Klasifikacijska stabla odlučivanja u kreditnom skoriranju

Miser, Bernarda

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike
Smjer: Financijska matematika i statistika

Bernarda Miser

Klasifikacijska stabla odlučivanja u kreditnom skoriranju

Diplomski rad

Osijek, 2023.

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike
Smjer: Financijska matematika i statistika

Bernarda Miser

Klasifikacijska stabla odlučivanja u kreditnom skoriranju

Diplomski rad

Mentor: prof. dr. sc. Nataša Šarlija
Komentor: izv. prof. dr. sc. Domagoj Matijević

Osijek, 2023.

Sadržaj

Uvod

Većina banaka pri odobravanju kredita stanovništvu koristi modele kreditnog skoriranja. S obzirom da je glavna djelatnost banke prikupljanje depozita i plasiranje sredstava u obliku kredita, pritom zadržavajući stopu obvezne rezerve propisane od strane središnje banke, od velike im je važnosti procijeniti kreditni rizik. Odluka o odobravanju kredita može biti podržana subjektivnom ocjenom kreditnih referenata ili modelima kreditnog skoriranja. Svrha modela kreditnog skoriranja je odrediti vjerojatnost da klijent neće ispuniti svoje obveze prema financijskoj ustanovi, tj. da će biti loš. S obzirom da modeli kreditnog skoriranja omogućuju kvalitetnije donošenje odluka i upravljanje kreditnim rizicima, financijske institucije se sve više okreću upotrebi modela. Neke metode koje se koriste za izgradnju modela kreditnog skoriranja su neuronske mreže, logistička regresija i stabla odlučivanja. Stabla odlučivanja se mogu prikazati i grafički te su time lakše objašnjiva. Cilj ovog rada je upravo napraviti model kreditnog skoriranja za stanovništvo pomoću klasifikacijskog stabla odlučivanja (*engl.* classification tree) kako bismo mogli procijeniti koji je klijent dobar, a koji loš.

U prvom poglavlju ćemo opisati što su stabla odlučivanja, kakva sve postoje, koje metode se koriste za njihovu izgradnju te kako pronaći najbolje stablo. Zatim će u sljedećem poglavlju biti objašnjeno što je kreditno skoriranje i kako se gradi model kreditnog skoriranja. Objasnit ćemo i kako primijeniti stabla odlučivanja kod izrade modela kreditnog skoriranja i opisati par prethodnih istraživanja u kojima su izrgrađeni takvi modeli. U trećem poglavlju bit će napravljena bivarijatna analiza podataka koje smo koristili za izradu klasifikacijskih stabala odlučivanja te prikazani i opisani dobiveni rezultati kod izradnje stabala. Na kraju ćemo interpretirati rezultate, testirati kvalitetu dobivenih modela te usporediti rezultate.

1 Stabla odlučivanja

Rudarenje podataka (*engl.* data mining) je proces u kojem se korištenjem raznih matematičkih tehnika, dobiju podaci relevantni za određenu organizaciju (npr. što ljudi najviše kupuju, u kojem obliku štede novac, da li otplaćuju kredit) i na taj način se podaci dalje koriste u marketingu, reklamama, kod određivanja kreditne sposobnosti klijenta, itd. Rudarenjem podataka se zapravo istražuju i analiziraju baze podataka s ciljem identificiranja uzorka i uspostavljanja veze za rješavanje problema. Najvažniji zadaci rudarenja podataka su stvoriti prediktivnu i opisnu moć. Prediktivnost dobijemo korištenjem metoda za proricanje buduće ili još nepoznate vrijednosti, dok opisnu moć dobijemo pronalaženjem iterpretabilnih uzoraka koji opisuju podatke. (vidi [10])

Metode rudarenja podacima možemo podijeliti na nadzirana i nenadzirana učenja. Nadzirana učenja su ona u kojima algoritmi koriste poznati skup ulaznih podataka i poznati skup izlaznih podataka te kao rezultat dobijemo model za predviđanje. U nadzirana učenja se ubrajaju klasifikacija i regresija. Klasifikacija je proces u kojem se ulazni podaci razvrstavaju u kategorije. Proces klasifikacije bavi se problemima u kojima se podaci mogu podijeliti na dvije ili više diskretne kategorije. Klasifikacija se često koristi za određivanje je li e-pošta neželjena (*engl.* spam) ili ne, u medicini kod određivanja je li tumor karcinogen ili dobroćudan, u financijskom svijetu kod odobravanja kredita klijentu i sl. Regresija je proces u kojem zavisnost varijable pokušavamo opisati neprekidnom funkcijom. Regresijom se predviđaju kontinuirane varijable, poput promjene temperature, fluktuacije potrošnje električne energije, sastav i kvaliteta proizvoda i sl. Kod nenadziranih učenja algoritam nema nikakve upute za rješavanje problema na početku osim ulaznih podataka i zadataka koje mora obaviti. Najčešća tehnika takvog učenja je grupiranje (*engl.* clustering). Kod grupiranja ulazni podaci se pokušavaju podijeliti u disjunktne klastere gdje se u svakom klasteru nalaze podaci sa sličnim određenim obilježjima.(vidi [1])

Stablo odlučivanja je, u matematičkom smislu, usmjerena povezana šuma[1] s fiksiranim korijenom. vidi([16]) Sastoji se od čvorova. Čvorovi su podskupovi skupa vrijednosti atributa, tj. varijabli. Početni čvor koji nema ulaznu granu i od kojeg se uvijek kreće naziva se još i korijen stabla. Svi ostali čvorovi imaju točno jednu ulaznu granu. Čvorovi koji imaju izlaznu granu nazivaju se unutarnji ili testni čvorovi dok su listovi čvorovi s jednom ulaznom granom i bez izlazne grane. Proces odlučivanja sastoji se od niza povezanih odluka. Gradi se na temelju dostupnih podataka kako bi puštanjem novih podataka kroz stablo mogli predvidjeti njihovu kategoriju.

Prema [17] prednosti stabla odlučivanja su:

- sposobnost generiranja razumljivih modela,
- relativno kratko vrijeme izvršavanja zadataka i zauzimanje malo memorije,
- sposobnost korištenja i numeričkih i kategoričkih atributa,
- jasno održavanje važnosti pojedinih atributa za konkretni predikcijski problem.

Neki nedostaci stabla odlučivanja su:

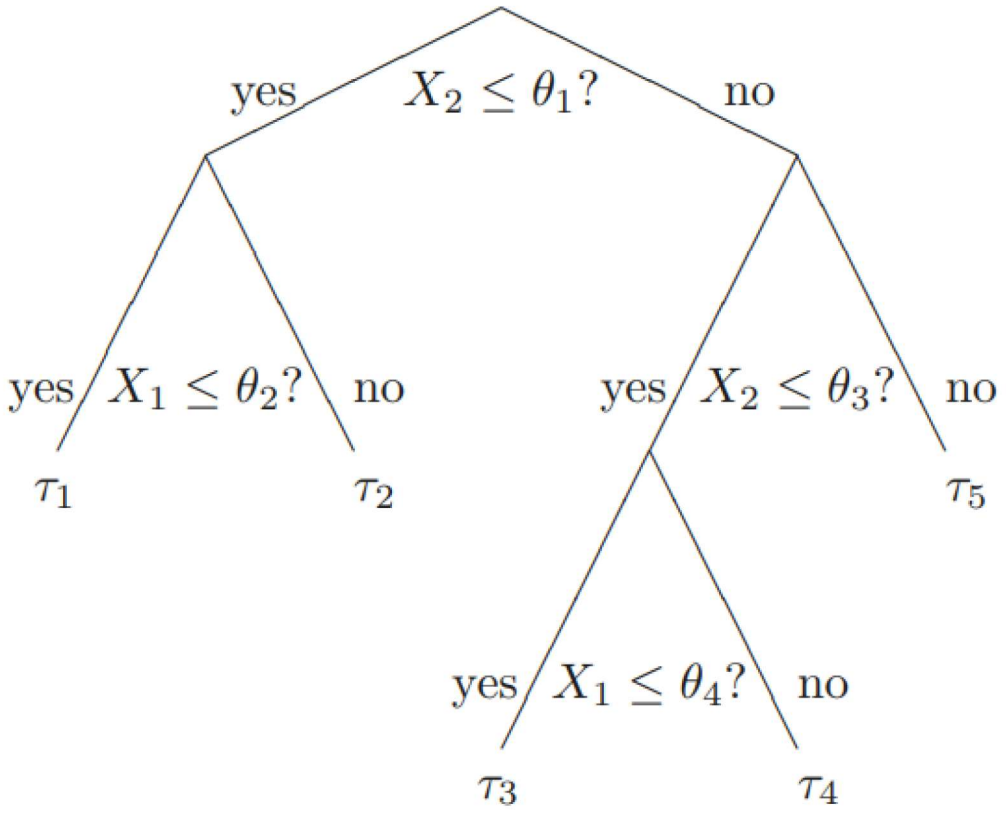
- sklonost greškama u višeklasnim problemima s relativno malim brojem podataka za treniranje,
- manja prikladnost za probleme kod kojih se traži predikcija kontinuiranih vrijednosti ciljanog, tj. zavisnog atributa.

[1] Šuma je graf bez ciklusa.

