View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by ePrints.FRI

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Benjamin Novak

Kako uspeti na Kickstarter-ju?

diplomsko delo

univerzitetni študijski program
prve stopnje
računalništvo in informatika

Mentor: prof. dr. Blaž Zupan

Ljubljana, 2016

Fakulteta za računalništvo in informatiko podpira javno dostopnost znanstvenih, strokovnih in razvojnih rezultatov. Zato priporoča objavo dela pod katero od licenc, ki omogočajo prosto razširjanje diplomskega dela in/ali možnost nadaljne proste uporabe dela. Ena izmed možnosti je izdaja diplomskega dela pod katero od Creative Commons licenc http://creativecommons.si

Morebitno pripadajočo programsko kodo praviloma objavite pod, denimo, licenco *GNU General Public License, različica 3.* Podrobnosti licence so dostopne na spletni strani http://www.gnu.org/licenses/.

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kako uspeti na Kickstarter-ju?

Tematika naloge:

V diplomski nalogi zgradite model, s katerim lahko napoveste uspeh predlogov za financiranje projektov na platformi za množično financiranje Kickstarter. Pridobite podatke za pretekle projekte, te opišite v atributni obliki in preučite napovedno uspešnost algoritmov strojnega učenja. Poročajte tudi o najpomembnejših atributih.

Iskreno se zahvaljujem mentorju prof. dr. Blažu Zupanu za strokovno vodenje, za vse predloge in pomoč pri izdelavi diplomske naloge. Posebna zahvala gre tudi punci Evi za podporo, motivacijo in pomoč pri pregledovanju diplomske naloge ter družini za podporo pri študiju.

Kazalo

Povzetek

Platforme za množično financiranje, kot je Kickstarter, v zadnjih letih postajajo vedno bolj priljubljene. Na njih poskušajo razvojne ekipe s kreativnimi projekti pridobiti bodoče kupce in podpornike. Vendar uspeh ni zagotovljen, saj je skoraj dve tretjini predlogov neuspešnih. V nalogi smo iz portala za množično financiranje Kickstarter pridobili podatke o opisu in uspešnosti projektov. Naš cilj je bil zgraditi model, ki bi iz opisa projekta znal napovedati uspešnost kampanje in bi skladno z rezultati v sorodnih delih dosegel točnost napovedi AUC vsaj 0,85. V delu predstavimo našo rešitev in tehnike strojnega učenja, ki smo jih uporabili. Zgrajene modele smo vrednotili s prečnim preverjanjem in na novih projektih. Ugotovili smo, da so pri napovedovanju uspešnosti najpomembnejši število projektov, ki jih je avtor podprl, ciljna vsota, število slik v opisu projekta in število ponujenih nagrad. Na testnih podatkih novih projektov smo dosegli točnost AUC = 0,93.

Ključne besede: Kickstarter, množično financiranje, napovedovanje uspešnosti, strojno učenje, klasifikacija, iskanje značilk.

Abstract

Crowdfunding platforms such as Kickstarter are becoming increasingly popular. These platforms are widely used by development teams which are trying to get new buyers and supporters using different creative projects. However, success is not guaranteed since two thirds of the project suggestions fail to achieve their goal. In our thesis, we gathered descriptions and success of different projects on Kickstarter. Our goal was to create a model that could predict success of project compaigns. With this model, we also wanted to reach prediction accuracy AUC = 0,85 that could be compared with the results of other related studies. In the thesis, we present our solution and techniques of machine learning that were used to gather data. These models were later assessed with cross validation and new projects. The results showed that the most important attributes are the number of the projects supported by the author, the goal, the number of pictures in the description of the project and the award number. AUC score accomplished on the test data of the new projects was 0,93.

Keywords: Kickstarter, crowdfunding, machine learning, classification, feature engineering.

Poglavje 1

Uvod

Spletne strani in okolja za množično financiranje v zadnjih letih pridobivajo na svoji priljubljenosti in prepoznavnosti. Med njimi izstopa predvsem Kickstarter[1], ki temelji na modelu financiranja vse ali nič. Na ta način projekti umetnikov in podjetnikov, ki se preizkusijo na platformi dobijo celoten znesek samo, če zberejo donacije višje od postavljenega cilja v zadanem časovnem roku. Drugače je predlog projekta označen za neuspešnega in ne pridobi doniranih sredstev.

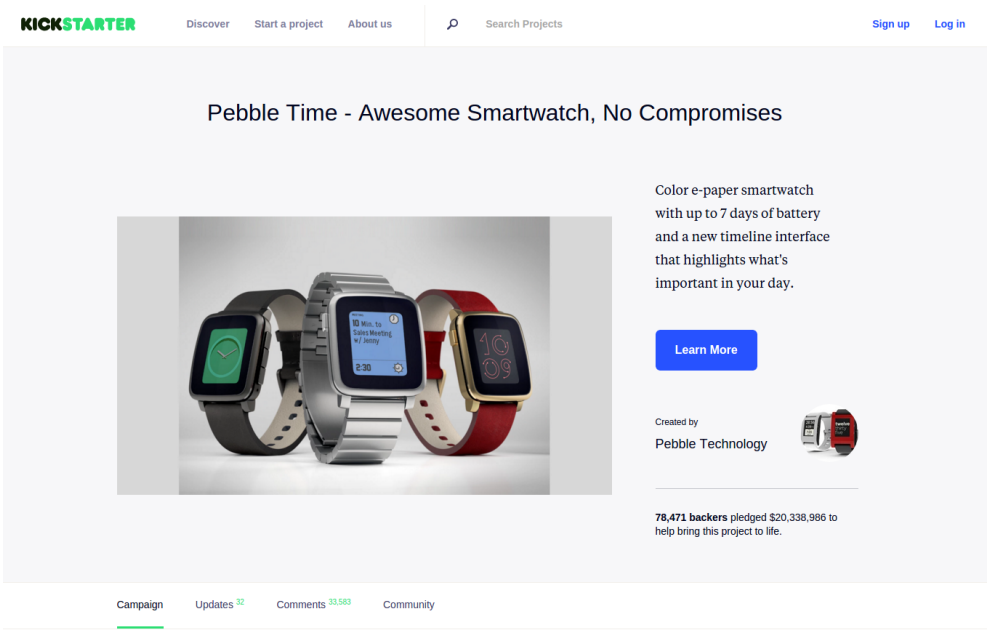
Vsak projekt na Kickstarter-ju ima svojo spletno stran (primer na sliki 1.1), preko katere ustvarjalci sporočajo svojim potencialnim donatorjem in podpornikom podrobnosti projekta. Predlagatelji projekta lahko v opisu navedejo različne zneske donacij in njim pripadajoče nagrade za podpornike. Poleg podrobnih podatkov o samem projektu so na strani vidni tudi podatki o ustvarjalcih, ki vsebujejo število dosedanjih projektov, doniranih sredstev, podatek o povezanosti na socialna omrežja in druga.

Do 13. marca 2016 je bilo na Kickstarter-ju vloženih 287.807 predlogov projektov, ki so skupaj zbrali več kot 2 milijardi dolarjev doniranega denarja. Vendar pa je le 36,11% vseh predlogov doseglo cilj in bilo uspešnih, kar poročajo podatki na uradni strani[2]. Zato se nam postavljata vprašanji:

[1] `https://www.kickstarter.com`
[2] `https://www.kickstarter.com/help/stats`

1. Lahko uspeh predloga projekta napovemo že pred začetkom kampanije iz opisa ali pa iz podatkov o številu objav in njihovih ogledov projekta na socialnih omrežjih?

2. Od katerih dejavnikov je najbolj odvisen uspeh (ali neuspeh) kampanije na Kickstarter-ju?



Slika 1.1: Stran najbolj uspešnega projekta na Kickstarter-ju - Pebble.

Cilj in čas kampanije, ki se določita na začetku vsake kampanije, močno vplivata na uspešnost projekta. Manjši projekti oz. projekti z nižjimi cilji v povprečju dosegajo boljše rezultate. Portal Mashable[3] poroča, da je do junija leta 2012 povprečje cilja uspešnih kampanij na Kickstarter-ju znašalo okoli \$5.000, neuspešnih pa kar \$16.000. Podobno velja tudi za trajanje zbiranja sredstev, kategorijo projekta in znesek cilja, kot kaže tabela 1.1. Čeprav podatki niso aktualni, prikazujejo osnovne zakonitosti in vpliv posameznih atributov glede na rezultat kampanije. Torej se že iz osnovnih atributov pro-

[3] `http://mashable.com/2012/06/12/kickstarter-failures`

jekta lahko učimo in novim projektom pomagamo do boljše izbire lastnosti predlaganega projekta in s tem boljših rezultatov.

Tabela 1.1: Podatki iz portala Mashable iz leta 2012.

