjer takšni modeli obstajajo, vendar je, kot bomo v nadaljevanju videli, tudi ta možnost težko izvedljiva.

Pri izbiri parametrov učenja je za optimizacijo modela zelo pomembna ocena izgube in donosa v primeru propadlega oziroma uspešno zaključenega posla, zato je ključno, da se ugotovi čim boljša ocena izgube ob neplačilu (v nadaljevanju LGD angl. *Loss given default*) in bodočega donosa (v nadaljevanju Y – angl. *yield*) posla.

Pri točkovnem modelu imamo vedno dve vrsti napak: napaka tipa I je zavrnitev dobremu komitentu, napaka tipa II pa je odobritev posla slabemu. V prvem primeru ima finančna institucija s poslom določene oportunitetne stroške, saj je ob donos, ki bi ga bila deležna v primeru odobritve, v drugem pa izgubi zaradi neplačila dela glavnice. Pri točkovnih modelih je število prvih napak obratno sorazmerno s številom drugih, kar pomeni, da se ob spreminjanju modela za zmanjševanje števila prvih napak povečuje število drugih in obratno. Katero razmerje je za nas optimalno, je odvisno od višine škode, ki nastopi pri obeh vrstah napak. Če je povprečna izguba ob neplačilu LGD', povprečen preostanek vrednosti ob neplačilu RV' (RV – angl. *Residual value*), povprečni donos Y' in povprečna vrednost financiranja FV' (FV – angl. *Finance value*), lahko zapišemo enačbo ocene celotnega stroška napak za izbrani model na naslednji način:

$$\textit{Ocena stroška napak} = (\textit{št. napak I} \times Y' \times FV') + (\textit{število napak II} \times LGD' \times RV') \quad (1)$$

Optimalni točkovni model je tisti, ki med vsemi možnimi modeli vrne najmanjšo oceno stroška napak.

Pri optimizaciji modela je potrebno upoštevati še dodatno dejstvo, da je pričakovana vrednost LGD (kakor tudi povprečni donos) lahko odvisna tudi od parametrov sklenjenega posla in ni nujno enaka za vse enote vzorca, kar poveča kompleksnost optimalnega modela. Hartman-Wendels, Miller in Tows (2014) so pri iskanju najprimernejšega modela za izračun LGD ugotovili, da je pričakovana vrednost te spremenljivke zelo odvisna od vrste opreme, ki se financira. Testirali so podatke skoraj 15.000 lizing poslov, ki so prišli v neplačilo (angl. *default*) med leti 2000 in 2010. Financirano opremo so razvrstili na vozila, mehanizacijo, informacijsko ali komunikacijsko tehnologijo, opremo in ostalo. Po njihovih empiričnih izračunih so imele pogodbe, ki so financirale vozila, najnižje izgube ob neplačilu in bi za takšen tip opreme dobili drugačen optimalen model kot pri ostalih.

V tej nalogi se ne bomo spuščali v izračun LGD spremenljivke in primerjali modelov v celotnem spektru razmerij napak, je pa pomembno vedeti, da sta za praktično uporabo točkovnega modela ključnega pomena tudi pravilen izračun povprečnega LGD in poznavanje morebitnih korelacij med povprečno vrednost LGD in neodvisnimi spremenljivkami modela.

Ker so povprečne vrednosti LGD pri lizing poslih drugačne kot pri bančnih, je zelo otežena tudi možnost uporabe bančnih točkovnih modelov za odobravanje lizing poslov. Razlike

povprečnih vrednosti sta Schmit in Stuyck (2002) analizirala v obsežni raziskavi, kjer sta analizirala 37.259 lizing poslov v neplačilu za finančne institucije različnih držav. Ugotovila sta, da je pri lizing poslih LGD nižji kot pri bančnem kreditu ter bistveno nižji kot pri obveznicah. Razlika je predvsem posledica dejstva, da ima lizing podjetje lastništvo nad predmetom lizinga in posledično možnost odvzema predmeta ter njegove prodaje na sekundarnem trgu. Dodatno sta ugotovila, da je čas izterjave pri lizing poslih krajši od časa izterjave bančnih kreditov, kar tudi vpliva na končni rezultat.

Dodatni razlog, ki oteži prevzem bančnih točkovnih modelov za odobravanja lizing poslov, je v razliki domene vrednosti LGD. Pri bančnih kreditih je lahko ta vrednost med 0 in 1, kar pomeni, da je stopnja izterjave po nastopu neplačila lahko od nič do celotne neodplačane glavnice. Pri lizing poslih je ta domena drugačna, saj je lizingodajalec lastnik predmeta lizinga in lahko pri odvzemu in prodaji predmeta pridobi tudi več sredstev od neplačanega dela. Schmit in Stuyck (2002) sta v analizi ugotovila, da je imelo do 59 % propadlih lizing poslov stopnjo izterjave nad 100 %, kar se pri bančnih kreditih ne more zgoditi. Kasneje sta Laurent in Schmit (2005) na podlagi drugih podatkovnih zbirk ugotovila, da ima takšno stopnjo izterjave sicer manj vendar še vedno 45 % primerov propadlih pogodb.

Zaradi omenjenih omejitev so lizing podjetja primorana vzpostaviti lasten točkovni model, za to pa imajo na voljo več možnih različnih pristopov.

Namen te naloge je ugotoviti, ali je možno s sodobnimi orodji, ki so na voljo, vzpostaviti takšen točkovni sistem, ki bo učinkovit in uporaben za praktično delo.

Cilji magistrskega dela so testiranje in primerjava logistične regresije, odločitvenih dreves ter metode naključnih gozdov z rezultati modela, ki se že uporablja v praksi, ter potrditi ali ovreči hipotezo, da je možno s katero od teh metod in trenutno tehnologijo producirati takšne točkovne modele, ki bodo uspešno klasificirali zahteve za financiranje glede tveganosti posla. Zato v tej nalogi skušamo odgovoriti na naslednja raziskovalna vprašanja:

1. Bodo dobljeni modeli po učinkovitosti res primerljivi z rezultati modela, ki se že uporablja v praksi, in s tem primerni tudi za praktično uporabo?
2. Bo postopek vzpostavitve modela s takšnimi orodji časovno in po obsegu dela bistveno manj zahteven od modela v uporabi?
3. Lahko s kombiniranjem omenjenih metod rezultate modelov še izboljšamo?

V prvem delu magistrskega dela se osredotočamo na teoretični pregled strokovne literature na področju kreditnega točkovanja, analiz najbolj uporabljenih algoritmov na tem področju ter primerjave njihovih rezultatov. Drugi, empirični del zajema konkretno uporabo teh algoritmov pri vzpostavitvi točkovnih modelov, in sicer dveh tradicionalnih in ene napredne metode ter primerjavo dobljenih rezultatov. Od tradicionalnih bomo testirali metodo logistične regresije in metodo odločitvenih dreves, od naprednih pa metodo naključnih gozdov (angl. *Random Forests*). Podatke, na katerih bomo gradili in testirali modele, smo v ta namen dobili pri slovenskem lizing podjetju, ki v praksi že dlje časa uporablja svoj

nestrukturirani točkovni model. Na voljo bomo imeli tudi njihove končne odločitve procesa odobritve zahtev za financiranje, kar nam bodo predstavljale ciljne vrednosti naših testiranih modelov. Podatke bomo najprej preverili, da ne bodo vsebovali nepopolnih ali napačnih vrednosti ter jih nato ločili v dve podmnožici, kjer bo prva namenjena učenju, druga pa testiranju vseh pridobljenih točkovnih modelov. Izbrane algoritme bomo izvajali nad istimi množicami podatkov, kar nam bo na koncu omogočilo primerjavo dobljenih rezultatov. Ker bomo prevzeli nabor vhodnih podatkov njihovega modela, lahko tudi rečemo, da bodo naši modeli do neke mere izhajali iz njihovega.

Podjetje, od katerega smo dobili podatke za našo raziskavo in primerjalni točkovni model, bomo poimenovali z izmišljenim imenom ABC Lizing, saj za namen naše raziskave pravo ime ni pomembno, poleg tega pa želimo preprečili možnost razkritja njihovih morebitnih poslovnih skrivnosti.

1 TEORETSKI OPIS MODELOV ZA OCENO KREDITNEGA TVEGANJA

Kreditno točkovanje, ki se izvaja v okviru nekega modela za oceno kreditnega tveganja (angl. *Scoring model*), lahko definiramo kot statistično metodo, ki se uporablja za napoved verjetnosti, da posojilojemalec v prihodnosti ne bo zmožen, ali pa ne bo želel, odplačati posojila (Mester, 1997). Kreditno točkovanje lahko definiramo tudi kot sistematično metodo za vrednotenje kreditnega tveganja, ki vrne skladno analizo kriterijev, za katere se je izkazalo, da povzročajo oziroma vplivajo na kreditno tveganje (Fensterstock, 2005). Razvoj modeliranja je prvotno temeljil na predpostavki, da je kreditna sposobnost prebivalstva neodvisna od časa in je zato mogoče pri razvrščanju novih vlog (glede na tveganje neplačila) uporabiti podatke o kreditojemalcih iz preteklosti. Predpostavka o časovni neodvisnosti kreditne sposobnosti se uporablja še danes, vendar z zavedanjem, da ta drži le na kratek rok. Posledično je potrebno izvajati monitoring napovedane uspešnosti s postopki rednih validacij. Model je potrebno redno menjati z novim, zato je zbiranje podatkov, modeliranje in validiranje modelov postalo kontinuiran proces (Jagrič, 2012).

Kreditno točkovanje se torej uporablja kot pomoč pri ugotavljanju kreditne sposobnosti partnerja.

Posledično se s kreditnim točkovanjem zmanjšuje tudi potreben čas procesa odobritve in dodatno je omogočana avtomatizacija postopka procesa posojil (Rimmer, 2005). S tem se lahko tudi občutno zmanjša potreba po človeški intervenciji v postopku kreditne presoje in posredno strošek procesa dajanja posojil (Diana, 2005). Določene finančne institucije kreditno točkovanje uporabljajo tudi za namene določanja kreditnih limit, obrestnih mer, zavarovanj in podobno (Sandler, McGinn, & Barloon, 2000). Študije še dokazujejo, da je od razpoložljivih metod zadnjega časa kreditno točkovanje eno od najmočnejših pokazateljev tveganja, saj omogoča najbolj natančne napovedi izgub (Miller, 2003). Skratka, s pomočjo

kreditnega točkovanja se lahko finančne inštitucije odločajo hitreje, bolje in z večjo natančnostjo (Banasiak & Kiely, 2000).

Metode točkovanja v praksi temeljijo na matematičnih modelih, ki jih lahko klasificiramo na več možnih načinov:

1. Na tradicionalne (linearne diskriminantne analize, multivariantne analize, logistične regresije logit, probit, binarna drevesa in podobno), ki temeljijo na klasičnih statističnih metodah, ali napredne (ekspertni sistemi, nevronske mreže, generična programiranja in podobno), ki temeljijo na metodah umetne inteligence oziroma strojnega učenja (Zhang, Huang, Chen, & Jiang, 2007)**.** Napredne metode, kot so nevronske mreže in generično programiranje, sicer zagotavljajo boljše rezultate predvidevanj od statističnih, vendar imajo to slabost, da proces učenja traja bistveno dlje. Poleg tega so pridobljeni algoritmi odločanja tipa črne škatle (angl. *Black box*) zelo težko razumljivi in zato v praksi manj uporabljeni. Poleg tradicionalnih in naprednih metod se v zadnjem času pojavljajo še hibridni sistemi, ki so mešanica obeh (Šušteršič, Mramor, & Zupan, 2009);
2. Na parametrične (pri podatkih se zahteva določena porazdelitev spremenljivk) ali neparametrične, ki jih lahko ločimo na (1) razvrščanje v skupine, (2) metode glavnih komponent (3) faktorske analize in podobno;
3. Na linearne (kjer se predpostavlja linearna povezanost med odvisnimi in neodvisnimi spremenljivkami) ali nelinearne;
4. Na regresijske (obstaja zvezna porazdelitev spremenljivk) ali klasifikacijske (obstaja diskretna porazdelitev);
5. Na metode s površinskim učenjem, kjer se iščejo vzorci, na podlagi katerih poskušamo napovedovati pripadnost novih dogodkov, ali metode z globinskim učenjem, kjer se za pravilno napoved išče globlji pomen kriterijev ter njihove medsebojne odnose.

V podjetju ABC Lizing že dobro leto testno uporabljajo nestrukturiran točkovni model, ki so ga razvili sami, na podlagi izkušenj. Model so definirali tako, da so najprej opredelili vse pojasnjevalne spremenljivke (opisane v poglavju 3.3.1), ki se preverjajo v primeru individualne presoje, ter jih nato vsebinsko razčlenili v tri razrede. Za vsak razred so definirali število točk, ki ustreza pomembnosti oziroma moči razreda. Poleg opredelitve razredov in njihovih točk so na podlagi izkušenj določili še izključitvene (angl. *knockout*) kriterije, na podlagi katerih lahko pride do zavrnitve zahtevka ne glede na rezultat ocene modela. Bonitetna ocena je torej rezultat skupne ocene točkovanja vseh pojasnjevalnih spremenljivk ob upoštevanju izključitvenih kriterijev. Če je vsota točk nad določeno mejo, se zahtevek avtomatsko odobri, sicer je odločitev potrebno sprejeti na podlagi individualne presoje. Namen uporabe odločitvenega modela v njihovem primeru torej ni avtomatizacija odločanja v celoti, ampak le najbolj izrazitih primerov (zelo pozitivni) in s tem zmanjšanje dela na oddelku odobritev. Testiranje modela so izvajali dlje časa, vzporedno z neodvisno individualno presojo kreditne sposobnosti (ročna kategorizacija), kar pomeni, da imajo za vsako prošnjo za financiranje poleg vseh pojasnjevalnih spremenljivk še podatek o rezultatu

avtomatske in ročne kategorizacije. To nam bo omogočilo tudi primerjanje naših rezultatov z njihovimi.

Kot že omenjeno, bomo v tej nalogi testirali odločitvene modele logistične regresije, odločitvenih dreves in naključnih gozdov.

Za metodi logistične regresije in odločitvenih dreves smo se odločili, ker se od statističnih metod za podobne probleme v praksi največkrat uporabljata in kot rezultat zagotavljata dokaj razumljiv odločitveni model. Metodi sta v primerjavi z ostalimi bolj uporabljeni predvsem zato, ker je potrebnih manj predpostavk in sta statistično bolj robustni (Cataldo, 2010).

Poleg logistične regresije in odločitvenih dreves bomo testirali še v praksi vse bolj uporabljano metodo naključnih gozdov. Metodo je leta 2001 predstavil Leo Breiman (2001) in se uspešno uporablja na različnih področjih, kot so avtomatsko prepoznavanje neželene elektronske pošte (angl. *spam detecting*), prepoznavanje obrazov na digitalnih slikah in podobno. Sharma (2012) je primerjal logistično regresijo kot eno najuporabnejših metod pri kreditnem točkovanju z naključnimi gozdovi in ugotovil, da bi se moralo pri določanju modela za oceno kreditnega tveganja uporabljati naključne gozdove. Ti predstavljajo močno orodje za vrednotenje pomena spremenljivk, ki ga ni možno pridobiti pri standardnih regresijskih modelih, in to odpira nove možnosti za bolj robustne rešitve.

1.1 Logistična regresija

V primerih, kjer nas zanima verjetnost pripadnosti enote neki skupini, se v praksi največ uporablja logistična regresija (angl. *Logistic regression*). Eden od primerov uporabe metode pri nas je model za določanje bonitetnih ocen gospodarskih družb (Agencija Republike Slovenije za javnopravne evidence in storitve, 2014). Metoda se za razliko od drugih statističnih orodij (npr. navadne linearne regresije, multiple regresije ali diskriminantne analize) uporablja tudi v primerih, ko distribucija spremenljivk ni normalno porazdeljena in je zato ena od najprimernejših orodij za kreditno točkovanje (Cataldo, 2010). Najbolj se uporablja binarna logistična regresija, kjer sta možni dve vrednosti odvisne spremenljivke (npr. da/ne), so pa možne tudi druge vrste, kot so ordinalne ali multinominalne regresije (Ong, Huang, & Tzeng, 2005). Zaradi narave problema ocene kreditnega tveganja se bomo v nalogi omejili na binarno logistično regresijo, ki predvideva dva možna rezultata (je kreditno sposoben/ni kreditno sposoben oziroma kar da/ne). Bistvo logistične regresije je napovedovanje verjetnosti nastanka določenega dogodka.

Za lažje razumevanje logistične regresije je potrebno pojasniti pojem obeti (angl. *Odds*), ki predstavlja razmerje števila primerov nekega dogodka in števila primerov, ko se dogodek ne zgodi (Subhash, 1995).

Za lažjo ponazoritev pojma obeti lahko navedemo hipotetičen primer (Subhash, 1995, str. 318), kjer imamo 24 podjetij, od katerih je enajst velikih (med njimi deset uspešnih in eno neuspešno) ter trinajst majhnih (od teh dve uspešni in enajst neuspešnih).

Vzorec lahko prikažemo tudi v kontingenčni tabeli (Tabela 1).

Tabela 1: Kontingenčna tabela uspešnosti podjetij

Tip podjetja	**Veliko**	**Majhno**	**Skupaj**
Uspešno	10	2	**12**
Neuspešno	1	11	**12**
Skupaj	**11**	**13**	**24**

V tabeli lahko vidimo, da je v primeru velikega podjetja razmerje med uspešnimi in neuspešnimi 10 : 1, v primeru majhnega podjetja pa 2 : 11. Obeti za uspešnost so v tem primeru pri velikih podjetjih bistveno večji kot pri majhnih, in sicer:

$$Obeti\ (uspešno \mid veliko) = 10 : 1 = 10 \tag{2}$$

$$Obeti\ (uspešno \mid malo) = 2 : 11 = 0{,}182 \tag{3}$$

Obeti in verjetnosti posedujejo enako informacijo, le da so v drugačni obliki. Pretvorba iz ene oblike v drugo in obratno je zato enostavna (enačbi 4 in 5):

$$p = \frac{Obeti(y\ /x1, x2, \ldots,\ xk)}{1 + Obeti(y\ /x1, x2, \ldots, xk)} \tag{4}$$

in

$$Obeti(y|x1,\ x2, \ldots,\ xk) = \frac{p}{1-p} \tag{5}$$

V našem primeru lahko verjetnost uspešnosti za velika podjetja torej izračunamo kot 90,9 % (enačba 6) oziroma za majhna podjetja kot 15,39 %.

$$p\ (\text{uspešno} \mid \text{veliko}) = \frac{Obeti(\text{uspešno} \mid \text{veliko})}{1 + Obeti(\text{uspešno} \mid \text{veliko})} = 10/11 = 90{,}9\ \% \tag{6}$$

Na splošno lahko pri modelu logistične regresije zapišemo enačbo naravnega logaritma obeta odvisne spremenljivke za k neodvisnih spremenljivk kot linearno funkcijo:

$$\ln[\,Obeti(y|x1,x2,\ldots,xk)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{7}$$

Temu pravimo tudi logit transformacija, ali na kratko logit, ki je linearna v parametrih. Parametri β so ocenjeni po metodi največjega verjetja (angl. *Maximum Likelihood Estimation*) in v tej enačbi predstavljajo uteži vseh neodvisnih spremenljivk. V primeru, da v enačbi obete pretvorimo v verjetnost, lahko enačbo zapišemo tudi kot

$$\ln[\frac{p}{1-p}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{8}$$

Verjetnost, da se bo dogodek dejansko zgodil, lahko potemtakem izračunamo kot:

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}} \quad (9)$$

Verjetnost dogodka ima lahko po enačbi 9 vrednosti med 0 (dogodek se zagotovo ne bo zgodil) in 1 (dogodek se bo zagotovo zgodil). Vsak model ima določeno še točko reza (angl. *Cut-off value*), ki je med 0 in 1 in določa, pri kateri vrednosti verjetnosti se dihotomna vrednost napovedi (Ne/Da) spremeni iz ene vrednosti v drugo. S premikanjem mejne vrednosti proti 0 povečujemo število napak napovedi tipa I in zmanjšujemo število napak tipa II. S premikanjem mejne vrednosti proti 1 pa je učinek obraten. Katera je optimalna mejna vrednost, je odvisno od modela in, kot smo že uvodoma ugotovili, ocene stroška napak (enačba 1).

Logistična regresija zna v model vključiti tudi dihotomne vhodne oziroma neodvisne spremenljivke s pomočjo slamnatih spremenljivk (angl. *Dummy variable*), vendar je načeloma bolj učinkovita pri zveznih spremenljivkah.

Pri logistični regresiji za razliko od ostalih statističnih metod ni potrebno zagotoviti že omenjene normalne porazdelitve pojasnjevalnih spremenljivk ali linearne povezave med neodvisnimi in odvisno spremenljivko. Prav tako ni potrebno, da so vse pojasnjevalne spremenljivke med sabo neodvisne. Ima pa ta metoda tudi določene pomanjkljivosti, med katerimi je (vsaj v naših primerih) najbolj problematična nezmožnost obravnave enot vzorca, ki nimajo opredeljenih vseh vrednosti pojasnjevalnih spremenljivk. To pomeni, da model, v katerem so bile vključene vse pojasnjevalne spremenljivke v fazi učenja, lahko za učenje upošteva samo tiste enote vzorca, ki vsebujejo vse možne vrednosti kriterijev. Poleg tega pridobljen odločitveni model ne more izračunati verjetnosti odvisne spremenljivke za vse tiste enote, ki ne vsebujejo vseh vrednosti in zato ne pokriva celotne populacije.

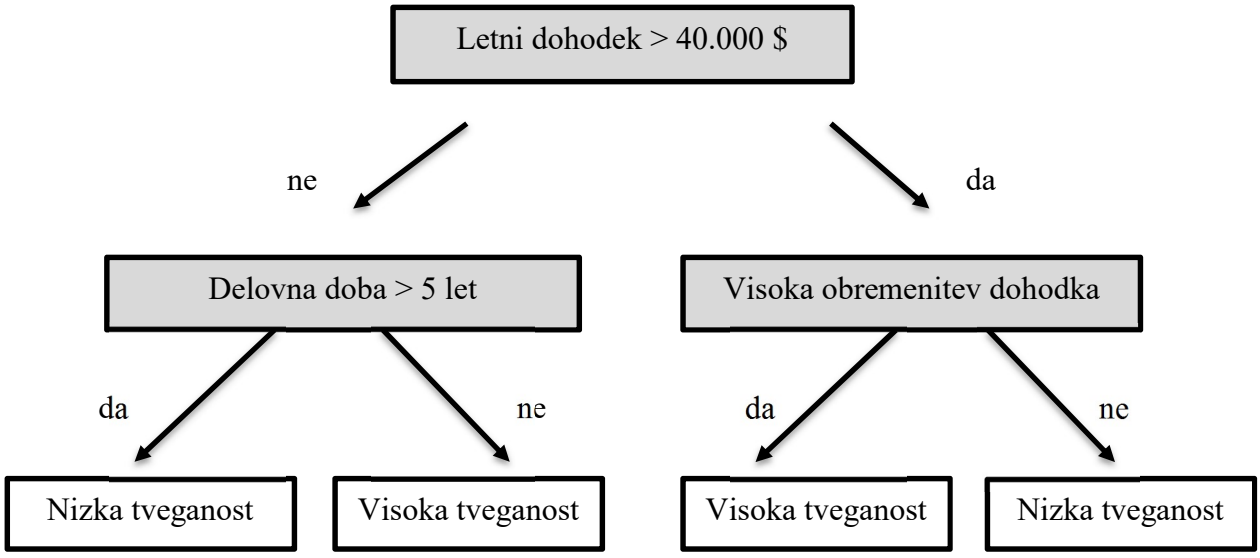
1.2 Odločitvena drevesa

Odločitvena drevesa (angl. *Decision tree*) so ena izmed metod podatkovnega rudarjenja (angl. *Data mining*), ki ga lahko opišemo kot proces odkrivanja zanimivih in pomembnih vzorcev oziroma znanja iz ogromne količine podatkov, ki so dostopni v podatkovnih bazah, podatkovnih skladiščih, na internetu in ostalih informacijskih odlagališčih (Han & Kamber, 2000). Metoda ima takšno ime zaradi posebne drevesne oblike diagrama, ki prikazuje pravila razvrstitve razredov oziroma vrednosti. Diagram je sestavljen iz vozlišč odločanja (angl. *Decision nodes*), vej drevesa (angl. *Tree branches*), ki predstavljajo povezave med vozlišči, in listov drevesa (angl. *Tree leaves*), ki predstavljajo končne razrede oziroma vrednosti izbire (Edelstein, 1999). Ideja odločitvenih dreves je, da se podatki rekurzivno delijo v vedno manjše podmnožice enot (celic), ki so vedno bolj čiste v smislu podobnosti spremenljivk. Algoritem odločitvenega drevesa vsako celico obravnava neodvisno. Da bi našli novo delitev, algoritem testira delitve, ki temeljijo na vseh razpoložljivih spremenljivkah in na koncu izbere najbolj pomembno oziroma najboljšo. Na koncu odločitveno drevo razdeli

podatke v segmente, definirane z delitvenim pravilom vsakega koraka. Skupek vseh pravil vseh segmentov tvori končni odločitveni model (Berry & Linoff, 1997).