Ako stablo odlučivanja gradimo pomoću klasifikacije tada se to stablo naziva klasifikacijsko stablo. Klasifikacijska stabla se često koriste u području financija, marketinga, medicine i inženjerstva. (vidi [7])

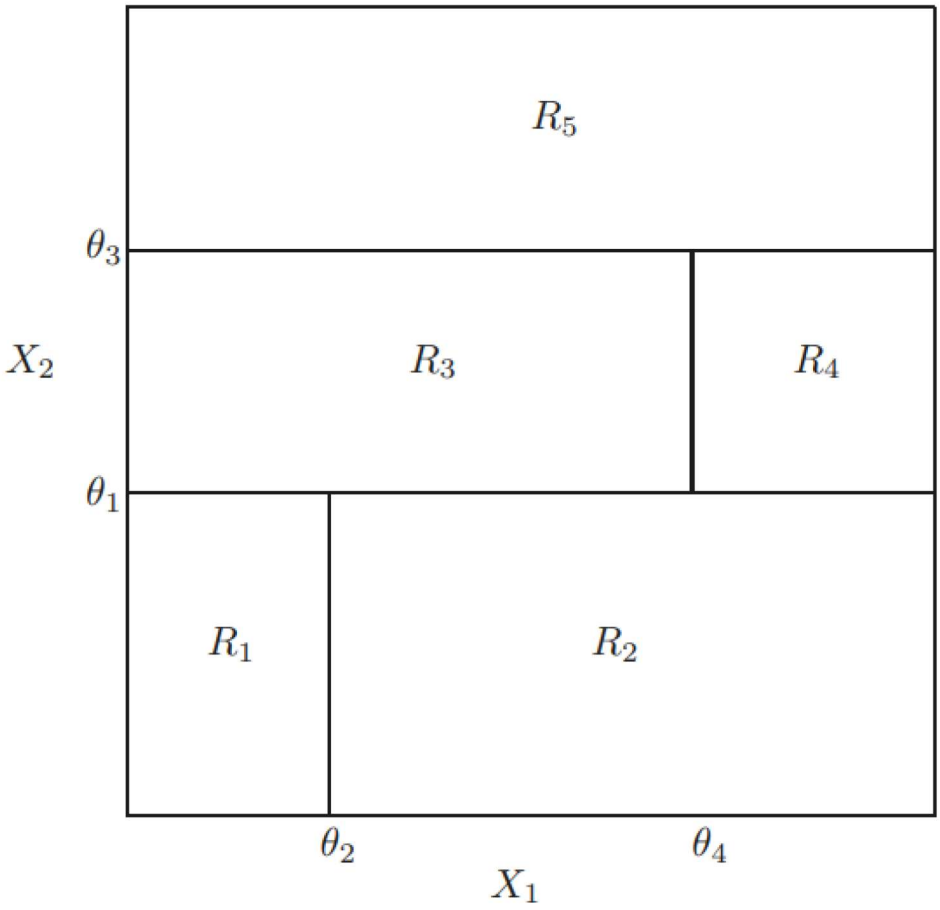
Klasifikacijsko stablo je binarno stablo, dakle u njemu se testni čvorovi dijele na dva čvora, u jednom je uvjet iz testnog čvora zadovoljen, a u drugom nije. Takvo dijeljenje određeno je Booleovim uvjetom na vrijednost jednog atributa.

U nastavku ćemo objasniti klasifikacijsko stablo na jednostavnom primjeru. Na *Slici 1* prikazano je klasifikacijsko stablo u kojem je podatak prikazan pomoću dva atributa (X_1, X_2). Krećemo od korijena stabla i prvo promatramo da li za promatrani podatak vrijedi da je X_2 manji ili jednak od θ_1. Ako upit vrijedi, spuštamo se u lijevu granu gdje dolazimo do upita da li je X_1 manji ili jednak od θ_2. Ako upit vrijedi, tad podatak dolazi do lista τ_1, a ako ne vrijedi tad podatak dolazi do lista τ_2. Za dobivanje rezultata iz ovog stabla moramo proći najviše tri testna čvora iz čega vidimo da je prikazano stablo dubine tri. (vidi [3])



Slika 1: Klasifikacijsko stablo [3]

Sliku 1 možemo prikazati i na sljedeći način prikazan na *Slici 2*.



Slika 2: Grafički prikaz particije skupa podataka [3]

Iz *Slike 2* možemo vidjeti da klase R_1, R_2, R_3, R_4 i R_5 čine jednu particiju skupa podataka za koju vrijedi

$$R_1 = \{X_1 \leq \theta_2, X_2 \leq \theta_1\}$$
$$R_2 = \{X_1 > \theta_2, X_2 \leq \theta_1\}$$
$$R_3 = \{X_1 \leq \theta_4, \theta_1 \leq X_2 \leq \theta_3\}$$
$$R_4 = \{X_1 > \theta_4, \theta_1 \leq X_2 \leq \theta_3\}$$
$$R_5 = \{X_2 > \theta_3\}$$

Što znači da se u skupu R_1 nalaze svi podaci koji su došli do lista τ_1, tj. za koje vrijedi da je X_1 manji ili jednak od θ_2 i X_2 manji ili jednak od θ_1.

1.1 Izgradnja klasifikacijskog stabla

Prije izgradnje klasifikacijskog stabla trebamo definirati skup svih podataka, ciljanog atributa te skup svih ostalih atributa. U nastavku je opisan algoritam za izgradnju klasifikacijskog stabla koje ćemo u empirijskom istraživanju zvati puno stablo.

1. Napravite stablo s jednim korijenskim čvorom.
2. Odaberite najbolji atribut i uvjet u kojem funkcija za odabir najboljeg atributa i uvjeta ima najmanju vrijednost.
3. Stavite najbolji atribut s najboljim uvjetom u čvor i skup podataka podijelite u dva podskupa tako da se u lijevoj grani dalje obrađuju samo podaci koji zadovoljavaju izabrani uvjet, a u desnoj koji ne zadovoljavaju.
4. Započnite izgradnju stabla rekurzivnim ponavljanjem prethodne dvije točke sve dok dijeljenjem skupa podataka u dva podskupa ne dobijete sve podatke u jednom skupu ili dok ne ponestane atributa za dijeljenje. (vidi [7])

Za odabir najboljeg atributa i uvjeta koriste se funkcije nečistoće. U nastavku ćemo objasniti dvije takve funkcije, gini indeks i entropiju, koje ćemo koristiti i u izgradnji stabala. Gini indeks je kriterij temeljen na 'nečistoćama' u čvoru koji mjeri vjerojatnost da će odabrani podatak u čvoru biti netočno klasificiran, a definiran je:

$$Gini(y, S) = 1 - \sum_{c_j \in dom(y)} (P(y = c_j))^2$$

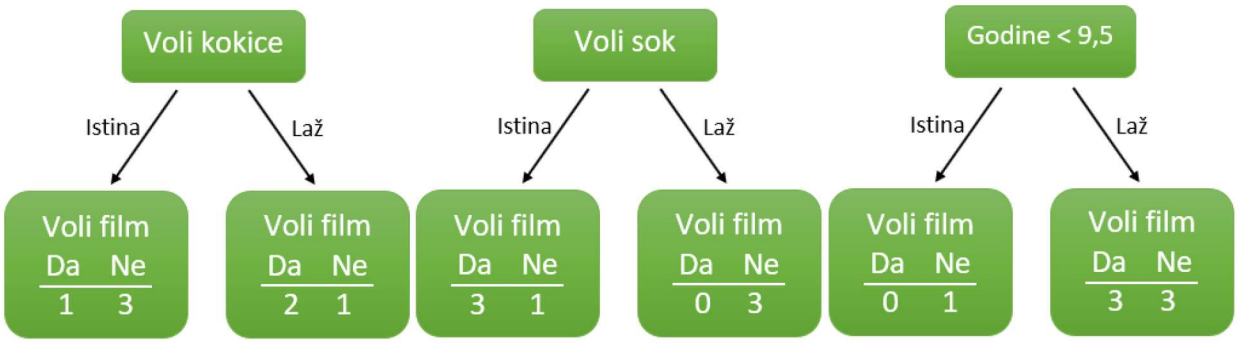
Gdje je

- y-ciljani atribut
- S-skup podataka
- $dom(y)$-skup svih vrijednosti ciljanog atributa y
- $P(y = c_j) = \frac{\text{broj podataka u kojima vrijedi } y=c_j}{\text{broj podataka u } S}$ -vjerojatnost da je u podatku atribut y jednak c_j

U nastavku ćemo na primjeru objasniti kako se gini indeks koristi. U ovom primjeru koristit ćemo mali skup podataka zbog lakšeg objašnjenja. Ulazni podaci prikazani su u Tablici 1 i odnose se na sedam osoba. Varijabla 'Voli film' je ciljana varijabla.

Voli kokice	Voli sok	Godine	Voli film
Da	Da	7	Ne
Da	Ne	12	Ne
Ne	Da	18	Da
Ne	Da	35	Da
Da	Da	38	Da
Da	Ne	50	Ne
Ne	Ne	83	Ne

Tablica 1: Ulazni podaci



Slika 3: Ilustracija stabala za lakši izračun gini indeksa

Prvo trebamo odrediti što će nam biti u korijenu. Promatramo varijablu 'Voli kokice'. Iz *Slike 3* vidimo da u lijevi čvor idu osobe za koje vrijedi da vole kokice. Od tih osoba koje vole kokice jedna voli film, a tri ne. Izračunajmo gini indeks za taj čvor.

$$Gini = 1 - (\text{vjerojatnost od 'Da' u čvoru})^2 - (\text{vjerojatnost od 'Ne' u čvoru})^2$$

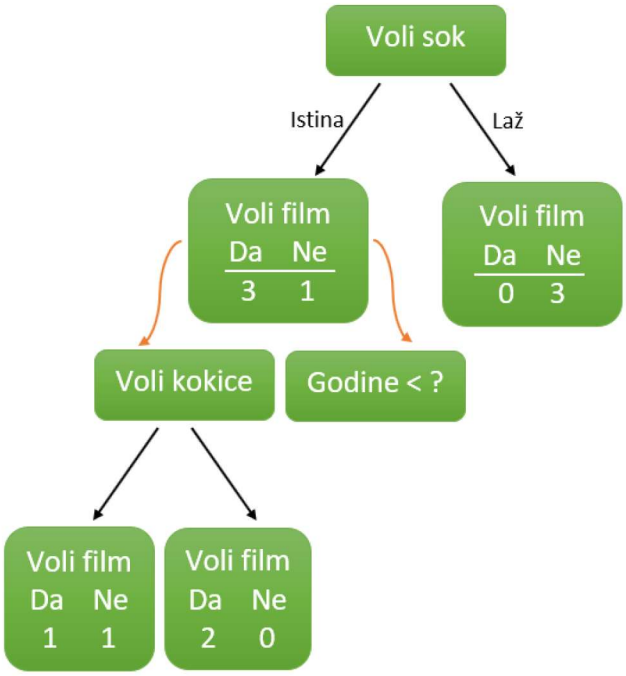
$$Gini = 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2 = 0,375$$

Na isti se način izračuna gini indeks i u desnom čvoru u kojem dobijemo da je gini indeks jednak 0,444. Pa je ukupna Gini nečistoća za varijablu 'Voli kokice' jednaka

$$\text{Gini nečistoća} = \frac{\text{broj podataka u lijevom čvoru}}{\text{ukupan broj podataka u lijevom i desnom čvoru}} \cdot \text{gini indeks u lijevom čvoru}$$
$$+\frac{\text{broj podataka u desnom čvoru}}{\text{ukupan broj podataka u lijevom i desnom čvoru}} \cdot \text{gini indeks u desnom čvoru}$$

$$\text{Gini nečistoća} = \frac{4}{4+3} \cdot 0,375 + \frac{3}{4+3} \cdot 0,444 = 0,405$$