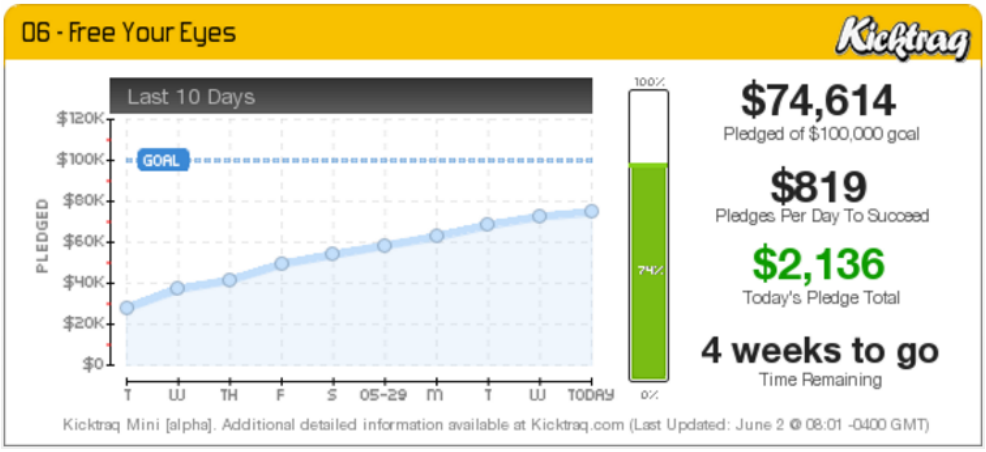
	Projekti	Donacije (USD)	Zbiranje sredstev	Cilj (USD)
Uspešen	22.902	197.592.643	38 dni	5.487
Neuspešen	18.939	16.995.691	43 dni	16.365
Skupaj	45.815	214.588.334	40 dni	10.388

1.1 Obstoječi pristopi

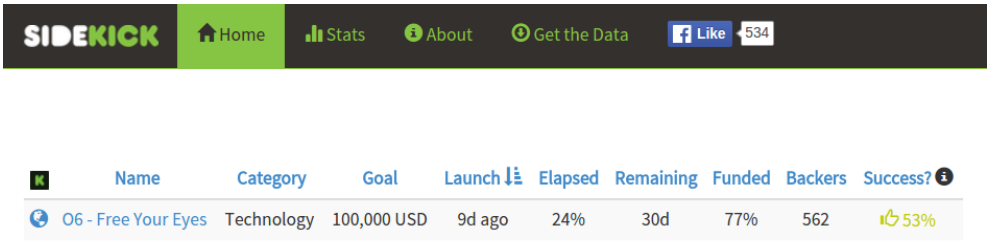
Napoved uspešnosti projekta poleg ustvarjalcev zanima tudi podpornike. Če je verjetnost uspeha majhna, ne bodo donirali svojega denarja ali pa bodo, v kolikor v idejo verjamejo, poskušali preko socialnih omrežij spodbuditi svoje prijatelje, da tudi sami podprejo projekt. V pomoč jim je lahko spletna stran Kicktraq[4], ki jo prikazuje slika 1.2. Stran sledi projektom na Kickstarter-ju in objavlja trend ter pričakovano vrednost donacij, ki jo bo posamezen projekt zbral do konca svoje kampanije. Slika 1.3 prikazuje spletno stran Sidekick[5], ki uporablja podatke o sledenju projektov iz Kicktraq-a in uspešnost projektov napoveduje z verjetnostjo.

Napovedni portal Sidekick je plod raziskave skupine znanstvenikov iz Lausanne [5], ki poleg podatkov iz spletne strani projekta na Kickstarter-ju uporablja tudi podatke iz socialnih omrežij in vzorčnih podatkov o spreminjanju donacij skozi čas. Izmed podatkov iz socialnih omrežij so uporabili podatke iz Twitterja, saj so javni in ni težav z zasebnostjo kot pri drugih omrežjih. S pomočjo uporabniškega vmesnika Twitter (angl. *Twitter Streaming API*) so pridobili število čivkov, njihovih odgovorov in število sledilcev. Tem atributom so dodali želeno ciljno vsoto, trajanja kampanije, podatke o avtorju in številu donatorjev in podatke o spreminjanju vsote donacij. Za

[4] `http://www.kicktraq.com`
[5] `http://sidekick.epfl.ch/`



Slika 1.2: Podatki kampanije O6 - Free Your Eyes na Kicktraq-u.



Slika 1.3: Napovedana uspešnost kampanije O6 - Free Your Eyes na portalu Sidekick.

napovedovanje uspešnosti so uporabili metodo k najbližjih sosedov in napovedovanje z Markovskim modelom, kjer so dobili boljše rezultate. Poročajo o 76% klasifikacijski točnosti štiri ure po začetku kampanije in 85% po 15% časa trajanja projekta.

Za projekt KickPredict iz Kalifornije [1] so uporabili podatke iz strani projekta na Kickstarter-ju in podatke iz socialnih omrežij (Twitter in Youtube). Za izgradnjo modela so uporabili naslednje atribute:

- prisotnost avtorja na Facebook-u,
- število dosedanjih projektov avtorja,
- število kampanij, ki jih je avtor podprl,

- ciljna vsota donacij,
- trajanje zbiranja denarja,
- število predlaganih nagrad, največja in najmanjša vsota,
- dolžina besedila opisa, število slik in prisotnost videoposnetka,
- prisotnost Youtube videoposnetka,
- število ogledov Youtube videoposnetka,
- število čivkov kampanije.

V raziskavi so ugotovili, da najpomembnejši atributi za napovedovanje uspešnosti izhajajo iz predstavitve projekta in ne iz podatkov socialnih omrežij. Uporabili so metodo podpornih vektorjev in zgradili model iz podatkov okoli 19.000 zaključenih projektov. Pridobili so še dodatnih 1.000 projektov, ki jih niso vključili v učenje, ampak so jih uporabili za testiranje. Klasifikacijska točnost, ki so jo dosegli po 40% trajanja projekta, znaša 90%, ob začetku kampanije pa 67%.

Raziskovalci iz Tehnološkega inštituta iz George [12] so za napovedovanje uspešnosti uporabili najpogostejše fraze v besedilu, ki so prisotne v projektih iz vseh kategorij. Pridobili so podatke 45.000 projektov, analizirali devet milijonov fraz in jim dodali 11 značilk projekta. Kot atribute so uporabili število komentarjev in število objav avtorja o napredku kampanije. Atributa sta močno odvisna od rezultata kampanije, zato so z logistično regresijo dobili visokih 83% pri napovedi z uporabo atributov projekta. Ko so dodali še najpogostejše fraze, so pri napovedovanju uspešnosti projektov dosegli 97,6% natančnost.

Na projektu univerze v Stanfordu [2] so za izgradnjo modela uporabili sedem atributov na strani projekta in dodali analizo besedila. Pridobili so podatke nekaj več kot 26.000 projektov. Najboljše rezultate so dobili z logistično regresijo in z metodo podpornih vektorjev, kjer so za prvi dan kampanij uspešnost ocenili z mero F1 0,80. Pokazali so [2], da uporaba značilk besedila

izboljša napovedno uspešnost (mero F1 izboljša za 0,05). Da bi izluščili zanimive jezikovne značilke, so uporabili program za analiziranje besedila LIWC (angl. *Linguistic Inquiry and Word Count*)[6].

1.2 Vsebina diplomske naloge

Diplomska naloga vsebuje uvod, dve metodološki poglavji in zaključek. V uvodnem poglavju predstavimo tematiko naloge in obstoječe pristope k napovedovanju uspešnosti na Kickstarter-ju. V drugem poglavju opišemo zajem podatkov, metode strojnega učenja, ki smo jih uporabili, izbiro parametrov metod uvrščanja in mere ter pristope uspešnosti, s katerimi smo vrednotili zgrajene klasifikacijske modele. V tretjem poglavju opišemo pridobljene podatke in njihove atribute, eksperimente in diskusijo. V razdelku eksperimentov opišemo nabor in tipično izbiro parametrov, rezultate vrednotenja s prečnim preverjanjem in na novih podatkih. V diskusiji predstavimo pomembne značilke, jih primerjamo s sorodnimi deli in komentiramo porazdelitve verjetnosti napačno napovedanih primerov. V zaključku predstavimo glavna opažanja, kaj je novega in predloge za nadaljno delo.

[6] `liwc.wpengine.com/`

Poglavje 2

Metode

Napovedovanja uspešnosti projektov iz podatkov podanih v atributni obliki se lahko lotimo z metodami strojnega učenja oz. pristopi klasifikacije. Uvrščanje v skupine ali klasifikacija predpostavlja, da poleg množice atributov primeri v učni množici vsebujejo tudi razred, zato spada v nadzorovano atributno učenje.

Za napovedovanje uspešnosti kampanij na Kickstarter-ju smo morali pridobiti podatke projektov in iz njih izluščiti atribute, s katerimi smo opisali posamezen primer v učni množici. Za pridobivanje podatkov, gradnjo klasifikacijskih modelov in testiranja uspešnosti smo uporabili programski jezik Python, ki ima veliko uporabnih knjižnic za podatkovno rudarjenje. Uporaba knjižnic nam omogoča pisanje enostavnejše kode in časovno učinkovitih programov, saj so algoritmi in metode v knjižnicah dobro optimizirani.

2.1 Zajem podatkov

Na strani Kickstarter[1], kjer lahko iščemo projekte po različnih parametrih, smo najprej izluščili internetne naslove projektov na podlagi spreminjanja podkategorije. Z uporabo Python knjižnice urllib[2] smo pridobili vsebino posamezne strani v formatu HTML (angl. *Hyper Text Markup Language*).