Na Sliki 1 vidimo enostaven primer drevesnega odločitvenega modela za ugotavljanje tveganja pri posojilu (Edelstein, 1999). Odvisna spremenljivka modela je tveganost posojila (Nizka tveganost/Visoka tveganost), atributi so pa »Letni dohodek«, »Zaposlitev v letih« ter »Trenutna obremenjenost dohodka«. Iz odločitvenega drevesa lahko razberemo, da je v primeru letnega dohodka pod 40.000 $ tveganost posla nizka v primeru delovne dobe nad 5 let in visoka pri delovni dobi, nižji od 5 let. V primeru letnega dohodka nad 50.000 $ je tveganost nizka v primeru majhne trenutne obremenitve dohodka ter visoka v primeru visoke obremenitve.

Slika 1: Primer odločitvenega drevesa tveganosti posojila



Prva komponenta je krovno vozlišče odločanja ali koren drevesa (angl. *Root node*), ki predstavlja prvi test, ki ga je potrebno izvesti (Letni dohodek > 50.000 $). Glede na vrednost atributa se odločimo za eno od možnih povezav na naslednje vozlišče in tako naprej do končnega zunanjega vozlišča (lista), ki predstavlja rešitev problema. Odvisno od algoritma ima lahko vsako vozlišče dve ali več vej oziroma na dnu drevesa dva ali več listov.

Odločitvena drevesa lahko delimo na več načinov, in sicer:

1. glede na naravo kriterijev:

    a. regresijska drevesa (angl. *Regession trees*), kjer imamo za kriterije zvezne numerične spremenljivke (primer: Letni dohodek: ≤ 40.000 $ / > 40.000 $),
    b. klasifikacijska drevesa (angl. *Classification trees*), kjer imamo za kriterije nominalne neodvisne spremenljivke (primer: Visoka obremenitev dohodka: Da/Ne);

2. glede na strukturo:

    a. binarna odločitvena drevesa, kjer imamo v vseh vozliščih natanko dve veji,
    b. splošna odločitvena drevesa, kjer je v vozliščih možnih tudi več vej;

3. glede algoritma učenja:

    a. skupina AID (angl. *Automatic Interaction detector*), v katero spada več različnih algoritmov: THAID, MAID, CHAID in XCHAID, ki se razlikujejo glede na svoje karakteristike (Ritschard, 2010) – algoritmi so nastajali razvojno, ni pa namen te naloge, da bi se podrobneje spuščali v njihovo razlago,
    b. ostali, od katerih so pomembnejši CART (*angl. Classification and Regression Tree*), Quest in C.50 (Edelstein, 1999);

4. glede delitvenega kriterija oziroma algoritem izbire najboljšega atributa (Rokach & Maimon, 2005);

5. glede načina klestenja (angl. *Pruning*):

    a. na omejevanje rasti (angl. *Forward stepwise regression* ali *Pre-pruning*), kjer se uporablja hi-kvadrat ali F test za preverjanje kritične meje za nadaljevanje, sproti med izgradnjo,
    b. obrezovanje drevesa (angl. *Post-pruning*), ki temelji na merilu cene kompleksnosti (angl. *Cost of complexity*) in se izvaja na koncu, ko je odločitveno drevo že zgrajeno (Wilkinson, 1992);

6. glede metode klestenja (Rokach & Maimon, 2005):

    a. kriterij cene kompleksnosti (angl. *Cost-Complexity Pruning*),
    b. klestenje na osnovi zmanjševanja napake (angl. *Reduced error Pruning*),
    c. klestenje na osnovi minimalne napake (angl. *Minimum error Pruning*),
    d. pesimistično klestenje,
    e. optimalno klestenje,
    f. klestenje na osnovi minimalne dolžine opisa (angl. *Minimum Description Length*) in ostali.

Odločitvena drevesa, podobno kot logistične regresije, ne zahtevajo normalne porazdelitve in medsebojne neodvisnosti vhodnih parametrov modela, kot tudi ne linearne povezave z odvisno spremenljivko. Za razliko od logistične regresije je ta metoda bolj učinkovita pri dihotomnih pojasnjevalnih spremenljivkah, predvsem pa se njena prednost izkaže pri manjkajočih podatkih, saj so v vsakem primeru vse enote vključene v obdelavo, še več, manjkajoča vrednost je lahko za model dodatna informacija, ki vpliva na končno odločitev.

Glavne prednosti odločitvenih dreves pred ostalimi metodami so berljivost in razumljivost dobljenega modela ter enostavnost pretvorbe modela v strukturirana pravila odločanja (programske ukaze). Odločitvena drevesa tudi dobro obvladujejo podatkovne baze, ki vsebujejo napake (Rokach & Maimon, 2005).

Imajo pa tudi nekaj slabosti:

- Prva je preobčutljivost v fazi učenja na nepomembne atribute in šume, kar privede do prevelikega prileganja (angl. *Overfitting*) (Quinlan, 1993);
- Druga je neuravnotežena moč atributov v izbiri delitve drevesa. Če imamo nominalne spremenljivke z različnim številom možnih vrednosti, pride pri merjenju informacijskega prispevka (*angl. Information gain)* do favoriziranja tistih spremenljivk, ki imajo več možnih vrednosti (Deng, Runger, & Tuv, 2011). Možna načina rešitve te slabosti sta:

    1. Ročna prilagoditev tovrstnih atributov z zmanjšanjem nabora njihovih možnih vrednosti. To lahko dosežemo tako, da vse vrednosti atributa, ki se v vzorcu pojavijo malokrat (npr. manj kot 1 % primerov), združimo v eno vrednost »Ostalo« in s tem zmanjšamo njen nabor vrednosti, vendar pa s tem izgubimo tudi nekaj informacij;
    2. Bonferronijeva prilagoditev, ki smo jo uporabili tudi v svojih analizah. Prilagoditev deluje na principu ustreznega ponderiranja kriterijev v fazi določanja najugodnejšega za delitev v vozlišču. Ponderji so določeni tako, da nižjo moč dobijo tisti kriteriji, ki imajo več možnih različnih delitev in s tem se izniči slabost favoriziranja spremenljivk z več različnimi vrednostmi (Ritschard, 2010).

Poleg omenjenih različnih pristopov so pri gradnji dreves pomembni še parametri, ki določajo mejo gradnje in s tem končno razvejenost odločitvenega drevesa. Parametra sta dva, in sicer maksimalna globina drevesa ter minimalno število enot na vozliščih ali listih.

1.3 Naključni gozdovi

V zadnjem času se z razvojem informacijske tehnologije povečuje dostopnost do informacij in s tem se odpirajo možnosti obdelav večjih količin podatkov. Da bi lahko izkoristili prednosti velikosti modernih podatkovnih zbirk, potrebujemo učne algoritme, ki omogočajo pokritje tovrstnih količin informacij in sočasno zagotavljajo zadostno statistično učinkovitost. Metoda naključnih gozdov (angl. *Random Forests*), ki jo je zasnoval Breiman (2001), spada med najbolj učinkovite metode, ki so trenutno na voljo za takšne primere (Biau & Scornet, 2016). Naključni gozdovi imajo tudi določene izpeljanke, kot so jedrni naključni gozdovi (angl. *Kernel Random Forest*), razvrstitveni gozdovi (angl. *Ranking Forests*) (Clemencon, Depecker, & Vayatis, 2013) in podobno, vendar jih v tej nalogi ne bomo podrobneje obravnavali.

Nazeeh Ghatasheh (2014) pri analizi razpoložljivih algoritmov za ocenjevanje kreditnega tveganja kot najbolj učinkovito metodo izpostavi prav naključne gozdove. V empirični analizi, kjer je najbolj poudarjal testiranje različnih variant naključnih gozdov, med drugim tudi ugotovi, da so najboljši rezultati pri uporabi razmerja v številu enot za učenje in testiranje 1 : 1, torej 50 % enot vzorca za učenje in 50 % za testiranje. Enako razmerje smo uporabili tudi v naši analizi.

Naključni gozdovi se izkažejo kot eni izmed najučinkovitejših metod tudi v primerjavah metod pri kreditnem točkovanju. Anne Kraus (2014) v doktorski disertaciji analizira več različnih algoritmov za namen kreditnega točkovanja in pride do zaključka, da naključni gozdovi prekašajo ostale algoritme v večini testov.

Poleg naključnih gozdov je analizirala še:

- logistične regresije,
- posplošene aditivne modele (angl. *Generalized Additive Model*),
- odločitvena drevesa,
- nekaj primerov različnih Boosting metod.

V analizi je bilo uporabljeno razmerje učnih in testnih enot 1 : 1. Empirična analiza je zajemala dva naključno generirana vzorca podatkov nemške komercialne banke za potrošniške kredite, kjer je prvi vseboval 20 %, drugi pa 50 % slabih pogodb. Oba vzorca sta vsebovala 65.818 enot in bila generirana iz podatkovne baze, ki je vsebovala približno 2,6 % slabih pogodb.

Naključni gozdovi tako kot Bagging in Boosting spadajo med ansambelske metode in predstavljajo nadgradnjo Bagging metode. Pri ansambelskih metodah je značilno, da imamo namesto enega odločitvenega drevesa (klasifikatorja) za model kombinacijo več odločitvenih dreves. Enote vzorca vrednotimo z vsemi klasifikatorji in jim na osnovi glasovanja dodelimo tisto vrednost, ki je dobila največ glasov. Izkaže se, da se varianca modela pri gradnji več odločitvenih dreves zelo zmanjša, ne da bi se povečala njegova pristranskost (angl. *Bias*). Čeprav so posamezna odločitvena drevesa občutljiva na šum v učnih podatkih, to ne velja za povprečje kombiniranega modela, in sicer samo pod pogojem, da odločitvena drevesa niso v korelaciji. V povezavi s tem je pomembna še ugotovitev, da je stopnja konvergence pri učenju odvisna od števila kvalitetnih atributov in ne od količine šuma učnih podatkov (Biau, 2012).

Bagging metoda (krajše za angl. *Bootstrap aggregating*) temelji na osnovi glasovanja več individualnih klasifikatorjev različnih učnih množic (Breiman, 1996). Iz osnovne množice enot (ϵ_0) se generira več novih učnih množic (ϵ_i'), tako da se enote iz osnovne množice izberejo naključno z vračanjem, dokler ne izberemo enakega števila enot, kot jih je v osnovni množici. To pomeni, da so določene enote v novi množici lahko izbrane večkrat, določene pa nikoli. Statistično gledano je v novi množici v povprečju izbranih 63,2 % enot osnovne

množice. Preostale enote predstavljajo tako imenovano množico OOB (angl. *Out Of Bag*). Na vsaki novi množici se kreira ločen klasifikator (odločitveno drevo), ki predstavlja del končnega kombiniranega modela.

Pri Boosting metodi (Freund & Schapire, 1997) je pristop interaktiven, saj učenje individualnih klasifikatorjev ni več neodvisno, ampak ugotovljene napake predhodnega vplivajo na učenje novega klasifikatorja. Za razliko od Bagging metode, kjer se enote vsake množice določajo naključno, se pri tej metodi pri kreiranju i-tega odločitvenega drevesa upoštevajo vse enote osnovne učne množice, vendar napačno klasificirane enote predhodnega odločitvenega drevesa dobijo večjo težo od pravilnih. Znan algoritem iz te družine je AdaBoost.

Naključni gozdovi, enako kot Bagging metoda, temeljijo na naključno generiranih množicah enot z vračanjem. Razlika je le v kreiranju odločitvenih dreves, saj je pri naključnih gozdovih dodana še mera naključnosti. To se doseže tako, da se ne izbira optimalen atribut za delitev vozlišča med vsemi možnimi atributi, temveč samo izmed določenega števila naključno izbranih atributov. Takšen postopek izbire optimalnega atributa lahko imenujemo tudi atributni bagging (angl. *Feature bagging*). Strukturo naključnih gozdov določata samo dva parametra, in sicer (1) število dreves v gozdu ter (2) število naključno izbranih pojasnjevalnih spremenljivk pri izbiri optimalnega atributa (Bernard, 2003).

Prednosti naključnih gozdov pred ostalimi algoritmi so predvsem učinkovitost oziroma točnost napovedi, robustnost ter enostavnost algoritma, saj strukturo modela določata samo dva parametra.

Glavna slabost algoritma je kljub njegovi enostavnosti predvsem velika kompleksnost končnega odločitvenega modela, kar zmanjšuje njegovo razumevanje ter oteži izvajanja raznih analiz (Misha, David, & Nando, 2013). Zaradi velike kompleksnosti je problematična tudi implementacija odločitvenih pravil pri praktični uporabi. Za omilitev tega problema sta Breiman in Lecture (2002) določila tri različne načine izračuna mere pomembnosti atributov, s pomočjo katerih lahko ocenimo, kateri atributi imajo večji in kateri manjši vpliv v odločitvenem modelu. Osnova za izračun pomembnosti i-tega atributa A_i je pri vseh treh načinih razlika rezultatov klasifikacije med izhodiščno OOB množico ter njeno transformacijo $\widehat{OOB}$. Množico transformiramo tako, da enotam naključno premešamo vrednosti atributa A_i ob nespremenjenih vrednostih ostalih atributov. Z drugimi besedami, vsem enotam določimo atributu A_i naključne vrednosti in s tem povzročimo, da atribut ne pojasni več nobene variance. To v nadaljevanju omogoča, da na osnovi razlik obeh ocen ugotovimo, koliko variance je imel na začetku. Podrobnejša pojasnila glede izračuna pomembnosti atributov v praksi z orodjem R so razvidna v prispevku Classification and regression by randomForest (Liaw & Wiener, 2002).

1.4 Validacija in merjenje uspešnosti učenja

Pri raziskavah na področju modeliranja kreditnih tveganj se poskuša doseči boljše klasifikacijske sposobnosti, boljšo interpretacijo rezultatov ter odkrivanje novih možnosti uporabe modelov. Za vse te aktivnosti je potrebno imeti možnost ocene klasifikacijske točnosti modela ter možnost primerjave dveh modelov.

Za primerjavo klasifikacijske točnosti modelov se v praksi večinoma uporabljata dve metodi primerjave oziroma analize, in sicer kontingenčne matrike (napovedanih in dejanskih vrednosti odvisnih spremenljivk) ter ROC analize, ki jih bomo podrobneje predstavili v nadaljevanju.

1.4.1 Kontingenčne tabele

Kontingenčne tabele (angl. *Contingency table*) ali matrike razvrstitev (angl. *Confusion matrix*) uporabljamo za prikaz rezultatov odločitvenega modela. V primeru binarne klasifikacijske tabele (Tabela 2) se vnesejo števila pravilno in nepravilno napovedanih primerov, iz katerih se zaradi lažjih primerjav izračunajo določena njihova razmerja oziroma metrike vrednotenja.

Kot je razvidno v Tabeli 2, so tu prikazane različne vsebinske vrednosti rezultata modela, in sicer:

- število pravilno napovedanih zavrnitev (v nadaljevanju TN – *angl. True Negative*),
- število nepravilno napovedanih zavrnitev (v nadaljevanju FN – *angl. False Negative*),
- število pravilno napovedanih odobritev (v nadaljevanju TP – *ang. True Positive*),
- število nepravilno napovedanih odobritev (v nadaljevanju FP – *angl. False Positive*).

Tabela 2: Kontingenčna tabela napovedanih in dejanskih vrednosti

		Napovedane vrednosti		
		Zavrnjeno	Odobreno	
Dejanske vrednosti	Zavrnjeno	TN	FP	*Specifičnost*
	Odobreno	FN	TP	*Občutljivost*
			Natančnost	*Klasifikacijska točnost*

Zaradi lažje primerjave rezultatov različnih modelov pa iz kontingenčne tabele računamo naslednje metrike vrednotenja:

- pravilno pozitivna mera ali mera občutljivosti (v nadaljevanju TPR – angl. *True Positive Rate* ali *Sensitivity*), ki predstavlja delež pravilno napovedanih odobritev glede na vse dejanske odobritve:

$$TPR = \frac{TP}{FN+TP} \tag{10}$$

- napačno pozitivna mera (v nadaljevanju FPR – angl. *False Positive Rate* ali *false alarm rate)*, ki predstavlja delež vseh napačno napovedanih odobritev glede na vse dejanske zavrnitve:

$$FPR = \frac{FP}{TN+FP} \quad \text{oz.} \quad FPR = 1\text{-}SPC \tag{11}$$

- mera specifičnosti (v nadaljevanju SPC – angl. *specificity)*, ki predstavlja delež vseh pravilno napovedanih zavrnitev glede na vse dejanske zavrnitve:

$$SPC = \frac{TN}{TN+FP} \quad \text{oz.} \quad SPC = 1 - FPR \tag{12}$$

- mera natančnosti (v nadaljevanju PPV – angl. *Positive Predictive Value ali Precision*), ki predstavlja delež pravilno napovedanih odobritev glede na vse napovedane odobritve:

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

- mera klasifikacijske točnosti (angl. *Accuracy*), ki predstavlja delež vseh pravilnih napovedi glede na vse napovedi:

$$Klasifikacijska\ točnost = \frac{TN+TP}{TN+TP+FN+FP} \tag{14}$$

1.4.2 ROC analiza

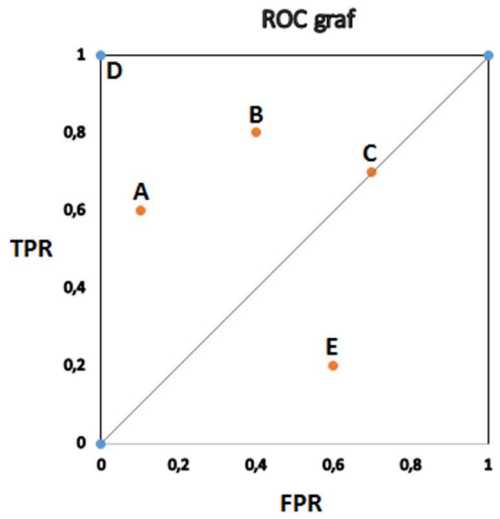
ROC analiza (angl. *Receiver Operating Characteristics*) predstavlja tehniko za vizualizacijo in izbor klasifikatorjev glede na njihove izvedbe in s tem omogoča izbiro optimalnih modelov (Swets, 1988). Poleg tega je ROC analiza tudi neposredno povezana z analizo stroškov/prednosti pri izdelavi diagnostičnih modelov (Swets, Dawes, & Monahan, 2000). V zadnjih letih se vse bolj uporablja na področju strojnega učenja, predvsem pri tehnikah, kot so naivni Baysov klasifikator (angl. *Naive Bayes classifier*), odločitvena drevesa, nevronske mreže (angl. *Artificial Neural Network*), k-bližnji sosedi (angl. k-Nearest Neighbours) ter podpora vektorskih strojev (angl. *Support Vector Machine*) (Majnik & Bosnić, 2013).

ROC graf je dvodimenzionalen graf; njegova ordinatna os je pravilno pozitivna mera (TPR), abscisna pa napačno pozitivna mera (FPR). Točka na grafu torej prikazuje razmerje med pravilnimi in nepravilnimi pozitivnimi razvrstitvami. Med vsemi variantami grafa imamo naslednje tri skrajne in pomembne točke (Fawcett, 2006):

1. Točka (0,0), ki predstavlja strategijo zavrnitve vseh zahtev. V tem primeru dobimo 0 % delež pravilnih odobritev in 0 % delež napačnih odobritev (saj nimamo nobene odobritve). Hkrati takšen model zagotavlja 100 % delež pravilnih zavrnitev in 100 % delež nepravilnih zavrnitev;
2. Točka (1,1), ki predstavlja strategijo odobritve vseh zahtev. Pri takšni klasifikaciji dobimo analogno prejšnji 100 % delež pravilnih in nepravilnih odobritev ter 0 % delež pravilnih in nepravilnih zavrnitev;
3. Točka D (0,1) predstavlja optimalno klasifikacijo, saj predstavlja 100 % delež pravilnih odobritev, 0 % delež nepravilnih odobritev, 100 % delež pravilnih zavrnitev in 0 % delež napačnih zavrnitev.

Cilj vsakega modela razvrstitve je pozicioniranje na ROC grafu čim bližje točki D (0,1). Poljubna točka ROC prostora je boljša od ostalih, če se nahaja bližje točki D.

Slika 2: Osnovni ROC graf s petimi diskretnimi klasifikatorji

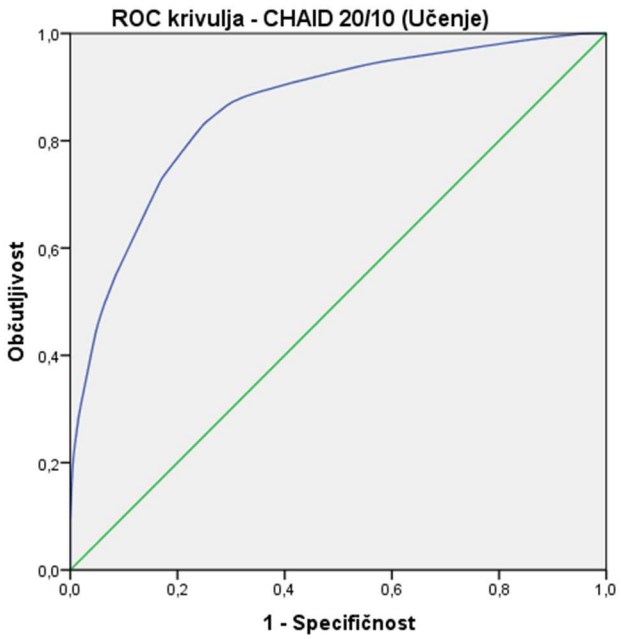


Vir: T. Fawcett, An introduction to ROC analysis,2006, str. 862.

Na Sliki 2 vidimo, da je A bolj konservativen kot B, saj ima samo 60 % pravilno pozitivnih vrednosti v primerjavi z B, kjer je ta delež 80 %. A ima v primerjavi z B tudi bistveno manjši odstotek slabih razvrstitev/napovedi, saj je tu razmerje 10 %, pri B pa 40 %. Katera klasifikacija je boljša, je odvisno od cene stroška/prednosti, obe pa sta približno enako oddaljeni od optimalne točke D. Točke na diagonali y = x predstavljajo strategijo naključnega ugibanja razvrstitve. Če bi metoda klasifikacije temeljila na naključnem ugibanju, ki bi polovico primerov klasificirala z DA in polovico z NE, bi po daljšem času imeli približno enako število pravilnih in nepravilnih klasifikacij. V tem primeru bi imeli kot rezultat točko (0,5, 0,5). Primer takšne naključne klasifikacije predstavlja tudi točka C (0,7, 0,7), za katero bi lahko rekli, da se je pozitivno ugibalo v 70 % in negativno v 30 % primerov (Fawcett, 2006).

Modeli, ki smo jih testirali v nalogi, v prvem koraku izračunajo zvezno vrednost verjetnosti klasifikacije (med 0 in 1) in na podlagi v naprej opredeljene točke reza (angl. *cut value*) določijo ustrezno končno skupino. V primeru nizke verjetnosti (od 0 do točke reza) model primer zavrne oziroma klasificira z NE, v primeru izračunane visoke verjetnosti (večje od točke reza) pa odobri oziroma klasificira z DA. S premikanjem točke reza od 0 proti 1 dobimo vsa možna razmerja števil primerov obeh množic in s tem za vsako točko reza drugo koordinato na ROC grafu. V primeru zelo velikega števila majhnih premikov točke reza dobimo povezano ROC krivuljo, ki pove, kako se opazovan model obnaša. Na Sliki 3 vidimo primer ROC krivulje CHAID odločitvenega drevesa na podatkih učenja.

Slika 3: Primer ROC krivulje odločitvenega drevesa CHAID 20/10



Površina pod krivuljo (v nadaljevanju AUC – angl. *Area Under the Curve*) v celoti pove, kako dober je model. S tem omogoča primerjavo različnih modelov v celotnem spektru razmerij. Če vemo, kakšna sta stroška napake tipa I in II, lahko na podlagi enačbe (1) ugotovimo optimalni (najcenejši) odločitveni model, ki predstavlja eno od točk na ROC krivulji. Površina AUC se v splošnem giblje od 0,5 do 1, kjer površina 0,5 predstavlja strategijo naključnega ugibanja, površina 1 pa popolni model, ki pravilno napove pripadnost vseh primerov.

Poleg AUC se pri oceni uspešnosti modela lahko primerja tudi Gini koeficient, ki pove, kje je umeščen opazovani model (med naključnim in popolnim), le da Gini indeks predstavlja ploščino med ROC krivuljo in diagonalo.

Med ocenama obstaja tudi povezava:

$$Gini = (AUC \times 2) - 1 \tag{15}$$

oziroma, če je AUC < 0,5:

$$Gini = 1 - (AUC \times 2) \tag{16}$$

2 EMPIRIČNA PRIMERJAVA METOD ZA VZPOSTAVITEV TOČKOVNEGA SISTEMA

2.1 Namen in cilji raziskave

Osrednji namen magistrskega dela je ugotoviti, ali je možno s sodobnimi orodji vzpostaviti takšen točkovni model, ki bo dovolj učinkovit pri ocenjevanju kreditnega tveganja, da se ga bo lahko uporabilo v praksi in hkrati izvedljiv v časovno spremenljivih okvirih. Cilj dela je proučiti, testirati in medsebojno primerjati v praksi najbolj uporabljene metode, kot so logistična regresija, odločitvena drevesa ter metoda naključnih gozdov, in ugotoviti, katera od njih je najbolj primerna za naše konkretne podatke. Učinkovitost algoritmov je namreč zelo odvisna od primera do primera, zato je v prvem koraku potrebno preveriti, katera izmed teh metod je najprimernejša za naše konkretne podatke.

