UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 1. stopnja

Jure Zukanovič

Analiza tveganj pri spletnih posojilih

Delo diplomskega seminarja

Mentor: doc. dr. Alen Orbanić

Ljubljana, 2017

Kazalo

Analiza tveganj pri spletnih posojilih

Povzetek

Rdeča nit tega dela diplomskega seminarja je predstavitev in primerjava različnih metod strojnega učenja za identificiranje "dobrega izposojevalca", na podlagi podatkov priljubljene spletne platforme za posojanje denarja Lending Club (kratica LC). V delu so predstavljeni rezultati ugotovitev članka z naslovom Risk assessment in social lending via random forests, avtorjev Milada Malekipirbazarija in Vurala Aksakallija, ki sta z uporabo programa WEKA prikazala, da je matoda naključnih gozdov pri identificiranju "dobrega izposojevalca" boljša od točk FICO in ocene s strani LC. To sta metodi, ki ju za identificiranje "dobrega izposojevalca" dandanes uporabljajo agencije, ki se ukvarjajo z računanjem kreditne ocene posameznika za spletne posojilnice.

V zaključku dela je predstavljena rekonstrukcija in potrditev navedb članka na podlagi rezultatov, ki sem jih dobil, ko sem simulacije v programu WEKA zagnal še sam.

Metode strojnega učenja, ki so predstavljene v tem delu diplomskega seminarja, so: metoda naključnih gozdov (ang. random forests), logistična regresija (ang. logistic regression), metoda podpornih vektorjev (ang. support vector machines) in metoda k-najbližjih sosedov (ang. k-nearest neighbors).

Risk assessment in social lending

Abstract

The cohesive thread of the following seminar thesis is the presentation and comparison of different methods of machine learning to identify a "good borrower", according to the popular online lending platform Lending Club (acronym LC).
The paper presents the results of the findings of an article entitled Risk assessment in social lending via random forests by Milad Malekipirbazari and Vural Aksakalli, who, with the help of WEKA program, showed that the random forests method (slo. metoda naključnih gozdov) in identifying a "good borrower" is better than FICO score and LC grades. These methods for identifying a "good borrower" are used by today's agencies to identify the credit rating of an individual for social lending.
In the conclusion of the thesis's seminar, a reconstruction and confirmation of results of the article is presented, based on results I have obtained after running the simulations in WEKA program first-hand.
Machine learning methods presented in this thesis's seminar are: random forests, logistic regression, support vector machines, and k-nearest neighbours method.

Math. Subj. Class. (2010): 15A80

Ključne besede: posojanje denarja prek spleta, ocena tveganja, spletna posojilnica, posojevalec, izposojevalec, strojno učenje, naključni gozd, logistična regresija, metoda podpornih vektorjev, metoda k-najbližjih sosedov

Keywords: peer-to-peer (P2P) lending, risk assessment, social lending platform, lender, borrower, machine learning, random forests, logistic regression, support vector machines, k-nearest neighbors

1. Uvod

Z razvojem spletnega poslovanja in socialnih omrežij se je pojavila možnost za nastanek spletnih posojilnic. Le-te posojajo in si izposojajo denar prek spleta. Spletno posojanje, bolj znano pod angleškim izrazom peer-to-peer (P2P) lending, kar lahko v slovenščini prevedemo kot neposredno posojanje, se je razvilo kot alternativa bankam, kjer si lahko posameznik izposodi ali posodi denar z uporabo trgovalne spletne strani, brez pomoči finančnega posrednika, kot je na primer banka.
Prednost spletnega posojanja je v večjem potencialu skupne profitabilnosti udeležencev. Izposojevalci lahko pridobijo posojilo po nižji obrestni meri, kot če bi si denar izposodili pri banki. Prav tako imajo korist tudi posojevalci denarja, ki pri neposrednem posojanju posodijo denar izposojevalcem po višji obrestni meri, kot če bi ga posodili banki.
Spletne posojilnice delujejo na način, da lahko posojevalec, ki bi rad posodil denar, na spletni platformi najde skupino potencialnih izposojevalcev, ki bi si radi izposodili določeno količino denarja, in se nato sam odloči, komu izmed njih želi denar posoditi. Ker so pri spletnih posojilnicah posojevalci in izposojevalci v večini primerov preprosti potrošniki, ni potrebe za povečanje likvidnosti posojil z zavarovanjem le-teh.
Najpopularnejše spletne posojilnice na svetovnem spletu so: Prosper in Lending Club Corp. (krat. LC) s sedežem v Združenih državah Amerike, Zopa Ltd. s sedežem v Veliki Britaniji in Smava GmbH s sedežem v Nemčiji. Ker sta stabilnost in potencialna širitev spletnih posojilnic močno odvisni od zanesljivega izračuna tveganja vrnitve posojila za posameznega izposojevalca, se zgoraj naštete spletne platforme za izračun kreditne ocene posameznika zatekajo k za to uposobljenim agencijam. Primeri takih agencij so: Experian, TransUnion LLC, Equifax Inc. in Schufa Holding AG.
Popularnost spletnih posojilnic v zadnjih letih vedno bolj narašča. Za primer lahko vzamemo največjo spletno posojilnico Lending Club. Od ustanovitve leta 2006 pa do konca leta 2014 je spletna posojilnica izdala za 7,5 milijard ameriških dolarjev posojil, do konca septembra 2016 pa se je število izdanih posojil spletne posojilnice povzpelo na več kot 22,5 milijard ameriških dolarjev.
Motivacija za delo mojega diplomskega seminarja izvira in dejstva, da se večina spletnih posojilnic za identificiranje "dobrega izposojevalca" zateka k agencijam, ki za izračun kreditne ocene posameznika uporabljajo klasični metodi točk FICO in ocene s strani LC. Z izrazom "dober izposojevalec" poimenujemo posameznika, ki je sposoben v celoti vrniti posojilo v za to določenem času.
Čeprav sta metodi kreditnih točk FICO in ocene s strani LC ustrezni za izračun kreditne ocene posameznika pa ne upoštevata različnih lastnosti obnašanja uporabnikov spletnih posojilnic. Na primer, ponudba posojil, ki se navezuje na določen seznam prošenj prosilcev za posojilo, s časom narašča v skladu s potenčnim zakonom. To nakazuje na čredno obnašanje posojevalcev. Predstavljajmo si, da so posojevalci in seznami prošenj prosilcev za posojilo predstavljeni z vozlišči. Povezava med posojevalcem in seznamom prošenj predstavlja, da je posojevalec zainteresiran posoditi denar skupini ljudi s seznama. V praksi lahko opazimo povezavo med ponudbo posojil za posamezno vozlišče, ki predstavlja nek seznam, in stopnjo tega vozlišča. To pomeni, da ko ima določen seznam sto ali več ponudb posojevalcev, je bolj verjetno, da bo ta seznam potencialnih izposojevalcev skozi čas privabil vedno več ponudb. To pa pomeni, da ima tak seznam prošenj večjo verjetnost sklenitve posojila zaradi

večjega interesa posojevalcev.
Analiza Lending Cluba Emekter-ja, Tu-ja, Jirasakuldech-a in Lud-a iz leta 2015 [2] razkriva dve ključni ugotovitvi:

- Upoštevati je potrebno, da si posamezniki z najvišjimi točkami FICO ne izposojajo denarja pri spletni posojilnici. Še več, zgornja tretjina potrošnikov glede na točke FICO ne zaprosi za posojilo pri Lending Clubu.
- Višja obrestna mera, ki se navezuje na bolj tveganega izposojevalca, ni vredna tveganja. Bolj natančno povedano, višja obrestna mera za posameznika z nizko oceno s strani LC ni dovolj visoka, da bi lahko nadomestila dodatno tveganje, ki ga ima posojevalec.

Zgornji dve ugotovitvi pojasnjujeta, da je z vidika profitabilnosti posojevalca denarja identifikacija "dobrega izposojevalca" ključnega pomena. Profitabilnost investitorjev pa je ključna za ponovno vračanje k spletnim posojilnicam, kar prinaša stabilnost trga spletnih posojilnic. Glede na zgoraj opisano tveganje povračila naložbe sledi, da bi bilo optimalno za posojevalce posoditi denar le "najvarnejšim izposojevalcem", z najvišjo oceno s strani LC. V nadaljevanju mojega dela diplomskega seminarja bom pokazal, da ta predpostavka ne drži in da posamezniki z največ točkami FICO in najvišjo oceno s strani LC niso nujno "dobri izposojevalci". To pa nakazuje, da tradicionalni metodi za ugotavljanje tveganj pri spletnih posojilnicah nista najbolj ustrezni.

2. Opis delovanja Lending Cluba

V delu bom uporabil podatke s spletne platforme Lending Club, zato sledi kratek opis delovanja. Spletna posojilnica Lending Club deluje na naslednji način:

- Posamezniki, ki si želijo izposoditi denar, so po principu razvrščanja Lending Cluba umeščeni na določen seznam prošenj za posojila.
- Lending Club določi obrestno mero posojila, ki je odvisna od izposojevalčeve ocene s strani LC.
- Za ocenjevanje in lažjo izbiro potencialnih izposojevalcev imajo posojevalci dostop do finančnih informacij izposojevalcev kot so točke FICO, ocena s strani LC, razmerje med dolgom in dohodkom, podatek o lastništvu nepremičnine in število odprtih računov. Prav tako seznam prošenj za posojila vključuje ostale podrobnosti kot so razlog za izposojo denarja in pomembnejše demografske podatke izposojevalcev, ki so ponujene vpogled posojevalcem.
- Posojevalci prevzamejo tveganje nevračila posojila. Zato je z namenom zmanjšanja tveganja najbolj pogosta strategija investiranja razporeditev denarja na večje število naložb.
- Da je prošnja za posojilo posameznika odobrena, mora izposojevalec pridobiti dovolj investicij, da je s tem pokrita celotna vsota denarja, za katero je zaprosil.
- Lending Club prejme od posojevalcev provizijo za vsako plačilo, ki ga posojevalci prejmejo s strani izposojevalcev.