Na isti način izračunamo i za varijablu 'Voli sok' i dobijemo da gini nečistoća iznosi 0,214. Varijabla 'Godine' je numerička varijabla pa podatke prvo moramo sortirati tako da vrijednosti varijable 'Godine idu' od najmanje prema najvećoj. U Tablici 1 je to već napravljeno. Uzmemo svaka dva susjedna podatka i izračunamo njihovu aritmetičku sredinu i taj iznos stavimo kao uvjet u čvor za koji izračunamo gini nečistoću. Ako promatramo prva dva podatka, aritmetička sredina 7 i 12 je 9,5. Na *Slici 3* vidimo kako se podaci dijele u ovom slučaju. Za gini nečistoću dobije se 0,429. Na kraju gledamo koji čvor ima najmanju gini nečistoću i njega uzimamo. U našem slučaju najmanju nečistoću ima čvor u kojem je varijabla 'Volim sok' i nju stavljamo u korijen stabla. Spuštamo se iz korijena u lijevi čvor u kojem se nalaze podaci kod kojih vrijedi da varijabla 'Voli sok' ima vrijednost 'Da'. Na isti način kao i za korijen odabiremo što će biti u tom čvoru samo što sada promatramo samo podatke u kojima varijabla 'Voli sok' ima vrijednost 'Da'. Ako se iz korijena spustimo desnom granom dolazimo do čvora u kojem se nalaze podaci u kojima varijabla 'Voli sok' ima vrijednost 'Ne'. Iz *Slike 4* vidimo da imamo samo tri takva podatka pri čemu za sva tri vrijedi da varijabla 'Voli film' ima vrijednost 'Ne' što znači da su nam svi podaci u jednom skupu pa u ovoj grani ne možemo više dijeliti podatke. (vidi [18])



Slika 4: Ilustracija stabla za lakši izračun gini indeksa

Druga funkcija nečistoće koju ćemo koristiti je entropija koja je definirana sa:

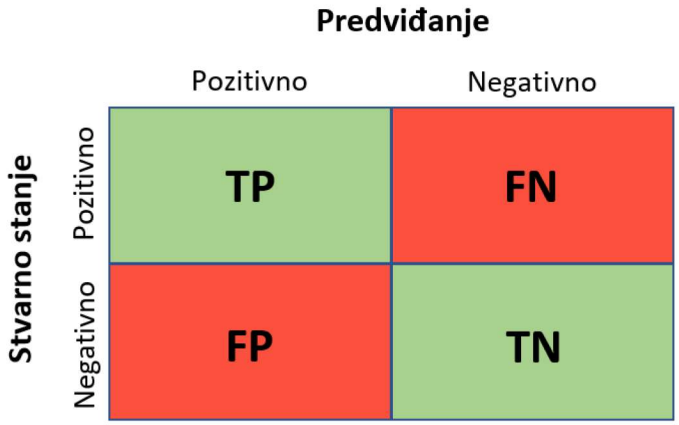
$$Entropy(y, S) = -\sum_{c_j \in dom(y)} P(y = c_j) \cdot \log_2 P(y = c_j)$$

Za obje funkcije nečistoće vrijedi da što je vrijednost manja to je čvor 'čistiji.' (vidi [7])

1.2 Preciznost klasifikacije

Prije izgradnje stabla skup podataka se obično dijeli u nekom omjeru u dva dijela, dio u kojem su podaci za treniranje, odnosno podaci za izradu stabla i dio u kojem su podaci za testiranje klasifikacije dobivene stablom. Postoji niz mjera kojima se može ocijeniti dobiveno stablo odlučivanja. Prvo ćemo navesti oznake, koje su prikazane i na *Slici 5*, pomoću kojih ćemo lakše objasniti mjere za ocjenjivanje dobivenog stabla. Pretpostavit ćemo da ciljani atribut poprima vrijednosti "pozitivno" i "negativno".(vidi [14])

- TP (*engl.* true positive)- broj pozitivnih podataka koji su ispravno klasificirani
- FP (*engl.* false positive)- broj pozitivnih podataka koji nisu ispravno klasificirani
- FN (*engl.* false negative)- broj negativnih podataka koji nisu ispravno klasificirani
- TN - (*engl.* true negative)- broj negativnih podataka koji su ispravno klasificirani



Slika 5: Matrica zabune (*engl.* confusion matrix)

Klasifikacijska točnost

Klasifikacijska točnost ili, skraćeno, točnost (*engl.* accuracy) je omjer broja podataka kojima je točno predviđena klasa i ukupnog broja podataka.

$$\text{Točnost} = \frac{\text{TP} + \text{TN}}{\text{Ukupan broj podataka}}$$

Klasifikacijska greška

Klasifikacijska greška ili, skraćeno, greška je omjer broja podataka kojima je krivo predviđena klasa i ukupnog broja podataka.

$$\text{Greška} = \frac{\text{FP} + \text{FN}}{\text{Ukupan broj podataka}}$$

Preciznost

Preciznost je omjer broja pozitivnih podataka koji su ispravno klasificirani i ukupnog broja pozitivnih podataka

$$\text{Preciznost} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Osjetljivost

Osjetljivost je omjer broja pozitivnih podataka koji su ispravno klasificirani i broja podataka koji su stvarno pozitivni.

$$\text{Osjetljivost} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specifičnost

Specifičnost je omjer broja negativnih podataka koji su ispravno klasificirani i broja podataka koji su stvarno negativni.

$$\text{Specifičnost} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

Greška tipa 1

Greška tipa 1 se često označava s α. To je omjer broja negativnih podataka koji su krivo klasificirani i broja podataka koji su stvarno negativni.

$$\alpha = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

Greška tipa 2

Greška tipa 2, često i s oznakom β, je omjer broja pozitivnih podataka koji su krivo klasificirani i broja podataka koji su stvarno pozitivni.

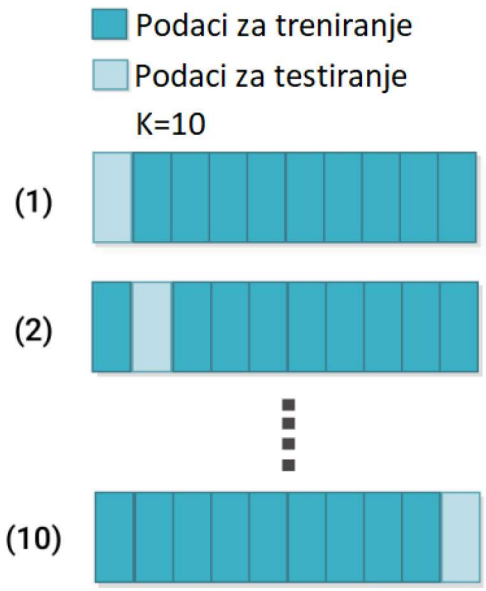
$$\beta = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}$$

ROC krivulja

Prema [2] ROC (*engl.* receiver operating characteristic) krivulja se često koristi kod ocjenjivanja klasifikacijske sposobnosti modela. Ona prikazuje kumulativne frekvencije pozitivnih podataka na x-osi i kumulativne frekvencije negativnih podataka a y-osi. Što je ona konkavnija to je model bolji. Ako je dijagonalna to bi značilo da model uopće ne razlikuje pozitivne i negativne podatke. AUC je površina područja ispod ROC krivulje. Što je AUC veći to je model bolji. Ako je AUC vrijednost manja od 0,5 tada je model loše prilagođen. Kod AUC vrijednosti veće od 0,7 model je dobro prilagoden, a za AUC vrijednost veću od 0,8 kažemo da je model izvrsno prilagođen. AUC vrijednost još možemo interpretirati kao vjerojatnost da će nasumično izabran loš klijent iz uzorka imati lošiji rejting od nasumično odabranog dobrog klijenta.

Unakrsna validacija

Prije gradnje stabla, kao što smo već naveli, podatke obično podijelimo na dva dijela, dio koji koristimo za treniranje i dio za testiranje. Kod malog seta podataka oni se često ne dijele u dva dijela već se koristi unakrsna validacija (*engl.* cross validation). Ona se može koristiti i u slučajevima u kojima podatke dijelimo na dio za treniranje i na dio za testiranje. Kod nje prvo skup podataka za trening podijelimo na K dijelova. Zatim se obavlja K iteracija gdje u svakoj iteraciji jedan dio služi za testiranje stabla, a preostalih $K - 1$ za izradu stabla. U svakoj iteraciji izračunamo klasifikacijsku točnost i kao rezultat unakrsne validacije uzimamo prosječnu klasifikacijsku točnost svih K iteracija. (vidi [15])



Slika 6: Ilustracija unakrsne validacije

1.3 Podrezivanje stabla

Kod nekih klasifikacijskih stabala može doći do "overfitting"-a podataka, tj. generira se dovoljno kompleksno stablo koje kod podataka za treniranje klasificira točno svaki podatak. Prema [13] takvo bi stablo bilo vrlo osjetljivo jer bi male promjene podataka za treniranje uzrokovale veliku promjenu u predviđenim klasama. Tada model ne bi dobro radio na novim podacima. Za izbjegavanje "overfitting"-a postoje dvije mogućnosti. Prva je da se proces izgradnje stabla zaustavi prije nego se postigne savršena klasifikacija dok je drugi način da se generira takvo stablo i zatim se određene grane stabla skraćuju prema prethodno definiranom kriteriju. Taj proces zove se podrezivanje. (vidi [17])

Podrezivanje stabla znači uklanjanje čvora i pripadnog podstabla koje ima korijen u tom čvoru, zamjenjujući ga s listom tako da se listu pridruži najčešća vrijednost ciljanog aributa u tom podčvoru.

Podrezivanjem punog stabla dobiveno stablo će biti manje točno na podacima s kojima smo kreirali puno stablo, ali potencijalno točnije u fazi predviđanja. (vidi [4])

Podrezano stablo će biti manje dubine nego puno stablo, što znači da ima manje pravila, a time je i računalno manje zahtjevnije.

Za podrezivanje stabla koristi se formula

$$R_\alpha(T_t) = R(T_t) + \alpha \cdot |T_t|, \ t \in \mathbb{N}_0$$

gdje je $R(T_t)$ ukupna klasifikacijska greška u listovima, $|T_t|$ broj listova, a α parametar složenosti. Ako je $\alpha = 0$ tada će se odabrati puno stablo. (vidi [13])

U nastavku je opisan proces podrezivanja stabla kod kojeg se prvo izgradi puno stablo, odnosno stablo gore navedenim algoritmom i zatim se skraćuje.

1. Izradite puno stablo.

2. Izračunajte $R_\alpha(T_t)$ svakog podstabla koje sadrži barem jedan list stabla koje promatramo.

3. Podrežite promatrano stablo od listova prema korijenu tako da se u svakom koraku podrezivanja $R_\alpha(T_t)$ minimizira.