2.2 Raziskovalna vprašanja in hipoteze

V magistrskem delu želimo potrditi ali ovreči hipotezo, da je možno s temi metodami in trenutno tehnologijo uspešno producirati točkovne modele, ki pravilno klasificirajo zahteve za financiranje glede tveganosti posla. Zato skušamo odgovoriti na naslednja raziskovalna vprašanja:

1. Bodo dobljeni modeli po učinkovitosti primerljivi z rezultati modela v uporabi v podjetju ABC Lizing in s tem primerni za praktično uporabo?
2. Bo postopek vzpostavitve modela s takšnimi orodji časovno in po obsegu dela bistveno manj zahteven od modela v uporabi?
3. Lahko s kombiniranjem omenjenih metod rezultate modelov še izboljšamo?

2.3 Metodologija

V analizi smo uporabili statistična programa SPSS in R. Z orodjem SPSS smo analizirali različne variante logistične regresije in odločitvenih dreves, naključne gozdove pa smo zaradi lažje dostopnosti funkcijskih modulov izvajali z orodjem R, kjer smo uporabili knjižnico »randomForest« (Breiman & Cutler, 2015).

2.3.1 Podatki

Podatki, ki jih imamo na voljo, predstavljajo 7.578 prošenj fizičnih oseb za financiranje osebnega vozila, prejetih v času od 1. 1. 2015 do 1. 12. 2015. Prošnje so se na oddelku za upravljanje s tveganjem na koncu odobrile ali zavrnile. Proučevana enota je torej prošnja za financiranje. Vsaka enota ima eno odvisno spremenljivko, tj. končno odločitev (ročna kategorizacija), in 30 neodvisnih oziroma pojasnjevalnih spremenljivk. Odvisna spremenljivka je dihotomna in ima dve možni vrednosti: Odobreno ali Zavrnjeno. Med vsemi zahtevki je bilo odobrenih 4.602 (60,7 %) in zavrnjenih 2.976 (39,3 %) zahtevkov. Neodvisne spremenljivke lahko vsebinsko klasificiramo v tri skupine, in sicer v:

1. Spremenljivke, ki opisujejo predmet financiranja:

    a. znamka in model vozila: blagovna znamka vozila ter model (spremenljivka je nominalna in statistično značilna);
    b. letnik vozila: leto prve registracije (spremenljivka je numerična, razmernostna in statistično značilna);
    c. maloprodajna cena (spremenljivka je numerična, razmernostna in statistično značilna);
    d. ali je dobavitelj na črni listi: ali dobavitelj dolguje povprečno tri obroke na pogodbo pri kateremkoli lizing podjetju, ki je v sistemu izmenjave podatkov; določena lizing podjetja si namreč med sabo izmenjujejo podatke dolžnikov, ki niso fizične osebe, zaradi zmanjšanja tveganja pri sklepanju novih poslov (spremenljivka je opisna, dihotomna);
    e. število pogodb dobavitelja: število vseh živih pogodb v sistemu, ki imajo predmet financiranja pri istem dobavitelju (spremenljivka je numerična, razmernostna in statistično značilna);
    f. eurotax procent: za koliko odstotkov se maloprodajna vrednost razlikuje od eurotax vrednosti v primeru rabljenega vozila; v primeru pozitivne vrednosti odstotka je prodajna vrednost večja, sicer je manjša (spremenljivka je numerična, razmernostna in statistično značilna).

2. Spremenljivke, ki opisujejo posojilojemalca:

    a. ali je posojilojemalec na črni listi: ali posojilojemalec dolguje povprečno tri obroke na pogodbo (spremenljivka je opisna, dihotomna in statistično značilna);
    b. državljanstvo posojilojemalca (spremenljivka je opisna in statistično značilna);
    c. državljan EU: ali je posojilojemalec državljan EU (spremenljivka je opisna, dihotomna in statistično značilna);
    d. pošta posojilojemalca: poštna številka stalnega prebivališča (spremenljivka je opisna in statistično značilna);

e. obligo: kakšen je bil trenutni dolg posojilojemalca v trenutku odobravanja v okviru drugih, obstoječih pogodb (spremenljivka je numerična, razmernostna in statistično značilna);
f. zapadlo neplačano: kakšen je bil trenutni znesek zapadlih neplačanih terjatev posojilojemalca v trenutku odobravanja v okviru drugih, obstoječih pogodb (spremenljivka je numerična, razmernostna);
g. neto plača: mesečna neto plača iz plačilne liste (spremenljivka je numerična, razmernostna in statistično značilna);
h. razpoložljivi OD: mesečna neto plača, povečana za malico in prevoz na delo (spremenljivka je numerična, razmernostna in statistično značilna);
i. starost posojilojemalca (spremenljivka je numerična, razmernostna in statistično značilna);
j. plačilni indeks: povprečna zamuda pri plačevanju v dnevih za obstoječe pogodbe (spremenljivka je numerična, razmernostna in statistično značilna);
k. tip delodajalca: gospodarske združbe, javne ustanove, državna uprava ali ZPIZ, odvetniki, notarji ali zasebne ordinacije, svobodni poklici ali delodajalci v tujini (spremenljivka je opisna in statistično značilna);
l. tip zaposlitve: zaposlitev za določen čas, nedoločen čas, upokojenec (spremenljivka je opisna in statistično značilna);
m. ali je delodajalec posojilojemalca na črni listi: enako kot ali je dobavitelj na črni listi, le da to velja za delodajalca (spremenljivka je opisna, dihotomna in statistično značilna);
n. blokada računa delodajalca: ali je imel delodajalec v preteklosti že blokiran TRR račun (spremenljivka je opisna, dihotomna in statistično značilna);
o. insolventnost delodajalca: ali je uveden insolventni postopek nad delodajalcem (spremenljivka je opisna, dihotomna in statistično značilna);
p. število zaposlenih pri delodajalcu (spremenljivka je numerična, razmernostna in statistično značilna);
q. doba poslovanja delodajalca v letih (spremenljivka je numerična, razmernostna in statistično značilna);
r. prihodki delodajalca (spremenljivka je numerična, razmernostna);
s. odstotek kapitala v bilančni vsoti delodajalca (spremenljivka je numerična, razmernostna in statistično značilna);
t. poslovni izid delodajalca (spremenljivka je numerična, razmernostna);
u. število odobritev: število že odobrenih prošenj za financiranje za posojilojemalca v preteklosti (spremenljivka je numerična, razmernostna in statistično značilna).

3. Spremenljivke, ki opisujejo način financiranja:

a. znesek pologa (spremenljivka je numerična, razmernostna in statistično značilna);
b. trajanje financiranja v mesecih (spremenljivka je numerična, razmernostna in statistično značilna);

c. obrestna mera – OM (spremenljivka je numerična, razmernostna in statistično značilna).

2.3.2 Priprava podatkov in postopek priprave testnih modelov

Pred začetkom priprave testnih modelov smo opravili statistično analizo podatkov. Na ta način lahko preverimo parametre ter lociramo in popravimo napačne vrednosti, ki so nastale zaradi napak ob vnosu. S tem smo sicer povzročili razlike med vhodnimi podatki svojih modelov in modelom primerjave ter do določene mere verjetno tudi razlike v rezultatih, vendar je bilo to nujno potrebno, saj smo želeli dobiti čim bolj realne rezultate svojih modelov točkovanja. Za vsako pojasnjevalno spremenljivko posebej smo še dodatno preverili, ali je statistično značilna pri testu odvisnosti na končno odločitev (ročna kategorizacija). Poleg poprave napačnih vrednosti podatkov smo pri preslikavi podatkov tudi nekoliko skrčili šifrant modelov vozil, saj je v njihovem odločitvenem modelu nabor vrednosti zelo razdrobljen (več kot 300 variant), kar bi za nas pomenilo ogromno slamnatih spremenljivk. Šifrant smo omejili samo na znamke vozil in ga s tem zmanjšali na 43 različnih vrednosti.

Pri multivariantnih regresijah ni nujno, da so pojasnjevalne spremenljivke v modelu posamično statistično značilne, saj lahko na odvisno spremenljivko zelo vpliva tudi določena kombinacija neznačilnih spremenljivk. Da bi ugotovili, kakšne so odvisnosti za značilne in neznačilne pojasnjevalne spremenljivke na končni rezultat modela, smo pri določenih testih naredili po dva odločitvena modela:

- na vseh pojasnjevalnih spremenljivkah,
- za primerjavo še samo na značilnih.

Za učenje vseh modelov smo vzorec primerov razdelili v dve naključni podmnožici, od katerih smo zaradi bolj natančne primerjave pri vseh testih uporabili isto podmnožico enot za učenje in testiranje.

2.3.3 Testiranje modelov logistične regresije

Logistične regresije smo testirali z orodjem SPSS, kjer smo v vseh primerih uporabili stepwise metodo s pragom verjetnosti za dodajanje spremenljivk 0,05 in pragom 0,1 za odvzemanje. Točka reza je bila na 0,5, maksimalno številom iteracij 20, v model pa smo vključili tudi konstanto.

Podatki, ki smo jih imeli na voljo, so za določene pojasnjevalne spremenljivke vsebovali veliko število enot brez vrednosti. To se je v primeru logistične regresije izkazalo kot problematično, saj metoda ne zna kategorizirati enot z manjkajočimi vrednostmi. V primeru, ko smo v modelu uporabili samo 20 izbranih pojasnjevalnih spremenljivk (za ta nabor imamo opredeljene vrednosti na celotnem vzorcu), smo obdelali vseh 7.578 enot vzorca. Če izberemo vse pojasnjevalne spremenljivke, jih upoštevamo samo 1.818, saj imamo pri

določenih kriterijih na voljo samo 24 % enot vzorca z vrednostmi. Z drugimi besedami to pomeni, da bi dobljen model nad vsemi pojasnjevalnimi spremenljivkami v povprečju lahko uporabljali za kategorizacijo samo četrtine populacije, kar pa v praksi nima uporabne vrednosti. Razlog za takšno veliko število manjkajočih vrednosti našega vzorca je v večini primerov vsebinski, saj je določenim pojasnjevalnim spremenljivkam možno opredeliti vrednost samo pri določeni podskupini enot.

Spremenljivke, ki zaradi svoje vsebine ne vsebujejo vrednosti za vse enote, so:

1. plačilni indeks in trenutna obveznost stranke (vrednost je prisotna samo v primeru že obstoječih strankah);
2. podrobni podatki delodajalca, kot so blokada računa, insolventnost, število zaposlenih, doba poslovanja, prihodki in odstotek kapitala v bilančni vsoti (podatki so na voljo le v primeru, ko je delodajalec profitna organizacija).

Da bi se izognili problemu manjkajočih podatkov in hkrati v modelu upoštevali maksimalni nabor pojasnjevalnih spremenljivk, smo en test enote vzorca razdelili v štiri vsebinsko skladne podskupine, ki so imele različen nabor pojasnjevalnih spremenljivk in zgradili model na vsaki podskupini.

Podskupine vsebinsko opredelimo kot:

1. nova stranka, delodajalec ni profitna organizacija;
2. nova stranka, delodajalec je profitna organizacija;
3. obstoječa stranka, delodajalec ni profitna organizacija;
4. obstoječa stranka, delodajalec je profitna organizacija.

Rezultate vseh štirih modelov smo na koncu združili zaradi lažje primerjave z ostalimi testi. Pri vsaki izvedbi logistične regresije smo izračunano verjetnost rezultata modela shranili v novo spremenljivko, pri kateri smo nato opravili še ROC analizo.

Zaradi velikega števila opisnih kriterijev imajo končni modeli logistične regresije veliko število slamnatih spremenljivk in zanimalo nas je, ali bi se z reduciranjem njihovega števila izboljšal rezultat. V ta namen smo pri opisni spremenljivki z največjim naborom vrednosti (znamka vozila) popravili vse maloštevilne vrednosti na vrednost »Ostalo«. S tem se je število slamnatih spremenljivk za kriterij Znamka zmanjšalo iz 43 na 7.

Poleg testiranja vseh možnih pojasnjevalnih spremenljivk smo pri vseh podskupinah testirali še varianto s samo značilnimi pojasnjevalnimi spremenljivkami, da bi videli, ali lahko s tem izboljšamo rezultate. Rezultati, ki smo jih dobili na samo značilnih pojasnjevalnih spremenljivkah, so bili brez izjeme v testiranjih malenkost slabši od rezultatov nad vsemi spremenljivkami. Zaradi velikega števila vseh različnih testov jih zaradi preglednosti primerjav v rezultatih ne bomo navajali.

2.3.4 Testiranje modelov odločitvenih dreves

V tem delu smo analizirali dva v praksi najbolj uporabljena algoritma odločitvenih dreves, in sicer algoritma CHAID ter CART. Odločitvena drevesa CHAID se od CART dreves razlikujejo v več delih. Prva razlika je v metodi izbire kriterija za delitev, saj se pri CHAID drevesih pri izgradnji uporablja hi-kvadrat metoda, pri CART pa Gini indeks, ki temelji na metodi najmanjših kvadratov (Breiman, Freidman, Stone, & Olsten, 1984). Druga razlika je v strukturi, saj so CART drevesa binarna, CHAID pa splošna. Zadnja večja razlika med metodama je način klestenja. Pri CHAID metodi se uporablja omejevanje rasti, pri CART pa obrezovanje drevesa.

Metode odločitvenih dreves smo testirali z orodjem SPSS, kjer smo imeli pri CHAID algoritmih naslednje predlagane nastavitve:

- meja značilnosti 0,05 za delitev vozlišč,
- meja značilnosti 0,05 za zlitje vozlišč,
- maksimalno število iteracij 100,
- *Pearsonov* hi-kvadrat model,
- *Bonferronijeva* prilagoditev pri določanju vpliva spremenljivk,
- maksimalno število intervalov zveznih spremenljivk 10.

Pri modelih CART smo uporabili Gini indeks z mejo minimalne spremembe za napredovanje 0,0001 ter brez uporabe klestenja.

Pri analizi odločitvenih dreves smo preverjali tri variante različnih razvejenosti, in sicer:

1. minimalno število enot notranjih vozlišč 100, minimalno število enot v listih 50 (100/50) ter maksimalna globina drevesa 3 v primeru CHAID in 5 v primeru CART;
2. minimalno število enot notranjih vozlišč 20, minimalno število enot v listih 10 (20/10) ter maksimalna globina drevesa 15 – za obe metodi;
3. minimalno število enot notranjih vozlišč 1, minimalno število enot v listih 1 (1/1) ter neomejena maksimalna globina drevesa.

Prva možnost razvejenosti je prednastavljena v SPSS orodju in kot rezultat vrne najmanj razvejeno odločitveno drevo v primerjavi z ostalimi. Največje in najbolj razvejeno odločitveno drevo je pri tretji varianti, saj praktično nima omejitev. Pri vseh treh možnostih smo opravili šest preverjanj, in sicer analizo tri CHAID in tri CART metode:

1. analizo metode nad vsemi nespremenjenimi pojasnjevalnimi spremenljivkami;
2. analizo metode nad samo značilnimi pojasnjevalnimi spremenljivkami ter
3. analizo metode nad vsemi pojasnjevalnimi spremenljivkami, kjer smo kriteriju Znamka vozila pretvorili izjeme v kategorijo Ostalo.

Pri vseh testih smo, podobno kot pri logistični regresiji, v novo spremenljivko shranili izračunano verjetnost rezultata ter nad njo opravili še ROC analizo.

Podobno kot pri logistični regresiji, tudi pri odločitvenih drevesih zaradi preglednosti ne bomo navajali rezultatov modelov nad samo značilnimi spremenljivkami, saj so bili vsi brez izjeme malenkost slabši kot pri osnovnih.

2.3.5 Testiranje modelov naključnih gozdov

Pri učenju naključnih gozdov smo uporabili prevzete nastavitve algoritma, ki med drugim določajo 500 odločitvenih CART dreves, maksimalno razvejenost (minimalno število enot v vozliščih in listih je 1) ter velikost podmnožice naključnih kriterijev $\sqrt[2]{n}$, kjer je n število vseh kriterijev vzorca. Naključne gozdove smo testirali z orodjem R, in sicer na štiri različne načine:

1. nad vsemi pojasnjevalnimi spremenljivkami (enak nabor kot pri testu odločitvenih dreves);
2. nad vsemi pojasnjevalnimi spremenljivkami, vključno z rezultatom logistične regresije z minimalnim naborom spremenljivk (v tabelah s pripono +LR);
3. nad vsemi pojasnjevalnimi spremenljivkami, vključno z rezultatom logistične regresije nad podmnožicami (nova/obstoječa stranka, je/ni delodajalca; v tabelah s pripono +LR1234);
4. nad vsemi pojasnjevalnimi spremenljivkami, vključno z obema rezultatoma logističnih regresij (v tabelah s priponama +LR in +LR1234).

Enako kot pri prvih dveh metodah, smo tudi pri naključnih gozdovih opravili vse teste nad celotnim naborom kriterijev, na samo značilnih kriterijih ter na podatkih s popravljeno znamko vozil pri izjemah v Ostalo. Upoštevanje samo značilnih kriterijev v modelih naključnih gozdov ni izboljšalo rezultatov in jih enako kot pri ostalih metodah ne bomo navajali.

2.3.6 Testiranje kombiniranih modelov

Poleg analize omenjenih treh modelov smo v fazi učenja preverili še nekaj primerov njihovih kombinacij. S tem smo želeli ugotoviti, ali lahko s kombiniranjem različnih metod dodatno izboljšamo odločitveni model. Kombinacije smo testirali na dva različna načina.

Pri prvem načinu smo obstoječim pojasnjevalnim spremenljivkam dodali končni rezultat razvrstitve druge metode kot dodatno pojasnjevalno spremenljivko (vrednost izračunane verjetnosti napovedi), in sicer:

1. logistični regresiji (Minimalni nabor) smo kot dodatni pojasnjevalni spremenljivki dodali še rezultata verjetnosti klasifikacije odločitvenega modela CHAID 100/50 in CART 100/50;

2. odločitvenim drevesom CHAID 100/50, CHAID 20/10, CHAID 1/1, CART 100/50, CART 20/10 in CART 1/1 smo dodali pojasnjevalno spremenljivko, ki je rezultat verjetnosti klasifikacije logistične regresije (Minimalni nabor);
3. naključnim gozdovom smo kot dodatne pojasnjevalne spremenljivke dodali rezultate verjetnosti klasifikacije ostalih metod (točke 2, 3 in 4 poglavja 2.3.5)

Kot drugi način kombiniranja različnih metod smo testirali še primer, ki sta ga izvedla Antipov in Pokryshevskaya (2010). Na podlagi delitvenih kriterijev vozlišč CHAID odločitvenega drevesa sta razdelila osnovno množico enot na ustrezne podmnožice in nad njimi izvedla logistične regresije. Na takšen način sta združila obe metodi in s tem nekoliko izboljšala rezultate modela.

3 REZULTATI

3.1 Logistična regresija

Analizo logistične regresije lahko generalno kategoriziramo v pet različnih variant, od katerih sta zadnji dve kombinirani z rezultati odločitvenih dreves:

1. logistična regresija na vseh enotah vzorca z minimalnim naborom pojasnjevalnih spremenljivk;
2. logistična regresija na vseh enotah vzorca z minimalnim naborom pojasnjevalnih spremenljivk in reduciranim številom slamnatih spremenljivk;
3. razdelitev podatkov na štiri vsebinsko ločene podmnožice (nova stranka, obstoječa stranka, obstaja delodajalec stranke, ne obstaja delodajalec stranke) in izvedba ločenih logističnih regresij nad podmnožicami z namenom maksimiranja števila pojasnjevalnih spremenljivk pri učenju;
4. logistična regresija v kombinaciji z odločitvenimi drevesi, kjer se poleg minimalnega nabora kriterijev v fazi učenja upoštevata tudi rezultata napovedane verjetnosti CHAID in CART odločitvenih dreves;
5. razdelitev podatkov na štiri vsebinsko ločene podmnožice, le da je tokrat delitev opravljena na osnovi delitvenih kriterijev CHAID odločitvenega drevesa; test je opravljen po že izvedenem postopku, ki sta ga opravila in analizirala Antipov & Pokryshevskaya (2010).

3.1.1 Minimalni nabor pojasnjevalnih spremenljivk

Pri testiranju logistične regresije z minimalnim naborom pojasnjevalnih spremenljivk smo v primerjavi z ostalimi modeli dobili slabše rezultate. Pri točki reza 0,5 je bila klasifikacijska točnost na testnih podatkih 69,5 %.

V Tabeli 3 vidimo ostale podatke testiranja modela, in sicer:

- v levem delu so prikazani rezultati modela na podatkih učenja,

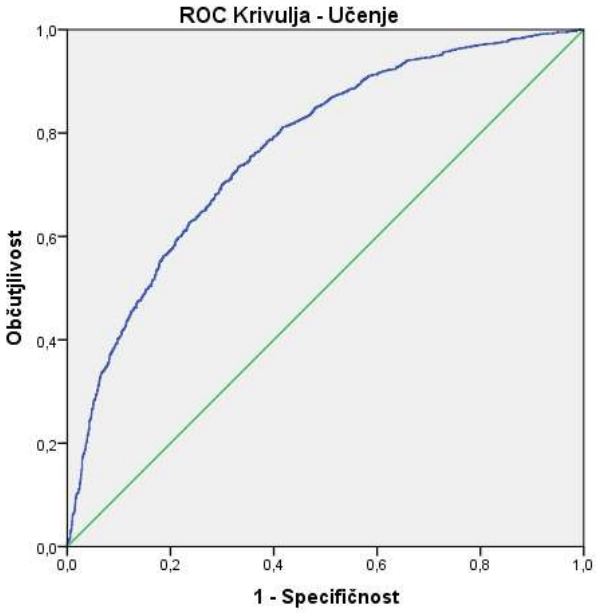
- v desnem na podatkih testnih primerov.

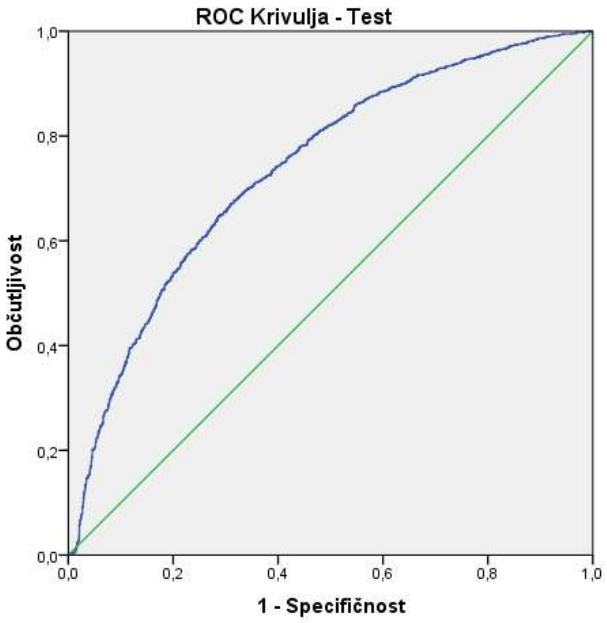
Tabela 3: Kontingenčna tabela logistične regresije z minimalnim naborom spremenljivk pri točki reza 0,5

		Napoved					
		Podatki učenja			Podatki testiranja		
		Zavrnitev	Odobritev	%	Zavrnitev	Odobritev	%
Opazovanja	Zavrnitev	733	754	49,3	678	806	45,7
	Odobritev	310	1986	86,5	348	1952	84,9
				71,9			69,5

Iz Tabele 3 je razvidno, da so rezultati na podatkih testnih primerov le malenkost slabši kot na podatkih učenja, kar pomeni, da je metoda dokaj robustna in da tu ni prisoten primer prevelikega prileganja. Pri ROC analizi (Slika 4) vidimo, podobno kot pri kontingenčni tabeli, da se krivulji le malenkost razlikujeta. Pri podatkih učenja je bila AUC površina 76,9 %, pri testnih pa 73,8 %.

Slika 4: ROC krivulji logistične regresije z minimalnim naborom spremenljivk





Če pogledamo dobljeno logit transformacijo (PRILOGA), vidimo, da imajo v modelu največjo moč slamnate opisne spremenljivke in to tiste, pri katerih se določene vrednosti pojavijo zelo redko:

1. slovaško državljanstvo (Exp(β) ≈ 0; zelo negativno vpliva na odločitev);
2. albansko državljanstvo (Exp(β) ≈ 0; zelo negativno vpliva na odločitev);
3. bahrajnsko državljanstvo (Exp(β) ≈ 0; zelo negativno vpliva na odločitev);
4. znamka vozila SMART (Exp(β) ≈ 0; zelo negativno vpliva na odločitev);
5. znamka vozila CRZ (Exp(β) ≈ 0; zelo negativno vpliva na odločitev);
6. znamka vozila JEEP (Exp(β) = 3701018269; zelo pozitivno vpliva na odločitev);
7. …

Šele na 33. in 34. mestu se pojavita prvi bolj pogosti vrednosti »Zaposlitev za nedoločen čas« ter »Zaposlitev za določen čas », kjer po pričakovanjih prva na končni rezultat vpliva pozitivno in druga negativno.

Zanimivo je, da ima spremenljivka »Razpoložljiv osebni dohodek z malico in prevozom« (v nadaljevanju Razpoložljiv OD) zelo majhen vpliv na končni rezultat. Exp(β) te spremenljivke je v tem modelu 1,001.

3.1.2 Minimalni nabor z reduciranim številom slamnatih spremenljivk

Pri testiranju logistične regresije smo poleg originalnega nabora kriterijev, zaradi zelo velikega števila slamnatih spremenljivk, preverili še možnost, ali lahko z reduciranjem njihovega števila izboljšamo model. V testu smo podatkom spremenili znamke vozila na vrednost Ostalo vsem, ki predstavljajo manj kot 5 odstotkov vzorca. Na ta način smo zreducirali nabor vrednosti kriterija na končnih 7 vrednosti, kot je razvidno v Tabeli 4.