3. Finančne lastnosti posameznikov v algoritmih strojnega učenja

Za definiranje primerjalnih modelov v strojnem učenju je ključna izbira atributov posameznikov, po katerih se med seboj razlikujejo. Finančne lastnosti izposojevalcev Lending Cluba, uporabljene v algoritmih, so:

- **status posojila:** Binarna spremenljivka, ki pove, ali je izposojevalec vrnil posojilo v celoti. Posojila, ki so bila odplačana v celoti, bomo poimenovali "dobra posojila", tista posojila, ki niso bila vrnjena, pa "slaba posojila".
- **letni prihodek:** Podatek, ki ga posameznik poda sam med registracijo na spletno platformo Lending Club.
- **kreditna starost:** Datum odprtja prvega kreditnega računa izposojevalca, zapisan v mesecih.
- **zamujena plačila:** Število zamujenih plačil izposojevalca v zadnjih dveh letih. Če je število večje od 2, ga postavimo kot 2.
- **zaposlitvena doba:** Doba zaposlitve posameznika v letih. Možne vrednosti so med 0 in 10, kjer 0 pomeni manj kot eno leto in 10 pomeni več kot 10 let.
- **lastništvo nepremičnine:** Podatek, ki ga posameznik poda sam med registracijo na spletno platformo Lending Club. Možna stanja so najem, lastništvo ali hipoteka.
- **število prošenj:** Število prošenj za posojilo izposojevalca v zadnjih šestih mesecih.
- **višina posojila:** Višina posojila izposojevalca. Znesek ne mora presegati 35000 ameriških dolarjev.
- **namen posojila:** Podatek, ki ga posameznik poda, ko zaprosi za posojilo. Nekatere od možnosti so: vračilo dolgov, renovacija nepremičnine, pokritje dolga kreditnih kartic, selitev, financiranje majhnih podjetij, avto, večji nakup, počitnice, zdravstvene storitve, nepremičnina, poroka itd.
- **odprti računi:** Število odprtih kreditnih računov izposojevalca.
- **število računov:** Število vseh kreditnih računov, ki jih je izposojevalec odprl.
- **število obrokov posojila:** Število mesečnih obrokov posojila. Vrednost je lahko 36 ali 60 mesecev.

Pomembne so tudi finančne lastnosti, ki so preprosta razmerja ostalih lastnosti. Ta razmerja podajajo nekatere karakteristike posameznikov, ki v splošnem ne bi bile upoštevane. To so:

- **razmerje med dolgom in dohodki:** Razmerje med mesečnim dolgom in mesečnim prihodkom izposojevalca.
- **razmerje med dohodki in dolgom iz posojil:** Razmerje med mesečnimi prihodki in mesečnimi obroki plačil dolga iz posojil izposojevalca. To razmerje ni standardna finančna lastnost posameznika, je pa pomembna z vidika primerjave posameznikov. Ideja je v tem, da 500 dolarjev mesečnega obroka posojila za nekoga, ki zasluži 10000 dolarjev mesečno, ni veliko, za nekoga, ki pa mesečno zasluži 1000 dolarjev, pa to predstavlja velik finančni zalogaj. Kljub pomembnosti je to razmerje težko vključiti v algoritme strojnega učenja.
- **višina nefiksiranega kredita:** Višina nefiksiranega kredita, ki ga ima posameznik. Le to je kredit, ki nima vnaprej določenega števila plačil. Primer so kreditne kartice.
- **razmerje med nefiksiranim kreditom in prihodki:** Razmerje med stanjem nefiksiranega kredita in izposojevalčevimi mesečnimi prihodki. To je še ena od nestandardnih finančnih lastnosti posameznikov.

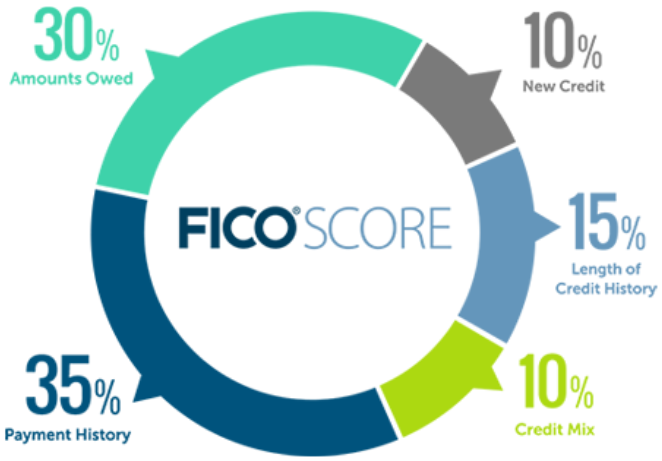
Izmed 15 finančnih lastnosti je 12 numeričnih, 3 pa so podane nominalno: lastništvo nepremičnine (3 možnosti), namen posojila (LC ponuja 13 možnosti) in število

obrokov posojila (2 možnosti). Za definiranje algoritmov k-najbližjih sosedov, logistične regresije in metode podpornih vektorjev, nominalnim vrednostim priredimo binarno vrednost. Pri metodi naključnega gozda nominalne atribute ohranimo kot dane.

4. Točke FICO in ocena s strani LC

Ker bo v nadaljevanju dela prikazana primerjava rezultatov dobljenih s principom točk FICO in ocene s strani LC z algoritmi strojnega učenja, sledi kratek opis teh dveh metod.

4.1. **Točke FICO.** Točke FICO so standardne kreditne točke, ki se uporabljajo pri večini odločitev o posojilih v Združenih državah Amerike. Izračunajo se iz različnih finančnih atributov izposojevalčevih kreditnih zapisov. Pri Lending Clubu so točke FICO posameznika predstavljene z dvema številkama, FICO low in FICO high. Povprečje teh dveh vrednosti imenujemo točke FICO posameznika. Spodnja slika [8] prikazuje delež posameznega finančnega atributa posameznika pri izračunu točk FICO. Vidimo, da ima pri izračunu točk FICO največji vpliv zgodovina posameznikovega plačevanja s 35 % (ang. payment history), nato s 30 % višina posameznikovega dolga (ang. amounts owed), s 15 % dolžina posameznikove zadolžitve (ang. length of credit history) in z 10 % nov kredit posameznika (ang. new credit) in vrste računov, ki sestavljajo posameznikovo kreditno poročilo (ang. credit mix).



Slika 1. Delež posameznega finančnega atributa posameznika pri izračunu točk FICO

4.2. **Ocena s strani LC.** Lending club z uporabo algoritmov, ki temeljijo na lastnostih posojila in tveganju posojanja posamezniku, določi vsakemu posojilu oceno od A do G. Vsaka ocena se deli še na 5 podocen, kar pomeni, da so ocene za posojila od A1 do G5. A1 pomeni, da je posojilo najbolj varno, G5 pa, da je posojilo najbolj tvegano. Glede na oceno s strani LC se posojilu dodeli obrestno mero med 5,32 % in 30,99 %. Vse ocene in pripadajoče obrestne mere prikazuje spodnja tabela [11].

Tabela 1. LC ocene in pripadajoče obrestne mere za posojilo

Ocena posojila	Obrestna mera	Ocena posojila	Obrestna mera
A1	5.32%	E1	22.91%
A2	7.07%	E2	23.88%
A3	7.21%	E3	24.85%
A4	7.35%	E4	25.82%
A5	7.97%	E5	26.30%
B1	9.44%	F1	28.72%
B2	9.93%	F2	29.69%
B3	10.42%	F3	30.17%
B4	10.91%	F4	30.65%
B5	11.99%	F5	30.75%
C1	12.62%	G1	30.79%
C2	13.59%	G2	30.84%
C3	14.08%	G3	30.89%
C4	15.05%	G4	30.94%
C5	16.02%	G5	30.99%
D1	17.09%		
D2	18.06%		
D3	19.03%		
D4	20.00%		
D5	21.45%		

5. Strojno učenje

Izraz strojno učenje v splošnem pomeni pridobivanje znanja na podlagi izkušenj. To ni učenje na pamet, ampak iskanje pravil v učnih podatkih.
Strojno učenje je podpodročje računalništva, ki daje računalnikom možnost učenja, ne da bi bili za to posebej sprogramirani. Razvilo se je iz študije o prepoznavanju vzorcev in teorije učenja v umetni inteligenci. Strojno učenje se ukvarja s konstruiranjem algoritmov, ki omogočajo učenje in podajajo napovedi na množici podatkov. Pri sistemih za strojno učenje razlikujemo med učnimi algoritmi, s pomočjo katerih se sistem uči, kar pomeni, da iz množice primerov tvori novo oziroma popravlja staro znanje, in izvajalnimi algoritmi, ki to znanje uporabljajo za reševanje problemov. Strojno učenje se pojavlja v številnih računalniških nalogah kot so na primer: filtriranje elektronske pošte, odkrivanje omrežnih vsiljivcev, optično prepoznavanje znakov, itd.

6. Metodologija algoritmov strojnega učenja pri spletnih posojilnicah

V tem poglavju dela je uporabljena notacija avtorjev članka "An introduction to statistical learning" avtorjev James-a, Witten-a, Hasti-ja, in Tibshirani-ja iz leta 2013 [3].
Problem napovedi izposojevalčevega statusa, se pravi ali gre za "dobrega" ali za "slabega izposojevalca", modeliramo na naslednji način: Naj bo binarna spremenljivka Y definirana kot:

$$Y = \begin{cases} 0; & \text{če je izposojevalec slab,} \\ 1; & \text{če je izposojevalec dober.} \end{cases}$$