[1] `www.kickstarter.com/discover`

[2] `docs.python.org/2/library/urllib.html`

Primer prikazuje pridobivanje i-te strani, ki vsebuje dvajset osnovnih opisov in povezav na projekte:

```
url = 'https://www.kickstarter.com/discover/advanced?'
values = {'category_id' : categories[index_id],
                'woe_id': '0', 'sort': 'end_date',
                'seed' : '2431954', 'page': i}
data = urllib.parse.urlencode(values)
reqUrl = url + str(data)
req = urllib.request.Request(reqUrl)
resp = urllib.request.urlopen(req)
respData = resp.read()
```

Spremenljivka `respData` hrani vsebino strani in iz nje smo izluščili internetne naslove projektov. Vsakega smo obiskali in pridobili vsebino na način, kot smo opisali zgoraj. Željene podatke na posamezni strani smo izluščili s pomočjo Python knjižnice Beautiful Soup[3] in knjižnice za regularne izraze. Pridobljene podatke smo shranili v datoteko tipa CSV (angl. *comma-separated values*). Za pripravo podatkov smo uporabili knjižnico Pandas[4], da smo pripravili pridobljene podatke za uporabo v algoritmih strojnega učenja.

Podatke projektov kampanij, ki so nastali med izdelavo diplomske naloge, smo pridobili na podoben način. Pri teh primerih smo najprej preverili čas njihovega nastanka.

2.2 Tehnike modeliranja

Za izgradnjo klasifikacijskih modelov obstaja cela vrsta algoritmov, ki se med seboj močno razlikujejo. Med preproste algoritme spadajo naivni Bayesov klasifikator, izgradnja klasifikacijskih dreves in metoda k najbližjih sosedov [8]. Čeprav modeli pridobljeni z naivnim Bayesom in klasifikacijskimi drevesi tipično ne dajejo najboljše točnosti, se ti metodi v praksi veliko uporabljata, saj drevesa omogočajo hitro in učinkovito interpretacijo. Podobno velja tudi za metodo k najbližjih sosedov, ki služi za primerjavo z drugimi metodami. Najboljše rezultate v praksi daje logistična regresija, naključni

[3] `www.crummy.com/software/BeautifulSoup/`

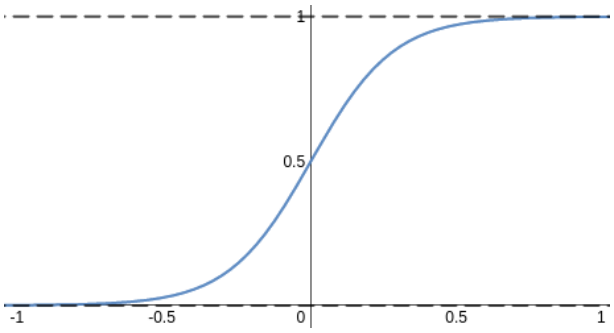
[4] `pandas.pydata.org`

gozd in različice metode podpornih vektorjev (angl. *Support Vector Machine, SVM*) [6].

V diplomski nalogi smo ovrednotili uporabo različnih klasifikacijskih algoritmov. Najboljše rezultate smo dobili pri logistični regresiji in pri ansambelskih (angl. *ensemble*) metodah - stacking, naključni gozd, Gradient Boosting in Extra trees. Vse metode razen pristopa stacking so že implementirane v knjižnici scikit-learn in pripravljene za uporabo.

2.2.1 Logistična regresija

Logistična regresija spada v skupino regresijskih modelov in je razširitev linearne regresije. Razred y ima pri klasifikaciji diskretne vrednosti, v našem primeru 0 ali 1. Za napovedovanje binarnega razreda model linearne regresije ni primeren. Zato iščemo model, ki se čim bolje prilega učnim podatkom. Izkaže se, da je za model logistične regresije prava funkcija sigmoida, ki jo prikazuje slika 2.1.



Slika 2.1: Sigmoida ali logistična funkcija, ki jo logistična regresija uporablja za hipotezo.

Logistična funkcija ima predpis:

$$g(z) = \frac{1}{1 + e^{-z}}$$

Zaloga vrednosti sigmoidne funkcije g(z) se nahaja na intervalu $Zf = [0, 1]$, njena vrednost za primer pa predstavlja verjetnost pripadnosti razredu 1.

Ker metoda predpostavlja, da je med razredoma linearna meja, gradi enostavne modele. Ti skupaj z regularizacijo zmanjšujejo verjetnost prevelikega prilagajanja učnim podatkom. Metoda se izvaja zelo hitro, vendar ne zna zgraditi kompleksnejših modelov.

2.2.2 Naključni gozd

Naključni gozd (angl. *Random Forest*) spada v ansambelske metode kot metoda Gradient Boosting in stacking. Te metode za učenje uporabijo več med seboj različnih klasifikatorjev. Njihove napovedi združijo v skupno napoved, ki v splošnem daje boljše rezultate kot enostavni klasifikatorji [3].

Naključni gozd za enostavne klasifikatorje uporablja klasifikacijska drevesa in primeru napove razred, kamor ga uvršča večina klasifikacijskih dreves, ki gozd sestavljajo. Raznolikost dreves dosežemo z naključnim izbiranjem podmnožic atributov za vsako vozlišče posebej in naključnim izbiranjem primerov z vračanjem. Metoda se pretirano ne prilagaja učni množici, je robustna na šum ter število atributov in tipično daje najboljše rezultate v praksi [6].

Uporaba Python knjižnice scikit-learn omogoča vzporedno gradnjo naključnega gozda in s tem hitro izgradnjo napovednega modela. Klic metode, ki zgradi napovedni model iz učnih podatkov (X - matrika z atributi primerov, y - vektor razreda primerov) je zato precej enostaven:

```
rfc = RandomForestClassifier()
rfc.fit(X, y)
```

2.2.3 Metoda Gradient Boosting

Algoritmi tipa Boosting v končni model iterativno dodajajo posamezne klasifikatorje. Pri algoritmu AdaBoost [13] na podlagi zadnjega razvitega klasifikacijskega drevesa izberemo novo učno množico. Če smo primer v začetnem klasifikacijskem drevesu napovedali napačno, mu povečamo utež, pravilnim pa utež zmanjšamo. Naslednje iteracije potekajo na enak način. Napačno napovedane primere dodatno vključimo in jim tako dodamo večji pomen. Učni

algoritem se na ta način osredotoča na težavne primere. Končna klasifikacija predstavlja uteženo glasovanje osnovnih klasifikatorjev.

Podoben pristop uporablja tudi Gradient Boosting [7], ki zgradi napovedni model in izračuna njegovo napako. Drugi model se poskuša naučiti napovedati to napako. Z dodajanjem novih modelov postopek ponavljamo, dokler je napaka previsoka. Napoved primera predstavlja seštevek napovedi posameznih modelov. Algoritem kljub temu, da ne vsebuje naključnosti, daje dobre rezultate, saj so učne množice med posameznimi iteracijami dovolj različne.

2.2.4 Metoda stacking

Stacking je še ena v vrsti skupinskih metod, ki združi napovedi več klasifikatorjev. Z razliko od naključnega gozda metoda stacking združuje različne učne algoritme. Metoda je sestavljena iz dveh nivojev. Na prvem nivoju za izbrane klasifikacijske algoritme izračuna verjetnost pripadnosti primera pozitivnemu razredu. S pomočjo prečnega preverjanja pridobi verjetnosti za vsak primer v učni množici. Vse napovedi primerov za izbrane klasifikatorje nato združi v novo učno množico, ki ji doda znano vrednost razreda. Iz nove učne množice se uči klasifikator na drugem nivoju. Na tem nivoju se za klasifikacijske probleme pogosto uporablja logistična regresija.

Pri napovedovanju razreda vsi modeli na prvem nivoju napovedo verjetnost, da primer pripada pozitivnemu razredu. Model na drugem nivoju iz ocenjenih verjetnosti posameznih klasifikatorjev izračuna končno napoved.

Stacking se uporablja tako za klasifikacijske kot tudi regresijske probleme. Metoda je časovno zahtevnejša od prej omenjenih, saj izgradnja modele poteka na dveh nivojih z uporabo več algoritmov. Rezultat stackinga je odvisen od velikosti učne množice, ki mora biti čim večja [11]. Vendar pa v splošnem daje boljše rezultate [4] kot najboljša metoda na prvem nivoju.

Za implementacijo te metode smo porabili kodo iz programskega paketa Orange in jo prilagodili uporabi algoritmov iz knjižnice scikit-learn. Za napovedovanje verjetnosti na prvem nivoju smo uporabili logistično regresijo,

naključni gozd ter metodo Gradient Boosting. Napovedi smo na drugem nivoju združili z logistično regresijo.

2.3 Izbira parametrov metod uvrščanja

Pri klasifikaciji je poleg izbire primerne metode pomembno tudi nastavljanje parametrov. Rezultati logistične regresije so odvisni na primer od stopnje regularizacije, pri naključnem gozdu pa od števila dreves. Izbor optimalnih parametrov metode in gradnja modela sta del razvoja napovednega modela, zato je potrebno za postopka uporabiti neodvisni učni množici. V nasprotnem primeru lahko pride do prevelikega prilagajanja učni množici in s tem slabega rezultata na novih primerih.