Tabela 4: Spremenjen nabor vrednosti kriterija Znamka

Znamka	Frekvenca	Delež
BMW	388	5,1
FORD	1086	14,3
HYUNDAI	550	7,3
KIA	1027	13,6
OPEL	708	9,3
OSTALO	3122	41,2
RENAULT	697	9,2
Skupaj	**7578**	**100,0**

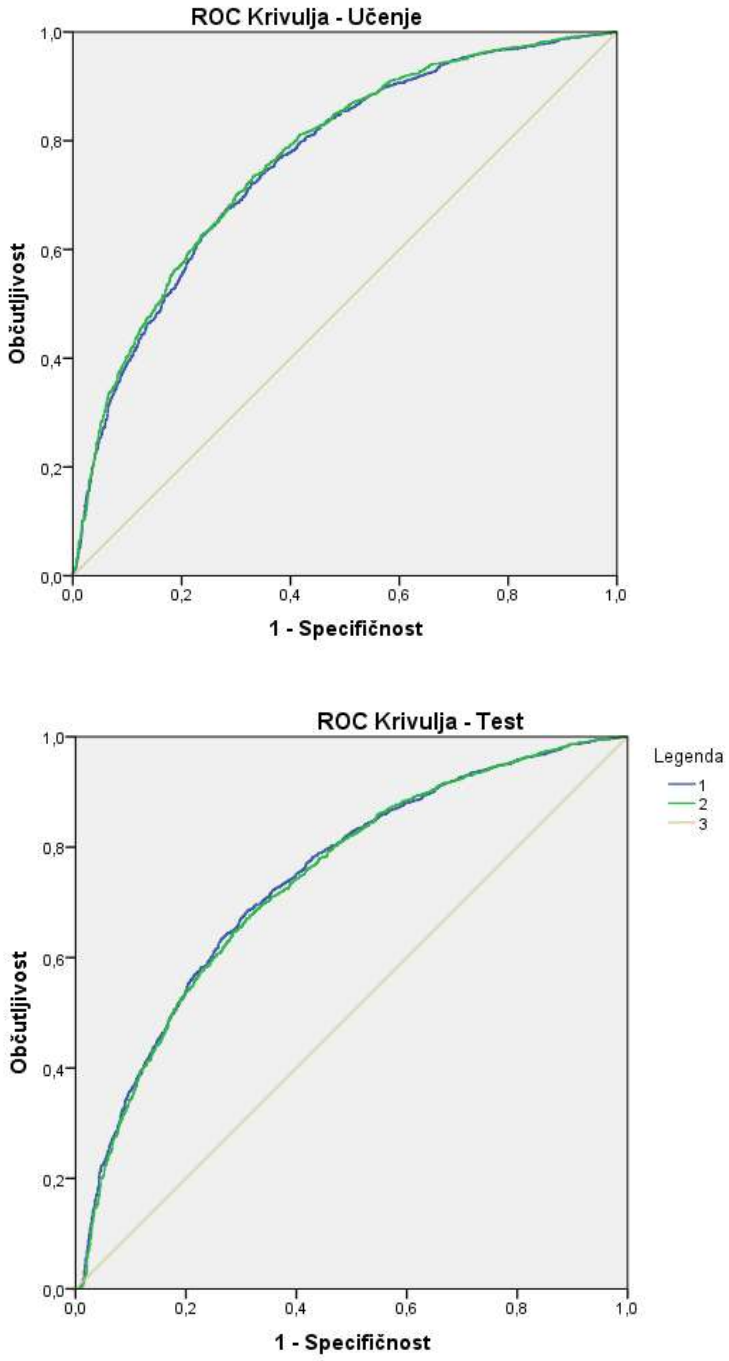
Rezultati v kontingenčni tabeli (Tabela 5) se od osnovnega modela razlikujejo samo za malenkost pri podatkih učenja. Pri podatkih testiranja opazne razlike ni.

Tabela 5: Kontingenčna tabela logistične regresije z reduciranim številom slamnatih spremenljivk

		Napoved					
		Podatki učenja			**Podatki testiranja**		
		Zavrnitev	Odobritev	%	Zavrnitev	Odobritev	%
Opazovanja	Zavrnitev	704	783	47,3	673	815	45,2
	Odobritev	303	1993	86,8	340	1962	85,2
				71,3			69,5

Če primerjamo rezultate ROC analize, vidimo (Slika 5), da je s to spremembo vrednost AUC krivulje malenkost nižja pri podatkih učenja (76,3 %), je pa v primeru testiranja prišlo do malenkostnega izboljšanja, saj je AUC površina 74,2 %.

Slika 5: Primerjava ROC krivulj osnovne in zreducirane logistične regresije



Legenda:

1 – LR minimalni nabor z reduciranimi znamkami vozil

2 – LR minimalni nabor

3 – Referenčna črta

3.1.3 Razdelitev podatkov na štiri vsebinsko ločene podmnožice

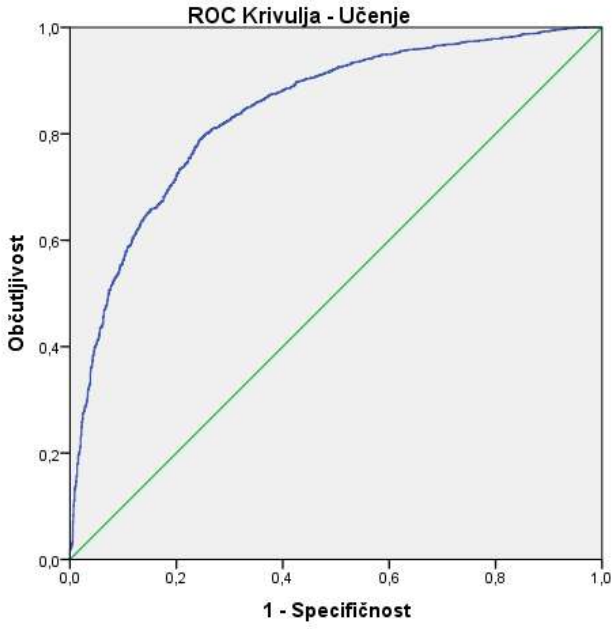
V tem testu smo podatke vsebinsko razdelili na štiri podmnožice in s tem izkoristili vse možne razpoložljive pojasnjevalne spremenljivke, ne da bi za to izgubili razpoložljive podatke za učenje. Model, ki smo ga dobili, je sestavljen iz štirih logit transformacij, kjer vsaka ustreza točno določeni podskupini enot ocenjevanja (nova stranka, obstoječa stranka ...). V Tabeli 6 so prikazani rezultati modela logistične regresije, kjer vidimo, da smo s to optimizacijo nekoliko izboljšali rezultate.

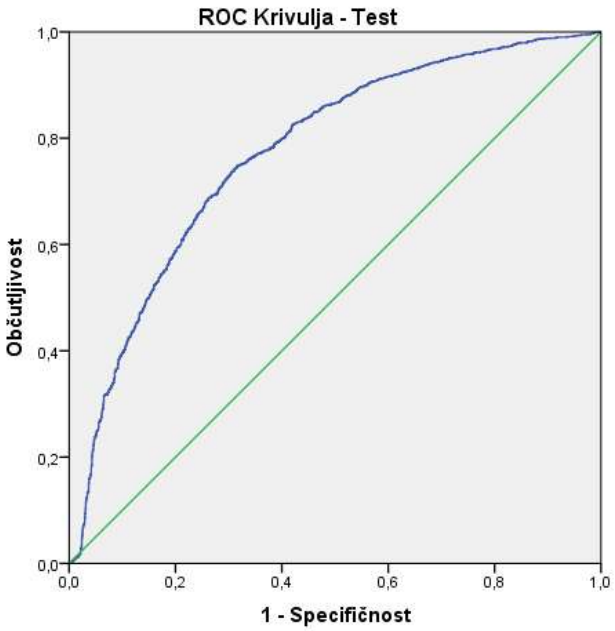
Tabela 6: Kontingenčna tabela logistične regresije nad podmnožicami

		Napoved					
		Podatki učenja			**Podatki testiranja**		
		Zavrnitev	Odobritev	%	Zavrnitev	Odobritev	%
Opazovanja	Zavrnitev	891	596	59,9	794	672	54,1
	Odobritev	270	2026	88,2	354	1937	84,5
				77,1			72,6

Če pogledamo rezultat ROC analize (Slika 6), vidimo, da so rezultati pri optimizaciji nekoliko izboljšani. Dodatno lahko vidimo, da je učinek optimizacije bolj izrazit pri podatkih učenja, saj je AUC površina povečana na 84,1 %. Pri testnih podatkih je AUC površina 77,2 %.

Slika 6: ROC krivulji logistične regresije nad podmnožicami





Strukture logit transformacij so podobne kot v minimalnem naboru spremenljivk in se razlikujejo le v pojasnjevalnih spremenljivkah, ki so dodane. To so spremenljivke, kot so indeks plačilne discipline, trenutna izpostavljenost stranke in podobno in nimajo večjega vpliva na rezultat, saj skupaj klasifikacijsko točnost povečajo le za 3,2 %.

3.1.4 Logistična regresija v kombinaciji z odločitvenimi drevesi

Pri tem testu smo v prvem koraku kreirali dve odločitveni drevesi (CHAID 100/50 in CART 100/50) in rezultat verjetnosti klasifikacije obeh metod shranili v novi spremenljivki (CHAID_100 in CRT_100). Logistično regresijo smo izvedli na enak način kot pri prvem testu, le da smo minimalnemu naboru spremenljivk dodali novi pojasnjevalni spremenljivki.

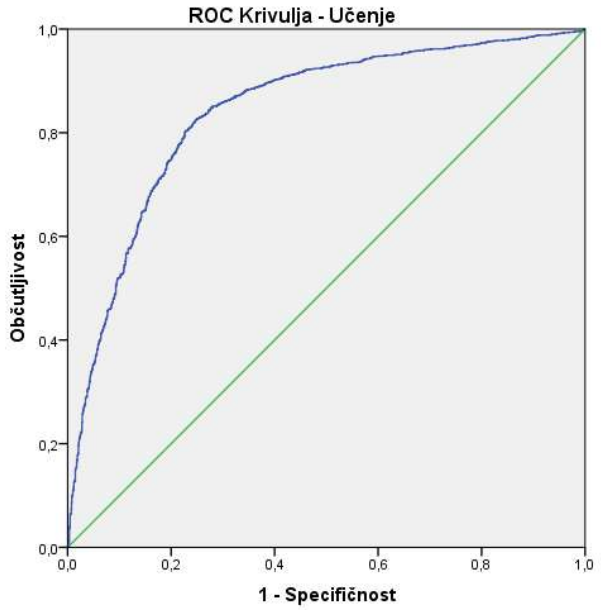
Rezultati testa so, kot vidimo v Tabeli 7, še boljši od predhodnih logističnih regresij in celo boljši od vseh treh metod, ki so vključene v model (Logistična regresija minimalni nabor, CHAID 100/50 in CART 100/50).

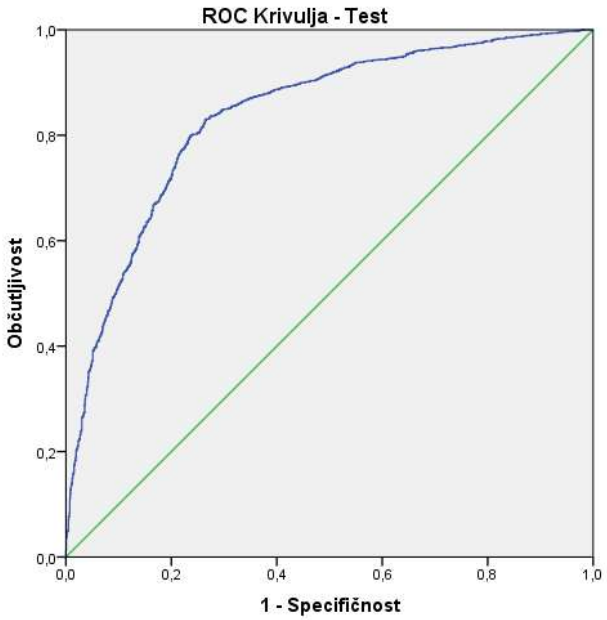
Tabela 7: Kontingenčna tabela logistične regresije s CHAID in CART metodi

		Napoved					
		Podatki učenja			**Podatki testiranja**		
		Zavrnitev	Odobritev	%	Zavrnitev	Odobritev	%
Opazovanja	Zavrnitev	1063	424	71,5	1031	453	69,5
	Odobritev	312	1984	86,4	370	1930	83,9
				80,5			78,3

Na sliki 7 sta prikazani krivulji ROC analize, kjer je AUC površina v primeru učenja 86,1 % ter v primeru testiranja 81,5 %.

Slika 7: ROC krivulji logistične regresije v kombinaciji s CHAID in CART metodo



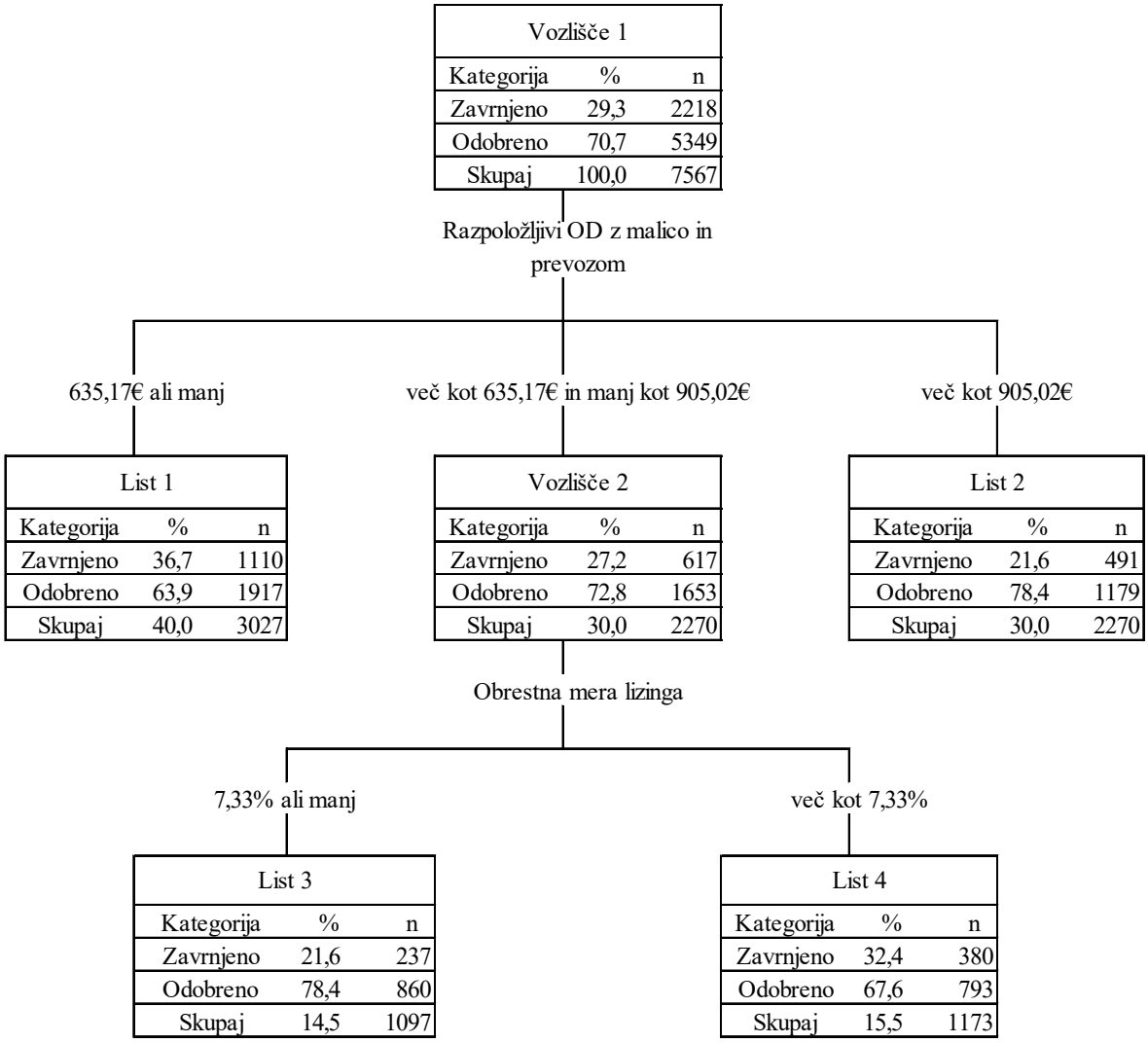


Če pogledamo Logit transformacijo, vidimo (PRILOGA 3), da imamo tokrat med pomembnejšimi spremenljivkami poleg številnih slamnatih spremenljivk tudi novi dve, in sicer na 23. mestu CRT_100 (Exp(β) = 42,767) ter na 27. mestu CHAID_100 z (Exp(β) = 5,483). Vseh spremenljivk v modelu je 87.

3.1.5 Primer »Antipov & Pokryshevskaya« – kombinacija logistične regresije s CHAID metodo

Za ta test smo v prvem koraku izvedli več CHAID odločitvenih dreves, tako da smo s spreminjanjem nastavitev (minimalno število enot v vozliščih/minimalno število enot v končnih vozliščih) iskali takšno odločitveno drevo, ki bo imelo točno štiri končne liste. To smo dosegli z nastavitvijo CHAID (1500/1000) – model lahko vidimo na Sliki 8.

Slika 8: CHAID odločitveno drevo 1500/1000



Iz modela lahko razberemo, da je v tem primeru glavni delitveni kriterij »Razpoložljiv OD« z naslednjimi mejami:

- 635,17 € ≤ Razpoložljiv OD;
- 635,17 € < Razpoložljiv OD ≤ 905,02 €;
- 905,02 € < Razpoložljiv OD.

Drugi razred Razpoložljivega OD se še dodatno deli na primere, ko je obrestna mera posojila manjša oziroma večja od 7,33 %.

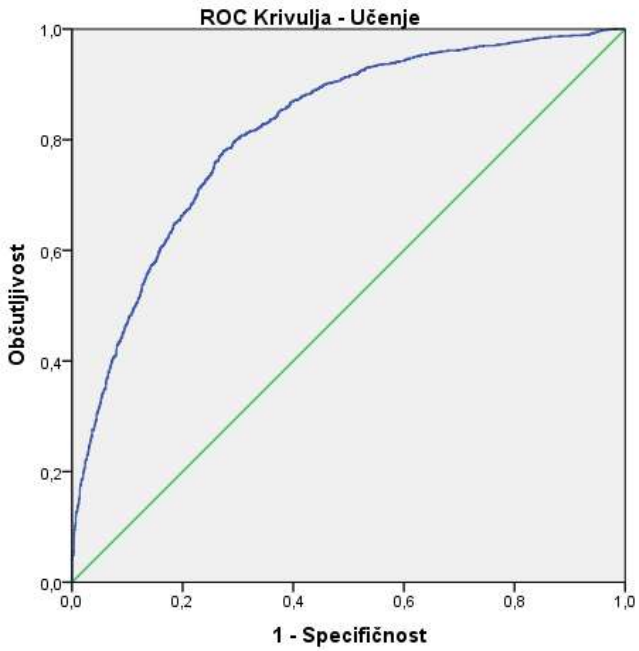
Na podlagi delitvenih kriterijev dobljenega CHAID modela smo v nadaljevanju osnovno množico primerov razdelili na štiri podmnožice, ki smo jih na koncu ločeno obdelali z logistično regresijo. Skupni rezultat klasifikacijske točnosti lahko vidimo v kontingenčni tabeli (Tabela 8).

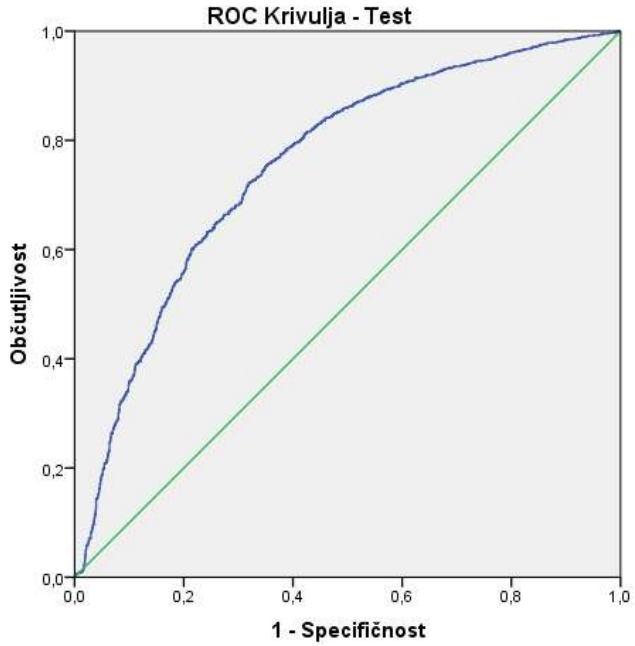
Tabela 8: Kontingenčna tabela logistične regresije v kombinaciji s CHAID 1500/1000

		Napoved					
		Podatki učenja			**Podatki testiranja**		
		Zavrnitev	Odobritev	%	Zavrnitev	Odobritev	%
Opazovanja	Zavrnitev	926	561	62,2	814	655	55,4
	Odobritev	308	1988	86,5	396	1896	82,7
				77,0			72,1

Na sliki 9 sta prikazani krivulji ROC analize, kjer je AUC površina pri podatkih učenja 82,5 % ter pri podatkih testiranja 74,8 %.

Slika 9: ROC krivulji logistične regresije v kombinaciji s CHAID 1500/1000





Vidimo, da je v tem primeru klasifikacijska točnost sicer za 2,3 % boljša od izhodiščne logistične regresije, vendar manj, kot je bila v primeru delitve na nove/obstoječe stranke in z/brez podatkov delodajalca. Log transformacije v tem primeru nimajo posebnosti (enako kot za ostale logistične regresije) in tudi tukaj prednjačijo slamnate spremenljivke opisnih kriterijev, predvsem z manj frekventnimi vrednostmi.

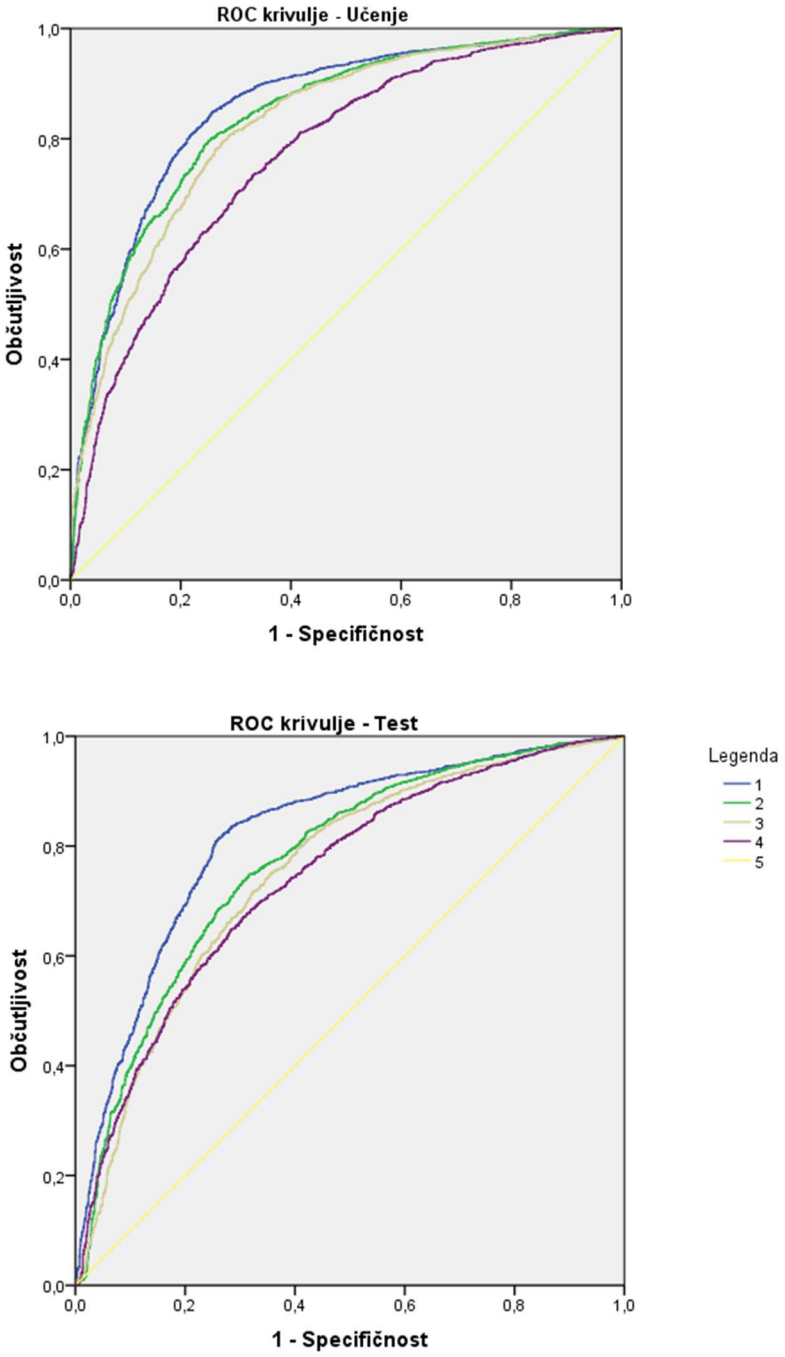
Če primerjamo rezultate vseh petih metod logistične regresije, vidimo (Tabela 9), da je k izboljšanju najbolj pripomogla tretja metoda, kjer smo pri učenju upoštevali rezultate CHAID in CART metod. Zanimivo je, da je v tem primeru rezultat celo boljši od ločenih metod CHAID 100/50 in CART 100/50. Torej model ni samo prevzel del pojasnjene variance modelov, ki sta v osnovi boljša, temveč je na nek način njuno znanje nadgradil v kombinaciji z metodo logistične regresije.