Matematični model metodologije napovedi izposojevalčevega statusa je dan kot

$$Y = f(X) + \epsilon,$$

kjer je $X = (X_1, ..., X_p)$ vektor finančnih lastnosti posameznika, p število finančnih lastnosti posameznika in ϵ nepopravljiva napaka modela, ki predstavlja mersko napako in ostale nevšečnosti neke skupine podatkov (ang. data noise). Funkcija f za vhodni podatek vzame nek vektor finančnih lastnosti in poda napoved izposojevalčevega statusa. Vedeti je potrebno, da kljub predpostavki, da funkcija f obstaja, v praksi skoraj nikoli ni znana. Še več, četudi bi funkcija f bila znana, bi še vedno obstajala napaka pri napovedi. Zato funkcijo f na novo definiramo kot

$$f(x) := E(Y|X = x),$$

kjer je na desni strani enačbe pogojno matematično upanje spremenljivke Y pri pogoju, da je vektor X enak neki realizaciji finančnih lastnosti. Sedaj vidimo, da funkcija f za vhodni podatek vzame vektor finančnih lastnosti, ki je enak neki konkretni realizaciji finančnih lastnosti posameznika in vrne pogojno pričakovano vrednosti izposojevalčevega statusa. Tako definirana funkcija f po definiciji pogojnega matematičnega upanja sedaj namesto binarne vrednosti podaja verjetnost, da je izposojevalec "dober", kar lahko zapišemo kot

$$f(x) := E(Y|X = x) = P(Y = 1|X = x).$$

Pomembna lastnost funkcije $f(x)$ je, da je to funkcija, ki minimizira pogojno matematično upanje kvadratične napake med napovedjo izposojevalčevega statusa Y in funkcijo g

$$E[(Y - g(X))^2|X = x],$$

za katerokoli funkcijo g in za vsako realizacijo vektorja finančnih lastnosti $X = x$, torej

$$f(x) = \arg\min_g E[(Y - g(X))^2|X = x].$$

Dokaz. Dokazujemo, da funkcija $f(x)$ minimizira $E[(Y - g(X))^2|X = x]$ za katero koli funkcijo g in za vsak $X = x$ [4].
Naj bo $g(x)$ napoved za Y pri $X = x$ in $(Y - g(X))^2$ kvadratična napaka napovedi. Tveganje pojavitve napake označimo z $R(g) = E[(Y - g(X))^2]$.
Naj bo sedaj $g(x)$ katerakoli funkcija spremenljivke x. Potem velja:

$$\begin{aligned} R(g) &= E[(Y - g(X))^2] = E[(Y - f(X) + f(X) - g(X))^2] \\ &= E[(Y - f(X))^2] + E[(f(X) - g(X))^2] + 2E[(Y - f(X))((f(X) - g(X))] \\ &\geq E[(Y - f(X))^2] + 2E[(Y - f(X))((f(X) - g(X))] \\ &= E[(Y - f(X))^2] + 2E[E[(Y - f(X))((f(X) - g(X))|X]] \\ &= E[(Y - f(X))^2] + 2E[(E(Y|X) - f(X))((f(X) - g(X))] \\ &= E[(Y - f(X))^2] + 2E[(f(X) - f(X))(f(X) - g(X))] \\ &= E[(Y - f(X))^2] = R(f) \end{aligned}$$

S tem smo dokazali, da je tveganje za pojav kvadratične napake za f manjše od tveganja za katero koli funkcijo g. Od tod sledi, da f minimizira
$E[(Y - g(X))^2|X = x]$ za katero koli funkcijo g za vsak $X = x$. □

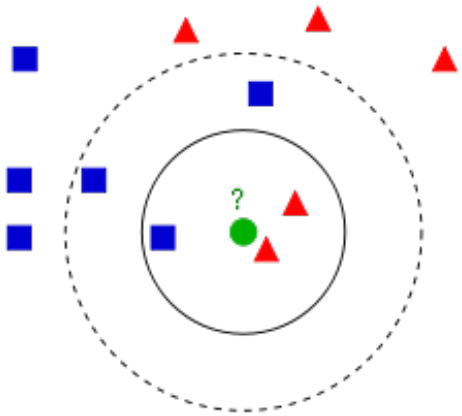
Končni cilj algoritmov strojnega učenja je najti približek funkcije $f(x)$, $\hat{f}(x)$, za katerega velja

$$E[(Y - \hat{f}(X))^2 | X = x] = [f(x) - \hat{f}(x)]^2 + Var(\epsilon),$$

pri čemer $[f(x) - \hat{f}(x)]^2$ predstavlja popravljivo napako, $Var(\epsilon)$ pa nepopravljivo napako, ki se pojavi pri določanju Y. Torej nas zanima, kako najti dober približek $\hat{f}(x)$, da bomo minimizirali napako $[f(x) - \hat{f}(x)]^2$.
V nadaljevanju so predstavljeni štirje modeli za reševanje tega problema.

7. Metode strojnega učenja

7.1. **Metoda k-najbližjih sosedov.** Metoda k-najbližjih sosedov (k-NN) je zelo preprosta, a zelo popularna in močna neparametrična metoda, pogosto uporabljena za klasifikacijo primerov. Klasifikacija je proces prepoznavanja, definiranja in razumevanja objektov.
Klasifikacija novega primera se opravi na podlagi večine glasov k najbližjih primerov iz učne množice v prostoru atributov. Izbran je razred, kateremu pripada največ glasov. Parameter k je naravno število in je tipično med 1 in 10. Tej vrsti učenja pravimo tudi leno učenje, saj učenja pri tej metodi strojnega učenja skorajda ni. Poseben primer te metode se pojavi, ko je $k = 1$. Takrat se novemu primeru pripiše razred najbližjega soseda.



Slika 2. Ugotavljanje k najbližjih sosedov [12]

Učne množice pri tej metodi so predstavljene kot vektorji v prostoru atributov. Za določanje k najbližjih sosedov se pogosto uporablja Evklidska razdalja [22], ki je najprimernejša za zvezne atribute. Faza učenja poteka tako, da se v njej shranijo le vektorji atributov in oznake razredov. V fazi klasifikacije pa se določijo razdalje med novim vektorjem in vektorji, ki so bili shranjeni v fazi učenja, na podlagi katerih algoritem nato izbere k najbližjih sosedov, z uporabo katerih se opravi klasifikacija novega vektorja. Najpogostejši način klasifikacije novega vektorja je klasifikacija glede na najbolj zastopan razred med k najbližjimi sosedi. Slabost te tehnike je, da pri klasifikaciji prevladujejo razredi z največ primeri. Zato je v pomoč, da otežimo vse glasove iz množice k najbližjih sosedov z razdaljo do novega primera, nato pa upoštevamo, da imajo bližji sosedi večji vpliv pri končni odločitvi kot bolj oddaljeni. Parameter k izberemo v odvisnosti od množice podatkov in problema, ki ga rešujemo. Parameter k ne določa velikosti okolice novega primera, iz katere izberemo

učne primere, ampak se okolica dinamično spreminja, glede na gostoto učnih primerov v danem podprostoru primerov. S tem se izognemo problemu gostejših, oziroma redkejših delov prostora.
V našem primeru lahko približek $\hat{f}_{k-NN}(x)$, dobljen po metodi k-najbližjih sosedov, zapišemo kot :

$$\hat{f}_{k-NN}(x) := \text{Majority}(Y|X \in \mathcal{N}_k(x)),$$

kjer Majority definira funkcijo večine glasov in $\mathcal{N}_k$ predstavlja k najbližje sosede realizacije finančnih lastnosti, glede na Evklidsko razdaljo v p-dimenzionalnem prostoru.

7.2. **Logistična regresija.** Regresija v splošnem je proces, ki ugotavlja razmerja oziroma povezavo med spremenljivkami. Predvsem se osredotoča na povezavo med odvisnimi in neodvisnimi spremenljivkami. Natančneje, regresija nam pomaga prikazati, kako se vrednost ene odvisne spremenljivke spremeni, če se spremeni vrednost ene neodvisne spremenljivke, medtem ko ostale neodvisne spremenljivke ohranijo svoje vrednosti [25]. Ena od oblik regresije je logistična regresija.
Logistična regresija (LR) konstruira linearni model na podlagi transformirane ciljne spremenljivke. Kot pri linearni regresiji je transformirana spremenljivka aproksimirana z uporabo linearne funkcije. Dobimo model:

$$P(1|a_1, a_2, \ldots, a_k) = \frac{1}{1 + e^{-w_0 - w_1 a_1 - \ldots - w_k a_k}},$$

kjer so w_i vrednosti uteži in a_i vrednosti atributov. Uteži morajo biti izbrane tako, da se dobljeni model dobro prilega učni množici.
Za naš problem je približek linearne regresije $\hat{f}_L(x)$ definiran kot linearna kombinacija posameznih finančnih lastnosti:

$$\hat{f}_L(x) := \sum_{i=0}^{p} \beta_i x_i = \beta' x, (x_0 = 1),$$

kjer približke $\hat{\beta}_i$ za β_i dobimo z metodo najmanjših kvadratov [23] za $i = 0, \ldots, p$ na podatkih iz učne množice. Bolj primerna matoda za binarno klasifikacijo je logistična regresija. Približek funkciji je sedaj oblike:

$$\hat{f}_L(x) := \frac{e^{\beta' x}}{1 + e^{\beta' x}}.$$

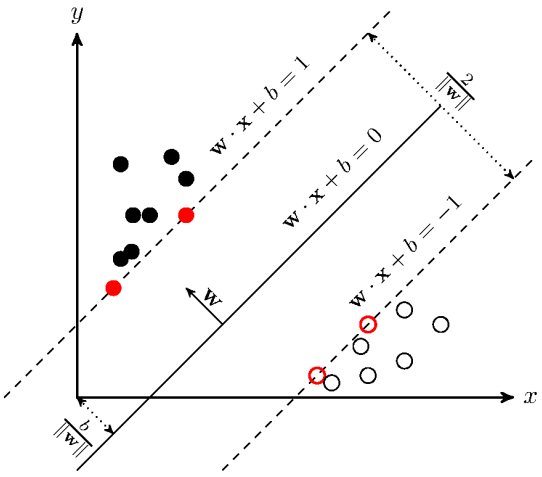
V metodi logistične regresije so sedaj približki $\hat{\beta}_i$ dobljeni z metodo največjega verjetja [24]. Največja koristnost logistične regresije je, da podaja rezultat med 0 in 1. Le tega pa si lahko interpretiramo kot razredno pogojno verjetnost v problemih klasifikacije. Z drugimi besedami, primer nam pokaže, kako se lastnosti razreda ujemajo z x.

7.3. **Metoda podpornih vektorjev.** Metoda podpornih vektorjev (SVM) je ena najuspešnejših metod klasifikacije in regresije. Metoda je primerna za učenje na velikih množicah podatkov.
Imamo neko populacijo primerov, ki so predstavljeni z vektorji iz $\mathbb{R}^p$ in dva razreda, pozitivnega in negativnega. Učno množico predstavljajo pari (x_i, y_i) za $i \in 1, \ldots, l$, kjer je $x_i \in \mathbb{R}^p$ vektor, $y_i \in \{1, -1\}$ pa njegova oznaka razreda. Radi bi dobili klasifikator, ki bo razločeval primera med seboj. Metoda je tako namenjena ločevanju razredov med seboj.
Če imamo opravka z več razredi, ponovimo postopek za vsak razred, ki ga želimo

ločiti od ostalih. Nov primer klasificiramo v razred z najvišjo vrednostjo odločitvene funkcije [6].
Metoda v p-dimenzionalnem prostoru atributov dva razreda loči s postavitvijo hiperravnine [26]. Hiperravnina je enako in hkrati najbolj oddaljena od najbližjih primerov obeh razredov. Te najbližje primere imenujemo podporni vektorji. Razdaljo med hiperravnino in podpornimi vektorji imenujemo rob hiperravnine (ang. margin). Hiperravnina, uporabljena pri metodi podpornih vektorjev, ima tako maksimalen rob. Z maksimiranjem širine roba zmanjšamo kompleksnost modela in splošno tveganje za pojavljanje napak.
Ravnino lahko zapišemo kot množico točk x, ki zadošča enačbi:

$$w \cdot x - b = 0,$$

kjer $\cdot$ predstavlja skalarni produkt [27], w pa normalo hiperravnine [28]. Parameter b dobimo iz parametra $\frac{b}{||w||}$, kjer ulomek določa odmik hiperravnine od izhodišča vzdolž normale. Parametra w in b izberemo tako, da maksimirata rob. Zgornje enačbe ravnine prikazuje spodnja slika 3 [14], ki prikazuje dvodimenzionalen primer metode podpornih vektorjev.