4. Ponavljaj prethodna dva koraka sve dok $R_\alpha(T_t)$ promatranog stabla ne bude manje od svih ostalih stabala. (vidi [3])

2 Kreditno skoriranje

Modeli kreditnog skoriranja služe nam za predviđanje kreditne sposobnosti zajmotražitelja. Kod odobravanja kredita zajmotražitelju koji neće na vrijeme izvršavati svoje obaveze, kreditori snose gubitak dok odbijanje kredita potecijalno dobrog zajmotražitelja kreditorima nosi manju zaradu. Model kreditnog skoriranja je sistem koji nam daje informaciju koliko je vjerojatno da zajmotražitelj kasni u otplati kredita.

Model kreditnog skoriranja obuhvaća skor-karticu i skup statističkih pokazatelja. Koraci u izgradnji skor-kartice su:

1. Studija provedivosti - provjeravanje da li uopće ima potrebe i mogućnosti za izgradnju skor-kartice

2. Definicija uzorka i prikupljanje podataka - prikupljanje podataka o zajmotražiteljima i određivanje o kakvom se zajmotražitelju radi

3. Analiza karakteristika - analiziranje karakteristika zasebno za dobre i zasebno za loše klijente

4. Zaključivanje o odbijenim komitetima - zaključivanje o tome kakvi bi bili odbijeni klijenti da su bili prihvačeni

5. Modeliranje skor-kartice - analiziranje povijesnih podataka prethodno odobrenih kredita te određivanje značajki zajmotražitelja koje su važne u predviđanju individualnog rizika

6. Validacija skor-kartice - testiranje modela kreditnog skoriranja primjenom različitih kvalitativnih i kvantitativnih testova

7. Postavljanje strategije i implementacija (vidi [9])

Modeli kreditnog skoriranja mogu se podijeliti na aplikativne modele, koji se odnose na nove klijente, i na bihevioralne, koji se odnose na postojeće klijente. Modeli bihevioralnog skoriranja su nastavak razvoja kreditnog skoriranja. Kod njih se za izradu modela uz dobivene početne podatke dobivene kod podnošenja zahtjeva za kredit, koriste i informacije o ponašanju klijenta pri otplati kreditnih obveza u prošlom periodu, koji se zove period promatranja. U većini slučajeva informacije o povijesnim performansama i trenutne informacije o otplati kredita klijenta više su kvalitetnije nego informacije dobivene kod podnošenja zahtjeva kredita. (vidi [6])

2.1 Stabla odlučivanja u kreditnom skoriranju

Modeli kreditnog skoriranja mogu se graditi pomoću različitih metoda kao što su neuronske mreže, logistička regresija, ali i stabla odlučivanja. Primjeri takvih istraživanja biti će navedeni u nastavku.

Prethodna istraživanja

Vojtek i Kočenda su u [5] iz podataka jedne Češke banke, imali cilj izgraditi moćan model kreditnog skoriranja za nova tržista Europske unije. Modele su gradili na dvjema grupama varijabli. Prva grupa varijabli su sociodemografske varijable. Zanimljivo je da banka ne evidentira informacije o prihodima i rashodima već samo omjere. Prvi omjer je postotak utrošenog dohotka podijeljen sa rashodima (Credit ratio 1), a drugi omjer je dohodak klijenta podijeljen sa službenom minimalnom plaćom u trenutku davanja zahtjeva za kredit (Credit ratio 2). Ostale varijable su: spol, bračni status, datum rođenja, sektor u kojem je klijent zaposlen, vrsta zaposlenja, obrazovanje, broj dosadašnjih poslova, br. godina rada kod trenutnog zaposlenja i mjesto stanovanja. U drugoj grupi varijabli bilježi se odnos između klijenta i banke. Zanimljiva je varijabla „Points" koja opisuje odnos klijenta prema vlastitom računu u toj banci. Ostale varijable su: vrsta kredita, broj jamaca, svrha kredita, osiguranje kredita, vlastiti resursi, iznos kredita, godina otvaranja računa, datum zahtjeva za kredit i duljina isplata prijašnjih kredita. Kod izgradnji modela koristili su dvije metode. Prva je bazirana na logističkoj regresiji, a druga na CART (*engl.* classification and regression tree) metodi. Obje metode se često koriste u razvoju modela. Konstruirali su tri različita modela koristeći logističku regresiju i jedan koristeći CART te ih usporedili. Zatim su promatrali koje varijable imaju značajan utjecaj na rezultat, a koje ne. Obje metode potvrdile su slične najutjecajnije varijable, a to su: količina resursa koje klijent posjeduje, stupanj obrazovanja, bračni status, svrha zajma i godina otvaranja računa u banci.

J. Zurada i M. Zurada su u [11] gradili modele na temelju podataka jedne financijske institucije. Kod izgradnje modela koristili su tri različite metode. To su stabla odlučivanja, neuronske mreže i logistička regresija. Cilj im je bio usporediti navedene metode. Napravili su dva scernarija. U prvom su koristili izvorne podatke u kojima su sadržani podaci za 3064 dobrih klijenata i 300 loših klijenata dok su u drugom scenariju stvorili uravnotežen skup podataka u kojem su nasumično izabrali 300 dobrih klijenata pa je u drugom scenariju skup podataka sadržavao podatke o 600 klijenata. Prvo su koristili izvorni skup podataka dan u scenariju, a u drugom slučaju su izbacili varijable koje su bile u slaboj korelaciji s ciljanom varijablom. Kod trećeg slučaja su izvornih 13 varijabli grupirali u četiri varijable. Kod svakog slučaja izgradili su tri različita modela pomoću navedenih metoda te četvrti model koji je bio kombinacija svih triju metoda. Zaključili su da su najbolje rezultate imali kod korištenja neuronskih mreža i kod modela koji je kombinacija svih triju tehnika.

Satchidananda i Simha su u [8] koristili podatke o poljoprivrednim zajmovima dvaju indijskih banaka. Cilj im je bio usporediti učinkovitost stabla odlučivanja i logističke regresije u predikciji dobrih i loših klijenata, odnosno zajmotražitelja. Za izradu modela koristili su varijable socio-odemografskih, financijskih i poljoprivrednih obilježja. Zaključili su da stabla odlučivanja bolje klasificiraju dobre i loše klijente nego model logističke regresije.

Kvesić je u [6] koristila podatke jedne poslovne banke u kojima je bilo 100 dobrih i 100 loših klijenata. Cilj ovog rada je bio razviti model kreditnog skoriranja pomoću stabla odlučivanja koji će biti primjenjiv u poslovanju hrvatskih financijskih institucija, a time bi i ovaj rad popunio prazninu u ovom području u teoriji i praksi. Za izgradnju modela korištene su varijable spol, dob, status klijenata, minimalno dozvoljeno prekoračenje tijekom 6 mjeseci, prosječan iznos svih plaćanja čekovima, prosječan iznos plaćanja karticom, prosječan saldo za svih 6 mjeseci (prosječno stanje na računu), prosječan iznos sredstava koji klijent još smije potrošiti i broj nedozvoljenih prekoračenja (koliko je puta klijent bio u nedozvoljenom prekoračenju). Stablo odlučivanja je dobiveno primjenom iscrpnog CHAID algoritma. Ovim modelom 92,5% klijenata je bilo dobro klasificirano što je zadovoljavajuće. Analizom rezultata pokazalo se da se važnim prediktorima mogu smatrati varijable starost klijenta i broj nedozvoljenih prekoračenja.

3 Empirijsko istraživanje: Primjena stabla odlučivanja u izgradnji modela za procjenu kreditnih rizika stanovništva

Na raspolaganju za izgradnju modela su podaci, tj. uzorak podataka o kreditima stanovništva jedne banke u Hrvatskoj. Cilj je izgraditi dobar model kreditnog skoriranja za stanovništvo pomoću klasifikacijskog stabla odlučivanja.

Istraživanje ćemo započeti opisom uzorka i varijabli koje ćemo koristiti za izgradnju klasifikacijskih stabala. Izgradit ćemo klasifikacijska stabla pomoću gini indeksa i pomoću entropije, a zatim na oba dobivena stabla primijeniti metodu podrezivanja. Za sva četiri dobivena stabla provest ćemo validaciju i usporediti ih kako bi mogli zaključiti koje je najbolje, tj. pomoću kojeg stabla dobivamo najbolji model kreditnog skoriranja.

3.1 Opis uzorka i varijabli

Baza podataka sastoji se od 15 varijabli i podacima o 1763 klijenata od kojih je njih 899 označeno kao loši klijenti, a 864 kao dobri klijenti. Za kreiranje modela korišteno je 70% ukupnih podataka, tj. 614 loših klijenata i 620 dobrih. Preostalih 30% podataka (od kojih je 285 loših i 244 dobrih) korišteno je za validaciju modela.

3.2 Deskriptivna statistika varijabli

Varijabla 'dobarlos'

Varijabla 'dobarlos' je kategorijalna varijabla koja ima dvije kategorije. '1' označava loše klijente, tj. one koji su u promatranom vremenu kasnili u plaćanju barem jedne rate kredita tri mjeseca ili više, a '0' označava dobre klijente, tj. one koji nisu u promatranom vremenu kasnili u plaćanju barem jedne rate kredita 3 mjeseca ili više. U Tablici 2 su prikazane frekvencije i relativne frekvencije varijable 'dobarlos'.

	Frekvencija	Relativna frekvencija
0	864	0,490
1	899	0,510

Tablica 2: Frekvencija i relativna frekvencija varijable 'dobarlos'

Varijabla 'spol'

Varijabla 'spol' je kategorijalna varijabla.

	% dobrih	% loših	Ukupno
1-žene	46%	54%	807
2-muškarci	52%	48%	953
nema podataka	33%	67%	3

Tablica 3: Frekvencija i omjer dobrih i loših klijenata varijable 'spol'

Iz Tablice 3 možemo uočiti da je kod žena malo veći postotak loših klijenata nego dobrih, dok ih je kod muškaraca podjednako. Također možemo primijetiti da je više muškaraca nego žena uzimalo kredit. Testiranjem zavisnosti između spola klijenta i urednosti vraćanja kredita upotrebom χ^2 - testa dobivamo $p = 0,03294 < 0,05$ iz čega možemo zaključiti da na razini značajnosti 0,05 postoji zavisnost, tj. ako je klijent dobar, veća je vjerojatnost da je klijent muškarac, a ne žena.