Tabela 9: Primerjava modelov Logistične regresije

	Kontingenčna tabela		**ROC krivulja**	
	Točnost		AUC (%)	
Metoda	Učenje	Test	Učenje	Test
LR + CHAID 100/50 + CART 100/50	80,5	78,3	86,1	**81,5**
LR podmnožice (nova/obstoječa stranka ...)	77,1	72,7	84,1	**77,2**
LR podmnožice glede na CHAID 1500/1000	77,0	72,1	82,5	**74,8**
LR z reduciranim številom slamnatih sprem.	71,3	69,5	76,3	**74,2**
LR – minimalni nabor	71,9	69,5	76,9	**73,8**

Na sliki 10 lahko vidimo ROC krivulje vseh štirih izpeljank logistične regresije.

Slika 10: ROC krivulje logističnih regresij učnih in testnih primerov



Legenda:

1 – Logistična regresija + CHAID 100/50 + CART 100/50

2 – Logistična regresija podmnožice združeno

3 – Primer Antipov & Pokryshevskaya

4 – Logistična regresija minimalni nabor

5 – Referenčna črta

3.2 Odločitvena drevesa

Pri odločitvenih drevesih smo testirali obe metodi (CHAID, CART) v treh različnih razmejenostih, tj. najmanjši (100/50), srednji (20/10) ter največji (1/1), z nastavitvijo, da se manjkajoče vrednosti obravnavajo kot posebna vrednost. Vse opravljene teste smo ponovili še z upoštevanjem dodatne pojasnjevalne spremenljivke rezultata verjetnosti klasifikacije pri logistični regresiji z minimalnim naborom kriterijev (pripona +LR) ter s pretvorbo znamk izjem v kategorijo Ostalo (pripona –ZI).

Tabela 10: Glavne karakteristike dobljenih modelov

Metoda	Število vseh vozlišč/končnih vozlišč	št. nivojev	Glavni delitveni kriteriji
CHAID 100/50	29/20	3	OD, POLOG, INDEKS, EUROTAX
CART 100/50	23/12	3	INDEKS, OD, EUROTAX,ZNAMKA
CHAID 20/10	124/76	10	OD, POLOG, INDEKS, EUROTAX
CART 20/10	119/60	15	INDEKS, OD, EUROTAX,ZNAMKA
CHAID 1/1	342/196	15	OD, PROC. KAP, POLOG, INDEKS
CART 1/1	941/471	27	INDEKS, OD, EUROTAX,ZNAMKA
CHAID 100/50 - ZI	28/19	3	OD, POLOG, LETNIK, EUROTAX
CART 100/50 - ZI	17/9	5	OD, EUROTAX, POLOG, INDEKS
CHAID 20/10 - ZI	120/72	9	OD, INDEKS, EUROTAX, POLOG
CART 20/10 - ZI	93/47	15	OD, EUROTAX, OD, POLOG
CHAID 1/1 - ZI	363/206	14	OD, PROC. KAP, POLOG, INDEKS
CART 1/1 - ZI	1009/505	34	OD, EUROTAX, POLOG, INDEKS
CHAID 100/50 + LR	38/24	3	RLR, POLOG, EUROTAX, OD
CART 100/50 + LR	19/10	5	RLR, OD, POŠTA, INDEX
CHAID 20/10 + LR	119/71	9	RLR, POLOG, INDEKS, OD
CART 20/10 + LR	91/46	13	RLR, OD, POŠTA, INDEX
CHAID 1/1 + LR	260/145	13	RLR, POLOG, EUROTAX, OD
CART 1/1 + LR	1031/516	29	RLR, INDEKS, POLOG, OD

V Tabeli 10 so prikazane glavne karakteristike odločitvenih dreves, kjer so prikazani velikost v številu vozlišč in nivojev ter prvi štirje delitveni kriteriji dreves. Med delitvenimi kriteriji vidimo, da se na prvih štirih mestih pojavljajo iste vrednosti, in sicer Razpoložljivi OD, INDEKS (Indeks plačilne nediscipline), POLOG (odstotek pologa), EUROTAX (relativna razlika prodajne cene od Eurotax cene), ZNAMKA (znamka vozila), LETNIK (letnik vozila), PROC. KAP (odstotek kapitala v bilanci delodajalca) in v primeru kombiniranega modela vedno na prvem mestu rezultat verjetnosti klasifikacije pri logistični regresiji (RLR).

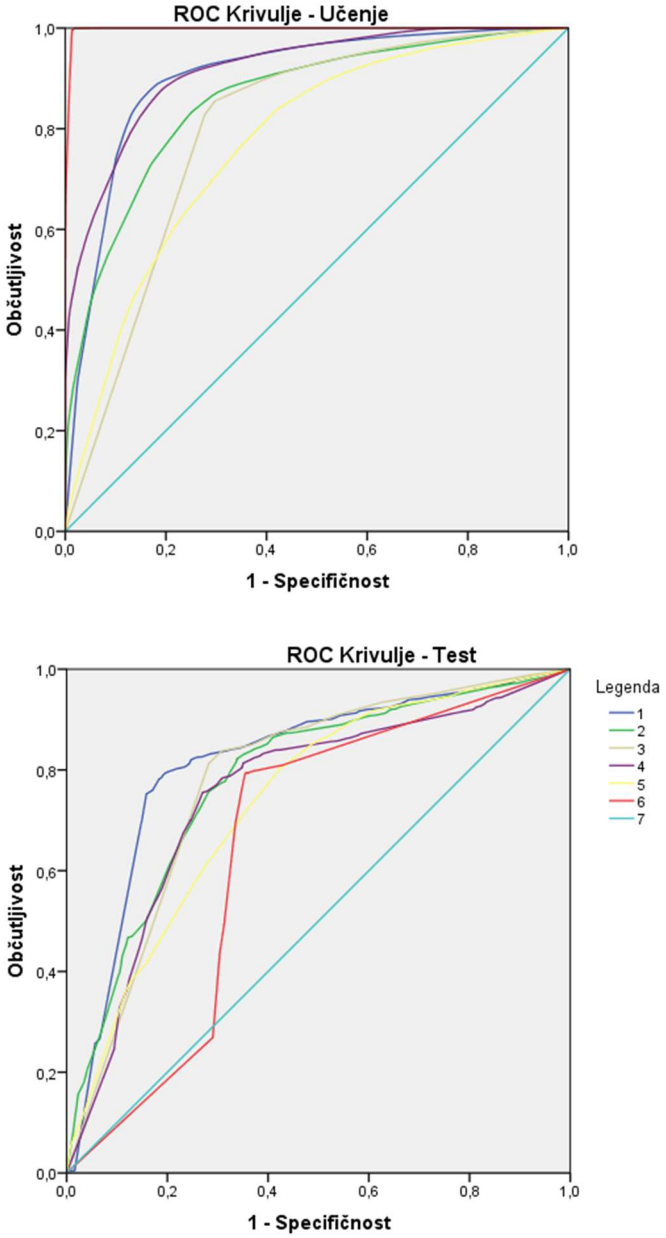
Tabela 11: Rezultati modelov odločitvenih dreves

	Kontingenčna tabela		ROC krivulja		
	Točnost		AUC		
Metoda	Učenje	Test	Učenje	Test	Razlika %
CART 20/10 – ZI	84,6	78,7	88,70	**82,60**	7,4
CART 20/10	86,1	78,9	90,30	**81,60**	10,7
CART 20/10 + LR	83,8	77,1	88,20	**79,70**	10,7
CHAID 20/10 + LR	79,6	73,1	86,80	**78,20**	11,0
CHAID 20/10	80,4	75,9	86,00	**78,20**	10,0
CHAID 20/10 – ZI	80,4	75,9	85,20	**77,70**	9,7
CART 100/50	79,3	78,0	80,00	**77,90**	2,7
CART 100/50 – ZI	79,3	78,0	78,60	**77,00**	2,1
CART 100/50 + LR	79,3	76,7	81,30	**77,00**	5,6
CHAID 1/1 + LR	83,7	72,9	91,10	**76,60**	18,9
CHAID 100/50 + LR	76,4	73,1	82,00	**76,00**	7,9
CHAID 1/1 – ZI	85,1	74,1	92,10	**75,40**	22,1
CHAID 1/1	85,1	74,4	91,80	**75,20**	22,1
CHAID 100/50	73,6	71,4	77,30	**73,90**	4,6
CHAID 100/50 – ZI	73,5	71,9	76,60	**73,80**	3,8
CART 1/1 – ZI	99,2	73,5	99,70	**66,30**	50,4
CART 1/1	99,2	73,5	99,80	**65,20**	53,1
CART 1/1 + LR	99,6	72,1	99,90	**65,20**	53,2

V Tabeli 11 je prikazana primerjava vseh variant metode odločitvenih dreves, ki smo jih testirali. Za naše podatke so najboljši rezultati pri srednjem CART modelu z minimalnim številom enot v vozliščih 20 in minimalnim številom enot v končnih vozliščih 10 (20/10) z reduciranim številom različnih znamk vozila ter najslabši pri najbolj razvejenih CART modelih. V zadnjem stolpcu je podatek o razliki v odstotkih med AUC površino na učnih in testnih primerih, kar predstavlja mero prevelikega prileganja. Kot vidimo, je najbolj prisotno pri večjih odločitvenih drevesih, saj je pokritost na primerih učenja najboljša (99,9 %) in hkrati najslabša (65,2 %) pri testnih primerih. Razlika je v tem primeru nad 50 %. Dodatno lahko vidimo, da so pri CART modelih rezultati kombiniranih metod slabši, kot bi bili brez upoštevanja rezultatov logistične regresije. Pri CHAID modelih je ta vpliv obraten. Če gledamo primere z reduciranim številom različnih znamk vozila, vidimo, da je vpliv spremembe pri CART modelih pozitiven, pri CHAID pa negativen.

Iz ROC krivulj (Slika 11) je razvidno, da je od osnovnih metod CART (1/1) najboljša na podatkih učenja in hkrati najslabša na testnih podatkih. V določenih primerih se izkaže celo slabše od strategije naključnega ugibanja, saj je na grafu pod referenčno linijo.

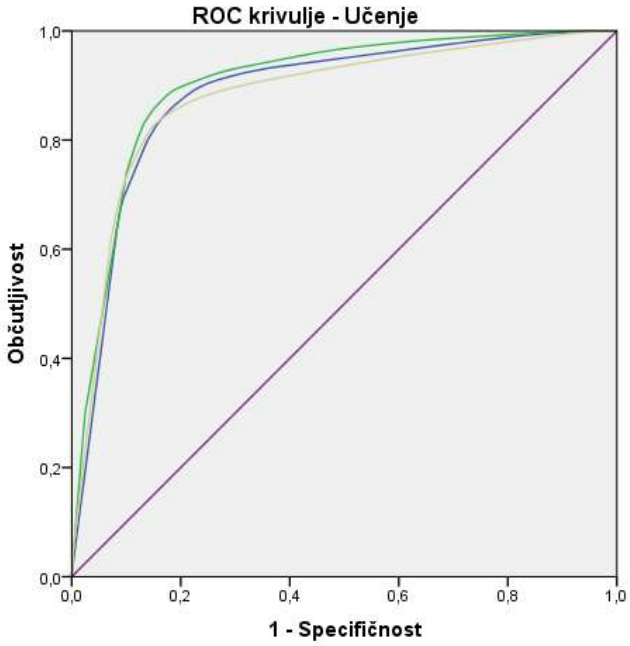
Slika 11: ROC krivulje osnovnih odločitvenih dreves

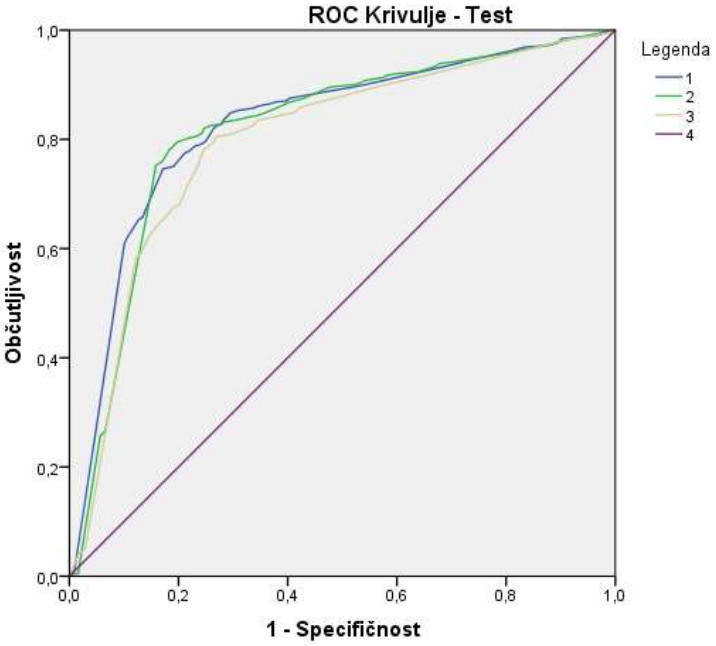


Legenda:
1 – CART 20/10
2 – CHAID 20/10
3 – CART 100/50
4 – CHAID 1/1
5 – CHAID 100/50
6 – CART 1/1
7 – Referenčna črta

Na sliki 12 vidimo primerjavo vseh treh ROC krivulj CART (20/10), ki so od vseh odločitvenih dreves najbolj učinkovite pri napovedi.

Slika 12: ROC krivulje treh izpeljank CART (20/10) odločitvenih dreves





Legenda:
1 – CART 20/10 – ZI
2 – CART 20/10
3 – CART 20/10 + LR
4 – Referenčna črta

3.3 Naključni gozdovi

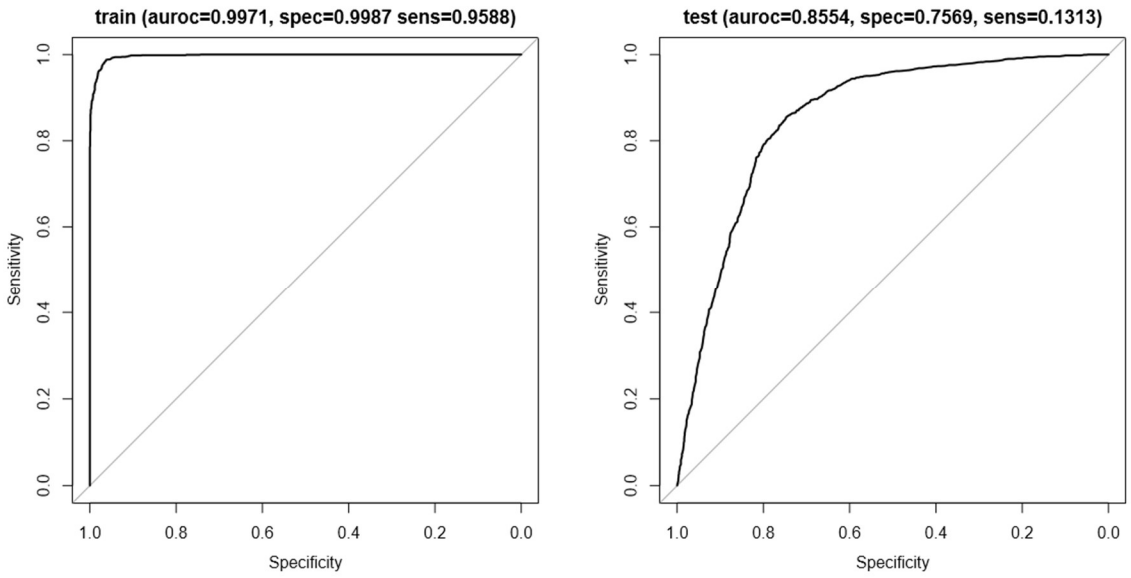
Pri primerjavi vseh osmih variant naključnih gozdov (Tabela 12) lahko vidimo, da so rezultati malenkost boljši, če smo pri znamkah vozil izjeme prekvalificirali v kategorijo Ostalo. Iz rezultatov je tudi razvidno, da modeli, ki smo jih kombinirali z logistično regresijo, dajo boljše rezultate v fazi učenja od osnovnega, kar pa ne velja pri testnih podatkih, saj so tu rezultat slabši. Z dodajanjem teh spremenljivk se torej ni povečala učinkovitost modela, temveč smo samo povzročili preveliko prileganje. Dodatno vidimo, da je učinek spremembe odvisen od modela logistične regresije. Modeli, ki imajo v osnovi boljšo AUC pokritost, povzročijo večjo razliko v rezultatu.

Tabela 12: Rezultati modelov Naključnih gozdov

	Kontingenčna tabela		**ROC krivulja**		
	Točnost		AUC		
Metoda	Učenje	Test	Učenje	Test	Razlika %
Naključni gozdovi – ZI	99,8	75,7	99,71	**85,54**	16,56
Naključni gozdovi	97,2	74,6	97,71	**85,04**	14,90
Naključni gozdovi + LR – ZI	99,9	73,8	99,80	**85,02**	17,38
Naključni gozdovi + LR	98,6	73,0	98,49	**84,68**	16,31
Naključni gozdovi + LR1234 – ZI	100,0	73,1	99,97	**84,75**	17,95
Naključni gozdovi + LR1234	98,9	71,3	98,85	**84,46**	17,04
Naključni gozdovi + LR + LR1234 -ZI	100,0	72,8	99,98	**84,60**	18,17
Naključni gozdovi + LR + LR1234	98,9	71,7	99,13	**84,57**	17,22

Na Sliki 13 sta prikazani ROC krivulji osnovne metode naključnih gozdov brez kombinacije z metodami logistične regresije na podatkih s popravkom znamke izjem v kategorijo Ostalo (–ZI). Vidimo, da je pokritost na podatkih učenja zelo visoka (97,71 %). Pri testnih podatkih je nekoliko nižja (85,54 %), vendar še vedno najvišja od vseh testiranj. Ostale ROC krivulje naključnih gozdov so prikazane v Prilogi 9.

Slika 13: ROC krivulji metode Naključnih gozdov – ZI

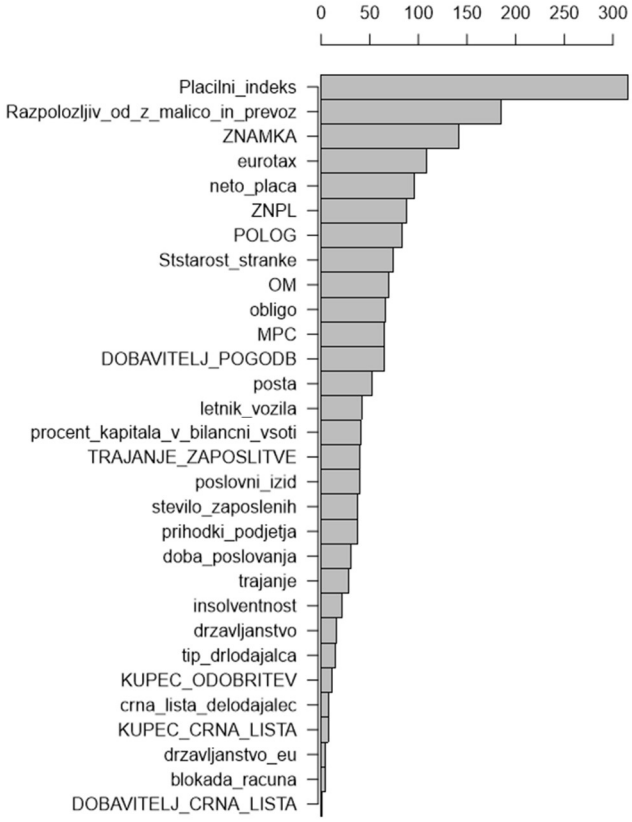


Legenda: Sensitivity – občutljivost; Specificity – specifičnost

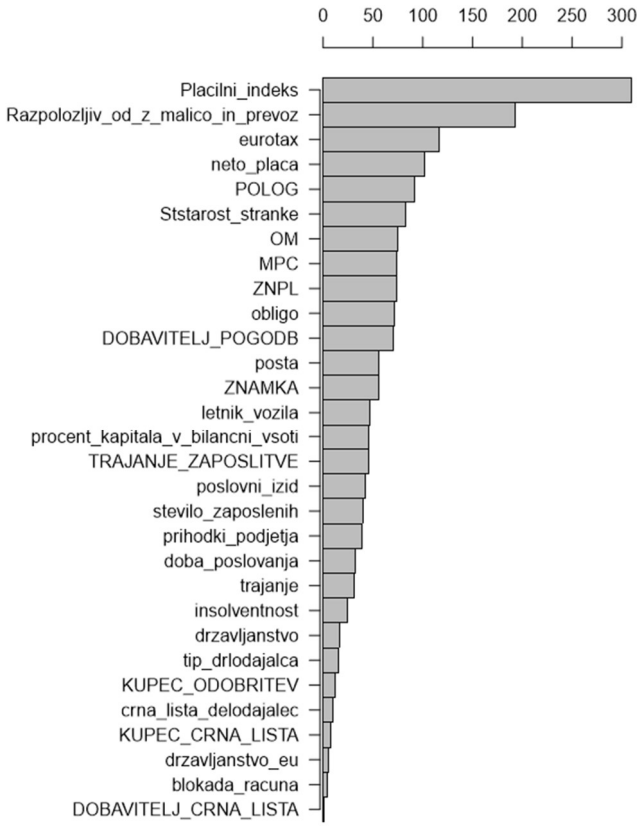
Pri analizi mere pomembnosti atributov modela lahko vidimo (Slika 14), da je indeks plačilne nediscipline med vsemi najbolj pomemben, na drugem mestu je razpoložljivi OD, znamka vozila je na tretjem in na četrtem relativna razlika prodajne cene z eurotax ceno. Kot vidimo, je seznam zelo podoben seznamom, ki smo jih dobivali pri testiranju odločitvenih dreves.

Če primerjamo najpomembnejše spremenljivke modela, kjer so popravljene znamke vozil pri izjemah v kategorijo Ostalo (Slika 15), lahko vidimo, da je s tem pomembnost kriterija Znamka padla iz tretjega na trinajsto mesto. Kljub temu, da smo s tem v podatkih izgubili nekaj informacij, so modeli naključnih gozdov vseeno pridobili na učinkovitosti napovedi.

Slika 14: Rezultat analize najpomembnejših spremenljivk modela z vsemi znamkami vozil



Slika 15: Rezultat analize najpomembnejših spremenljivk modela brez znamk



3.4 Primerjava vseh testiranih metod

Če primerjamo klasifikacijske točnosti vseh testiranih metod na podatkih testiranja (Tabela 13 in Tabela 14), vidimo, da so na naših konkretnih podatkih najbolj točne metode srednje razvejenih ter manj razvejenih CART odločitvenih dreves. Med vsemi se pokaže kot najboljša CART 20/10 metoda. V tabeli je še razvidno, da so razlike rezultatov minimalne, saj se vrednosti najboljših in najslabših rezultatov razlikujejo za manj kot 10%. Najslabše rezultate smo dobili pri logističnih regresijah z minimalnem naborom pojasnjevalnih spremenljivk.

Iz dobljenih rezultatov vidimo tudi, da smo s kombiniranjem algoritmov rezultate lahko nekoliko izboljšali, vendar ne v vseh primerih. Pri logistični regresiji je učinek zelo izrazit in pozitiven, saj se je rezultat povečal za 8,8 % in to z zmanjšanjem prevelikega prileganja, smo pa pri naključnih gozdovih s kombiniranjem z ostalimi algoritmi učinkovitost celo nekoliko poslabšali. Pri odločitvenih drevesih izrazite spremembe ni videti.

V zadnjem stolpcu rezultatov, kjer smo izračunali razmerje med klasifikacijsko točnostjo učenja in testiranja (relativno povečanje v %), vidimo, da so stopnje prevelikega prileganja različne, in sicer:

- najmanj izrazite pri manj razvejenih odločitvenih drevesih in logističnih regresijah (pod 10 %);
- nekoliko bolj izrazite pri bolj razvejenih CHAID odločitvenih drevesih (okrog 15 %);
- najbolj izrazite pri najbolj razvejenih CART odločitvenih drevesih ter naključnih gozdovih (od 30 do 40 %).

Tabela 13: Primerjava klasifikacijske točnosti vseh testiranih metod

Metoda	**Učenje**	**Test**	**Razlika %**
CART 20/10	86,1	**78,9**	9,13
CART 20/10 - ZI	84,6	**78,7**	7,50
LR - minimalni nabor + CHAID + CART	80,5	**78,3**	2,81
CART 100/50	79,3	**78,0**	1,67
CART 100/50 + LR	79,3	**78,0**	1,67
CART 20/10 + LR	83,8	**77,1**	8,69
CART 100/50 - ZI	79,3	**76,7**	3,39
CHAID 20/10	80,4	**75,9**	5,93
CHAID 20/10 – ZI	80,4	**75,9**	5,93
Naključni gozdovi - ZI	99,8	**75,7**	31,84
Naključni gozdovi	97,2	**74,6**	30,29
CHAID 1/1 – ZI	85,1	**74,4**	14,38

se nadaljuje

Tabela 14: Primerjava rezultatov kontingenčnih tabel vseh testiranih metod (nad.)