Slika 3. Enačbe hiperravnine

V praksi se pogosto pojavlja, da razrede ne moremo ločiti med seboj s hiperravnino. Posledično se primerom, ki ležijo na napačni strani roba, prišteje nek pozitiven pribitek. Pribitek se povečuje ali zmanjšuje v odvisnosti od oddaljenosti primerov na napačni strani roba. Cilj je minimizirati vsoto teh pribitkov, ob tem pa maksimirati širino roba.
Problem optimizacije metode podpornih vektorjev lahko predpostavimo kot [1]:

$$\max_{\alpha_l} \left(\sum_{l=1}^{n} \alpha_l + \frac{1}{2} \sum_{l,r=1}^{n} \gamma_l \gamma_r \alpha_l \alpha_r K(x_l, x_r) \right),$$

z omejitvami, da je za l-ti primer iz učne množice $l = 1 - n$, $0 \leq \alpha_l \leq C$ in $\sum_{l=1}^{n} \gamma_l \alpha_l = 0$, kjer je α_l Lagrange-ov multiplikator primera l. Parameter C v našem primeru predstavlja relativni strošek vsake naloge v procesu optimizacije in je določen z 1 v naši interpretaciji. V tem delu diplomskega seminarja predpostavimo, da je kvadratni polinom $K(x_l, x_r) = (1 + x_l x_r)^2$. Za definiranje funkcije $K(x_l, x_r)$

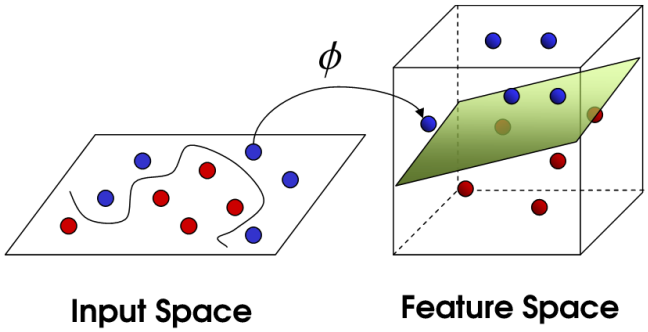
je potrebno poznati in razumeti pojem jeder. Za razumevanje jeder pa je potrebno razumeti naslednje [7]: Recimo, da bi pred učenjem vse učne vektorje preslikali v nek nov vektotrski prostor F (ta proces prikazuje slika 4 [14]):

$$\phi : \mathbb{R} \to F,$$
$$x \to \phi(x).$$

Če pišemo $\phi(x) = (\phi_1(x), \phi_2(x), \ldots)$, so $\phi_i(x)$ značilke (ang. features) primerka x, F pa prostor lastnosti (ang. feature space). Pri konstruiranju problema optimizacije nam ni treba nikjer eksplicitno delati s slikami $\phi(x)$. Dovolj je, če znamo izračunati skalarne produkte med njimi:

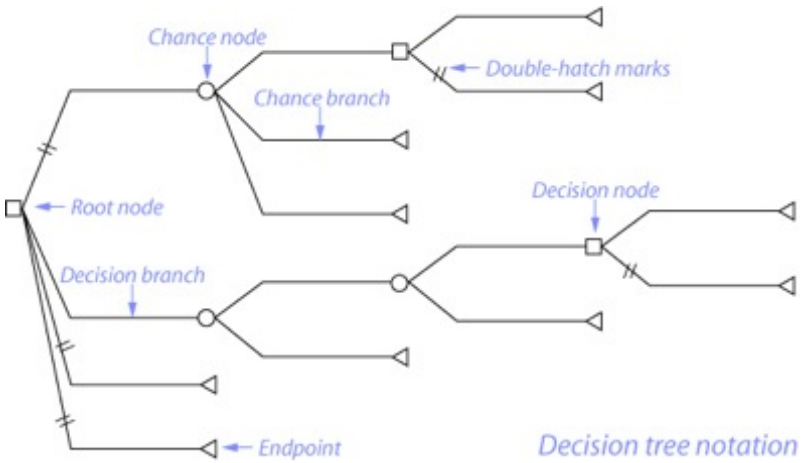
$$K(x, \hat{x}) = \langle \phi(x), \phi(\hat{x}) \rangle_F.$$

Taki funkciji rečemo jedro (ang. kernel). Lahko bi si najprej izbrali ϕ in F, ter glede na to potem poskušali najti primerno jedro K (takšno, ki jo bo preprosto izračunati). Pogosto pa vzamemo neko jedro K in se moramo le še prepičati, da obstajata nek ϕ in F, ki jima K ustreza.
Zadostni pogoj za K določa Mercerjev izrek [7]: Za vsako končno množico $\{x_1, \ldots, x_l\}$ mora biti matrika $K = (k_{ij})$ z elementi $k_{ij} := K(x_i; x_j)$ pozitivno semidefinitna.



Slika 4. Postavitev hiperravnine.

7.4. **Metoda naključnih gozdov.** Za razumevanje metode naključnih gozdov moramo najprej opredeliti pojem klasifikacijskega drevesa. Strojno učenje z uporabo klasifikacijskih dreves je priljubljena metoda klasifikacije, ki s pomočjo usmerjenih dreves z razredno pogojno verjetnostjo na koncu vej opravi klasifikacijo novega primera. Veja v tem primeru predstavlja vse povezave od korena do lista. Vsako klasifikacijsko drevo začnemo graditi s korenom in nato nadaljujemo z izgradnjo poddreves, dokler ne pridemo do lista. List je v primeru klasifikacijskih dreves atribut, na podlagi katerega se opravi klasifikacija primera. Vsako notranje vozlišče drevesa se navezuje na lastnost primera, ki ga klasificiramo (npr. izposojevalec je lastnik nepremičnine) in vsaka povezava določa lastnosti primera, ki ga želimo klasificirati. Proces konstruiranja klasifikacijskega drevesa poteka po principu deli in vladaj v smislu, da se koren navezuje na vse primere iz učne množice in vsako notranje vozlišče loči primere glede na njihove lastnosti. Obstajata dva problema strojnega učenja s klasifikacijskimi drevesi. Zanima nas, katere lastnost izbrati za notranja vozlišča, po katerih bomo primere ločili med seboj in koliko stopenj naj ima vsaka veja ter posledično, kdaj je smiselno prenehati ločevati.

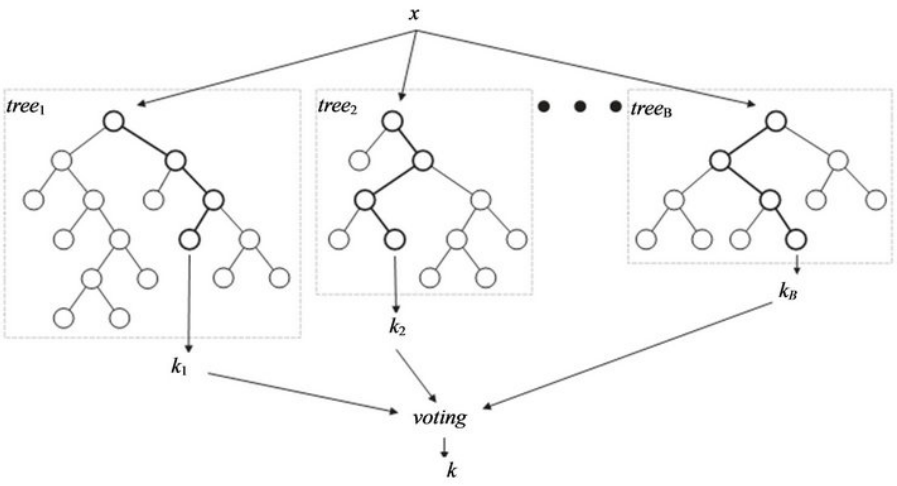


Slika 5. Klasifikacijsko drevo [16]

V primeru naključnih gozdov, ki so skupek klasifikacijskih dreves, se ločevanje v posameznem klasifikacijskem drevesu opravi na podlagi tako imenovanega *Gini indeksa*, število stopenj posamezne veje pa bomo omejili s parametrom d.
Gini indeks notranjih vozlišč drevesa je definiran na naslednji način: Za lastnost, ki je kandidat za ločevanje, označimo jo z X_i, definiramo možne stopnje, na katerih se lahko pojavi z $L_1, \ldots, L_J$. *Gini indeks* za tako lastnost se izračuna po formuli:

$$G(X_i) := \sum_{j=1}^{J} P(X_i = L_j)(1 - P(X_i = L_j)) = 1 - \sum_{j=1}^{J} P(X_i = L_j)^2.$$

Ko izračunamo *Gini indekse* za vse lastnosti, ki so kandidatke za ločevanje, ločevanje poteče po tisti lastnosti, ki ima najvišji *Gini indeks*.
Klasifikacija primerov s klasifikacijskimi drevesi pa ima tudi nekaj prednosti. Klasifikacijska drevesa si je lahko predstavljati, delujejo tako za numerične kot tudi za nominalne podatke in lahko jih je konstruirati.
Kljub vsemu pa uporaba klasifikacijskih dreves ni vedno najboljša metoda klasifikacije. Zato z namenom izboljšanja natančnosti dreves včasih uporabimo metodi kot sta boosting [29] (ponavljajoče se učenje za neuvrščene primere) in bagging [30] (izgradnja večih dreves in kombinacija rešitev), pod katero uvrščamo tudi metodo naključnih gozdov.
Metoda naključnih gozdov (RF) je metoda klasifikacije, ki klasifikacijo novega primera opravi na podlagi večine glasov, ki jih dobi iz končno mnogo klasifikacijskih dreves. Metoda konstruira gozd tako, da primere iz učne množice naključno razdeli v B različnih množic, nato pa za vsako množico zgradi klasifikacijsko drevo, ki primere iz te množice loči med seboj. Vendar metoda ne loči vseh primerov iz množice, ampak naključno izbere neko podmnožico. Prav tako metoda pri izgradnji dreves ne vzame vseh lastnosti primerov iz učne množice, ampak naključno izbere m od vseh p lastnosti. Na vsakem vozlišču, na katerem poteka delitev, se upošteva natanko ena izmed teh m lastnosti in na vsakem vozlišču ločitve se opravi nova izbira teh m atributov. V vsakem drevesu ločitve potekajo toliko časa, dokler ne pridemo do globine d. Klasifikacija novega primera poteka na način, da je primer, glede na glasove iz vseh dreves, ki jih imamo v gozdu, razporejen v razred z največ glasovi.