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'spol' ima smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'Depozit'

Varijabla 'Depozit' je numerička varijabla koja nam govori koliki je iznos depozita u HRK.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0-dobri	0,0	0,0	0,0	282,4	0,0	19726,0
1-loši	0,0	0,0	0,0	731,4	0,0	17567,7

Tablica 4: Elementarna statistika varijable 'Depozit'

	% dobrih	% loših	Ukupno
0	51%	49%	1610
0-5000	34%	66%	53
5000-8000	10%	90%	71
8000-13000	44%	56%	18
>13000	29%	71%	11
>0	24%	76%	153

Tablica 5: Frekvencija i omjer dobrih i loših klijenata varijable 'Depozit' s obzirom na urednost vraćanja kredita prema varijabli 'dobarlos'

Iz Tablice 5 se može uočiti da mali broj klijenata ima depozit. Među onima koji imaju depozit samo otprilike četvrtina klijenata je svrstana u dobre klijente. T-test je pokazao da postoji statistički značajna razlika u očekivanju za depozit između onih klijenata koji uredno vraćaju kredit i onih koji su loši ($p = 2,352e^{-06} < 0,05$) iz čega zaključujemo da je manjim depozitom veća vjerojatnost da klijent bude dobar.

S obzirom na provedenu analizu u kojoj vidimo da više od 90% klijenata ima depozit nula, tj. nema depozit u banci, možemo zaključiti da varijablu 'Depozit' nema smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'NazivRadnogMj'

Varijabla 'NazivRadnogMj' je kategorijalna varijabla koja ima 12 kategorija.

	% dobrih	% loših	Ukupno
0 - direktori	34%	66%	32
1 - vozači	37%	63%	92
2 - medicinsko, farmaceutsko i veterinarsko osoblje neuključujući doktore	39%	61%	69
3 - radnici u obrazovanju	44%	56%	75
4 - radnici s uredskim poslovima	42%	58%	456
5 - umirovljenici	48%	52%	33
6 - osobe u uslužnim djelatnostima	45%	55%	250
7 - svi ostali radnici	53%	47%	139
8 - radnici u proizvodnji	50%	50%	195
9 - automehaničari i radnici u metalnoj industriji i građevini	57%	43%	179
10 - radnici u policiji i vojsci	62%	38%	53
11 - doktori	69%	31%	13
12 - nema podataka	69%	31%	177

Tablica 6: Frekvencija i omjer dobrih i loših klijenata varijable 'NazivRadnogMj'

Iz Tablice 6 možemo primijetiti da je kod klijenata koji su na direktorskim pozicijama, vozači ili rade u medicini (neuključujući doktore) veći postotak loših klijenata nego dobrih dok je kod doktora najveći postotak dobrih klijenata. Testiranjem zavisnosti između radnog mjesta i urednosti vraćanja kredita upotrebom χ^2 - testa dobivamo $p = 7,489e^{-09} < 0,05$ iz čega možemo zaključiti da postoji zavisnost.

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'NazivRadnogMj' ima smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'StrucnaSpremaId'

Varijabla 'StrucnaSpremaId' je kategorijalna varijabla koja ima sedam kategorija.

	% dobrih	% loših	Ukupno
17 - NKV	63%	37%	121
18 - PKV	60%	40%	20
19 - NSS	56%	44%	18
20 - KV	56%	44%	85
21 - SSS	44%	56%	996
22 - VŠS	50%	50%	115
23 - VSS,mr.,dr.	43%	57%	173
nema podataka	64%	36%	235

Tablica 7: Frekvencija i omjer dobrih i loših klijenata varijable 'StrucnaSpremaId'

Iz Tablice 7 možemo vidjeti da najveći postotak dobrih klijenata ima kod nekvalificiranih i polukvalificiranih radnika, dok najmanji postotak dobrih klijenata je kod klijenata koji imaju srednju stručnu spremu i kod visokokvalificarnih klijenata. Testiranjem zavisnosti između naziva radnog mjesta i urednosti vraćanja kredita upotrebom χ^2 - testa dobivamo $p = 6,397e^{-08} < 0,05$ iz čega možemo zaključiti da postoji zavisnost između tih dviju varijabli.

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'StrucnaSpremaId' ima smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'RadniStazUkGOD'

Varijabla 'RadniStazUkGOD' je numerička varijabla koja nam govori koliko radnog staža ima klijent.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0-dobri	0,000	9,127	18,000	17,669	26,000	39,000
1-loši	0,000	7,000	15,000	16,200	25,000	43,330

Tablica 8: Elementarna statistika varijable 'RadniStazUkGOD' s obzirom na urednost vraćanja kredita prema varijabli 'dobarlos'

	% dobrih	% loših	Ukupno
0-15	43%	57%	725
15-30	50%	50%	644
>30	53%	47%	178
nema podataka	63%	37%	216

Tablica 9: Frekvencija i omjer dobrih i loših klijenata varijable 'RadniStazUkGOD'

Ako iz Tablice 9 izuzmemo klijente koji nemaju radnog staža, možemo reći da se povećanjem radnog staža klijenta povećava i omjer dobrih klijenata naspram loših. T-test je pokazao da

postoji statistički značajna razlika u očekivanju za ukupan broj godina radnog staža između onih koji uredno vraćaju kredit i onih koji su loši ($p = 0,0003857 < 0,05$).

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'RadniStazUkGOD' ima smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'BracnoStanjeId'

Varijabla 'BracnoStanjeId' je kategorijalna varijabla koja nam govori kakav je bračni status klijenta.

	% dobrih	% loših	Ukupno
2-samac	39%	61%	459
3-u braku	54%	46%	1031
4-rastavljen	43%	57%	161
5-udovac	52%	48%	67
6-izvanbračna zajednica	47%	53%	30
nema podataka	40%	60%	15

Tablica 10: Frekvencija i omjer dobrih i loših klijenata varijable 'BracnoStanjeId'

Iz Tablice 10 vidimo da je najbolji omjer dobrih klijenata kod klijenata koji su u braku, a najlošiji kod samaca. Testiranjem zavisnosti između bračnog stanja i urednosti vraćanja kredita upotrebom χ^2 - testa dobivamo $p = 8,131e^{-06} < 0,05$ iz čega možemo zaključiti da postoji zavisnost između tih dviju varijabli.

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'BracnoStanjeId' ima smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'BrojClanovaKucanstva'

Varijabla 'BrojClanovaKucanstva' je kategorijalna varijabla koja nam govori koliko članova živi u kućanstvu klijenta.

	% dobrih	% loših	Ukupno
1	40%	60%	178
2	50%	50%	326
3	45%	55%	437
4	54%	46%	511
≥ 5	54%	46%	269
nema podataka	33%	67%	42

Tablica 11: Frekvencija i omjer dobrih i loših klijenata varijable 'BrojClanovaKucanstva'

Iz Tablice 11 vidimo da je kod klijenata koji žive u kućanstvu sa 2 člana jednak omjer dobrih i loših klijenata. Sa 4 i više članova veći je postotak dobrih klijenata nego loših, dok je kod ostalih obratno. Testiranjem zavisnosti između broja članova kućanstva i urednosti vraćanja kredita

upotrebom χ^2 - testa dobivamo $p = 0,00703 < 0,05$ iz čega možemo zaključiti da postoji zavisnost između tih dviju varijabli.

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'BrojClanovaKucanstva' ima smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'BrojClanovaKucanstvaStalnaPrimanja'

Varijabla 'BrojClanovaKucanstvaStalnaPrimanja' je kategorijalna varijabla koja nam govori koliko članova u kućanstvu klijenta ima stalna primanja.

	% dobrih	% loših	Ukupno
0-1	48%	52%	428
2	51%	49%	813
3	49%	51%	310
≥ 4	45%	55%	112
nema podataka	43%	57%	100

Tablica 12: Frekvencija i omjer dobrih i loših klijenata varijable 'BrojClanovaKucanstvaStalnaPrimanja'

Iz Tablice 12 možemo uočiti da je jedino kod klijenata koji u kućanstvu imaju 2 člana sa stalnim prihodima veći postotak dobrih klijenata dok je kod ostalih obratno. Testiranjem zavisnosti između broja članova kućanstva sa stalnim primanjima i urednosti vraćanja kredita upotrebom χ^2 - testa dobivamo $p = 0,4954 > 0,05$ iz čega ne možemo zaključiti da postoji zavisnost između tih dviju varijabli.

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'BrojClanovaKucanstvaStalnaPrimanja' nema smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'StanovanjeId'

Varijabla 'StanovanjeId' je kategorijalna varijabla koja nam govori kakav je stambeni status klijenta.

	% dobrih	% loših	Ukupno
12-vlastiti stambeni prostor	52%	48%	1126
13-unajmljeni stambeni prostor	36%	64%	58
14-kod roditelja	46%	54%	436
15-ostalo	34%	66%	48
nema podataka	40%	60%	95

Tablica 13: Frekvencija i omjer dobrih i loših klijenata varijable 'StanovanjeId'

Iz Tablice 13 vidimo da je najveći postotak dobrih klijenata onih koji žive u vlastitom prostoru, dok je najmanji kod osoba koji žive u unajmljenom prostoru ili u nečem trećem. Testiranjem

zavisnosti između vrste smještaja i urednosti vraćanja kredita upotrebom χ^2 - testa dobivamo $p = 0,0004189 < 0,05$ iz čega možemo zaključiti da postoji zavisnost između tih dviju varijabli.

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'StanovanjeId' ima smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'NetoPlaca'

Varijabla 'NetoPlaca' je numerička varijabla koja pokazuje kolika je neto plaća klijenta.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0-dobri	935,2	2659,0	3789,3	6764,3	5050,0	2224087,0
1-loši	1174	2898	4002	4367	5153	24548

Tablica 14: Elementarna statistika varijable 'NetoPlaca' s obzirom na urednost vraćanja kredita prema varijabli 'dobarlos'

	% dobrih	% loših	Ukupno
0-2500	58%	42%	312
2500-5000	48%	52%	950
5000-10000	45%	55%	458
>10000	52%	48%	33
nema podataka	70%	30%	10

Tablica 15: Frekvencija i omjer dobrih i loših klijenata varijable 'NetoPlaca'

Iz Tablice 15 vidimo da je veći postotak dobrih klijenata nego loših kod klijenata čija je neto plaća manja ili jednaka 2500 ili veća od 10000, dok je kod klijenata čije su neto plaće između 2500 i 5000 te 5000 i 10000 veći postotak loših klijenata. T-test je pokazao da ne možemo reći da postoji statistički značajna razlika u očekivanju neto plaće između onih koji uredno vraćaju kredit i onih koji su loši ($p = 0,3445 > 0,05$).