Metoda	**Učenje**	**Test**	**Razlika %**
CHAID 1/1	85,1	**74,4**	14,38
Naključni gozdovi + LR - ZI	99,9	**73,8**	35,37
CART 1/1 – ZI	99,2	**73,5**	34,97
CART 1/1	99,2	**73,5**	34,97
Naključni gozdovi + LR1234 - ZI	100,0	**73,1**	36,80
CHAID 20/10 + LR	79,6	**73,1**	8,89
CHAID 100/50 + LR	76,4	**73,1**	4,51
Naključni gozdovi + LR	98,6	**73,0**	35,07
CHAID 1/1 + LR	83,7	**72,9**	14,81
Naključni gozdovi + LR + LR1234 - ZI	100,0	**72,8**	37,36
LR podmnožice (nova/obstoječa stranka,..)	77,1	**72,7**	6,05
LR podmnožice glede na CHAID 1500/1000	77,0	**72,1**	6,80
CART 1/1 + LR	99,6	**72,1**	38,14
CHAID 100/50 - ZI	73,5	**71,9**	2,23
Naključni gozdovi + LR + LR1234	98,9	**71,7**	37,94
CHAID 100/50	73,6	**71,4**	3,08
Naključni gozdovi + LR	98,9	**71,3**	38,71
LR – minimalni nabor - ZI	71,3	**69,5**	2,59
LR - minimalni nabor	71,9	**69,5**	3,45

Če pogledamo primerjavo AUC na podatkih testiranja (Tabela 15), vidimo, da so od vseh testiranih metod najbolj učinkoviti naključni gozdovi. Med njimi se kot najboljša obnese osnovna, torej metoda naključnih gozdov brez kombinacije z algoritmi logistične regresije s popravkom znamke izjem v kategorijo Ostalo. Nekoliko slabše rezultate od naključnih gozdov smo dobili pri srednje razvejenih odločitvenih drevesih, sledijo logistične regresije in najbolj razvejena odločitvena drevesa. Najslabše rezultate smo dobili pri najbolj razvejenem CART odločitvenim drevesu (CART 1/1) v kombinaciji z logistično regresijo.

Podobno kot pri primerjavi klasifikacijskih točnosti, se tudi pri ROC analizi vseh testiranih metod pokaže izboljšava logistične regresije v kombinaciji z drugimi metodami. Rezultat se je izboljšal za 7,7 % ob minimalni spremembi prevelikega prileganja. Enako se pri naključnih gozdovih s kombinacijo z ostalimi algoritmi učinkovitost nekoliko poslabša in pri odločitvenih drevesih bistveno ne spremeni.

Pri izračunu razmerja med AUC površino učenja in testiranja dobimo naslednje stopnje prevelikega prileganja:

- najmanj izrazite pri manj razvejenih odločitvenih drevesih in logističnih regresijah (pod 10 %);

- nekoliko bolj izrazite pri metodah naključnih gozdov ter srednje razvejenih odločitvenih drevesih (od 10 do 20 %);
- najbolj izrazite pri najbolj razvejenih odločitvenih drevesih (do 55 %).

Tabela 15: Primerjava rezultatov ROC analize vseh testiranih metod

Metoda	**Učenje**	**Test**	**Razlika**
Naključni gozdovi - ZI	99,71	**85,54**	16,56
Naključni gozdovi	97,71	**85,04**	14,90
Naključni gozdovi + LR - ZI	99,80	**85,02**	17,38
Naključni gozdovi + LR	98,49	**84,68**	16,31
Naključni gozdovi + LR1234 - ZI	98,97	**84,76**	17,94
Naključni gozdovi + LR + LR1234 - ZI	99,98	**84,6**	18,17
Naključni gozdovi + LR1234	98,85	**84,46**	17,04
Naključni gozdovi + LR + LR1234	99,13	**84,57**	17,22
CART 20/10 - ZI	88,7	**82,6**	7,38
CART 20/10	90,3	**81,6**	10,66
LR - minimalni nabor + CHAID + CART	86,1	**81,5**	5,64
CART 20/10 + LR	88,2	**79,7**	10,66
CHAID 20/10 + LR	86,8	**78,2**	11,00
CHAID 20/10	86	**78,2**	10,00
CART 100/50	80	**77,9**	2,70
CHAID 20/10 – ZI	85,2	**77,7**	9,65
LR podmnožice (nova/obstoječa stranka,..)	84,1	**77,2**	8,94
CART 100/50 - ZI	81,3	**77**	5,58
CART 100/50 + LR	78,6	**77**	2,07
CHAID 1/1 + LR	91,1	**76,6**	18,93
CHAID 1/1 – ZI	92,1	**75,4**	22,14
CHAID 1/1	91,8	**75,2**	22,07
LR podmnožice glede na CHAID 1500/1000	82,5	**74,8**	10,29
LR – minimalni nabor - ZI	76,3	**74,2**	10,26
CHAID 100/50	77,3	**73,9**	4,60
CHAID 100/50 - ZI	76,6	**73,8**	3,79
LR - minimalni nabor	76,9	**73,8**	4,20
CHAID 100/50 + LR	73	**69,5**	5,04
CART 1/1 – ZI	99,7	**66,3**	50,37
CART 1/1	99,8	**65,2**	53,07
CART 1/1 + LR	99,9	**65,2**	53,22

Iz Tabele 15 lahko tudi razberemo, da so najboljši rezultati pri podatkih učenja predvsem pri najbolj razvejenih odločitvenih drevesih in naključnih gozdovih. Oboji dosegajo AUC pokritost več kot 99 %. Pri podatkih testa vidimo še, da se pri odločitvenih drevesih CART (1/1) rezultati najbolj razlikujejo od rezultatov učenja, torej je tu najbolj izrazito preveliko prileganje.

Potrebni časi za kreiranje odločitvenega modela so bili pri metodah naslednji:

- do nekaj sekund pri logističnih regresijah,
- od nekaj sekund do 5 minut pri odločitvenih drevesih,
- od 10 do 20 minut pri metodah naključnih gozdov.

4 RAZPRAVA

Podobno kot v raziskavi Anne Kraus (2014), smo tudi mi dobili najboljše rezultate prav pri naključnih gozdovih. Če jih primerjamo z odločitvenimi drevesi CART (1/1), ki so njihova osnova, vidimo, da oboji na podatkih učenja dosegajo podobne, in to najboljše rezultate. Razlikujejo se v tem, da se naključni gozdovi pri podatkih testiranja pokažejo kot najboljši, odločitvena drevesa pa kot najslabša od vseh. Iz tega lahko sklepamo, da je pri algoritmu naključnih gozdov dodana komponenta naključnosti znatno pripomogla k zmanjšanju prevelikega prileganja, ki je najbolj izrazito prav pri odločitvenih drevesih CART (1/1). Manj razvejena odločitvena drevesa v fazi učenja nekoliko izgubijo učinkovitost, vendar se s tem preveliko prileganje toliko zmanjša, da so rezultati testnih podatkov še vedno boljši od bolj razvejenih.

Rezultati logističnih regresij so bili v osnovi med slabšimi, so pa pridobili največ v kombinaciji z drugimi metodami. Iz analize še ugotovimo, da je ta metoda logistične regresije bistveno bolj učinkovita pri podatkih z večjim številom spremenljivk kot pri večjem številu enot vzorca.

Pri vseh testiranih algoritmih (logistične regresije, odločitvenih dreves in naključnih gozdov) se je pri testiranju pokazalo, da če se model gradi na samo značilnih neodvisnih spremenljivkah, dobimo (brez izjeme) slabše rezultate, kot če bi upoštevali vse neodvisne spremenljivke. Razlike so bile v rezultatih AUC površine ROC analize minimalne, večinoma pod enim odstotkom.

V primerih, ko smo v podatkih prekvalificirali redko prisotne znamke vozil (v deležu, manjšem od 5 %) v kategorijo Ostalo, smo kljub izgubi delčka informacije v podatkih pri določenih modelih pridobili na učinkovitosti in s tem tudi dobili izpeljanko modela naključnih gozdov, ki je z ROC analizo na testnih podatkih dobila najboljši rezultat AUC površine.

Na raziskovalno vprašanje, ali so dobljeni modeli primerni tudi za praktično uporabo, lahko odgovorimo delno pritrdilno oziroma skoraj pritrdilno. Doseženi rezultati klasifikacijske

sposobnosti so bili namreč zelo visoki, če jih primerjamo z rezultati v literaturi, saj so se AUC koeficienti na testnih podatkih gibali od 65 pa vse do 85 %. Vendar ta primerjava ni vedno popolnoma objektivna, saj je učinkovitost metod pri različnih vzorcih lahko zelo različna. Da je različne pristope možno primerjati med sabo, je potrebno izvajati obdelave na istem vzorcu, čemur smo sledili tudi v naši raziskavi.

Če primerjamo rezultate naših modelov z rezultatom avtomatskega točkovanja podjetja ABC lizing, vidimo, da so naši malenkost slabši, vendar je bilo to pričakovano, saj je bil namen njihovega ročnega točkovanja kontrola avtomatskega v prehodnem obdobjem vpeljave v prakso. Nepravilne napovedi njihovega modela so bile posledica še nedokončanega procesa finih nastavitev uteži in zato izjeme. Po končanem prilagajanju uteži niso več preverjali vseh avtomatskih točkovanj, temveč samo še zavrnjene.

Slabši rezultati naših modelov v primerjavi z njihovim tudi nimajo dejanske teže, torej ne pomenijo, da so naši modeli slabši. Z našimi modeli smo se namreč poskušali približati njihovim končnim odločitvam, ki pa niso nujno vedno pravilne. Tako imamo lahko tudi primer različne napovedi, ki se nanaša na njihovo napačno in našo pravilno določitev. Koliko je takšnih primerov, v konkretnem primeru ne bomo nikoli ugotovili, saj je v prihodnosti sicer možno preverili, kateri od njihovih odobrenih poslov bodo slabi, ne bo pa možno ugotoviti, kateri od zavrnjenih poslov bi bili dobri, če ne bi prišlo do zavrnitve. Poleg tega imamo lahko tudi primere, kjer je avtomatska klasifikacija negativna, na oddelku odobravanja pa jo vseeno odobrijo pod pogojem dodatnih zavarovanj. Za takšne primere je bila naša zavrnitev smatrana kot pravilna, lahko pa je tudi napačna, saj ne vemo, ali se ne bi enako izteklo tudi v primeru, če dodatnih zavarovanj ne bi bilo.

Pomemben kriterij za praktično uporabo modela je poleg učinkovitosti napovedi še interpretacija rezultatov, ki je bila v naših primerih pod pričakovanji. Namreč, če primerjamo vse testirane modele, opazimo, da so modeli z boljšimi rezultati manj učinkoviti pri njihovi razlagi, saj so enačbe logističnih regresij preobsežne (vsebujejo tudi do 100 členov zaradi opisnih spremenljivk), odločitvena drevesa preveč razvejena oziroma preveč številčna v primeru naključnih gozdov. Zato pri ocenjeni enoti ni mogoče enostavno ugotoviti, kateri so bili tisti kriteriji, ki so privedli do končne odobritve/zavrnitve. Vprašanje pa je, ali to informacijo res potrebujemo, saj pri teh metodah ni smiselno ročno prilagajati odločitvenih modelov novi situaciji, ampak je obstoječe modele potrebno prilagajati avtomatsko. To lahko dosežemo tako, da v primeru ugotovitve napačne napovedi popravimo izhodiščne podatke učenja s spremenjeno napovedjo in ponovno pripravimo odločitveni model.

Pri drugem vprašanju, ali bo postopek vzpostavitve modela s takšnimi orodji časovno in po obsegu dela bistveno manj zahteven, lahko ponovno odgovorimo pritrdilno. Hitrosti vzpostavitve klasifikacijskega modela so bile pri vseh testiranih primerih zadovoljive, saj nismo nikoli čakali na zaključek obdelave več kot trideset minut, kar zagotovo ne bo problematično tudi v primeru ponavljajočih se obdelav, kadar se iščejo optimalni parametri testirane metode. To je bistveno manj časa kot v primeru vzpostavitve nestrukturiranega

odločitvenega modela. V slednjem je za postavitev/določitev modela potrebnih več dni dela. Tu seveda ni upoštevan porabljen čas za statistično analizo in pripravo podatkov, saj je to potrebno opraviti vedno, ne glede na izbran pristop.

Na zadnje vprašanje, ali je možno s kombiniranjem omenjenih metod rezultate še izboljšati, lahko odgovorimo pritrdilno, vendar samo v primeru logistične regresije. Pri odločitvenih drevesih učinka ni bilo zaznati, pri metodah naključnih gozdov pa je bil ta celo negativen. Zanimivo je še, da so se pri naključnih gozdovih z dodajanjem spremenljivk rezultati poslabšali pri testnih podatkih in izboljšali pri podatkih učenja. To pomeni, da smo z dodajanjem spremenljivk model privedli le do bolj izrazitega prevelikega prileganja. V primeru, ko smo v podatkih popravili nekatere vrednosti znamk vozil in s tem malenkost izgubili na količini informacij, pa smo rezultate na testnih primerih celo malenkost izboljšali. Pri ponovitvi poskusa Antipov in Pokryshevskaya (2010) smo sicer dobili malenkost boljše rezultate, vendar ne toliko kot pri ostalih kombinacijah logistične regresije.

SKLEP

Na podlagi testiranj lahko potrdimo, da sodobna orodja v dokaj kratkem času omogočajo kreiranje dovolj učinkovitih odločitvenih modelov, pod pogojem, da imamo dovolj podatkov ustrezne kvalitete.

Pri testih nismo dobili zadovoljivih interpretacij rezultatov, vsaj ne pri modelih, ki imajo najboljše rezultate napovedovanja. Pri naključnih gozdovih sicer lahko vidimo, kateri kriteriji imajo v modelu največji vpliv na odločitev, vendar iz te informacije ni možno sklepati, kako je potrebno spremeniti vrednost kriterija, da bi dosegli spremembo odločitve. Iskanje oziroma zagotavljanje dobre interpretacije rezultatov tudi ni bil namen naše raziskave, lahko pa je dobra iztočnica za nadaljnje raziskovanje.

V nadaljevanju bi bilo zanimivo testirati še modele, ki niso binarni in omogočajo napoved več možnih vrednosti odvisne spremenljivke, na primer tri. Tako bi imeli poleg odobreno ali zavrnjeno še tretjo, vmesno (recimo neopredeljeno) vrednost in bi lahko pri takšnem modelu avtomatizirali odločanje samo pri bolj izrazitih komitentih (zelo dobri ali zelo slabi). Neopredeljene komitente bi dodatno preverili in individualno ocenili njihovo kreditno sposobnost na podlagi dodatnih informacij. Pri takšnem modelu se verjetnost napake avtomatskega odločanja zmanjša ob povečevanju področja neopredeljenih enot in obratno. V robnem primeru, ko množico neopredeljenih predstavlja celotni vzorec, ni avtomatskih odločitev in s tem tudi ni napak. V primeru nične množice neopredeljenih pa imamo ponovno binarni model, kjer se izvede avtomatska napoved kreditne sposobnosti za vse enote vzorca, in s tem tudi maksimalno število napak. Kateri model od vseh možnih različnih variant bi bil optimalen, je odvisno od ocene stroška napak (enačba 1) in povprečnega stroška za dodatno preverjanje komitenta. Vsekakor bi moral imeti optimalni model čim večji delež avtomatskega odločanja ob čim nižji stopnji napake. Verjetno bi bili takšni modeli bolj

primerni za praktično uporabo kot binarni in bi bilo zato vredno raziskavo nadaljevati še na tem področju.

LITERATURA IN VIRI

1. Agencija Republike Slovenije za javnopravne evidence in storitve (2014). *Metodologija za določanje bonitetnih ocen gospodarskih družb (AJPES S.BON model)*. Najdeno 10. novembra 2016 na spletnem naslovu http://www.ajpes.si/doc/Bonitete/S.BON/ AJPES_S.BON-opis_metodologije-2014.pdf

2. Antipov, E., & Pokryshevskaya, E. (2010). Applying CHAID for logistic regression diagnostics and classification accuracy improvement. *Journal of Targeting, Measurement and Analysis for Marketing, 18*(2), 109–117.

3. Banasiak, M. J., & Kiely, G. L. (2000). Predictive collection score technology. *Business Credit 102.2*, 18–20.

4. Bernard, Ž. (2003). *Izboljšave metode skladanja klasifikatorjev* (magistrsko delo). Ljubljana: Fakulteta za računalništvo in informatiko.

5. Berry, M. J., & Linoff, G. (1997). *Data mining techniques: for marketing, sales, and customer support.* New York, USA: John Wiley & Sons, Inc.

6. Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13, Apr, 1063–1095.

7. Biau, G., & Scornet, E. (2016). A Random Forest guided Tour. *Test, 25(2), 197–227.* Springer.

8. Breiman, L. (1996). Bagging predictors. *Machine learning, 24*(2), 23–140).

9. Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

10. Breiman, L., & Cutler, A. (2015, 7. oktober). *Package 'randomForest'.* Najdeno 16. novembra 2016 na spletnem naslovu https://cran.r-project.org/web/packages/randomForest/ randomForest.pdf

11. Breiman, L., Freidman, J., Stone, C., & Olsten, R. (1984). *Classification and Regression Trees.* United States: Taylor & Francis.

12. Breiman, L., & Lecture, W. (2002). *Looking inside the black box.* Wald Lecture II. Najdeno 13. januarja 2017 na spletnem naslovu http://www.stat.berkeley.edu/~breiman/wald2002-2.pdf

13. Cataldo, Z. (2010). Classification and prediction in customer scoring. *Journal of Modelling in Managament*, 5(1), 38–53.

14. Clemencon, S., Depecker, M., & Vayatis, N. (2013). Ranking forests. *The Journal of Machine Learning Research, 14*(1), 39–73. JMLR, Inc., MIT Press, and Microtome Publishing (USA)

15. Deng, H., Runger, G., & Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. *Artificial Neural Networks and Machine Learning – ICANN 2011*, 293–300. Springer.

16. Diana, T. (2005). Credit risk analysis and credit scoring - now and in the future. *Business Credit 107*(3), 38–42.

17. Edelstein, A. H. (1999). *Introduction to Data Mining and Knowledge Discovery* (3rd ed.) Potomac: Two Crows Corporation.

18. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters 27*(8), 861–874. Elsevier.

19. Fensterstock, A. (2005). Credit scoring and the next step. *Business Credit 107*(3), 46–49.

20. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences, 55*(1), 119–139. Elsevier.

21. Ghatasheh, N. (2014). Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study. *International Journal of Advanced Science and Technology, 72*, 19–30. Science Publications.

22. Han, J., & Kamber, M. (2000). *Data mining: concepts and techniques (the Morgan Kaufmann Series in data management systems).* Morgan Kaufmann.

23. Hartman-Wendels, T., Miller, P., & Tows, E. (2014). Loss given default for leasing: Parametric and nonparametric estimations. *Journal of Banking & Finance*, 40, 364–375. Elsevier.

24. Jagrič, V. (2012). *Ocenjevanje verjetnosti neplačila za kredite prebivalstvu: nelinearen pristop s samoorganizirajočimi se mrežami* (doktorska disertacija). Maribor: Ekonomsko-poslovna fakulteta.

25. Koh, H. C. (2015). A two-step method to construct credit scoring models with data mining techniques. *International Journal of Business and Information, 1*(1), 96–118. ExcelingTech.

26. Kraus, A. (2014). *Recent methods from statistics and machine learning for credit scoring* (doktorska disertacija). München: Fakultät für Mathematik, Informatik und Statistik der Ludwig–Maximilians.

27. Laurent, M., & Schmit, M. (2005). *Estimating distressed LGD on defaulted exposures: a portfolio model applied to leasing contracts*, Bruxelles: ULB - Universite Libre de Bruxelles.

28. Leonard, K. J. (1995). The development of credit scoring quality measures for consumer credit applications. *International Journal of Quality \& Reliability Management, 12*(4), 79–85. MCB UP Ltd.

29. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news 2.3*, 18–22. R Foundation for Statistical Computing.

30. Majnik, M., & Bosnić, Z. (2013). ROC analysis of classifiers in machine learning: A survey. *Intelligent Data Analysis*, 531–558. IOS Press.

31. Mester, L. (1997). What's the point of credit scoring? *Business Review*, 3–16. Federal Reserve Bank of Philadelphia.

32. Miller, M. (2003). research confirms value of credit scoring. *National Underwriter 107*(42), 30. Property Casualty 360'.

33. Misha, D., David, M., & Nando, D. F. (2013). Narrowing the gap: Random forests in theory and in practice. *arXiv preprint arXiv:1310.1415*.

34. 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. Elsevier.

35. Prakash, S. (1995). Mortgage lenders see credit scoring as key to hacking through red tape. *American Banker 160*(161), 1–2.

36. Quinlan, J. R. (1993). *Programs for machine learning.* San Francisco: Morgan Kaufmann Publishers Inc.

37. Rimmer, J. (2005). Contemporary changes in credit scoring. *Credit Control 26*(4), 56–60. Trade Publication.

38. Ritschard, G. (2010). *CHAID and earlier supervised tree methods.* Juillet.

39. Rokach, L., & Maimon, O. (2005). *Data mining and knowledge discovery handbook.* Springer.

40. Sandler, A., McGinn, S., & Barloon, J. (2000). Fair lending scrutiny of credit score-based underwriting system. *ABA Bank Compliance*, 37–43. Proquest ABI/INFORM.

41. Schmit, M., & Stuyck, J. (2002). *Recovery Rates in the Leasing Industry*. Salzburg: Working Paper presented at Leaseurope's Annual Working Meeting, Salzburg

42. Sharma, D. (2012). Improving the art, craft and science of economic credit risk scorecards using random forests: why credit scorers and economists should use random forests. *Academy of Banking Studies Journal*, 93–116. Jordan Whitney Enterprises, Inc.

43. Subhash, S. (1995). *Applied multivariate techniques.* John Wiley & Sons, Inc.

44. Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science 240*(4857), 1285–1293. American Association for the Advancement of Science.

45. Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better DECISIONS through. *Scientific American 283*, 82–87.

46. Šušteršič, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. V *Expert Systems with Applications 36*(3), (str. 4736–4744).

47. Wilkinson, L. (1992). Tree structured data analysis: AID, CHAID and CART. *Retrieved February 1* Najdeno 16. novembra 2016 na spletnem naslovu https://datamining.bus.utk.edu/Documents/Tree-Structured-Data-Analysis-(SPSS).pdf

48. Zhang, D., Huang, H., Chen, Q., & Jiang, Y. (2007). A Comparision Study of Credit Scoring Models. *Natural Computation 1*, 15–18.