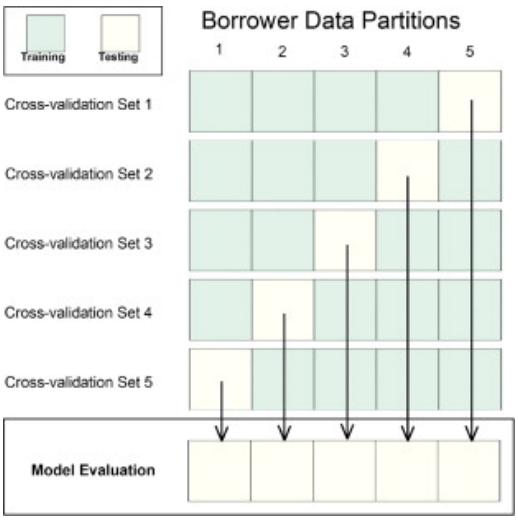
Slika 6. Naključni gozd [16]

8. Predstavitev rezultatov članka

V tem poglavju bodo predstavljeni postopki in rezultati primerjave zgoraj opisanih modelov strojnega učenja, točk FICO in ocene s strani LC za indentifikacijo "dobrega izposojevalca", na podlagi članka z naslovom Risk assessment in social lending via random forests avtorjev Milada Malekipirbazarija in Vurala Aksakallija. Avtorja sta ugotovila, da je metoda naključnih gozdov boljša od ostalih modelov strojnega učenja, kot tudi od točk FICO točk in ocene s strani LC.
V članku je primerjava prikazana z uporabo programa WEKA. V programu so bili uporabljeni podatki o posojilih spletne posojilnice Lending Club med januarjem 2012 in septembrom 2014, dostopni na uradni strani Landing Cluba [17]. Datoteko sestavlja okoli 350000 kreditnih zapisov izposojevalcev. Ker večina kreditov v datoteki ni bila odobrena ali v tistem času krediti še niso zapadli, in zato ne podajajo informacij o kreditni sposobnosti izposojevalcev, sta avtorja članka filtrirala podatke, da sta na koncu dobila le posojila s statusom: v celoti poplačano ali nepoplačano. Na koncu sta dobila datoteko s približno 68000 kreditnimi zapisi, ki so skupaj vsebovali za približno 1 milijardo ameriških dolarjev posojil.

8.1. **Analiza občutljivosti.** Pri spletnih posojilnicah odobritev posojila "slabemu izposojevalcu" prinaša veliko večje tveganje kot zavrnitev posojila "dobremu izposojevalcu". V ta namen se v modelih uporablja stroškovno utežena matrika. Le-ta se uporablja z namenom, da bi zmanjšali možnost napačne klasifikacije "slabega izposojevalca". Kot je priporočeno po [9] so bili v članku vsi eksperimenti opravljeni s stroški v razmerju 5 proti 1, kar pomeni, da ima napačna klasifikacija "slabega izposojevalca" 5-krat večji strošek kot napačna klasifikacija "dobrega izposojevalca". V ta namen je bil v programu WEKA uporabljen klasifikator $CostSensitiveClassifier$.

8.2. **Kriteriji za ustrezno izbiro klasifikacijske metode.** Za izbiro klasifikacijske metode strojnega učenja je bil v članku uporabljen popularen 5-stopenjski križni sistem preverjanja, ki zagotavlja kompromis med prevelikim in premajhnim prileganjem modela [10]. V ta namen je bila datoteka 68000 posojil razdeljena v 5 enako velikih skupin podatkov. Ena od petih skupin je bila namenjena testiranju, ostale štiri pa so bile uporabljene kot učne množice. Opisani proces prikazuje slika 7 [1].



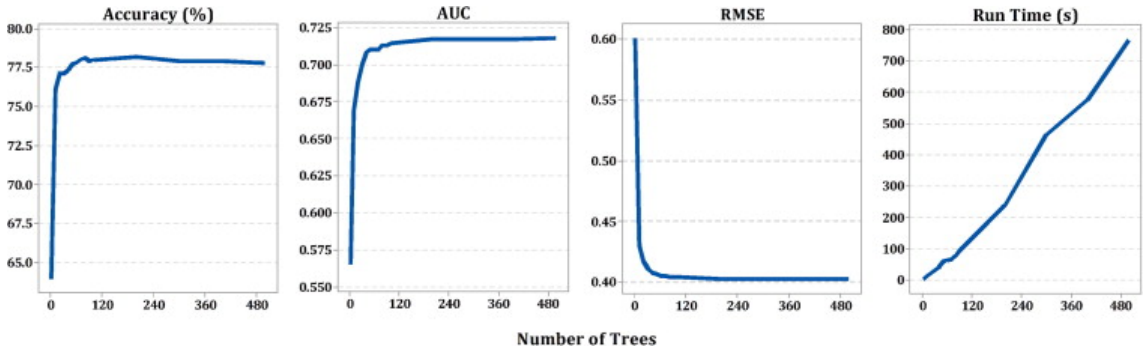
Slika 7. 5 stopenjski križni sistem preverjanja

8.3. **Metrike ustreznosti klasifikacijskih metod.** Za oceno ustreznosti klasifikacijskih metod v modelu so bile upoštevane naslednje metrike, ki izhajajo iz standardne metodologije strojnega učenja:

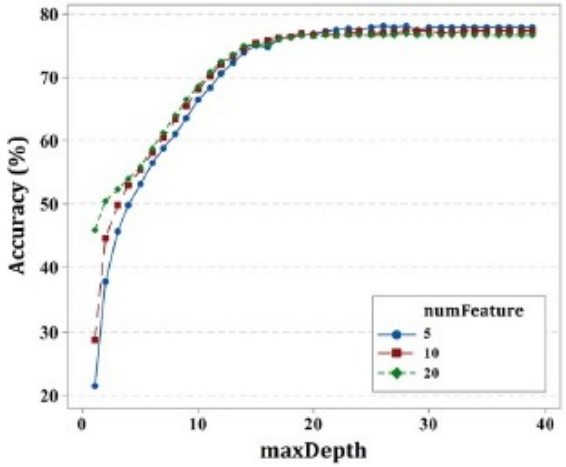
- splošna stopnja natančnosti razvrščanja testne množice (ang. overall classification accuracy rate on the test slice). To je stopnja natančnosti metode strojnega učenja, ki se uporabi za klasifikacijo. V našem primeru klasifikacije "dobrega" in "slabega izposojevalca" to pomeni odstotek pravilno prepoznanih izposojevalcev, "dobrih" za "dobre" in "slabih" za "slabe".
- območje pod sprejemno operativno karakteristično krivuljo - AUC (ang. the usual area under the Receiving Operating Characteristic - ROC) za "dober" razred v testni množici podatkov. ROC graf, ki prikazuje sposobnost oziroma natančnost napovedovanja binarnega klasifikacijskega sistema [18].
- koren povprečne kvadratne napake (ang. root mean square error - RMSE) [31];
- pravilno pozitivna (ang. true positive - TP) stopnja. To je stopnja pravilne klasifikacije za "dober" razred v testni množici. V statistiki TP predstavlja verjetnost pravilne zavrnitve ničelne hipoteze [19].
- napačno pozitivna (ang. false positive - FP) stopnja. To je stopnja pravilne klasifikacije za "slab" razred v testni množici. V statistiki FP predstavlja verjetnost nepravilne zavrnitve ničelne hipoteze [20].

8.4. **Konstrukcija naključnega gozda.** Pred začetkom primerjave metod strojnega učenja je bilo potrebno poiskati optimalno velikost gozda, število finančnih lastnosti, ki se bodo pojavljala v klasifikacijskih drevesih m in globino dreves d. V ta namen sta avtorja članka z uporabo opcije *auto* v programu WEKA za parametra m in d konstruirala gozd z velikostjo od 1 do 500 dreves s povečevanjem za 10 dreves. Slika 8 [1] prikazuje grafe metrik, opisanih v prejšnjem podpoglavju, za naključne gozdove velikosti od 1 do 500.

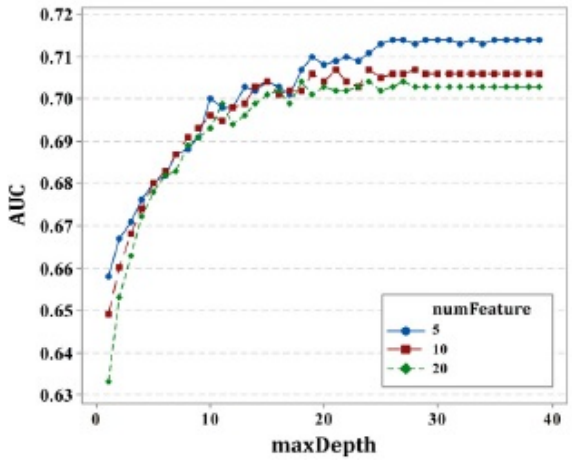
Iz grafov vidimo, da velikost gozda pod 80 drevesni ne zagotavlja optimalne izvedbe, čeprav se z večanjem števila dreves povečuje čas delovanja programa. Vezano na omenjena problema sta se avtorja članka odločila, da je velikost gozda z 80 drevesi primeren kompromis med časom delovanja programa in klasifikacijsko izvedbo.



Predlagano število lastnosti, po katerih se bodo opravljale delitve, je $log_2(p)$ po [21], kar je v našem primeru 5. Za določanje optimalnega števila m sta avtorja naredila simulacijo za $m = 5, 10, 20$. Za vsako število m sta vzela parameter d med 1 in 40 in velikost gozda 80. Rezultate simulacije prikazujeta sliki 9 in 10 [1].