V diplomski nalogi smo iz učne množice vzeli 20% vseh primerov in jih uporabili za izbiro parametrov. Izbrali smo tiste parametre, pri katerih smo s prečnim preverjanjem (opisan v poglavju 2.4.1) dobili v povprečju najboljše rezultate. Delo nam je poenostavila knjižnica scikit-learn, ki ima že implementirane potrebne metode za opisan postopek. Tako smo denimo izbrali najboljše parametre logistične regresije z uporabo metode *GridSearchCV*, ki vsebuje atribut *best_params_*:

```
lr_params = {'penalty': ['l1', 'l2'],
        'C': [0.1, 0.5, 1, 10, 100, 500, 1000]}

lr = GridSearchCV(LogisticRegression(), lr_params, cv=5,
scoring="roc_auc")
lr.fit(X_param, y_param)

print(lr.best_params_)
```

Programska koda nam izpiše parametre, ki so s 5-kratnim prečnim preverjanjem dosegli najvišjo mero AUC, ki je opisana v poglavju 2.4.4.

2.4 Vrednotenje

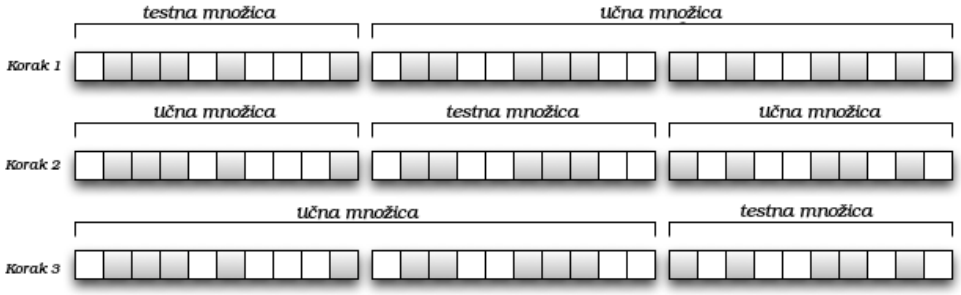
Klasifikacijske tehnike ocenjujemo z merami uspešnosti. Poznamo več pristopov ocenjevanja uspešnosti, pri katerih je zelo pomembno, da uspešnost nikoli

ne ocenjujemo na učnih primerih. Uporabiti je potrebno primere, ki še niso bili uporabljeni v postopku izdelave napovednega modela. Veliko množico primerov lahko razdelimo na učno in testno, vendar pa s tem zmanjšamo število učnih primerov in posledično dobimo manj točen napovedni model.

Eden od pristopov ocenjevanja uspešnosti, ki smo ga uporabili v diplomski nalogi, je prečno preverjanje [10] (angl. *cross validation, CV*). Uspešnost napovednih modelov smo ocenjevali s klasifikacijsko točnostjo (angl. *classification accuracy, CA*), mero F1 in AUC (angl. *area under the curve*).

2.4.1 Prečno preverjanje

Tehnika k-kratnega prečnega preverjanja (angl. *k-fold cross validation*) množico primerov razdeli na k enako velikih množic. Lahko se uporablja za testiranje ali izbiro optimalnih parametrov. Poteka v k korakih (kot prikazuje slika 2.2), kjer v vsakem koraku izberemo eno množico za testiranje, vse ostale pa za izgradnjo modela.



Slika 2.2: 3-kratno prečno preverjanje množice s 30 elementi.

V diplomski nalogi smo uporabili 5-kratno prečno preverjanje pri izbiri optimalnih parametrov posamezne klasifikacijske metode in 10-kratno prečno preverjanje pri testiranju rezultatov. Uporabili smo naključno izbiranje primerov iz celotne množice in ne zaporedne, kot prikazuje zgornja slika 2.2. Primer gradnje modela in napovedovanja znotraj prečnega preverjanja podaja programska koda, ki smo jo uporabili pri vrednotenju rezultatov:

```
kf_total = cross_validation.KFold(len(X),
        n_folds=10, shuffle=True, random_state=42)
for train, test in kf_total:
        rfc.fit(X[train], y[train])
        ypred = rfc.predict(X[test])
```

2.4.2 Klasifikacijska točnost

Klasifikacijska točnost predstavlja delež pravilno napovedanih primerov. Izračuna se po enačbi:

$$CA = \frac{TP + TN}{\check{s}t.\ vseh\ primerov}$$

kjer TP predstavlja število pravilno napovedanih pozitivnih (angl. *true positive*), TN pa pravilno napovedanih negativnih primerov (angl. *true negative*). Poznamo boljše mere, saj je uspešnost klasifikacijske točnosti odvisna od verjetnosti bolj verjetnega razreda. Vseeno smo jo uporabili, saj jo nekatere dosedanje raziskave na področju napovedovanja uspešnosti na Kickstarter-ju kot edino navajajo v svojih poročilih.

2.4.3 Mera F1

Mera F1 združuje dve meri - natančnost (angl. *precision*) in priklic (angl. *recall*), ki ju lažje razumemo z uporabo kontingenčne tabele, ki jo prikazuje tabela 2.1. Natančnost predstavlja delež pravilno napovedanih pozitivnih primerov ($Pr = \frac{TP}{TP+FP}$), priklic pa delež pravilno negativnih ($Re = \frac{TP}{TP+FN}$).

Mero F1 lahko interpretiramo kot uteženo povprečje natančnosti (označeno s Pr) in priklica (označeno z Re) in jo izračunamo po formuli:

$$F1 = 2 * \frac{Pr * Re}{Pr + Re}$$

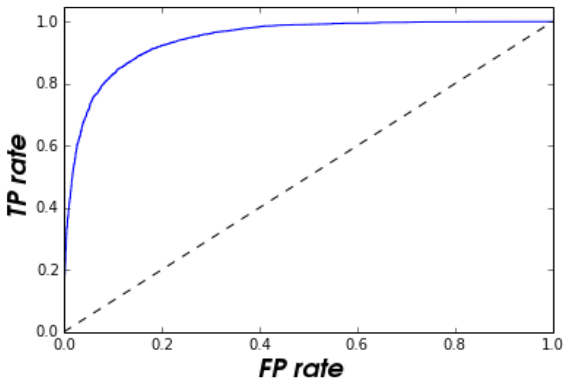
Najboljša možna klasifikacija dosega mero $F1 = 1$. Pri klasifikaciji, ki vse primere napove napačno, pa velja $F1 = 0$.

Tabela 2.1: Kontingenčna tabela (angl. *confusion matrix*).

		Napovedana vrednost	
		P	**N**
Dejanska vrednost	**P**	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

2.4.4 Mera AUC

Mera AUC nam pove, kako dobro nam napovedni model loči pozitivne in negativne napovedi. Izračuna se kot delež površine pod krivuljo ROC (angl. *area under ROC curve, AUC*). Predstavlja verjetnost, da model za naključno izbrani pozitivni primer napove večjo verjetnost pripadnosti pozitivnemu razredu kot naključno izbranemu negativnemu primeru. Boljši klasifikatorji imajo torej višji AUC. Naključni klasifikator ima $AUC = 0,5$ in je na sliki 2.3 označen s črtkano črto.



Slika 2.3: Primer krivulje ROC.

Krivulja ROC ima na vodoravni asimptoti delež napačno uvrščenih pozitivnih primerov (angl. *FP rate*), na navpični pa delež pravilno uvrščenih pozitivnih primerov (angl. *TP rate*).

Za klasifikacijsko točnost, mero F1 in AUC smo uporabili knjižnico scikit-learn, ki nam omogoča enostavno uporabo ocenjevanja točnosti napovednih modelov. Primer prikazuje izračun mere AUC za testne podatke znotraj prečnega preverjanja:

```
fpr, tpr, thresholds = metrics.roc_curve(y[test],
                rfc.predict(X[text]), pos_label=1)
auc = metrics.auc(fpr, tpr)
```

Poglavje 3

Eksperimenti in diskusija

Modele gradimo z različnimi tehnikami iz učnih podatkov, ki so predstavljeni z atributi. Za napovedovanje razreda je tako poleg izbire uspešnih tehnik ključnega pomena tudi izbira učnih podatkov. Podatki so sestavljeni iz primerov, ki jih opišemo z množico atributov X in vrednostjo razreda y, v katerega pripada. Za klasifikacijske probleme je za izgradnjo uspešnih napovednih modelov potrebno pridobiti čim več primerov in jih opisati z atributi od vrednosti katerih je odvisna vrednost razreda. V tem poglavju zato začnemo z opisom podatkov, potem pa nadaljujemo z izbiro parametrov učnih metod, uporabi teh na podatkih, vrednotenju in diskusiji o rezultatih.

3.1 Podatki

Pridobili smo podatke 25.250 zaključenih kampanij na Kickstarter-ju. V podatke smo vključi samo projekte, katerih kampanija je že zaključena in je moč določiti njeno uspešnost. Delež uspešnih kampanij se je v zadnjih letih zmanjšal za več kot 5%, zato smo želeli pridobiti aktualne podatke. Tako smo pridobili le projekte, ki so bili ustvarjeni v letu 2014 ali kasneje, kot prikazuje tabela 3.1.