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'NetoPlaca' nema smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'MjesecneObustave'

Varijabla 'MjesecneObustave' je kategorijalna varijabla koja nam govori kolike su mjesečne obustave klijenta na plaću.

	% dobrih	% loših	Ukupno
6 (0)	52%	48%	786
5 (0-300)	41%	59%	110
4 (300-600)	54%	46%	123
3 (600-1000)	52%	48%	186
2 (1000-2000)	49%	51%	231
1 (2000<)	52%	48%	124
7 (nema podataka)	33%	67%	203

Tablica 16: Frekvencija i omjer dobrih i loših klijenata varijable 'MjesecneObustave'

Iz Tablice 16 vidimo da je najveći postotak dobrih klijenata nego loših kod klijenata čije su mjesečne obustave od 300 do 600, dok je kod klijenata gdje su mjesečne obustave između 0 i 300 najveći postotak loših klijenata neuzimajući u obzir klijente gdje nemamo podataka. χ^2 - test je pokazao da postoji zavisnost između varijable 'MjesecneObustave' i urednosti vraćanja kredita ($p = 5,171e^{-05} < 0,05$).

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'MjesecneObustave' ima smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'suduznik'

Varijabla 'suduznik' je kategorijalna varijabla koja ima dvije kategorije.

	% dobrih	% loših	Ukupno
0-nema sudužnika	48%	52%	1899
1-ima sudužnika	78%	22%	40

Tablica 17: Frekvencija i omjer dobrih i loših klijenata varijable 'suduznik'

Iz Tablice 17 vidimo da je veći postotak dobrih klijenata kod klijenata koji imaju sudužnika nego kod onih koji nemaju sudužnika. Testiranjem zavisnosti između varijable 'suduznik' i urednosti vraćanja kredita upotrebom χ^2 - testa dobivamo $p = 0,0004896 < 0,05$ iz čega možemo zaključiti da postoji zavisnost između tih dviju varijabli.

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'suduznik' ima smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'jamac'

Varijabla 'jamac' je kategorijalna varijabla koja ima dvije kategorije.

	% dobrih	% loših	Ukupno
0-nema jamca	48%	52%	1676
1-ima jamca	60%	40%	87

Tablica 18: Frekvencija i omjer dobrih i loših klijenata varijable 'jamac'

Iz Tablice 18 vidimo da kod klijenata koji imaju jamca je veći postotak dobrih klijenata u odnosu na klijente koji nemaju jamca. Testiranjem zavisnosti između varijable 'jamac' i urednosti vraćanja kredita upotrebom χ^2 - testa dobivamo $p = 0,05122 > 0,05$ iz čega ne možemo zaključiti da postoji zavisnost između tih dviju varijabli.

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'jamac' nema smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Varijabla 'TipPoslodavcaId'

Varijabla 'TipPoslodavcaId' je kategorijalna varijabla koja nam govori kod kojeg tipa poslodavca je klijent zaposlen.

	% dobrih	% loših	Ukupno
29-trgovačko društvo	45%	55%	940
31-obrt	47%	53%	77
32-slobodna zanimanja	50%	50%	14
33-javna uprava	47%	53%	159
34-javne ustanove	53%	47%	339
35-financijske institucije	44%	56%	32
36-umirovljenik	67%	33%	167
37-ostalo	48%	52%	21
nema podataka	50%	50%	14

Tablica 19: Frekvencija i omjer dobrih i loših klijenata varijable 'TipPoslodavcaId'

Iz Tablice 19 vidimo da kod umirovljenika imamo najveći postotak dobrih klijenata, a najmanji je kod klijenata koji rade u trgovačkim društvima i financijskim ustanovama. Testiranjem zavisnosti između tipa poslodavca klijenta i urednosti vraćanja kredita upotrebom χ^2 - testa dobivamo $p = 0,0002464 < 0,05$ iz čega možemo zaključiti da postoji zavisnost između tih dviju varijabli.

S obzirom na provedenu analizu, možemo zaključiti da varijablu 'TipPoslodavcaId' ima smisla uključiti u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

3.3 Rezultati stabla odlučivanja

Iz bivarijatne analize dobili smo da varijable 'Depozit', 'BrojClanovaKucanstvaStalnaPrimanja', 'NetoPlaca' i 'jamac' nema smisla uključivati u razvoj modela kreditnog skoriranja za procjenu kreditnog rizika.

Za izgradnju klasifikacijskih stabala koristili smo sve varijable i u nastavku ćemo usporediti s bivarijatnom analizom da li se dolazi do istih zaključaka.

Za izradu stabala koristili smo program Python, točnije njegov paket NumPy. NumPy se koristi za znanstveno i numeričko računanje.

U *Kodu* 1 možemo vidjeti da je 70% podataka korišteno za treniranje, tj. za izradu stabala, a 30% za testiranje.

Tih 70% podataka koristiti ćemo za izgradnju punih klasifikacijskih stabala pomoću gini indeksa i pomoću entropije. Ta stabla ćemo zatim pomoću metode podrezivanja podrezati i za oba puna i podrezana stabla provesti unakrsnu validaciju.

```
y = podaci.dobarlos # Zavisna varijabla
X = podaci[feature_cols] # Ostale varijable
# Razdvanje podataka na set za treniranje i set za testiranje
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

Kod 1: Ulazni podaci

Gini indeks

U *Kodu* 2 prikazan je kod za izradu punog klasifikacijskog stabla pomoću gini indeksa.

```
### PUNO STABLO
import numpy as np
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
clf = DecisionTreeClassifier(criterion="gini", random_state=42)

clf = clf.fit(X_train,y_train)

y_pred = clf.predict(X_test)

# Točnost modela
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

Kod 2: Izrada klasifikacijskog stabla pomoću gini indeksa

Kod izgradnje punog klasifikacijskog stabla pomoću gini indeksa 61% klijenata je ispravno klasificirano. Dubina tog stabla je 25 što znači da je to stablo jako veliko i u njemu je zadano mnogo pravila.

U *Kodu* 3 je postupak provođenja unakrsne validacije u 10 koraka nad punim klasifikacijskim stablom dobivenim pomoću gini indeksa. Prosječni postotak točnosti primjenom unakrsne validacije iznosi 56%.

```
### UNAKRSNA VALIDACIJA
from sklearn import datasets
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import KFold, cross_val_score, cross_validate

k_folds = KFold(n_splits = 10)

scores = cross_val_score(clf, X, y, cv = k_folds)

print("Prosječna uspješnost svih K iteracija: ", scores.mean())
```

Kod 3: Provođenje unakrsne validacije nad punim klasifikacijskim stablom dobivenim pomoću gini indeksa

Način na koji se u programu radi podrezivanje stabla napisan je u *Kodu* 4. Podrezivanjem stabla dobili smo stablo točnosti 63%. Podrezano stablo je dubine 6, što je puno manje od dubine punog stabla. Kod podrezanog stabla korištene su samo varijable 'RadniStazUkGOD', 'MjesecneObustave' ,'TipPoslodavcaId', 'Depozit', "NazivRadnogMj' i 'NetoPlaca'. Provođenjem unakrsne validacije u 10 koraka prosječan postotak točnosti kod podrezanog stabla iznosi 61%.

```
### PODREZANO STABLO
from sklearn.model_selection import train_test_split, ParameterGrid, GridSearchCV
ccp_alphas=clf.cost_complexity_pruning_path(X_train, y_train)["ccp_alphas"]
```

```
ccp_alpha_grid=GridSearchCV(
estimator=DecisionTreeClassifier(criterion="gini",random_state=42),
scoring=make_scorer(accuracy_score),
param_grid=ParameterGrid({"ccp_alpha":[[alpha] for alpha in ccp_alphas]}
                         ),)
```

```
ccp_alpha_grid.fit(X_train,y_train)
```

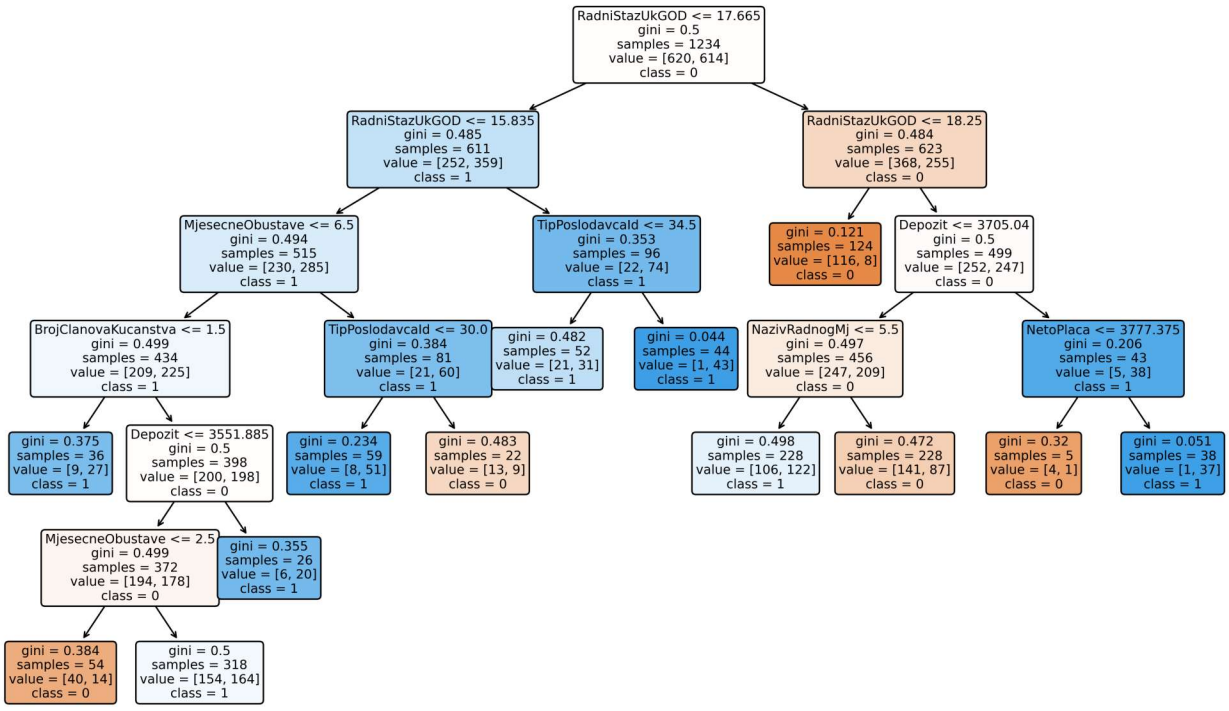
```
GridSearchCV(estimator=DecisionTreeClassifier(random_state=42),
             param_grid=<sklearn.model_selection._search.ParameterGrid object at 0x000001CE8001EFA0>,
             scoring=make_scorer(accuracy_score))
```

```
ccp_alpha_grid.best_params_
```

```
{'ccp_alpha': 0.002143164515726853}
```

```
best_ccp_alpha_tree=ccp_alpha_grid.best_estimator_
```