Slika 8. Graf stopnje natančnosti v odvisnosti od maksimalne globine dreves, glede na število finančnih lastnosti, uporabljenih v drevesih m



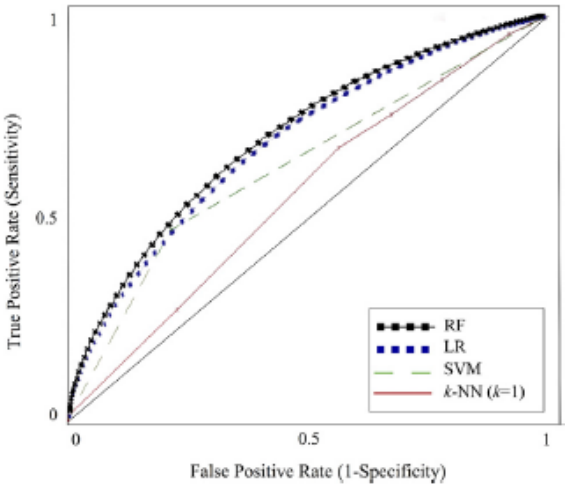
Slika 9. Fraf AUC v odvisnosti od maksimalne globine dreves, gledena na število finančnih lastnosti uporabjenih v drevesih m

Iz grafov lahko vidimo, da je $m = 5$ res optimalno število finančnih lastnosti in da sta stopnja natančnosti in AUC skoraj konstantni po $d = 25$. Zato sta se avtorja odločia, da bosta v primerjavi metod za metodo naključnega gozda uporabila velikost gozda 80, $m = 5$ in $d = 25$.

8.5. **Rezultati primerjave metod strojnega učenja.** Rezultati primerjave metod strojnega učenja avtorjev članka v programu WEKA prikazujeta slika 11 in slika 12 [1], tabele in grafa.

Rank	Classifier	Accuracy (%)	AUC	RMSE	TP Rate		FP Rate	
					Good	Bad	Good	Bad
1	Random forest	78.0	0.71	0.42	0.88	0.31	0.69	0.13
2	Nearest neighbor	70.1	0.53	0.55	0.82	0.25	0.74	0.18
3	Support vector machine	63.3	0.62	0.68	0.47	0.78	0.22	0.53
4	Logistic regression	54.5	0.68	0.51	0.49	0.77	0.23	0.51

Slika 10. Tabela primerjave metod strojnega učenja glede na metrike

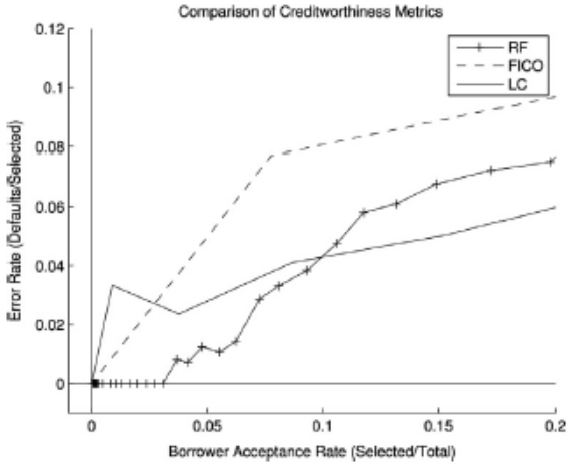


Slika 11. Primerjava ROC krivulje metod strojnega učenja

Iz tabele vidimo, da ima metoda naključnih gozdov najvišjo stopnjo natančnosti z 78.0 % in najvišjo vrednost AUC z 0.71 %, pri tem pa ima najnižjo vrednost RMSE z 0.42. Iz tega lahko povzamemo, da je metoda naključnih gozdov najbolj primerna metoda za klasifikacijo v našem primeru.

8.6. **Primerjava metode naključnih gozdov s FICO točkami in LC ocenami.** Metodologija avtorjev članka je bila naslednja. Izračunati stopnjo napačne klasifikacije za vse tri načine ugotavljanja kreditne ustreznosti pri enaki stopnji sprejema. Za točke FICO in oceno s strani LC sta avtorja primerjala stopnjo med nepoplačanimi posojili in številom posojil, ki se navezuje na stopnjo sprejema, za metodo naključnih gozdov pa sta gledala napako pri klasifikaciji. To pomeni, če bi sprejemali samo izposojevalce s točkami FICO nad 750, bi sprejeli približno 8 % izposojevalcev. V tem primeru bi bila stopnja nepoplačila posojila 8.2 %. To

je stopnja napačne klasifikacije, saj so posojila s točkami FICO nad 750 najmanj tvegana posojila, kar pomeni, da naj bi bila vsa poplačana.
Nato sta avtorja poiskala prag, ki ustreza točno tej 8% stopnji sprejema. Metoda naključnih gozdov je imela v tem primeru samo 3.1% stopnjo napačne klasifikacije. To pomeni, da je metoda 3.1% posojil klasificirala narobe, če vzamemo število, ki pripada 8 % kot celoto. Narobna klasifikacija je, če metoda "dobra" posojila prepozna kot "slaba" in "slaba" kot "dobra". Iz tega lahko sklepamo, da je metoda naključnih gozdov za identifikacijo "dobrega izposojevalca" boljša od točk FICO.
Rezultate primerjave metode naključnih gozdov, točk FICO in ocene s strani LC prikazuje slika 13 [1], slika grafa, ki prikazuje primerjavo omenjenih metod za stopnjo sprejema med 0.1% in 20%.



Slika 12. Graf stopnje med nepoplčanimi posojili in številom posojil, v odvisnosti od stopnje sprejema izposojevalca

Vidimo, da do 3 % metoda naključnih gozdov ne prepozna nobenega "slabega izposojevalca" kot "dobrega izposojevalca" in nobenega "dobrega izposojevalca" kot "slabega izposojevalca".
Opazimo tudi, da je do 10% stopnje sprejema metoda naključnih gozdov boljša od ostalih dveh metod za identifikacijo "dobrega izposojevalca". Rezultat za metodo naključnih gozdov nad 10% stopnjo sprejema je posledica napačne idnetifikacije "dobrega izposojevalca" kot "slabega izposojevalca". Kljub temu, da je metoda naključnih gozdov boljša od točk FICO, pa je za 10 % - 20 % najboljša metoda ocene s strani LC.

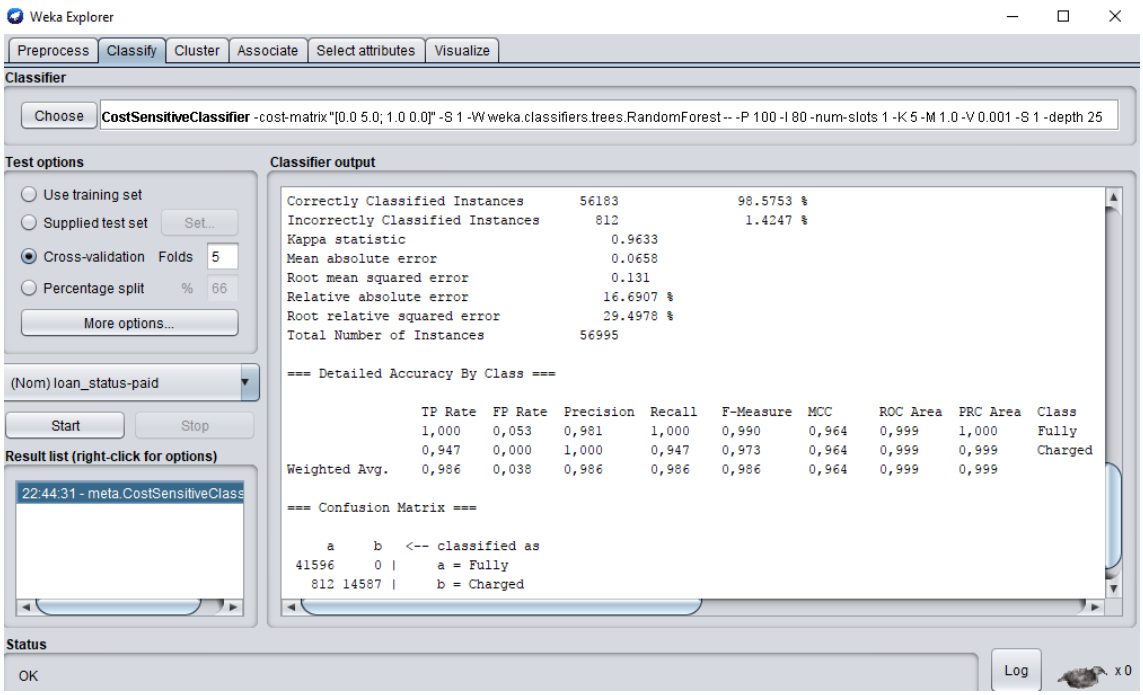
9. Predstavitev lastnih rezultatov

9.1. **Podatki.** Za samostojno analizo rezultatov članka sem uporabil program WEKA, različica 3.9, ki velja tudi za razvijalsko različico. Tako kot avtorja članka sem tudi sam podatke pridobil z uradne spletne strani Lending Cluba [17]. S spletne strani sem iz več excelovih dokumentov sestavil en excelov dokument, v katerem je bilo, tako kot pri avtorjih članka, približno 350000 kreditnih zapisov izposojevalcev za obdobje med januarjem 2012 in septembrom 2014. Nato sem podatke filtriral glede na odobrenost in zapadlost. Dobil sem 56995 kreditnih zapisov. Krediti, ki sem jih dobil, so imeli status "fully paid", kot "odplačani", kar pomeni, da lahko takega izposojevalca uvrstimo med "dobre izposojevalce" in status "charged-off", kot krediti, ki niso bili odplačani in za katere spletna posojilnica ne pričakuje nadaljnjih