Tabela 3.1 prikazuje, da je bilo 8.645 pridobljenih projektov uspešnih oz. so pridobili finančna sredstva. Delež uspešnih v naši učni množici znaša

Tabela 3.1: Število projektov v zbranih podatkih po letih.

	Uspešen	Neuspešen	Delež uspešnih(%)
2014	3695	5232	41.4
2015	4145	9690	30.0
2015	805	1683	32.4
skupaj	8645	16605	34.2

34,2% in se le za 2% razlikuje od celotne uspešnosti projektov na platformi, ki smo jo navedli v uvodnem poglavju.

Pri pridobivanju podatkov smo se omejili na 6 od skupno 15 kategorij projektov. Posamezna kategorija je razdeljena na podkategorije. Tabela 3.2 prikazuje teh 6 kategorih in njihovih 29 podkategorij.

Tabela 3.2: Kategorije in podkategorije naših projektov.

Kategorije in podkategorije		
Technology	**Art**	**Design**
3D Printing	Digital Art	Graphic Design
Apps	Illustration	Product Design
Camera Equipment	Painting	
DIY Electronics	Performance Art	**Games**
Fabrication Tools	Sculpture	Tabletop Games
Flight		
Gadgets	**Film and Video**	
Hardware	Animation	
Makerspaces	Documentary	
Robots	Narrative Film	
Sound	Shorts	
Software	Webseries	
Space Exploration		
Wearables	**Crafts**	
Web	Crafts	

Med izbrane kategorije smo uvrstili tudi tehnologijo, saj je za nas računalničarje najbolj zanimivo področje. Ostale smo izbrali naključno. Imamo kategorije

z velikim številom projektov kot tiste z malo. Tako smo izbrali kategorijo filma in videa (angl. *Film and Video*), ki je glede na podatke z uradne strani[1] najbolj popularna kategorija na Kickstarter-ju z več kot 50.000 projekti.

Uspešnost posameznega projekta, ki smo ga pridobili iz platforme, smo izračunali na podlagi vrednosti želene zbrane vsote in pridobljenega denarja. Razred projekta, kjer so pridobili več denarja od predhodno zastavljenega cilja, smo označili z vrednostjo $y = 1$. Kampanije, ki niso pridobile financiranja, so dobile vrednost razreda $y = 0$.

V času od pričetka dela na diplomski nalogi se je na platformi Kickstarter pojavilo več kot 10.000 novih projektov. Za končno testiranje napovednih modelov smo zato lahko pridobili testno množico novih 1.248 projektov, ki so svoje kampanije zaključili v mesecu aprilu ali maju 2016. Pri izboru projektov smo se omejili na kategorije in podkategorije, ki so predstavljene v tabeli 3.2. Delež uspešnih novo pridobljenih projektov znaša 40%.

3.1.1 Atributni opis projektov

Vsak projekt v naši učni in testni množici smo opisali z 20 atributi, katerih vrednosti smo pridobili s strani projektov na Kickstarter-ju. Izluščili smo jih skladno s postopkom opisan v poglavju 2.1 iz različnih mest na strani posamezne kampanije. Ciljna vsota, trajanje kampanije, podatki o nagradah, kategorije projekta so značilni tudi za druge portale množičnega financiranja (IndiGoGo[2], RocketHub[3]). Uporabili smo le podatke, ki so objavljeni pri vseh projektih, neodvisno od uspešnosti kampanije.

Nekateri podatki projektov (lokacija, kategorija, mesec in leto) vsebujejo več možnih vrednosti, zato smo za potrebe algoritmov strojnega učenja za vsako ustvarili nov atribut. Podatek o lokaciji vsebuje kratico države, v katerem projekt nastaja, in vsebuje 227 različnih vrednosti. Ker smo že ob prvih testiranjih ugotovili, da ta podatek ne izboljša rezultat napovedovanja,

[1] `https://www.kickstarter.com/help/stats`
[2] `www.indiegogo.com`
[3] `www.rockethub.com`

smo ga zaradi časovne neučinkovitosti izločili. Tako smo za gradnjo modelov uporabili 59 atributov, ki jih prikazuje tabela 3.12. Slike 3.1, 3.2 in 3.3 prikazujejo, kje na strani se nahajajo vrednosti pridobljenih atributov.

Tabela 3.3: Atributi, ki smo jih uporabili za gradnjo modelov.

Atributi		
ciljna vsota	trajanje	dolžina naslova
št. nagrad	najmanjša donacija	največja donacija
podkategorija (29)	prisotnost na FB	št. kampanij
št. podprtih kampanij	št. znakov v opisu avtorja	št. videoposnetkov
prisotnost videa	št. znakov v opisu	št. slik
mesec (12 atributov) in leto (3 atributi) nastanka kampanije		

1. **Ciljna vsota donacij:** Vsota denarja, ki jo ustvarjalec projekta skuša doseči v kampaniji. Podana je v različnih valutah. Vse atribute, ki so v določeni valuti, smo pretvorili v ameriške dolarje.

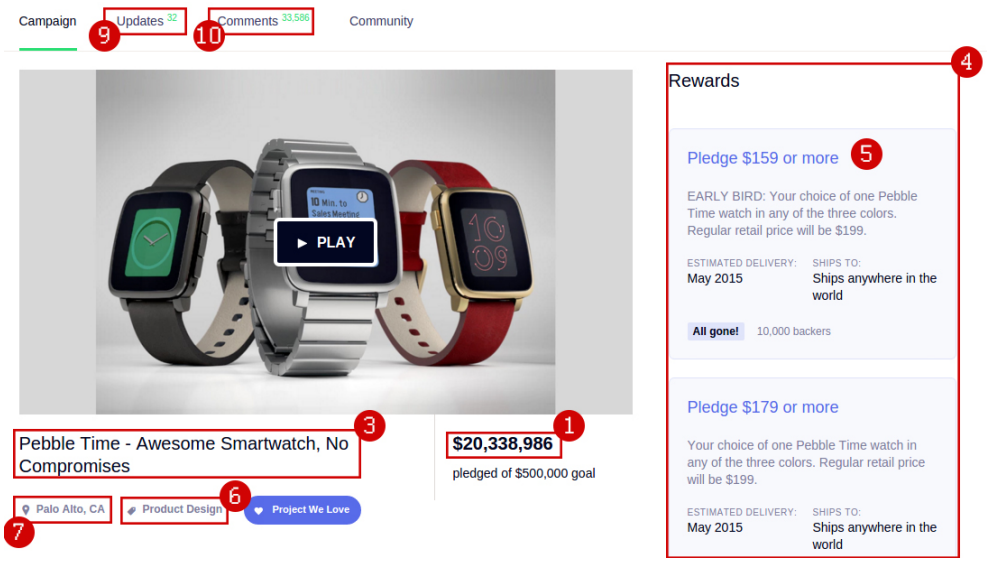
2. **Trajanje zbiranja denarja:** Čas zbiranja donacij je omejen in ga določi ustvarjalec projekta. Atribut je označen na sliki 3.3.

3. **Število znakov v naslovu:** Bolj brani članki na spletu imajo tipično krajše naslove [9]. S tem atributom smo preverili, če ta lastnost velja tudi za projekte na Kickstarter-ju.

4. **Število predlaganih nagrad:** Ustvarjalci na strani projekta objavijo različne možne zneske donacij. Za vsako obljubijo nagrado, ki pripada investitorju.

5. **Največja in najmanjša donacija:** Možne donacije imajo tipično različne vrednosti med \$1 in \$10000. Z atributoma največje in najmanjše donacije smo opisali razpon donacij, ki ga projekt ponuja.

6. **Podkategorija projekta:** Delež uspešnih projektov se razlikuje med podkategorijami. Zato smo ustvarili 29 atributov, kjer vsak predstavlja



Slika 3.1: Primer dela spletne strani projekta Pabble in nekateri osnovni podatki: (1) ciljna vsota, (3) naslov projekta, (4) ter (5) podatki o predlaganih nagradah, (7) podatki o lokaciji projekta, (9) število objav projekta in (10) število komentarjev.

eno podkategorijo. Ker posamezen projekt spada le v eno kategorijo, je le en atribut od 29 različen od 0 oz. enak 1.

7. **Lokacija projekta:** Na Kickstarter-ju so projekti iz različnih delov sveta in zanimalo nas je, če je uspeh kampanije odvisen od države, kjer projekt nastaja. Pri vsakem projektu smo izluščili ime države in ustvarili smo 227 atributov, saj smo pridobili projekte iz 227 držav sveta.

8. **Mesec in leto začetka kampanije:** Po poročanju portala Genius Games[4] je uspeh projekta na Kickstarter-ju odvisen od meseca, v katerem se odvija kampanija. Ker se skozi leta zmanjšuje delež uspešnih projektov, smo kot atribut vzeli tudi leto, v katerem se je odvijala kampanija. Lokacija teh dveh atributov na strani projekta prikazuje slika 3.3.