PRILOGE

KAZALO PRILOG

PRILOGA 1: Statistična analiza neodvisnih spremenljivk

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
DOBAVITELJ_POGODB	7578	0	18079	1950,59	2698,746	1,551	,028	1,520	,056
letnik_vozila	7578	0	13	3,18	3,408	,607	,028	-,937	,056
MPC	7578	1400,00	110043,01	13287,9592	7795,11311	2,865	,028	20,918	,056
eurotax	7578	-70,67	2984,63	-,9946	44,77069	46,120	,028	2754,041	,056
POLOG	7578	,00	86,66	27,0288	16,43074	1,136	,028	,965	,056
OM	7578	,00	20,74	7,0526	1,50262	,329	,028	4,243	,056
Placilni_indeks	3029	,00	827,65	8,4925	33,85556	14,391	,044	275,104	,089
ZNPL	3029	,00	16200,00	56,9599	642,68780	17,773	,044	354,920	,089
obligo	7578	892,80	100407,17	11700,3093	7756,91867	2,258	,028	11,477	,056
trajanje	7578	6,00	84,00	62,2126	21,33571	-,612	,028	-,701	,056
neto_placa	7578	,00	79153,00	1015,7408	1090,15461	50,073	,028	3499,236	,056
razpolozljiv_od_z_malico_in_prevozom	7578	-652,75	21038,95	827,6566	618,41960	11,651	,028	290,016	,056
Ststarost_stranke	7578	19,00	82,00	43,5546	12,50736	,409	,028	-,504	,056
stevilo_zaposlenih	7578	,00	9797,00	345,0641	1262,41318	5,902	,028	38,061	,056
doba_poslovanja	4472	1,00	62,00	18,3204	9,71591	,150	,037	-,324	,073
prihodki_podjetja	7578	,00	3327014201	61308809,14	256271505,0	7,337	,028	69,112	,056
procent_kapitala_v_bilancni_vsoti	7396	,00	100,00	24,7963	28,22404	,784	,028	-,634	,057
poslovni_izid	7578	-1466787000	328946637,0	-694613,2938	37847101,40	-31,121	,028	1199,313	,056
Valid N (listwise)	1744								

PRILOGA 2: Rezultat T-testa neodvisnih spremenljivk

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
KUPEC_ODOBRITEV	Equal variances assumed	,006	,937	-,087	7576	,931	-,002	,021	-,043	,039
	Equal variances not assumed			-,086	6124,309	,932	-,002	,021	-,043	,040
DOBAVITELJ_POGODB	Equal variances assumed	32,306	,000	-6,578	7576	,000	-416,399	63,306	-540,495	-292,303
	Equal variances not assumed			-6,665	6627,251	,000	-416,399	62,473	-538,866	-293,932
letnik_vozila	Equal variances assumed	67,243	,000	16,846	7576	,000	1,326	,079	1,172	1,480
	Equal variances not assumed			16,417	5798,171	,000	1,326	,081	1,168	1,484
MPC	Equal variances assumed	6,693	,010	-10,323	7576	,000	-1879,726	182,09823	-2236,689	-1522,7631
	Equal variances not assumed			-10,581	6855,562	,000	-1879,726	177,65484	-2227,985	-1531,4676
eurotax	Equal variances assumed	16,568	,000	-3,466	7576	,001	-3,64746	1,05236	-5,71039	-1,58454
	Equal variances not assumed			-3,096	4188,957	,002	-3,64746	1,17808	-5,95712	-1,33781
POLOG	Equal variances assumed	78,608	,000	-9,056	7576	,000	-3,48163	,38445	-4,23525	-2,72801
	Equal variances not assumed			-9,284	6858,584	,000	-3,48163	,37501	-4,21676	-2,74650
OM	Equal variances assumed	2,935	,087	8,773	7576	,000	,30853	,03517	,23959	,37747
	Equal variances not assumed			8,754	6305,422	,000	,30853	,03524	,23944	,37762
Placilni_indeks	Equal variances assumed	202,574	,000	11,314	3027	,000	13,88571	1,22728	11,47931	16,29210
	Equal variances not assumed			9,464	1282,962	,000	13,88571	1,46716	11,00741	16,76400
ZNPL	Equal variances assumed	11,978	,001	-1,707	3027	,088	-40,58946	23,77389	-87,20406	6,02515
	Equal variances not assumed			-1,985	2357,778	,047	-40,58946	20,44561	-80,68270	-,49621
obligo	Equal variances assumed	5,949	,015	-6,413	7576	,000	-1167,142	181,98251	-1523,878	-810,40572
	Equal variances not assumed			-6,552	6790,918	,000	-1167,142	178,13292	-1516,338	-817,94554

se nadaljuje

Rezultat T-testa neodvisnih spremenljivk (nad.)

trajanje	Equal variances assumed	8,787	,003	-2,098	7576	,036	-1,05278	,50176	-2,03637	-,06919
	Equal variances not assumed			-2,083	6193,595	,037	-1,05278	,50544	-2,04362	-,06194
neto_placa	Equal variances assumed	10,352	,001	-7,699	7576	,000	-196,6821	25,54535	-246,7581	-146,60615
	Equal variances not assumed			-8,791	7125,650	,000	-196,6821	22,37407	-240,5419	-152,82229
razpolozljiv_od_z_malico_in_prevozom	Equal variances assumed	13,797	,000	-14,022	7576	,000	-201,3902	14,36270	-229,5451	-173,23538
	Equal variances not assumed			-14,036	6370,877	,000	-201,3902	14,34849	-229,5181	-173,26237
Ststarost_stranke	Equal variances assumed	66,037	,000	-4,670	7576	,000	-1,37200	,29380	-1,94794	-,79606
	Equal variances not assumed			-4,539	5744,171	,000	-1,37200	,30224	-1,96451	-,77949
stevilo_zaposlenih	Equal variances assumed	17,614	,000	2,427	7576	,015	72,06099	29,68581	13,86858	130,25340
	Equal variances not assumed			2,383	5950,367	,017	72,06099	30,24416	12,77147	131,35051
doba_poslovanja	Equal variances assumed	18,843	,000	-5,973	4470	,000	-1,76166	,29494	-2,33988	-1,18343
	Equal variances not assumed			-5,922	3760,269	,000	-1,76166	,29747	-2,34487	-1,17844
prihodki_podjetja	Equal variances assumed	,982	,322	,518	7576	,605	3120690,8	6028493,9	-8696828	14938210
	Equal variances not assumed			,527	6725,916	,598	3120690,8	5920320,6	-8485013	14726394
procent_kapitala_v_bilancni_vsoti	Equal variances assumed	1,095	,295	-2,464	7576	,014	-2,85323	1,15817	-5,12356	-,58289
	Equal variances not assumed			-2,283	4801,734	,022	-2,85323	1,24955	-5,30292	-,40354
poslovni_izid	Equal variances assumed	,760	,383	,132	7576	,895	117851,60	890324,47	-1627431	1863134,3
	Equal variances not assumed			,153	6840,012	,878	117851,60	769916,83	-1391425	1627127,9

PRILOGA 3: Logit transformacija za LR minimalni nabor

	B	S.E.	Wald	df	Sig.	Exp(B)
ZNAMKA			32,468	42	,855	
ZNAMKA(01)	–,278	,560	,245	1	,620	,758
ZNAMKA(02)	–,003	,421	,000	1	,994	,997
ZNAMKA(03)	,047	,396	,014	1	,906	1,048
ZNAMKA(04)	–,016	,473	,001	1	,972	,984
ZNAMKA(05)	–21,062	40192,970	,000	1	1,000	,000
ZNAMKA(06)	,370	,401	,852	1	,356	1,448
ZNAMKA(07)	–22,595	40192,970	,000	1	1,000	,000
ZNAMKA(08)	–,052	,516	,010	1	,920	,950
ZNAMKA(09)	–20,284	40192,970	,000	1	1,000	,000
ZNAMKA(10)	,007	,420	,000	1	,986	1,007
ZNAMKA(11)	,069	,374	,034	1	,853	1,072
ZNAMKA(12)	–,104	,530	,039	1	,844	,901
ZNAMKA(13)	,088	,392	,050	1	,823	1,092
ZNAMKA(14)	–20,501	40192,970	,000	1	1,000	,000
ZNAMKA(15)	–1,568	1,482	1,119	1	,290	,208
ZNAMKA(16)	19,979	25559,902	,000	1	,999	475262767,009
ZNAMKA(17)	22,032	40192,969	,000	1	1,000	3701018269,666
ZNAMKA(18)	,383	,381	1,008	1	,315	1,466
ZNAMKA(19)	–1,541	1,416	1,184	1	,276	,214
ZNAMKA(20)	,437	1,201	,132	1	,716	1,548
ZNAMKA(21)	1,452	1,148	1,600	1	,206	4,270
ZNAMKA(22)	,016	,417	,001	1	,970	1,016
ZNAMKA(23)	,118	,471	,063	1	,802	1,126
ZNAMKA(24)	–,121	,770	,025	1	,875	,886
ZNAMKA(25)	,098	,663	,022	1	,883	1,103
ZNAMKA(26)	–,157	,440	,128	1	,721	,855
ZNAMKA(27)	,027	,381	,005	1	,943	1,028
ZNAMKA(28)	–,510	1,389	,135	1	,714	,601
ZNAMKA(29)	,189	,415	,207	1	,649	1,208
ZNAMKA(30)	–,861	1,097	,616	1	,433	,423
ZNAMKA(31)	1,370	,951	2,077	1	,150	3,936
ZNAMKA(32)	19,885	40192,969	,000	1	1,000	432367409,888
ZNAMKA(33)	,219	,383	,328	1	,567	1,245
ZNAMKA(34)	19,283	40192,969	,000	1	1,000	236859169,966

se nadaljuje

Logit transformacija za LR minimalni nabor (nad.)

	B	S.E.	Wald	df	Sig.	Exp(B)
ZNAMKA(35)	–,085	,532	,026	1	,873	,918
ZNAMKA(36)	–22,642	40192,970	,000	1	1,000	,000
ZNAMKA(37)	1,796	1,050	2,926	1	,087	6,028
ZNAMKA(38)	,316	1,396	,051	1	,821	1,371
ZNAMKA(39)	,381	,467	,665	1	,415	1,463
ZNAMKA(40)	,536	,541	,983	1	,321	1,710
ZNAMKA(41)	,637	,414	2,365	1	,124	1,890
ZNAMKA(42)	–,121	,395	,093	1	,760	,886
KUPEC_CRNA_LISTA(1)	1,829	,281	42,262	1	,000	6,225
KUPEC_ODOBRITEV	–,147	,068	4,654	1	,031	,863
DOBAVITELJ_CRNA_LISTA(1)	,606	,360	2,832	1	,092	1,833
DOBAVITELJ_POGODB	,000	,000	,200	1	,655	1,000
letnik_vozila	–,087	,020	19,768	1	,000	,917
MPC	,000	,000	2,833	1	,092	1,000
eurotax	,024	,004	43,031	1	,000	1,024
POLOG	,020	,004	31,530	1	,000	1,020
OM	,026	,028	,827	1	,363	1,026
obligo	,000	,000	4,808	1	,028	1,000
trajanje	–,001	,003	,303	1	,582	,999
neto_placa	,000	,000	51,131	1	,000	1,000
razpolozljiv_od_z_malico_in_prevozom	,001	,000	98,091	1	,000	1,001
Ststarost_stranke	,007	,004	2,834	1	,092	1,007
TRAJANJE_ZAPOSLITVE			155,857	2	,000	
TRAJANJE_ZAPOSLITVE(1)	–,882	,252	12,196	1	,000	,414
TRAJANJE_ZAPOSLITVE(2)	1,168	,171	46,815	1	,000	3,216
posta			11,794	8	,161	
posta(1)	22,018	40192,969	,000	1	1,000	3651827660,112
posta(2)	,340	,133	6,561	1	,010	1,405
posta(3)	,070	,118	,347	1	,556	1,072
posta(4)	,196	,157	1,559	1	,212	1,216
posta(5)	,384	,184	4,352	1	,037	1,469
posta(6)	–,004	,263	,000	1	,989	,996
posta(7)	,024	,191	,015	1	,901	1,024
posta(8)	,192	,194	,981	1	,322	1,212

se nadaljuje

Logit transformacija za LR minimalni nabor (nad.)

	B	S.E.	Wald	df	Sig.	Exp(B)
drzavljanstvo			48,644	16	,000	
drzavljanstvo(1)	–38,904	56834,468	,000	1	,999	,000
drzavljanstvo(2)	1,888	56834,467	,000	1	1,000	6,606
drzavljanstvo(3)	–38,848	56834,468	,000	1	,999	,000
drzavljanstvo(4)	–19,978	40183,104	,000	1	1,000	,000
drzavljanstvo(5)	–18,597	40183,104	,000	1	1,000	,000
drzavljanstvo(6)	–20,859	40183,104	,000	1	1,000	,000
drzavljanstvo(7)	2,770	56834,467	,000	1	1,000	15,952
drzavljanstvo(8)	–19,918	40183,104	,000	1	1,000	,000
drzavljanstvo(9)	–18,338	40183,104	,000	1	1,000	,000
drzavljanstvo(10)	1,421	48142,502	,000	1	1,000	4,143
drzavljanstvo(11)	3,123	49164,094	,000	1	1,000	22,713
drzavljanstvo(12)	–19,593	40183,104	,000	1	1,000	,000
drzavljanstvo(13)	1,875	56834,467	,000	1	1,000	6,518
drzavljanstvo(14)	–40,584	56834,468	,000	1	,999	,000
drzavljanstvo(15)	–18,017	40183,104	,000	1	1,000	,000
drzavljanstvo(16)	–19,934	40183,104	,000	1	1,000	,000
tip_drlodajalca			33,138	3	,000	
tip_drlodajalca(1)	–,699	,170	16,893	1	,000	,497
tip_drlodajalca(2)	–,602	,109	30,549	1	,000	,548
tip_drlodajalca(3)	–,115	,782	,022	1	,883	,891
Constant	13,983	40183,104	,000	1	1,000	1182573,308

PRILOGA 3: Logit transformacija za LR minimalni nabor + CHAID + CART

	B	S.E.	Wald	df	Sig.	Exp(B)
ZNAMKA			28,589	42	,943	
ZNAMKA(1)	–,312	,626	,248	1	,618	,732
ZNAMKA(2)	,090	,471	,036	1	,849	1,094
ZNAMKA(3)	,116	,440	,070	1	,791	1,124
ZNAMKA(4)	,023	,525	,002	1	,964	1,024
ZNAMKA(5)	–21,896	40192,970	,000	1	1,000	,000
ZNAMKA(6)	,452	,445	1,031	1	,310	1,571
ZNAMKA(7)	–22,624	40192,970	,000	1	1,000	,000
ZNAMKA(8)	–,179	,586	,093	1	,760	,836
ZNAMKA(9)	–19,736	40192,970	,000	1	1,000	,000
ZNAMKA(10)	,025	,466	,003	1	,958	1,025
ZNAMKA(11)	,105	,413	,065	1	,799	1,111
ZNAMKA(12)	,196	,610	,104	1	,748	1,217
ZNAMKA(13)	,214	,438	,240	1	,624	1,239
ZNAMKA(14)	–19,721	40192,970	,000	1	1,000	,000
ZNAMKA(15)	–1,361	1,900	,513	1	,474	,256
ZNAMKA(16)	20,499	26221,329	,000	1	,999	799093450,633
ZNAMKA(17)	22,695	40192,969	,000	1	1,000	7180209101,970
ZNAMKA(18)	,400	,422	,897	1	,344	1,491
ZNAMKA(19)	–,472	1,787	,070	1	,792	,624
ZNAMKA(20)	1,712	1,388	1,522	1	,217	5,538
ZNAMKA(21)	,917	1,263	,527	1	,468	2,502
ZNAMKA(22)	,107	,466	,052	1	,819	1,113
ZNAMKA(23)	–,069	,532	,017	1	,897	,933
ZNAMKA(24)	–,297	,852	,122	1	,727	,743
ZNAMKA(25)	–,211	,735	,083	1	,774	,810
ZNAMKA(26)	–,173	,494	,122	1	,727	,841
ZNAMKA(27)	,236	,424	,310	1	,578	1,266
ZNAMKA(28)	–1,220	1,458	,700	1	,403	,295
ZNAMKA(29)	,391	,459	,722	1	,395	1,478
ZNAMKA(30)	,010	1,190	,000	1	,993	1,010
ZNAMKA(31)	1,774	1,013	3,070	1	,080	5,894
ZNAMKA(32)	18,696	40192,969	,000	1	1,000	131668880,266
ZNAMKA(33)	,424	,423	1,005	1	,316	1,529
ZNAMKA(34)	19,685	40192,969	,000	1	1,000	354219203,874
ZNAMKA(35)	,219	,611	,129	1	,720	1,245
ZNAMKA(36)	–20,759	40192,970	,000	1	1,000	,000
ZNAMKA(37)	1,620	1,084	2,233	1	,135	5,053
ZNAMKA(38)	1,948	1,457	1,787	1	,181	7,014
ZNAMKA(39)	,829	,546	2,311	1	,128	2,292
ZNAMKA(40)	,640	,615	1,080	1	,299	1,896

se nadaljuje

Logit transformacija za LR minimalni nabor + CHAID + CART (nad.)

	B	S.E.	Wald	df	Sig.	Exp(B)
ZNAMKA(41)	,496	,457	1,175	1	,278	1,642
ZNAMKA(42)	,118	,438	,073	1	,788	1,125
KUPEC_CRNA_LISTA(1)	,670	,315	4,531	1	,033	1,955
KUPEC_ODOBRITEV	–,048	,053	,828	1	,363	,953
DOBAVITELJ_CRNA_LISTA(1)	,397	,418	,901	1	,343	1,487
DOBAVITELJ_POGODB	,000	,000	,455	1	,500	1,000
letnik_vozila	–,039	,022	3,112	1	,078	,962
MPC	,000	,000	,202	1	,653	1,000
eurotax	,001	,002	,453	1	,501	1,001
POLOG	,016	,004	15,781	1	,000	1,016
OM	,068	,033	4,256	1	,039	1,070
obligo	,000	,000	3,553	1	,059	1,000
trajanje	–,001	,003	,214	1	,644	,999
neto_placa	,000	,000	,516	1	,472	1,000
razpolozljiv_od_z_malico_in_prevozom	,000	,000	1,749	1	,186	1,000
Ststarost_stranke	,009	,005	3,650	1	,056	1,009
TRAJANJE_ZAPOSLITVE			19,193	2	,000	
TRAJANJE_ZAPOSLITVE(1)	,239	,285	,700	1	,403	1,270
TRAJANJE_ZAPOSLITVE(2)	,753	,203	13,791	1	,000	2,123
posta			6,107	8	,635	
posta(1)	21,338	40192,969	,000	1	1,000	1849154693,580
posta(2)	,243	,154	2,492	1	,114	1,275
posta(3)	,133	,139	,914	1	,339	1,142
posta(4)	,200	,184	1,187	1	,276	1,222
posta(5)	,434	,216	4,033	1	,045	1,544
posta(6)	–,102	,300	,115	1	,735	,903
posta(7)	,128	,220	,338	1	,561	1,136
posta(8)	,075	,225	,111	1	,739	1,078
drzavljanstvo			63,312	16	,000	
drzavljanstvo(1)	–38,826	56841,397	,000	1	,999	,000
drzavljanstvo(2)	,540	56841,397	,000	1	1,000	1,717
drzavljanstvo(3)	–40,715	56841,397	,000	1	,999	,000
drzavljanstvo(4)	–20,962	40192,905	,000	1	1,000	,000
drzavljanstvo(5)	–20,049	40192,905	,000	1	1,000	,000
drzavljanstvo(6)	–22,201	40192,905	,000	1	1,000	,000
drzavljanstvo(7)	,898	56841,397	,000	1	1,000	2,454
drzavljanstvo(8)	–20,687	40192,905	,000	1	1,000	,000
drzavljanstvo(9)	–17,530	40192,905	,000	1	1,000	,000

se nadaljuje

Logit transformacija za LR minimalni nabor + CHAID + CART (nad.)

	B	S.E.	Wald	df	Sig.	Exp(B)
drzavljanstvo(10)	,290	49183,852	,000	1	1,000	1,336
drzavljanstvo(11)	3,953	48426,048	,000	1	1,000	52,074
drzavljanstvo(12)	–21,032	40192,905	,000	1	1,000	,000
drzavljanstvo(13)	,379	56841,397	,000	1	1,000	1,461
drzavljanstvo(14)	–39,431	56841,397	,000	1	,999	,000
drzavljanstvo(15)	–18,796	40192,905	,000	1	1,000	,000
drzavljanstvo(16)	–21,195	40192,905	,000	1	1,000	,000
tip_drlodajalca			19,119	3	,000	
tip_drlodajalca(1)	–,606	,196	9,510	1	,002	,546
tip_drlodajalca(2)	–,532	,126	17,906	1	,000	,587
tip_drlodajalca(3)	–,374	,814	,211	1	,646	,688
CHAID_100	1,702	,262	42,086	1	,000	5,483
CRT_100	3,756	,217	300,318	1	,000	42,768
Constant	13,019	40192,905	,000	1	1,000	450964,519

PRILOGA 4: Seznam slamnatih spremenljivk

ZNAMKA(01)	ALFA
ZNAMKA(02)	AUDI
ZNAMKA(03)	BMW
ZNAMKA(04)	CHEVROLE
ZNAMKA(05)	CHRYSLER
ZNAMKA(06)	CITROEN
ZNAMKA(07)	CRZ
ZNAMKA(08)	DACIA
ZNAMKA(09)	DODGE
ZNAMKA(10)	FIAT
ZNAMKA(11)	FORD
ZNAMKA(12)	HONDA
ZNAMKA(13)	HYUNDAI
ZNAMKA(14)	INFINITI
ZNAMKA(15)	INFO
ZNAMKA(16)	JAGUAR
ZNAMKA(17)	JEEP
ZNAMKA(18)	KIA
ZNAMKA(19)	LANCIA
ZNAMKA(20)	LAND-ROV
ZNAMKA(21)	LEXUS
ZNAMKA(22)	MAZDA
ZNAMKA(23)	MERCEDES
ZNAMKA(24)	MINI
ZNAMKA(25)	MITSHUBI
ZNAMKA(26)	NISSAN
ZNAMKA(27)	OPEL
ZNAMKA(28)	OSEBNO
ZNAMKA(29)	PEUGEOT
ZNAMKA(30)	PRENOS
ZNAMKA(31)	RABLJENO
ZNAMKA(32)	RANGE-RO
ZNAMKA(33)	RENAULT
ZNAMKA(34)	SAAB
ZNAMKA(35)	SEAT
ZNAMKA(36)	SMART
ZNAMKA(37)	SSANGYON
ZNAMKA(38)	SUBARU
ZNAMKA(39)	SUZUKI
ZNAMKA(40)	ŠKODA
ZNAMKA(41)	TOYOTA
ZNAMKA(42)	VOLKSWAG
Brez	VOLVO

DRZAVLJANSTVO(01)	ALBANIA
DRZAVLJANSTVO(02)	AUSTRIA
DRZAVLJANSTVO(03)	BAHRAIN
DRZAVLJANSTVO(04)	BOSNIA A
DRZAVLJANSTVO(05)	BULGARIA
DRZAVLJANSTVO(06)	CROATIA
DRZAVLJANSTVO(07)	ETHIOPIA
DRZAVLJANSTVO(08)	GERMANY
DRZAVLJANSTVO(09)	HUNGARY
DRZAVLJANSTVO(10)	ITALY
DRZAVLJANSTVO(11)	KOSOVO
DRZAVLJANSTVO(12)	MACEDONI
DRZAVLJANSTVO(13)	NETHERLA
DRZAVLJANSTVO(14)	SLOVAKIA
DRZAVLJANSTVO(15)	SLOVENIA
DRZAVLJANSTVO(16)	SRBIJA I
Brez	UNITED K

Posta(1)	BA-7
Posta(2)	SI-1
Posta(3)	SI-2
Posta(4)	SI-3
Posta(5)	SI-4
Posta(6)	SI-5
Posta(7)	SI-6
Posta(8)	SI-8
Brez	SI-9

tip_delodajalca(1)	GOSP. DR
tip_delodajalca(2)	JAVNI SEKTOR
tip_delodajalca(3)	ODVETNIK
Brez	SVOBODNI POKLIC

TRAJANJE_ZAPOSLITVE(1)	DOLOČEN
TRAJANJE_ZAPOSLITVE(2)	NEDOLOČE
Brez	UPOKOJEN

PRILOGA 5: Tabelaričen prikaz odločitvenega drevesa CART 100/50

Sample		0		2		Total		Predicted Category	Parent Node	Primary Independent Variable		
		N	Percent	N	Percent	N	Percent			Variable	Improvement	Split Values
Training	0	1487	39,3 %	2296	60,7 %	3783	100,0 %	2				
	1	475	72,0 %	185	28,0 %	660	17,4 %	0	0	razpolozljiv_od_z_malico_in_prevozom	,045	<= 485,155
	2	1012	32,4 %	2111	67,6 %	3123	82,6 %	2	0	razpolozljiv_od_z_malico_in_prevozom	,045	> 485,155
	3	332	81,2 %	77	18,8 %	409	10,8 %	0	1	razpolozljiv_od_z_malico_in_prevozom	,005	<= 413,085
	4	143	57,0 %	108	43,0 %	251	6,6 %	0	1	razpolozljiv_od_z_malico_in_prevozom	,005	> 413,085
	5	245	73,6 %	88	26,4 %	333	8,8 %	0	2	eurotax	,033	<= -14,660
	6	767	27,5 %	2023	72,5 %	2790	73,8 %	2	2	eurotax	,033	> -14,660
	7	106	70,7 %	44	29,3 %	150	4,0 %	0	4	POLOG	,004	<= 25,195
	8	37	36,6 %	64	63,4 %	101	2,7 %	2	4	POLOG	,004	> 25,195
	9	172	83,1 %	35	16,9 %	207	5,5 %	0	5	POLOG	,003	<= 27,815
	10	73	57,9 %	53	42,1 %	126	3,3 %	0	5	POLOG	,003	> 27,815
	11	627	24,0 %	1987	76,0 %	2614	69,1 %	2	6	Placilni_indeks	,027	<= 10,465

se nadaljuje

Tabelaričen prikaz odločitvenega drevesa CART 100/50 (nad.)