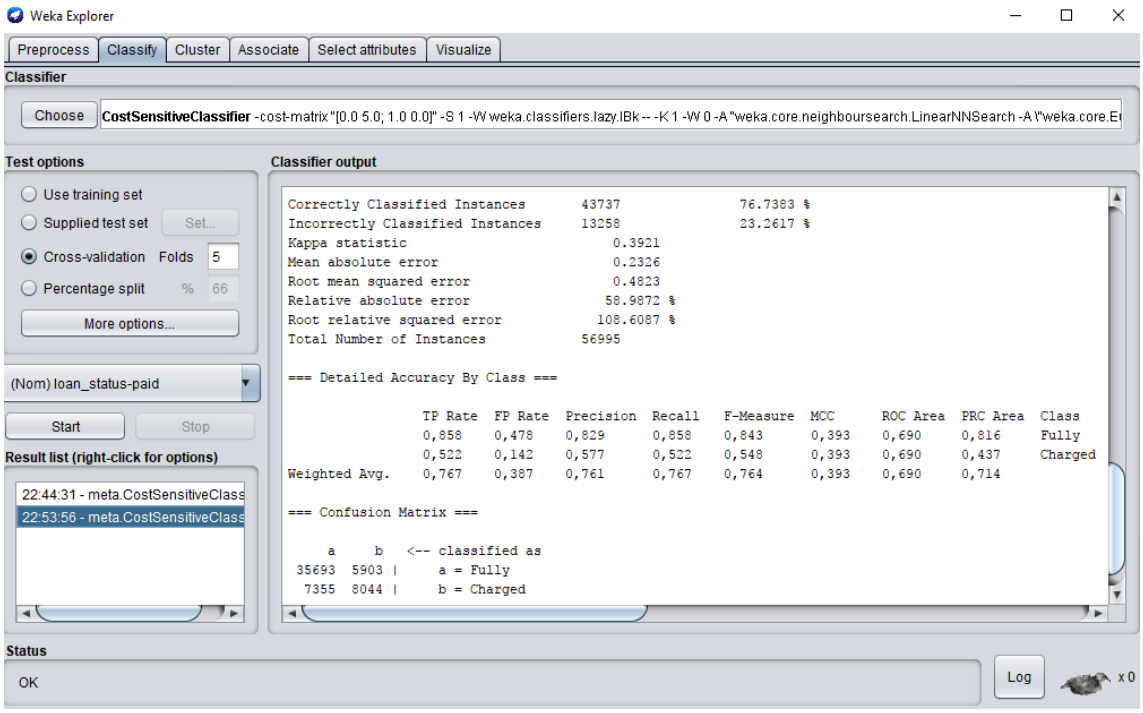
plačil. Take izposojevalce prepoznamo kot "slabe izposojevalce". Podatke sem nato pretvoril v obliko *arff*. To je oblika datoteke, ki jo bere program WEKA.

9.2. **Analiza s programom WEKA.** Kot je omenjeno v članku, sem tudi sam v programu WEKA uporabil klasifikator $CostSensitiveClassifier$ in s tem utežil napačno klasifikacijo slabega izposojevalca v primerjavi z napačno klasifikacijo dobrega izposojevalca z razmerjem 5 proti 1. Za grajenje tesne množice sem, kot je opisano v prejšnjem poglavju, uporabil 5-stopenjski križni sistem preverjanja.

9.3. **Primerjava metode naključnih gozdov in k-najbližjih sosedov.** Po rezultatih članka sta to edini metodi, ki imata več kot 70% natančnost. Na podlagi tega sem se odločil, da ju bom primerjal v programu WEKA in s tem pokazal, katera metoda je res najboljša za identificiranje "slabega izposojevalca".
Naključni gozd v programu WEKA najdemo med skupino klasifikatorjev, poimenovanih *trees*, pod imenom *RandomForests*. Za definiranje naključnega gozda sem uporabil enake lastnosti, kot je opisano v prejšnjem poglavju. Velikost gozda sem nastavil na 80, maksimalno globino posamezne veje d na 25 in število naključno izbranih lastnosti m na 5.
Metodo k-najbližjih sosedov pa v programu WEKA najdemo med skupino klasifikatorjev, poimenovanih *lazy*, pod imenom IBk. Za definiranje metode sem za število sosedov uporabil $k = 1$, kar je tudi privzeta opcija v programu, saj avtorja članka posebej ne navajata števila k, ki sta ga uporabila za svojo analizo.
Rezultate klasifikacij prikazujeta sliki 14 in 15, sliki konzol programa WEKA. V konzolah so odplačana posojila poimenovana kot "Fully" in nepoplačana kot "Charged".



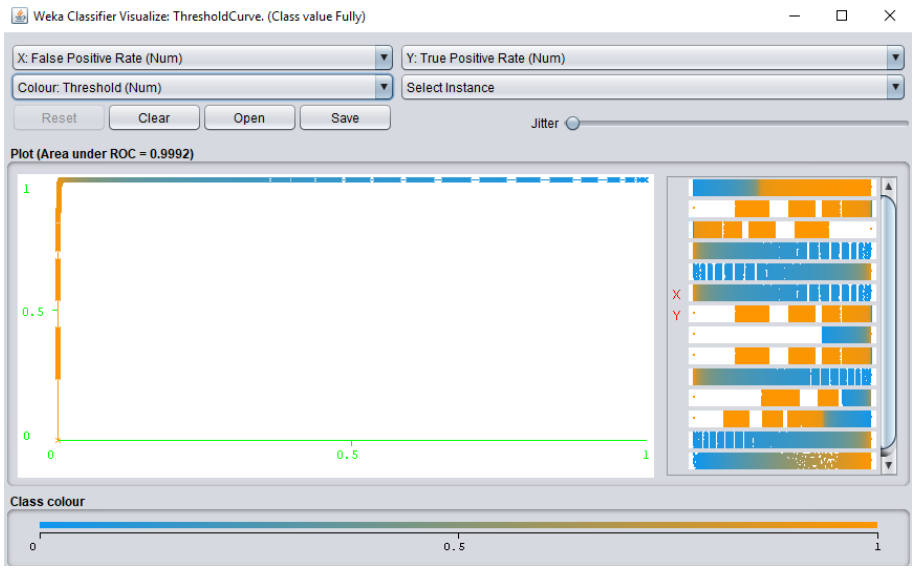
SLIKA 13. Konzola programa WEKA po opravljeni metodi naključnih gozdov z upoštevanjem stroškovne matrike



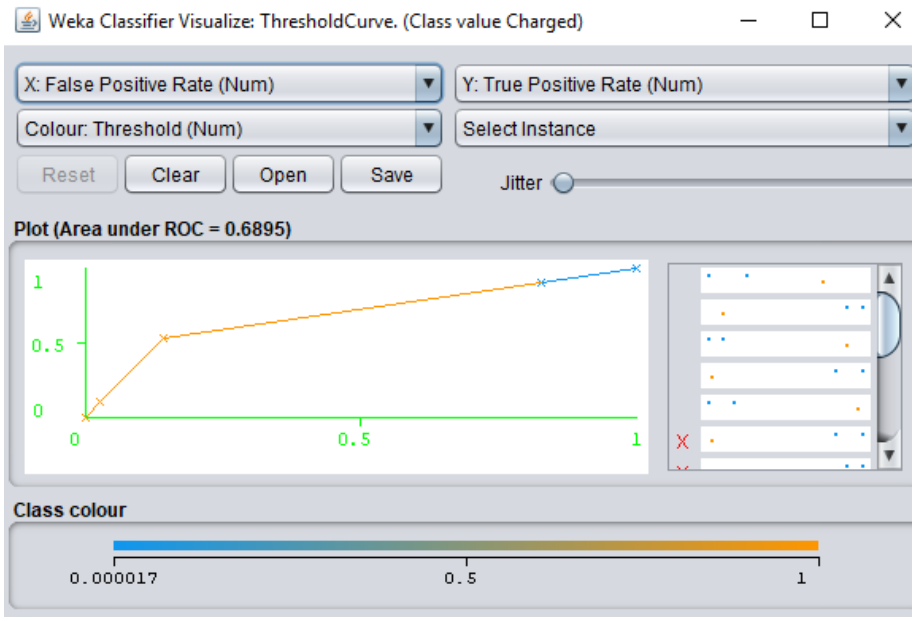
SLIKA 14. Konzola programa WEKA po opravljeni metodi k-najbližjih sosedov z upoštevanjem stroškovne matrike

Iz slike konzole za metodo naključnih gozdov vidimo, da ima metoda za moje podatke 98.5753% natančnost klasifikacije, pri čemer je bilo 41596 "dobrih izposojevalcev" prepoznanih za "dobre", 812 "dobrih izposojevalcev" prepoznanih za "slabe", 14587 "slabih izposojevalcev" prepoznanih za "slabe" in 0 "slabih izposojevalcev" prepoznanih za "dobre".
Razlika v natančnosti klasifikacije med dobrimi in slabimi izposojevalci je posledica uporabe klasifikatorja $CostSensitiveClassifier$, saj stroškovna matrika upošteva, da ima klasifikacija "slabega izposojevalca" kot "dobrega izposojevalca" 5-krat večji strošek, zato se program WEKA bolj osredotoči na pravilno klasifikacijo "slabega izposojevalca". Iz slike 14 vidimo, da ima za metodo naključnih gozdov koren povprečne kvadratne napake vrednost 0.131. Iz slike 15 pa lahko za metodo k-najbližjih sosedov razberemo naslednje podatke. Natančnost metode je 76.7383%, pri čemer je bilo 35693 "dobrih izposojevalcev" prepoznanih za "dobre", 7355 "dobrih izposojevalcev" prepoznanih za "slabe", 8044 "slabih izposojevalcev" prepoznanih za "slabe" in 5903 "slabih izposojevalcev" prepoznanih za "dobre". Koren povprečne kvadratne napake za metodo k-najbližjih sosedov ima vrednost 0.4823. ROC krivulji metod prikazujeta naslednji sliki 16 in 17.

Iz slik 16 in 17 lahko vidimo, da ROC krivulja za metodo naključnih gozdov dosega višje vrednosti kot ROC krivulja metode k-najbližjih sosedov.
Iz vseh zgoraj opisanih podatkov analize metode naključnih gozdov in metode k-najbližjih sosedov lahko pridemo do zaključka, da je metoda naključnih gozdov za klasifikacijo "dobrega" in "slabega" izposojevalca boljša od metode k-najbližjih sosedov in posledično najprimernejša metoda za klasifikacijo med metodami strojnega učenja opisanih v tem delu.