[4] `https://gotgeniusgames.com/kickstarter-stats-101-does`

9. **Število objav o dejavnostih projekta:** Ustvarjalci projektov imajo možnost, da objavijo stanje oz. napredek projekta. S tem obvestijo trenutne investitorje in skušajo pridobiti nove. Gre za celoštevilski atribut večji ali enak 0.

10. **Število komentarjev:** Trenutni in potencialni investitorji lahko s pomočjo komentarjev zastavijo vprašanje ustvarjalcem projekta. Obsega enake vrednosti kot atribut števila objav projekta.

Slika 3.2: Primer dela spletne strani s podatki o ustvarjalcu projekta Pabble: (11) podatek o prisotnosti na Facebook-u, (12) podatek o številu dosedanjih ter podprtih kampanij in (13) opis avtorja projekta.

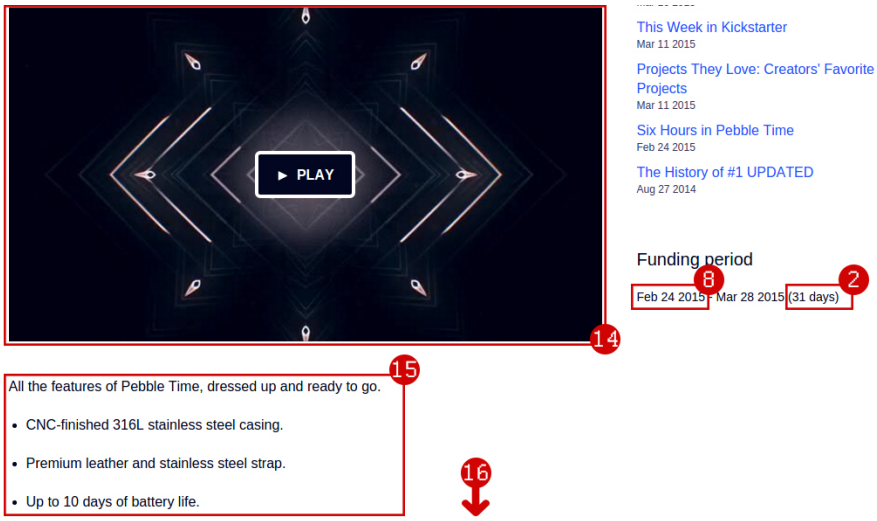
Za vsak projekt smo pridobili tudi atribute ustvarjalca. Podatki ustvarjalca se nahajajo na ločeni strani. Primer podatkov o avtorju najbolj donosnega projekta na Kickstarter-ju Pabble prikazuje slika 3.2.

11. **Prisotnost avtorja na Facebook-u:** Binarni atribut, ki določa, ali ima avtor oz. ekipa določenega projekta stran na Facebook-u.

12. **Število dosedanjih projektov in število projektov, ki jih je avtor podprl:** Iz podatkov o ustvarjalcu lahko pridobimo podatek o tem, koliko projektov je že imel na Kickstarter-ju in koliko projektov drugih avtorjev na Kickstarter-ju je avtor do sedaj že podprl. Oba atributa imate vrednosti iz množice naravnih števil $\mathbb{N}$.

13. **Število znakov v opisu ustvarjalca:** Dolžina opisa se med ustvarjalci precej razlikuje. Nekateri avtorji opisa sploh ne dodajo, kar je lahko vzrok za slabšo uspešnost kampanije.

Preko predstavitve projekta na Kickstarter-ju ustvarjalci navežejo stik s potencialnimi podporniki. Opis projekta je zelo pomemben za pridobivanje novih investitorjev, zato smo v izbor dodali atribute opisa projekta. To so storili tudi na projektu KickPredict [1]. Ugotovili so, da video in slike spadata med najpomembnejše atribute napovedovanja uspešnosti.



Slika 3.3: Primer dela spletne strani z opisom projekta Pabble: (2) trajanje kampanije, (8) podatek o mesecu in letu začetka kampanije, (14) videoposnetek, (15) besedilo projekta in (16) slike.

14. **Število videoposnetkov, prisotnost videa:** V opisu besedila lahko avtor projekta umesti več videoposnetkov, s katerimi predstavi svoj izdelek ali storitev. Na uradni strani Kickstarter poročajo o večji uspešnosti projektov, ki vsebujejo videoposnetke[5]. Atribut število videoposnetkov vsebuje vrednosti naravnih števil $\mathbb{N}$. Prisotnost videa ima za vse projekte, ki imajo videoposnetek, vrednost 1, za ostale 0.

15. **Število znakov opisa projekta:** Preverili smo, če dolžina besedila v opisu vpliva na uspešnost kampanije. Atribut predstavlja število zna-

[5] `http://kck.st/24VxT3H`

kov, iz katerih je sestavljen opis projekta, in vsebuje vrednosti naravnih števil $\mathbb{N}$.

16. **Število slik:** Poleg besedila se v opisu pojavijo slike. Ker se število slik med projekti razlikuje, bi lahko ta atribut vplival na uspeh projektov.

Pri poročanju o uspešnosti klasifikacijskih modelov smo izpustili dve značilki in vzeli le tiste, ki so definirane že pred začetkom kampanije. Podatek o številu komentarjev in o poročanju izvajanja projekta ni znan pred začetkom same kampanije. Značilki smo uporabili le pri gradnji modela, ki smo ga uporabili za primerjavo z raziskavo iz Tehnološkega inštituta iz George [12].

3.2 Eksperimenti

Iz pridobljenih podatkov smo gradili napovedne modele. Njihovo točnost smo ocenili na dva načina. V prvem, ki je opisan v razdelku 3.2.2, smo rezultate ocenili z 10-kratnim prečnim preverjanjem. Pri drugem pa smo za izgradnjo modela uporabili celotno učno množico, za testno množico pa vzeli projekte, ki so nastali med izdelavo projektne naloge. Rezultati testiranj z novimi projekti so opisani v razdelku 3.2.4. Pri obeh načinih testiranja smo izdelali napovedne modele z različnimi metodami strojnega učenja in jih primerjali med seboj.

3.2.1 Izbira parametrov

Pred izgradnjo modelov smo iz učne množice izločili 20% podatkov in jih uporabili za izbiro vrednosti parametrov. Teh 20% podatkov potem nismo uporabili ne v prečnem preverjanju ne pri gradnji končnih modelov. Nabor možnih parametrov po posameznih metodah strojnega učenja prikazuje tabela 3.4. Vrednosti ostalih parametrov nismo spreminjali.

Izvedli smo postopek, ki je opisan v poglavju 2.3. Postopek tipično izbere parametre, ki jih prikazuje tabela 3.5. Pri logistični regresiji se za najboljši tip regularizacije izkaže ridge regression (zaradi uporabe kvadratne fukcije

Tabela 3.4: Parametri metod in njihove možne vrednosti.

Metoda	Parameter	Nabor vrednosti
Logistična regresija	Tip regularizacije	$l1, l2$
	Stopnja regularizacije	$0.1, 1, 10, 100, 250,$ $500, 750, 1000$
Naključni gozd	Število dreves	$10, 100, 500, 1000,$ $2500, 5000, 7500, 10000$
Gradient Boosting	Število iteracij	$1, 10, 100, 250,$ $500, 750, 1000$

se označuje z $l2$) s stopnjo regularizacije $\lambda = 2000$. V našem primeru dobi naključni gozd najboljše rezultate, če je sestavljen iz 5.000 dreves, metoda Gradient Boosting pa se najbolje obnese po 100 iteracijah.

Tabela 3.5: Izbrani parametri.

Metoda	Parameter	Izbrana vrednost
Logistična regresija	Tip regularizacije	$l2$
	Stopnja regularizacije	2.000
Naključni gozd	Število dreves	5.000
Gradient Boosting	Število iteracij	100

Izbiro parametrov smo poskusili narediti tudi z večjim deležem podatkov, vendar se izbira parametrov in končen rezultat napovednih modelov ni spremenil.

3.2.2 Vrednotenje s prečnim preverjanjem

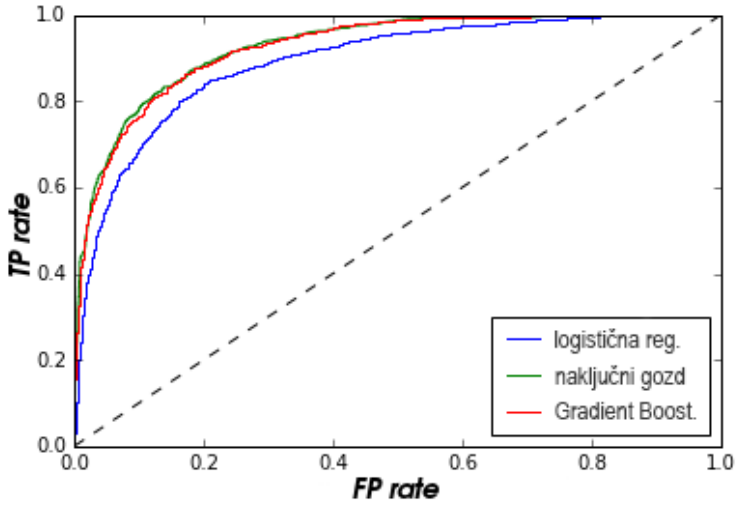
Preostalih 80% primerov iz učne množice, ki jih nismo uporabili za izbiro parametrov, smo uporabili za vrednotenje modelov z 10-kratnim prečnim preverjanjem. Nabor podatkov za testiranje s prečnim preverjanjem je tako vseboval 20.200 primerov. Dobili smo rezultate, ki jih prikazuje tabela 3.6.