Kod 4: Podrezivanje klasifikacijskog stabla izgrađenog pomoću gini indeksa



Slika 7: Podrezano klasifikacijsko stablo izgrađeno pomoću gini indeksa

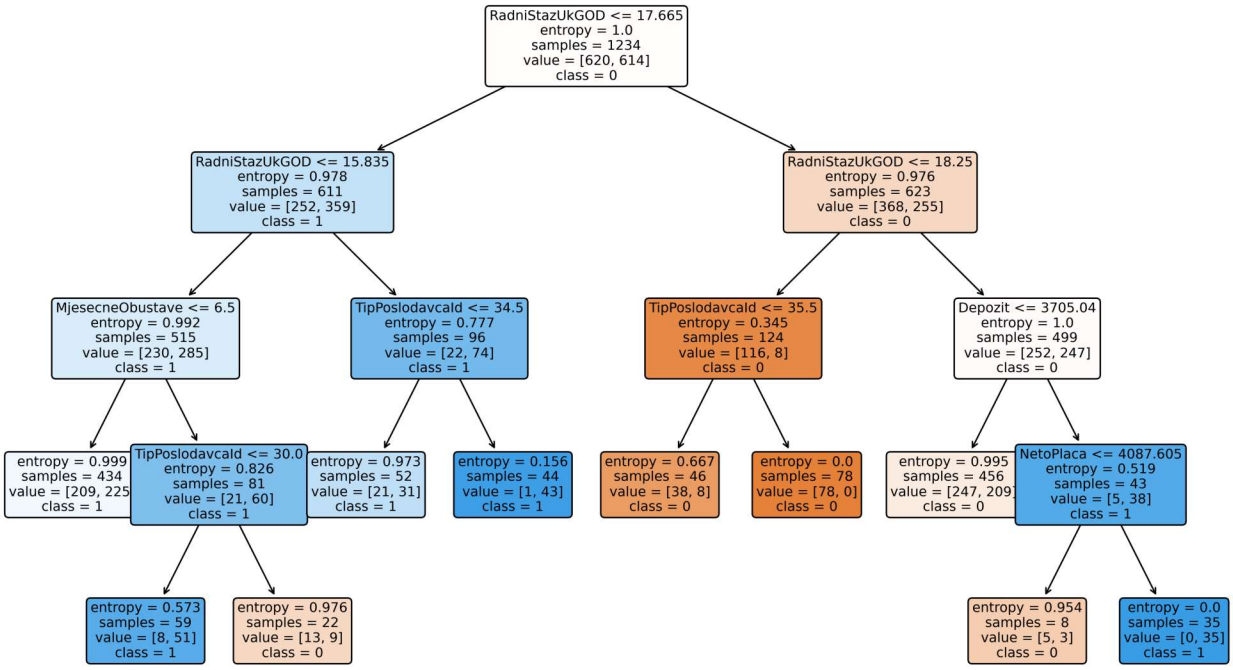
Promotrimo *Sliku 7* na kojoj je prikazano podrezano klasifikacijsko stablo izgrađeno pomoću gini indeksa. Iz ilustracijskog prikaza stabla možemo lako vidjeti da li će navedeno stablo klijenta svrstati u loše ili dobre klijente. U nastavku ćemo navesti jedan primjer za lošeg i jedan za dobrog klijenta. Krećemo od varijable 'RadniStazUkGOD'. Ako je njezina vrijednost promatranog klijenta manja ili jednaka od 17,665 tada se spuštamo u lijevu granu. Sada opet promatramo varijablu 'RadniStazUkGOD'. U ovom koraku gledamo ako je vrijednost manja ili jednaka od 15,835. Ako je, opet se spuštamo lijevom granom gdje je nam stoji uvjet da je varijabla 'MjesecneObustave' manja ili jednaka od 6,5. Ako uvjet vrijedi, opet se spuštamo u lijevu granu gdje dolazimo do uvjeta da li je varijabla 'BrojClanovaKucanstva' manja od 1,5. Ako je uvjet točan, klijent se svrstava u klasu '1' koja označava loše klijente. Ako pak imamo klijenta kod kojeg je vrijednost varijable 'RadniStazUkGOD' veća od 17,665, ali i manja od 18,25 tada klijent pripada klasi '0', tj. dobrim klijentima. Na ovaj način možemo za svakog klijenta provjeriti kojoj će klasi pripasti.

Entropija

Kod izgradnje punog stabla pomoću entropije 59% klijenata je ispravno klasificirano što je 2 postotna poena manje nego kod korištenja gini indeksa. Dubina ovog stabla je još veća i iznosi 27.

Kod podrezivanja stabla dobili smo stablo točnosti 61%, ali dubine 4 što znači da sad imamo stablo i manje dubine i veće točnosti. Podrezivanjem punog stabla dobivenog pomoću kriterija entropije dobili smo stablo u kojem se koriste varijable 'RadniStazUkGOD', 'MjesecneObustave', 'TipPoslodavcaId', 'Depozit' i 'NetoPlaca'.

Provođenjem unakrsne validacije u 10 koraka kod podrezanog stabla dobili smo da je prosječna točnost 61%.



Slika 8: Podrezano klasifikacijsko stablo izgrađeno pomoću entropije

Promatrajući *Sliku 8* na kojoj je prikazano podrezano klasifikacijsko stablo izgrađeno pomoću entropije vidimo da će klijent koji ima vrijednost varijable 'RadniStazUkGOD' veću od 17,665, ali i veću do 18,25 te vrijednost varijable 'Depozit' manju od 3705,04 biti svrtan u klasu '0', tj. u dobre klijente.

Kod izrade podrezanih stabala ni u jednom stablu se nisu koristile varijable 'BrojClanovaKucanstvaStalnaPrimanja' i 'jamac' kao ni varijable 'spol', 'StrucnaSpremaId', 'BracnoStanjeId', 'StanovanjeId' i 'suduznik' za koje smo bivarijatnom analizom zaključili da ih ima smisla uključiti u model. Za varijable 'Depozit' i 'NetoPlaca' bivarijatnom analizom smo zaključili da ih nema smisla uključiti u model, ali su se obje varijable koristile za izradu podrezanih stabala.

3.4 Interpretacija rezultata i diskusija

U ovom dijelu izračunat ćemo mjere pojedinog stabla kako bi odredili preciznost klasifikacije te ih zatim usporediti.

	Točnost	Greška	Prosječna točnost unakrsne validacije	AUC vrijednost
puno stablo	61%	39%	56%	0,61
podrezano stablo	63%	37%	61%	0,67

Tablica 20: Rezultati stabala dobivenih pomoću gini indeksa

Matrica zabune punog stabla dobivenog pomoću gini indeksa

	predviđanja	
stvarni	dobri	loši
dobri	151	93
loši	115	170

Matrica zabune podrezanog stabla dobivenog pomoću gini indeksa

	predviđanja	
stvarni	dobri	loši
dobri	115	129
loši	66	219

Slika 9: Matrica zabune stabala dobivenih pomoću gini indeksa

Pomoću *Slike 9* za puno stablo dobiveno gini indeksom izračunali smo da preciznost iznosi 57%, osjetljivost 62%, dok je specifičnost 60%. Greška tipa 1 iznosi 0,40, što znači da bi 40% loših klijenata dobilo kredit. Greška tipa 2 iznosi 0,48 iz čega možemo zaključiti da 48% dobrih klijenata ne bi dobilo kredit. Kod podrezanog stabla dobivenog gini indeksom preciznost iznosi 64%, osjetljivost 47%, dok je specifičnost 77%. Greška tipa 1 iznosi 0,23, a greška tipa 2 0,53.

Matrica zabune punog stabla dobivenog pomoću entropije

	predviđanja	
stvarni	dobri	loši
dobri	151	93
loši	123	162

Matrica zabune podrezanog stabla dobivenog pomoću entropije

	predviđanja	
stvarni	dobri	loši
dobri	157	87
loši	120	165

Slika 10: Matrica zabune stabala dobivenih pomoću entropije

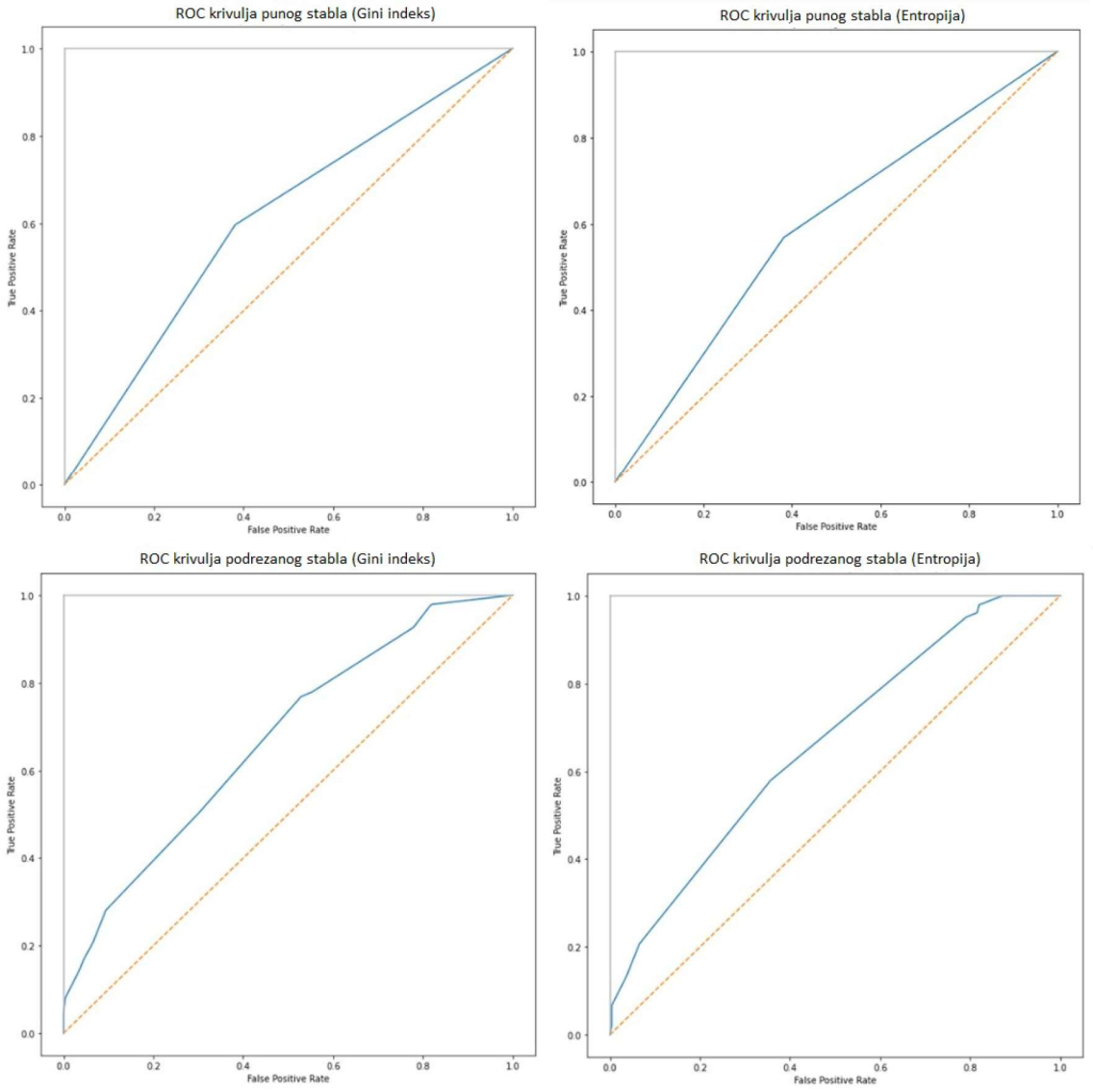
Iz *Slike 10* možemo izračunati da kod punog stabla dobivenog entropijom preciznost iznosi 55%, osjetljivost 62%, dok je specifičnost 57%. Greška tipa 1 iznosi 0,43, a greška tipa 2 iznosi 0,38. Kod podrezanog stabla dobivenog entropijom preciznost iznosi 57%, osjetljivost 64%, dok je specifičnost 58%. Greška tipa 1 iznosi 0,42, a greška tipa 2 iznosi 0,36.