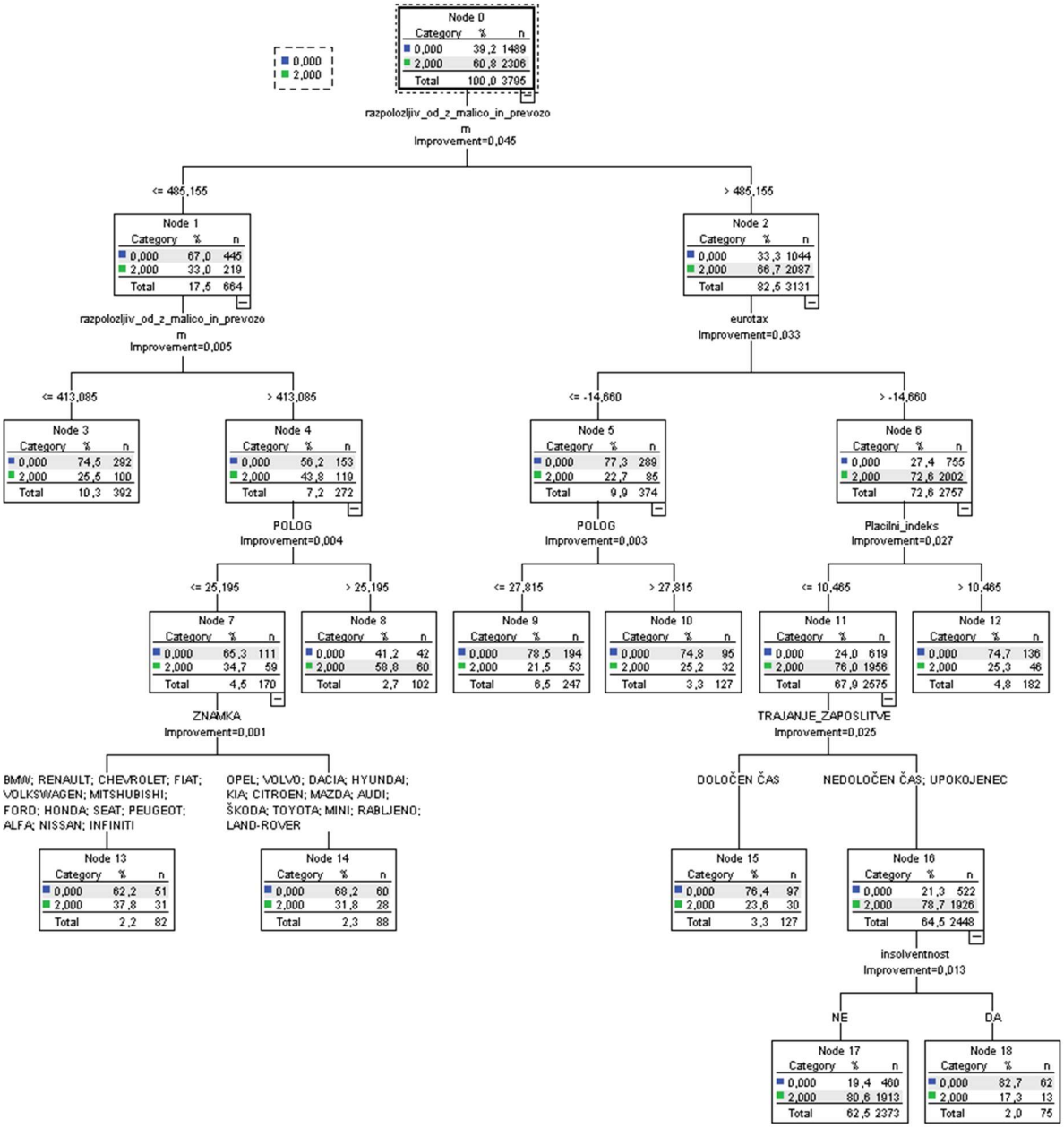
	12	140	79,5 %	36	20,5 %	176	4,7 %	0	6	Placilni_indeks	,027	> 10,465
	13	65	83,3 %	13	16,7 %	78	2,1 %	0	7	ZNAMKA	,001	BMW; RENAULT; CHEVROLET; FIAT; VOLKSWAGEN; MITSHUBISHI; FORD; HONDA; SEAT; PEUGEOT; ALFA; NISSAN; INFINITI
	14	41	56,9 %	31	43,1 %	72	1,9 %	0	7	ZNAMKA	,001	OPEL; VOLVO; DACIA; HYUNDAI; KIA; CITROEN; MAZDA; AUDI; ŠKODA; TOYOTA; MINI; RABLJENO; LAND-ROVER

se nadaljuje

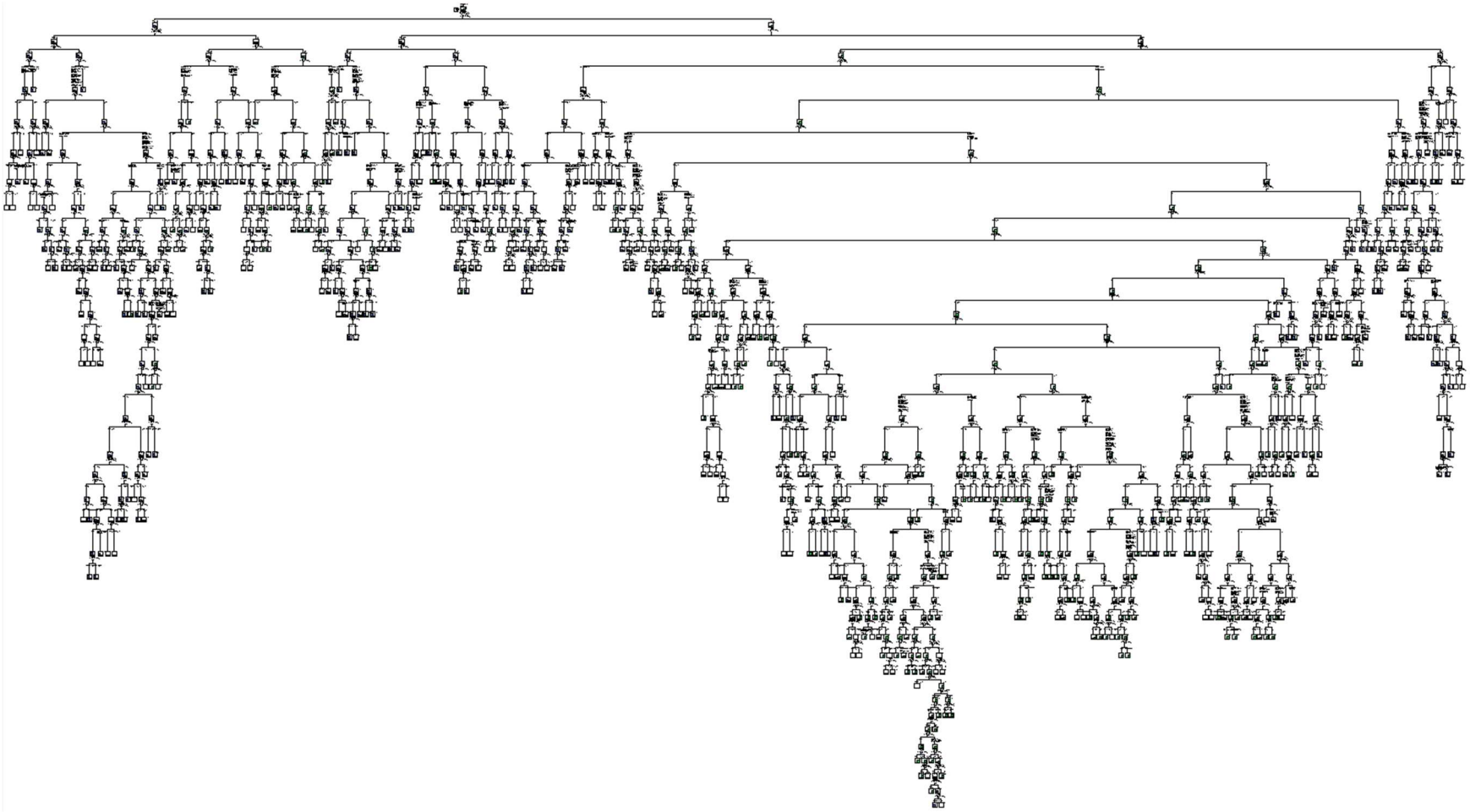
Tabelaričen prikaz odločitvenega drevesa CART 100/50 (nad.)

	15	116	78,4 %	32	21,6 %	148	3,9 %	0	11	TRAJANJE_ZAPOSLITVE	,025	DOLOČEN ČAS
	16	511	20,7 %	1955	79,3 %	2466	65,2 %	2	11	TRAJANJE_ZAPOSLITVE	,025	NEDOLOČEN ČAS; UPOKOJENEC
	17	452	18,9 %	1938	81,1 %	2390	63,2 %	2	16	insolventnost	,013	NE
	18	59	77,6 %	17	22,4 %	76	2,0 %	0	16	insolventnost	,013	DA

PRILOGA 6: Grafični prikaz odločitvenega drevesa CART (100/50)

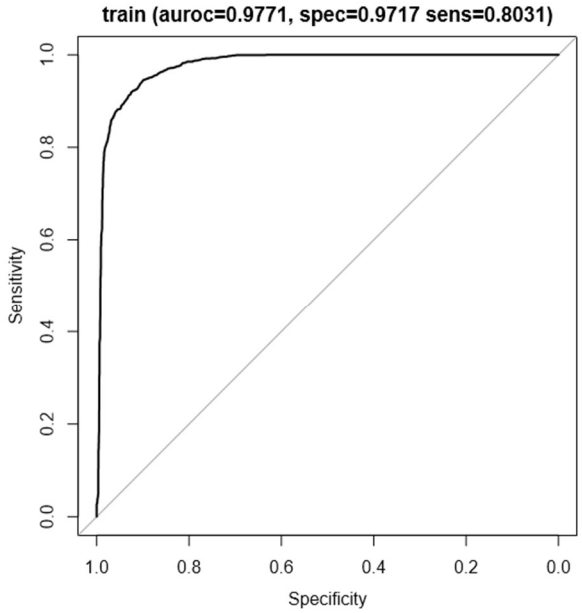


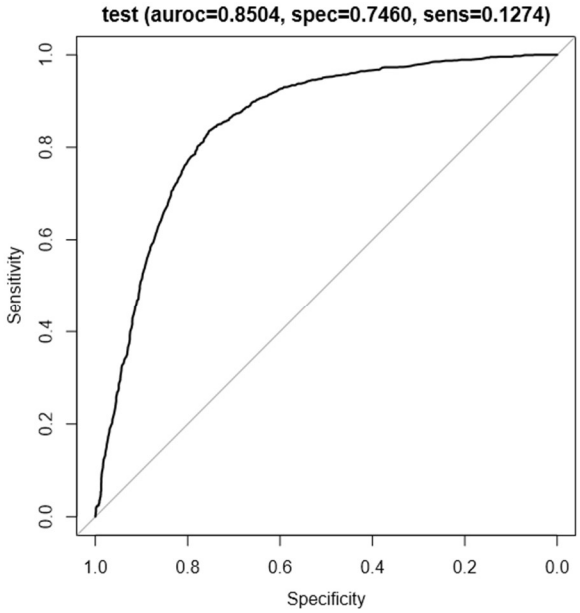
PRILOGA 7: Grafični prikaz odločitvenega drevesa CART (1/1)



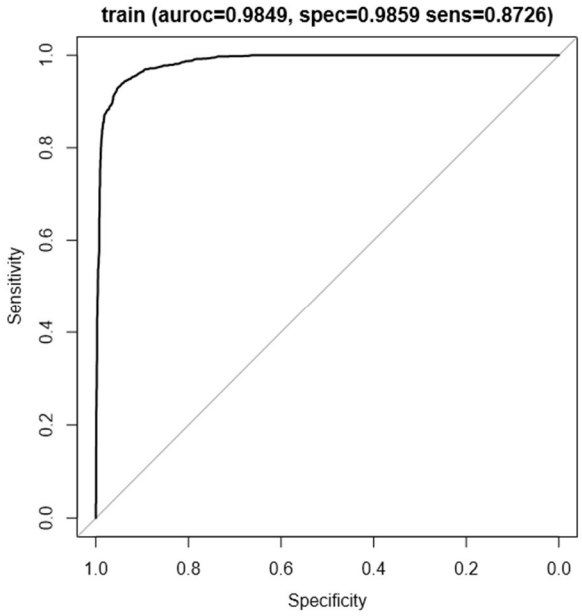
PRILOGA 8: ROC krivulje naključnih gozdov

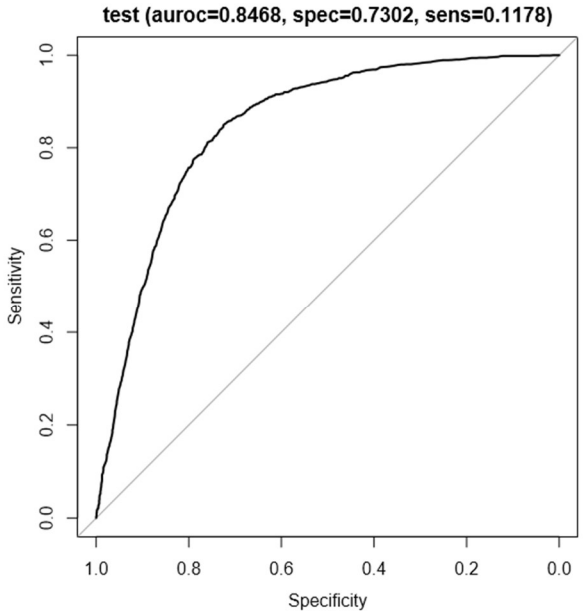
1. Nad vsemi pojasnjevalnimi spremenljivkami (enak nabor kot pri testu odločitvenih dreves)



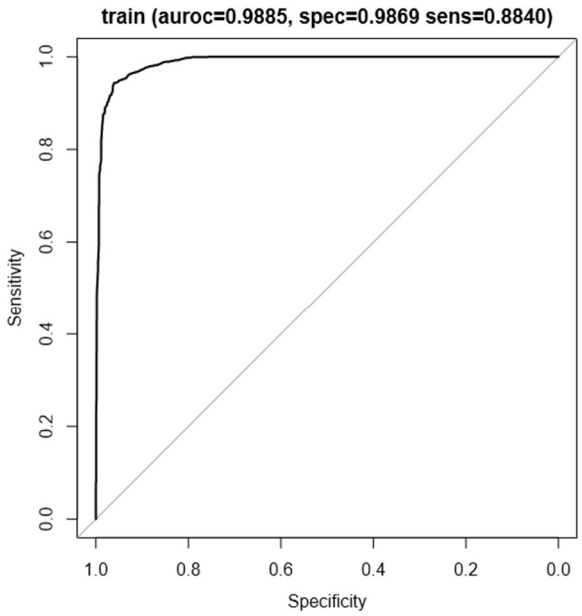


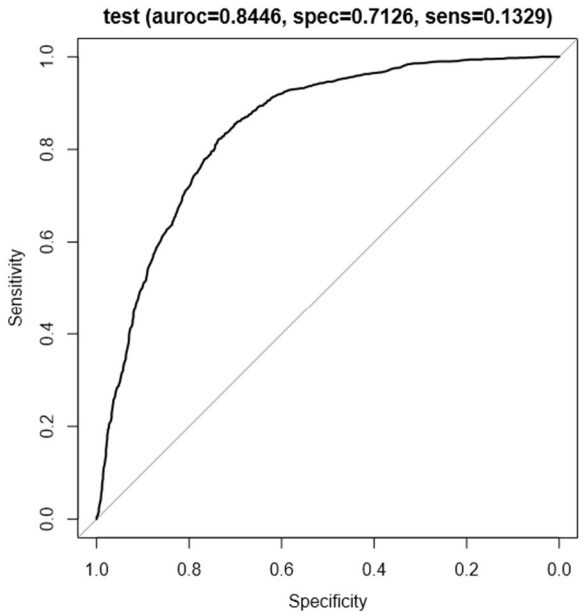
2. Nad vsemi pojasnjevalnimi spremenljivkami, vključno z rezultatom logistične regresije z minimalnim naborom spremenljivk



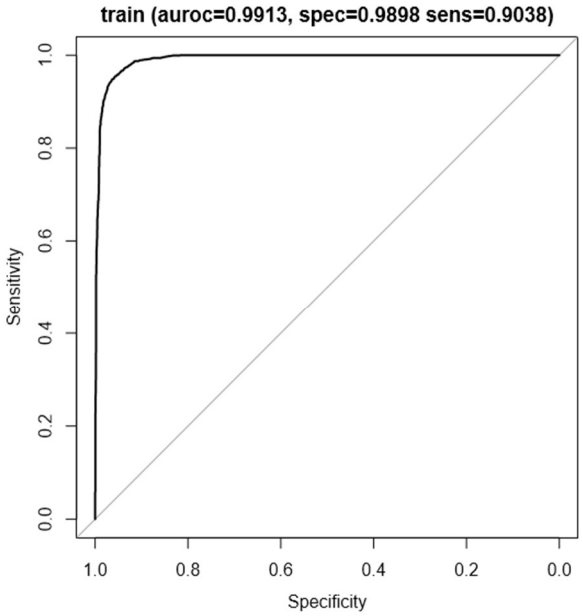


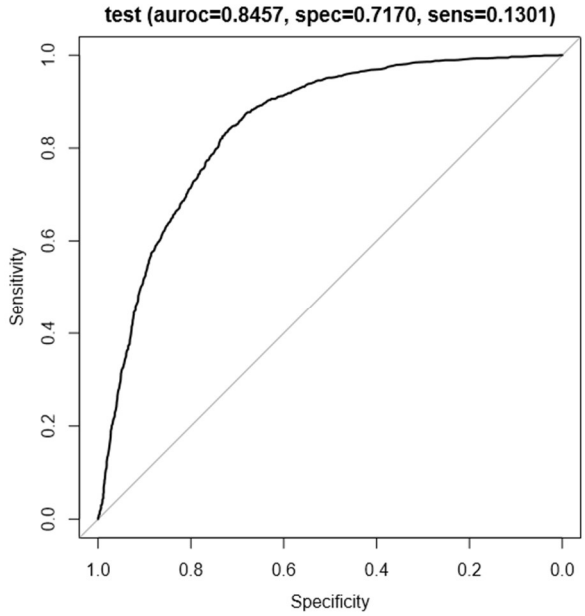
3. Nad vsemi pojasnjevalnimi spremenljivkami, vključno z rezultatom logistične regresije nad podmnožicami (nova/obstoječa stranka, je/ni delodajalca)





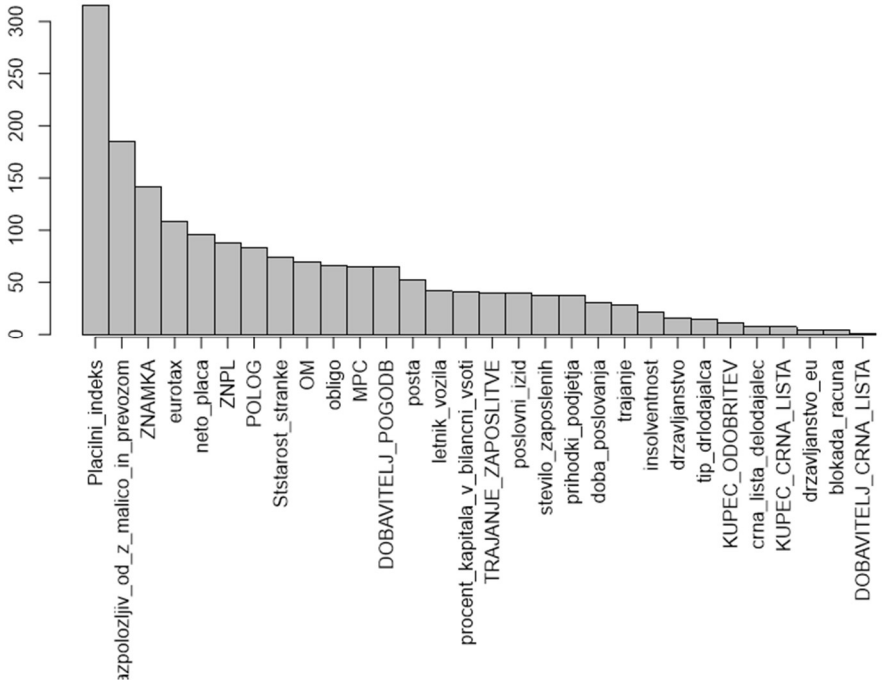
4. Nad vsemi pojasnjevalnimi spremenljivkami, vključno z obema rezultatoma logističnih regresij (+LR +LR1234)



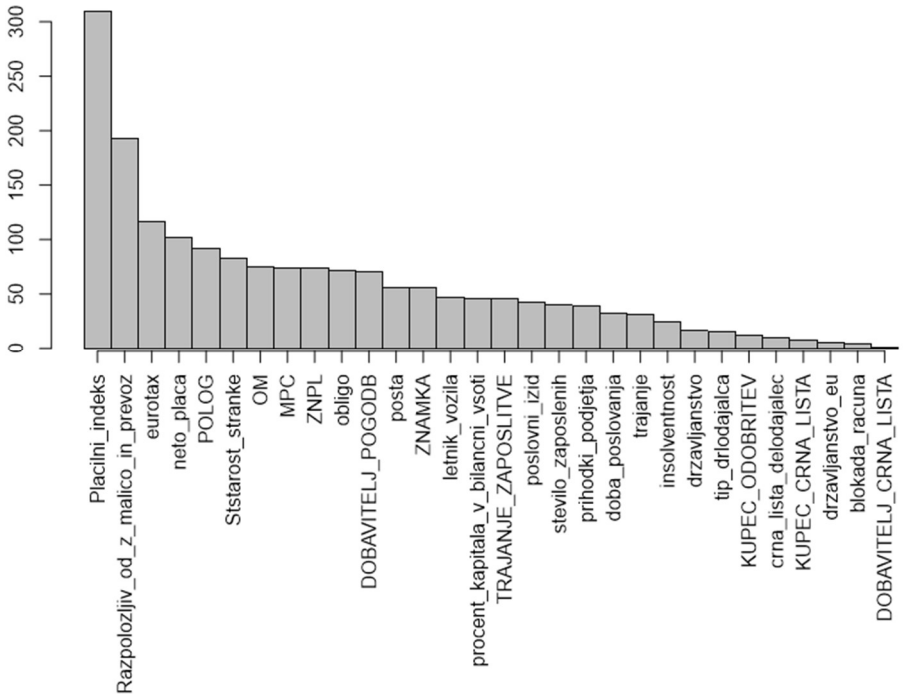


PRILOGA 9: Test pomembnosti spremenljivk naključnih gozdov

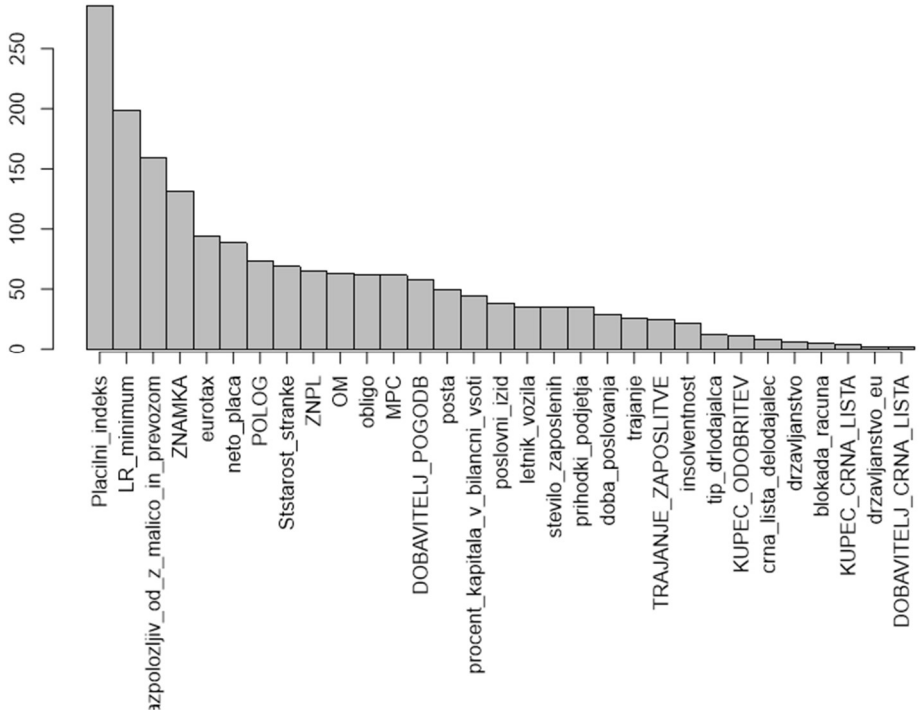
1. Nad vsemi pojasnjevalnimi spremenljivkami (enak nabor kot pri testu odločitvenih dreves)



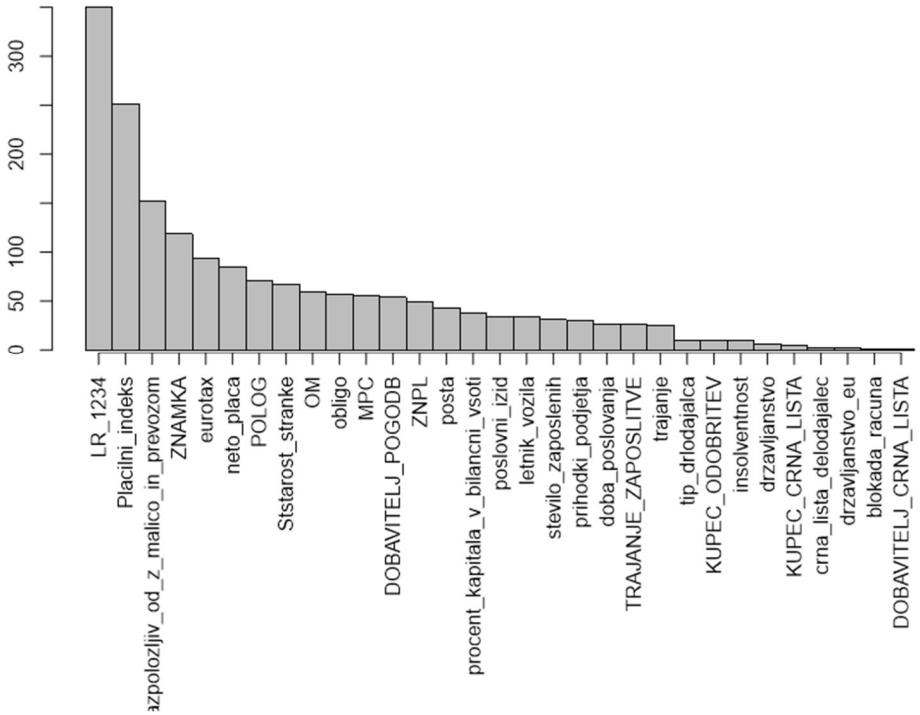
2. Nad vsemi pojasnjevalnimi spremenljivkami brez znamk izjem



3. Nad vsemi pojasnjevalnimi spremenljivkami, vključno z rezultatom logistične regresije z minimalnim naborom spremenljivk



4. Nad vsemi pojasnjevalnimi spremenljivkami, vključno z rezultatom logistične regresije nad podmnožicami (nova/obstoječa stranka, je/ni delodajalca)



5. Nad vsemi pojasnjevalnimi spremenljivkami, vključno z obema rezultatoma logističnih regresij (+LR +LR1234)