Rezultate testiranih metod smo primerjali med seboj. Pomagali smo si

Tabela 3.6: Rezultati z 10-kratnim prečnim preverjanjem.

Metoda	CA	F1	AUC
Logistična regresija	0.81	0.70	0.89
Naključni gozd	0.84	0.76	0.92
Gradient Boosting	0.84	0.76	0.92

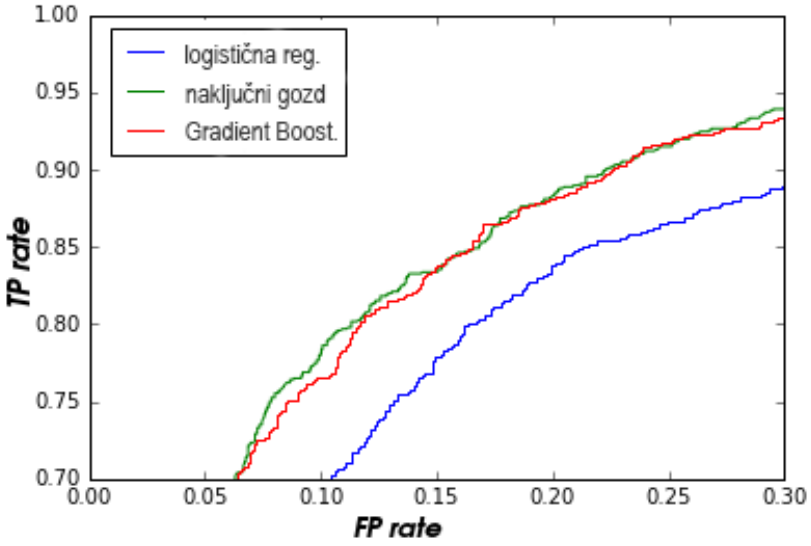
tudi z ROC krivuljami, ki jih prikazujeta sliki 3.4 in 3.5. Najboljše rezultate



Slika 3.4: Primerjava ROC krivulj med metodami.

smo dobili z metodo Gradient Boosting in metodo naključni gozd. Obe smo uporabili pri metodi stacking za napovedovanje verjetnosti na prvem nivoju. Dodali smo tudi metodo Extra trees, ki je podobna naključnemu gozdu. Razlikuje se pri gradnji dreves, saj ima metoda Extra trees večjo naključnost delitve v listih in je zato primerna za uporabo v metodi stacking. Napovedi smo združili z logistično regresijo na drugem nivoju in dobili rezultat, ki ga prikazuje tabela 3.7.

Metoda stacking v našem primeru le malenkost izboljša napovedni model in s tem mere, s katerimi ocenimo njegovo uspešnost. Nepravilno napove 2.385 od 20.200 primerov, kot prikazuje tabela 3.8. Opazimo, da je delež



Slika 3.5: Povečana ROC krivulja.

Tabela 3.7: Rezultati pri metodi stacking z 10-kratnim prečnim preverjanjem.

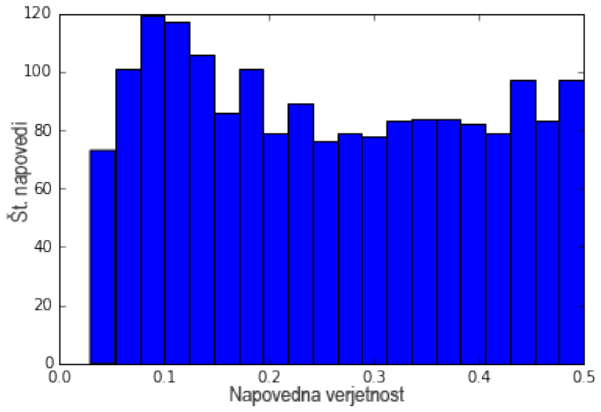
Metoda	CA	F1	AUC
Stacking	0.85	0.77	0.92

napačno napovedanih pozitivnih primerov večji v primerjavi z napačno napovedanimi negativnimi primeri. Slika 3.6 prikazuje verjetnost pripadnosti pozitivnemu razredu napačno negativnih primerov (angl. *false nagative, FN*). Povprečna vrednost napovedane verjetnosti v tem primeru znaša 26%, pri napačno pozitivnih primerih (angl. *false positive, FP*) pa 70%. Verjetnosti slednjih so prikazane na sliki 3.7.

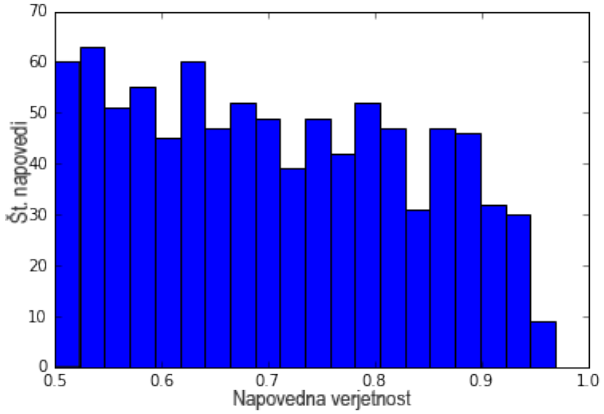
Delež napačno klasificiranih primerov se močno razlikuje glede na podkategorije primerov. Kot je prikazano v tabeli 3.9, je delež napačno uvrščenih primerov v kategorijah z velikim številom projektov tipično manjši. Na drugi strani pa so podkategorije z malo projekti manj uspešne.

Tabela 3.8: Kontingenčna tabela metode stacking.

		Napovedana vrednost	
		P	**N**
Dejanska vrednost	**P**	5.192	1.309
	N	1.793	11.906



Slika 3.6: Porazdelitev verjetnosti napačno negativnih primerov.



Slika 3.7: Porazdelitev verjetnosti napačno pozitivnih primerov.

Tabela 3.9: Napačne napovedi metode stacking po podkategorijah z več kot 400 pridobljenimi projekti.

Kategorija	Št. projektov	Napačne napovedi	Delež
Narrative Film	474	112	0.24
Illustration	589	119	0.20
DIY Electronics	414	75	0.18
Wearables	614	112	0.18
Hardware	1666	289	0.17
Painting	662	115	0.17
Webseries	807	138	0.17
Animation	433	70	0.16
Documentary	1570	230	0.15
Gadgets	1447	213	0.15
Crafts	1663	215	0.13
Product Design	2221	267	0.12
Shorts	862	70	0.08
Software	1463	96	0.07
Tabletop Games	1559	109	0.07
Apps	3018	136	0.05
Web	2460	127	0.05

3.2.3 Primerjava z referenčno študijo

V naši nalogi smo projekt opisali z atributi, katerih vrednosti so znane že pred pridobivanjem investitorjev. Pri pregledu področja smo opazili, da raziskava iz Tehnološkega inštituta iz George [12] uporablja dve značilki, ki se med kampanijo spreminjata. To sta podatka o številu komentarjev in številu objav o izvajanju projekta. Podatka pred začetkom same kampanije nista znana in njuna vrednost je zelo odvisna od uspeha kampanije. Atributa zato ne sodita v nabor ustreznih parametrov, vendar pa smo ju dodelili med podatke za primerjavo z omenjeno študijo. S tem smo pokazali, da se točnost napovednih modelov zelo poveča, kot prikazuje tabela 3.10.

Tabela 3.10: Rezultati z 10-kratnim prečnim preverjanjem z dodanima atributoma število komentarjev in število objav o izvajanju projekta.

Metoda	CA	F1	AUC
Logistična regresija	0.90	0.86	0.96
Naključni gozd	0.92	0.89	0.98
Gradient Boosting	0.93	0.90	0.98

Klasifikacijska točnost in mera F1 klasifikacijskih modelov sta se namreč povečale za več kot 10% in tako v omenjeni raziskavi posledično poročajo o previsoki klasifikacijski točnosti napovednih modelov, ki znaša 85% z uporabo osnovnih atributov projekta.