	Točnost	Greška	Prosječna točnost unakrsne validacije	AUC vrijednost
puno stablo	59%	41%	58%	0,59
podrezano stablo	61%	39%	61%	0,66

Tablica 21: Rezultati stabala dobivenih pomoću entropije

Iz Tablice 20 i 21 vidimo da se AUC vrijednost tj. površina ispod ROC krivulje kreće između 0,59 i 0,67 u svim modelima što i nije loše, ali nije niti pretežito dobro. Najveću AUC vrijednost ima podrezano stablo dobiveno pomoću gini indeksa koje je za 0,01 veće od AUC vrijednosti podrezanog stabla dobivenog pomoću entropije. Ako promatramo ROC krivulje na *Slici 11* ne možemo uočiti neke značajne razlike u ROC krivuljama podrezanih stabala kao niti kod punih stabala.

Uzimajući u obzir sve navedene parametre za određivanje preciznosti modela, podrezana stabla pokazuju se bolja od punih stabala, a i računalno su manje zahtjevnija. Ako bi promatrali samo podrezana stabla teško je reći koje je bolje jer većina pokazatelja preciznosti ima sličnu vrijednost. Bitno je naglasiti da je greška tipa 1 kod podrezanog stabla dobivenog pomoći gini indeksa 0,23, dok kod podrezanog stabla dobivenog pomoći entropije iznosi 0,42. Na grešku tipa 1 banke više paze jer je ona puno skuplja za banku s obzirom da je odobravanje kredita lošem klijentu direktna šteta banci. Ako grešku tipa 1 uzmemo kao jednu od bitnijih pokazatelja možemo reći da se podrezano stablo dobiveno pomoći gini indeksa pokazalo najboljim.



Slika 11: ROC krivulja

4 Zaključak

Kako je glavna djelatnost banke prikupljanje depozita i plasiranje viška rezervi u obliku kredita, a da pri tome zadrže stopu obvezne rezerve koju propisuje središnja banka nužno je da banke dobro procijene kreditni rizik. Odluka odobravanja kredita može biti podržana subjektivnom ocjenom kreditnih referenata ili modelima kreditnog skoriranja. Modeli kreditnog skoriranja omogućuju kvalitetnije donošenje odluka i upravljanje kreditnim rizicima pa se financijske institucije sve više okreću njima. Jedan način izrade modela je pomoću klasifikacijskih stabala odlučivanja. Stabla odlučivanja dobar su alat za jednostavniji i lakše objašnjivi prikaz dobivenih rezultata. U ovom smo radu izradili takav model pomoću klasifikacijskih stabala odlučivanja. Prije izrade stabala upoznali smo se s teorijskim dijelom izrade klasifikacijskih stabala odlučivanja i kreditnim skoriranjem te smo nad dobivenim podacima proveli bivarijatnu analizu. Klasifikacijska stabla odlučivanja izgradili smo pomoću gini indeksa i entropije. Prvo smo izgradili puna stabla koja smo zatim podrezali. Na kraju je napravljena interpretacija rezultata u kojoj smo vidjeli da podrezana stabla imaju bolje rezultate od punih stabala. Ako grešku tipa 1 uzmemo kao jednu od bitnijih pokazatelja možemo reći da se podrezano stablo dobiveno pomoći gini indeksa pokazalo najboljim. Rezultati analize pokazali su da su se varijable koje nam daju informacije o godinama radnog staža, iznosu mjesečnih obustava na plaću i tipu poslodavca klijenta pokazale značajnima za ocjenu rizičnosti stanovništva na primjeru jedne banke.

Literatura

[1] N. Bolf, *Osvježimo znanje: Strojno učenje*, Kemija u industriji, vol. 70, br. 9-10, 2021, str. 591-593 https://hrcak.srce.hr/263495

[2] M. Crnobrnja, *Modeliranje odljeva igrača u online klađenjima primjenom neuronskih mreža i logističke regresije*, diplomski rad, Odjel za matematiku, Osijek, 2020.

[3] A. J. Izenman, *Modern Multivariate Statistical Techniques*, Springer Texts in Statistics, USA, 2013.

[4] L. Kiseljak, *Implementacija i vrednovanje algoritma za izgradnju stabla odluke C4.5*, završni rad, Fakultet elektrotehnike i računarstva, Zagreb, 2018.

[5] E. Kočenda, M.Vojtek, *Default predictors and credit scoring models for retail banking*, CESifo Working Paper Series No. 2862, Češka, 2019.

[6] Lj. Kvesić, *Primjena stabla odlučivanja u kreditnom skoringu* ,Sveučilište Josipa Jurja Strossmayera u Osijeku, Ekonomski fakultet, Osijek, 2013., https://www.proquest.com/docview/1513233740

[7] L. ROKACH, O. MAIMOM, *Data maining with decision trees: theory and aplications*, Izrael, 2015.

[8] S. S. Satchidananda, J. B. Simha, *Comparing decision trees with logistic regression for credit risk analysis* , SAS APAUGC, Mumbai, 2006.,https://www.researchgate.net/profile/S-Satchidananda/publication/237356603_Comparing_decision_trees_with_logistic_regression_for_credit_risk_analysis/links/54dc0ad40cf2a7769d94beff/Comparing-decision-trees-with-logistic-regression-for-credit-risk-analysis.pdf

[9] N. Šarlija, *Predavanja za kolegij "Upravljanje kreditnim rizicima"*, Odjel za matematiku, Sveučilište J. J. Strossmayera u Osijeku, 2009.

[10] A. Šimec, D. Lozić, *Nove tehnologije u primjeni*, Zagrebačka škola ekonomije i managmenta, 2020.

[11] J. Zurada, M. Zurada, *How Secure Are "Good Loans": Validating Loan-Granting Decisions And Predicting Default Rates On Consumer Loans*, The Review of Business Information Systems, Vol. 6, No.3, 2002.

[12] N. Boskovic, *Klasifikacija korišćenjem stabla odlučivanja*, https://medium.com/@nemanja.boskovic17/klasifikacija-korišćenjem-stabla-odlučivanja-255456cf6a4d

[13] E. Krueger, *Build Better Decision Trees with Pruning*, https://towardsdatascience.com/build-better-decision-trees-with-pruning-8f467e73b107

[14] J. Mohajon, *Confusion Matrix for Your Multi-Class Machine Learning Model*, https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

[15] D. Shulga, *5 Reasons why you should use Cross-Validation in your Data Science Projects*, https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79

[16] *Grafovi - X.FER*, https://www.fer.unizg.hr/_download/repository/Osnovni_pojmovi-teorija_grafova.pdf

[17] *Stabla odlučivanja*, http://dms1.irb.hr/tutorial/hr_tut_dtrees.php

[18] *Decision and Classification Trees, Clearly Explained!*, https://www.youtube.com/watch?v=_L39rN6gz7Y

Sažetak

U ovom radu upoznajemo se s klasifikacijskim stablima odlučivanja te njihovom primjenom u kreditnom skoriranju. Na početku su opisana klasifikacijska stabla odlučivanja te njihova izrada. Zatim su objašnjene mjere za određivanje preciznosti klasifikacijskog stabla. Opisano je kreditno skoriranje i prethodna istraživanja u kojima su korištena stabla odlučivanja. Provedeno je empirijsko istraživanje nad stvarnim podacima te je napravljena interpretacija dobivenih rezultata.

Ključne riječi

stabla odlučivanja, klasifikacijska stabla odlučivanja, podrezivanje stabla odlučivanja, unakrsna validacija, kreditno skoriranje

Classification trees in credit scoring

Summary

In this work, we will present the methodology of using classification trees for credit scoring. First, we will present the theory of classification trees and methods used to build such trees. Second, we will describe the problem of credit scoring and provide some references in which decision trees are used for credit scoring. Finally, we will show and analyse the experimental results obtained using real life data.

Keywords
decision trees, classification trees, pruning decision trees, cross validation, credit scoring

Životopis

Rođena sam 6. ožujka 1997. godine u Čakovcu. Pohađala sam Osnovnu školu Donja Dubrava u Donjoj Dubravi. Nakon osnovne škole upisujem prirodoslovno-matematički smjer Gimnazije Josipa Slavenskog Čakovec. Srednju školu završavam 2015. godine, te iste godine upisujem preddiplomski studij Matematika na Odjelu za matematiku u Rijeci. S temom Lucasovi brojevi pod mentorstvom doc. dr. sc. Marine Šimac, završavam preddiplomski studij 2019. godine te na jesen upisujem diplomski studij Financijska matematika i statistika na Odjelu za matematiku u Osijeku. Trenutno sam zaposlena kao asistent za aktuarske poslove u službi aktuarskih poslova Uniqa osiguranja u Zagrebu.