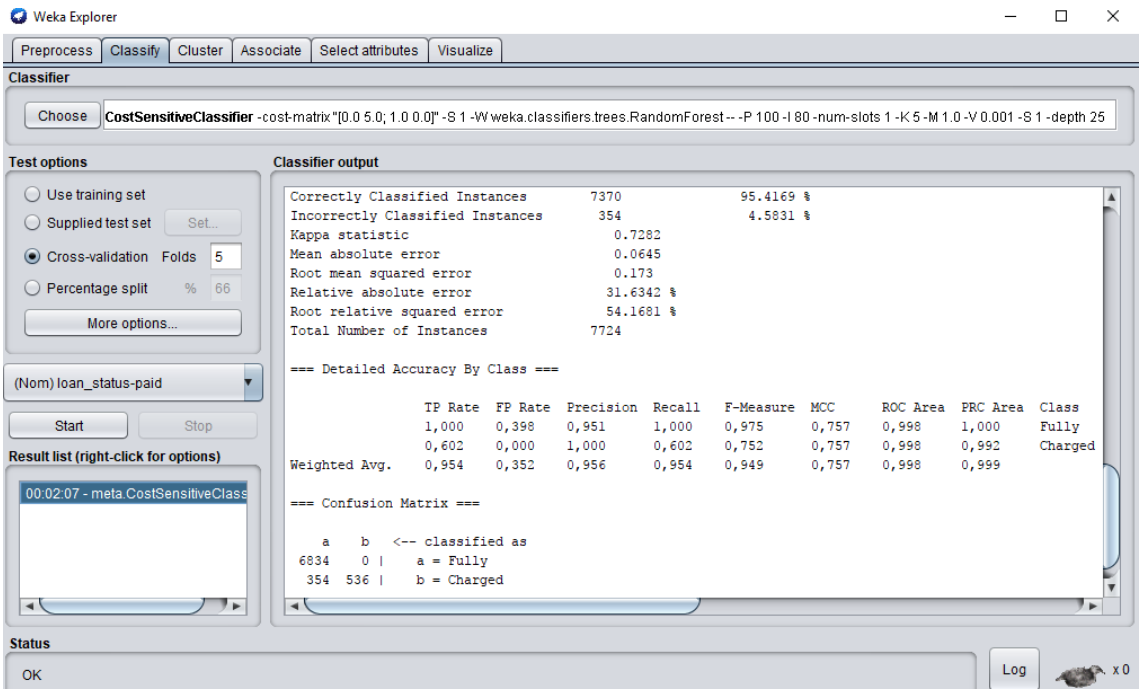


Slika 15. ROC krivulja za metodo naključnih gozdov



Slika 16. ROC krivulja za metodo k-najbližjih sosedov

9.4. **Primerjava modete naključnih gozdov z metodama točk FICO in ocene s strani LC.** Pri primerjavi metod sem uporabil enako metodologijo kot avtorja članka. Ta metodologija je opisana v prejšnjem poglavju. Edina razlika, ki sem jo bil primoran narediti, je ta, da sem moral prag sprejema gledati glede na oceno s strani LC, saj podatek o točkah FICO posojil ni bil podan v podatkih, ki sem jih pridobil iz spletne strani [17] in sem jih uporabil za analizo.
Podatke sem filtriral glede na oceno s strani LC "A", ki prinaša najnižjo obrestno mero. To pomeni, da naj bi ta ocena z največjo verjetnostjo zagotavljala vračilo posojila. S tem sem dobil 7724 posojil, kar predstavlja 13.55%. Od tega jih ima 890 status "Charged". To je 11.52%. Ko sem iste podatke pognal v programu WEKA, sem dobil naslednje rezultate. Rezultate prikazuje slika 18.



Slika 17. Konzola programa WEKA po opravljeni metodi naključnih gozdov za podatke z oceno s strani LC "A"

Iz konzole vidimo, da metoda naključnih gozdov 354 posojil med 7724 posojili klasificira narobe, kar predstavlja 4.58 %.
Če za stopnjo sprejema 13.55 % vzamem podatek, ki sta ga pridobila avtorja članka, to je približno 9 %, vidim, da ima metoda naključnih gozdov najmanjšo stopnjo napačne napovedi.
S tem sem pokazal, da je metoda naključnih gozdov boljša od metod točk FICO in ocene s strani LC za identifikacijo "slabega izposojevalca".

10. Zaključek

V delu diplomskega seminarja sem se osredotočil na opisovanje in uporabo nekaterih popularnih metode strojnega učenja, ki se uporabljajo za klasifikacijo primerov. Na kratko je opisano tudi, kaj izraz strojno učenje na splošno sploh pomeni. V delu so opisani pojmi, ki so se pojavili pri definiranju opisanih metod strojnega učenja. Če pojem v delu ni posebej opisan, je ob njem podana referenca vira, kjer se opis nahaja. Tako ima bralec skupaj z viri popoln vpogled v tematiko dela diplomskega seminarja.
V predzadnjem poglavju so opisani rezultati, ki sem jih dobil, ko sem klasifikacijo izposojevalcev s programom WEKA naredil tudi sam. Poskusil sem kar najbolje rekonstruirati postopek, ki sta ga izvedla avtorja večkrat omenjenega članka, ki je bil tudi povod tematike tega dela diplomskega seminarja. Kljub temu, da sem postopke naredil tako, kot je bilo napisano v članku, nisem dobil povsem identičnih rezultatov. Razlika je nastala že v podatkih, saj so se podatki, ki sem jih, tako kot avtorja, pridobil na spletni strani [17], razlikovali po številu posojil. Razlika je bila v okoli 10000 posojilih.
Na začetku dela sem si zadal cilj, da bom dokazal trditev avtorjev članka. Ta pravi, da je metoda naključnih gozdov za identifikacijo "dobrega" in "slabega" izposojevalca

boljša kot metodi, ki se za ugotavljanje kreditne ocene uporabljata v praksi. Kljub vsemu lahko trdim, da mi je to trditev uspelo dokazati.

risk assessment ocena tveganja
peer-to-peer (P2P) lending posojanje denarja prek spleta
social lending platform spletna posojilnica
lender posojevalec
borrower izposojevalec
machine learning strojno učenje
random forests naključni gozd
logistic regression gosta množica
support vector machines metoda podpornih vektorjev
k**-nearest neighbors** metoda k-najbližjih sosedov

Literatura

[1] M. Malekipirbazari, V. Aksakalli, *Risk assessment in social lending via random forests*, Expert Systems with Applications, **42** (2015) 4621–4631.

[2] R. Emekter, Y. Tu, B. Jirasakuldech, M. Lu *Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending*, Applied Economics, **47** (2015) 54–70.

[3] G. James, D. Witten, T. Hastie, R. Tibshirani *An introduction to statistical learning*,**1**, Springer, New York, Združene države Amerike, 2013.

[4] L. Wasserman, *Lecture Notes 15*, 20.4.2010, [ogled 4. 8. 2017], dostopno na `http://www.stat.cmu.edu/~larry/=stat705/Lecture15.pdf`.

[5] I.H. Witten,E. Frankl, *Data Mining: Practical Machine Learning Tools and Techniques*,**1**, Morgan Kaufmann, San Francisco, Združene države Amerike, 2005.

[6] N. Christianini, J. Shawe- Taylorl, *An introduction to Support Vector Machines and other kernel-based methods*,**1**, Cambridge University Press, Cambridge, Združeno kraljestvo Velike Britanije in Severne Irske, 2000.

[7] B. Scholkopf, A. J. Smola, *Learning with kernels*,**1**, MIT Press Cambridge, Cambridge, Združene države Amerike, 2001.

[8] How a FICO Score breaks down, [ogled 12. 11. 2016], dostopno na `http://www.myfico.com/credit-education/whats-in-your-credit-score/`.

[9] D. Baier, K.D. Wernecke,*Innovations in Classification, Data Science, and Information Systems*,**1**, Springer, Cottbus, Nemčija, 2003.

[10] C.L. Huang, M.C. Chen, C.J. Wang, *Credit scoring with a data mining approach based on support vector machines*, Expert Systems with Applications, **33** (2007) 847–856.

[11] Loan Grades, [ogled 12. 11. 2016], dostopno na `https://www.lendingclub.com/foliofn/rateDetail.action`.

[12] Example of k-NN classification, [ogled 16. 5. 2017], dostopno na `https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm`.

[13] Maximum-margin hyperplane, [ogled 17. 5. 2017], dostopno na `https://en.wikipedia.org/wiki/Support_vector_machine`.

[14] Support Vector Machine, [ogled 17. 5. 2017], dostopno na `https://www.linkedin.com/pulse/support-vector-machine-srinivas-kulkarni`.

[15] Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic, [ogled 20. 5. 2017], dostopno na `http://file.scirp.org/Html/6-9101686_31887.htm`.

[16] Decision Tree, [ogled 17. 5. 2017], dostopno na `http://decisiontree.kleinmediation.com/`.

[17] Lending Club Statistics, [ogled 20. 9. 2017], dostopno na `https://www.lendingclub.com/info/download-data.action`.

[18] Receiver operating characteristic, [ogled 20. 9. 2017], dostopno na `https://en.wikipedia.org/wiki/Receiver_operating_characteristic`.

[19] Sensitivity and specificity, [ogled 20. 9. 2017], dostopno na `https://en.wikipedia.org/wiki/Sensitivity_and_specificity`.

[20] False positive rate, [ogled 20. 9. 2017], dostopno na `https://en.wikipedia.org/wiki/False_positive_rate`.

[21] L. Breiman, *Random Forests*, Machine Learning, **45** (2001) 5–32.

[22] Euclidean distance, [ogled 26. 11. 2017], dostopno na `https://en.wikipedia.org/wiki/Euclidean_distance`.

[23] B. Plestenjak, *Numerične metode: delovna verzija*, 4.3.2010, [ogled 26. 11. 2017], dostopno na `https://www.fmf.uni-lj.si/~kozak/PedagoskoDelo/Gradiva/NumericneMetodeI_praktiki/Skripta/BorPlestenjakKnjigaNM.pdf`.

[24] M. Pohar Perme, *Verjetnost in statistika z nalogami*, 2014, [ogled 26. 11. 2017], dostopno na `http://maksi2.ef.uni-lj.si/zaloznistvoslike/432/Pohar%20Perme_Verjetnost%20in%20statistika%20z%20nalogami.pdf`.

[25] Regression analysis, [ogled 26. 11. 2017], dostopno na `https://en.wikipedia.org/wiki/Regression_analysis`.

[26] Hiperravnina, [ogled 26. 11. 2017], dostopno na `https://sl.wikipedia.org/wiki/Hiperravnina`.

[27] Skalarni produkt, [ogled 26. 11. 2017], dostopno na https://sl.wikipedia.org/wiki/Skalarni_produkt.
[28] Normala, [ogled 26. 11. 2017], dostopno na https://hr.wikipedia.org/wiki/Normala.
[29] Boosting, [ogled 27. 11. 2017], dostopno na https://en.wikipedia.org/wiki/Boosting_(machine_learning)
[30] Bagging, [ogled 27. 11. 2017], dostopno na https://en.wikipedia.org/wiki/Bootstrap_aggregating
[31] Root-mean-square deviation, [ogled 27. 11. 2017], dostopno na https://en.wikipedia.org/wiki/Root-mean-square_deviation