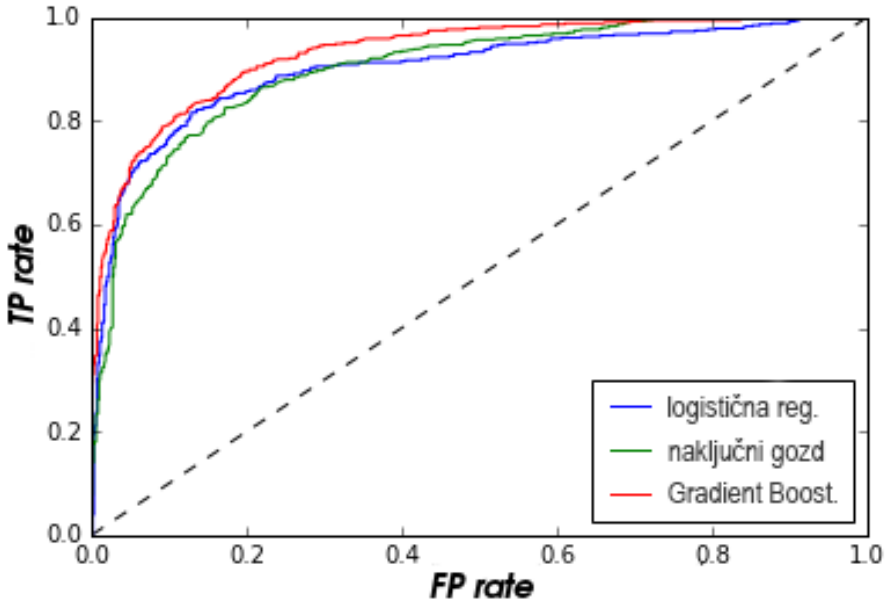
3.2.4 Vrednotenje na novih projektih

Napovedovali smo tudi uspešnost projektov, ki so nastali po začetku izdelave diplomske naloge. Pri gradnji modela smo uporabili podatke, ki so opisani v poglavju 3.1. Najboljše rezultate, ki jih prikazuje tabela 3.11 in slika krivulje ROC 3.8, smo dobili pri logistični regresiji, naključnem gozdu in metodi Extra trees.

Vse tri metode smo uporabili pri metodi stacking za napovedovanje verjetnosti na prvem nivoju in jih združili z logistično regresijo na drugem

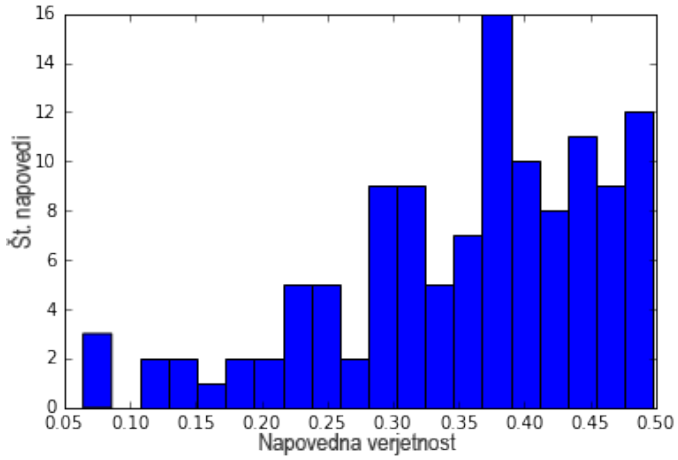
Tabela 3.11: Rezultati novih projektov na Kickstarter-ju.

Metoda	CA	F1	AUC
Logistična regresija	0.77	0.77	0.90
Naključni gozd	0.78	0.77	0.90
Extra trees	0.85	0.81	0.93

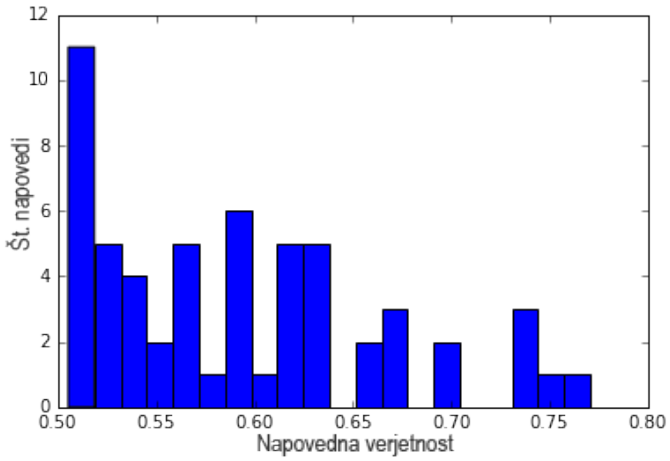


Slika 3.8: Primerjava ROC krivulj med metodami.

nivoju. Stacking ni izboljšal rezultate metode Extra trees, ki je od 1.248 napačno klasificirala 177 primerov. Sedeminpetdeset primerov je klasifikator napačno označil za pozitivne. Slika 3.9 prikazuje porazdelitev verjetnosti napačno napovedanih pozitivnih primerov. Ostalih 120 primerov pa je napačno označenih kot negativni. Napovedni model jim je napovedal verjetnosti pripadnosti pozitivnemu razredu, ki so prikazane na sliki 3.10.



Slika 3.9: Napovedane verjetnosti napačno negativnih primerov.



Slika 3.10: Napovedane verjetnosti napačno pozitivnih primerov.

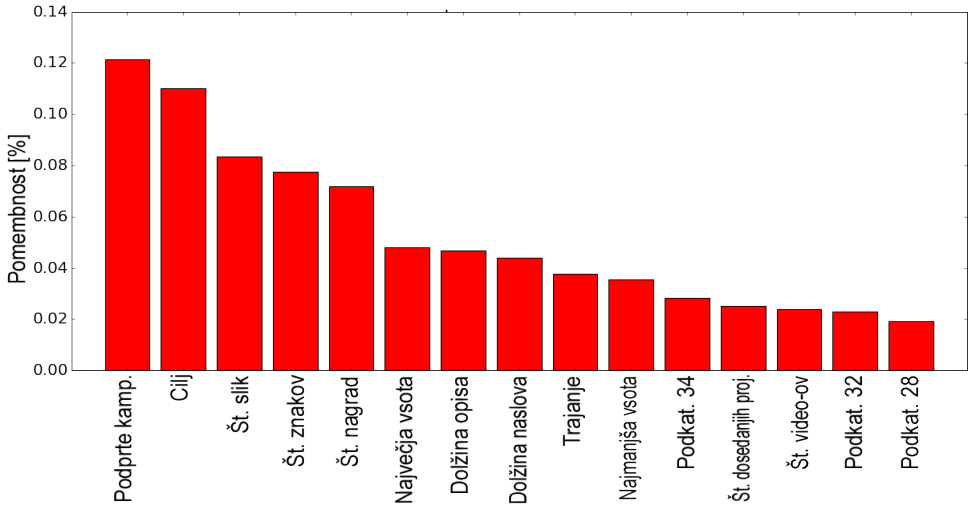
3.3 Diskusija

Za napovedovanje uspešnosti projektov na Kickstarter-ju smo uporabili algoritme strojnega učenja za klasifikacijo. Ti gradijo različne napovedne klasifikacijske modele, ki pri napovedovanju različno utežijo pasamezne atribute. Kot prikazuje tabela 3.12, je ciljna vsota donacij (opisan kot prvi atribut) najpomembnejši atribut pri logistični regresiji, saj ima največjo utež po izgradnji napovednega modela. Sledi število dosedanjih podprtih projektov, število slik v projektu, pripadnost posamezni kategoriji in atributi ponujenih donacij ter pripadajočih nagrad.

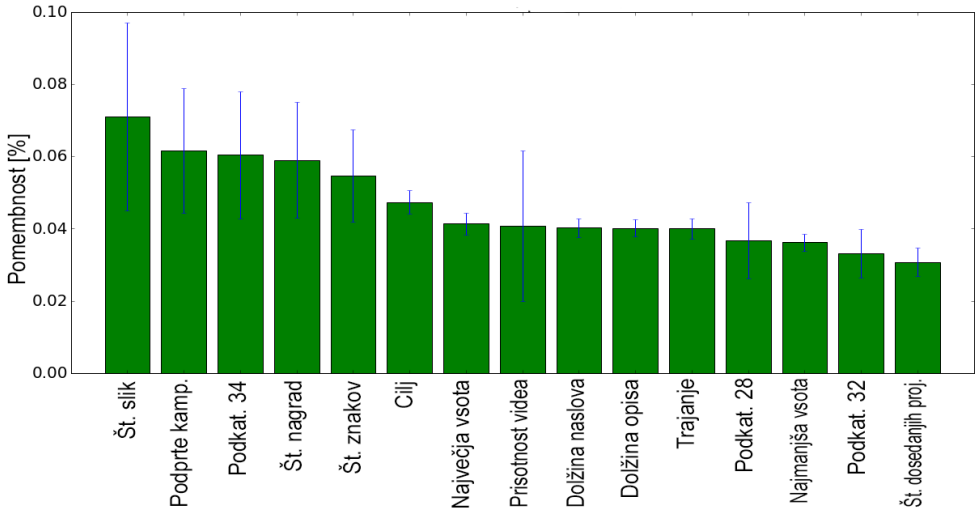
Tabela 3.12: Atributi z največjimi utežmi pri logistični regresiji.

Atribut	Utež
ciljna vsota	-23.40
št. podprtih projektov	0.74
št. slik v opisu	0.51
kategorija 332	-0.45
št. ponujenih nagrad	0.45
kategorija 342	-0.36
znesek najmanjše donacije	-0.35
prisotnost videa	0.35
kategorija 28	0.29
št. znakov v opisu	0.28
trajanje kampanije	-0.27

Naključni gozd in metoda Extra trees, implementirana v knjižnici scikit-learn, ponujata možnost pogleda najpomembnejših atributov. Najpomembnejše atribute naključnega gozda prikazuje slika 3.11. Najpomembnejši atribut je število dosedanjih podprtih projektov avtorja in je pomembnejši od ciljne vsote. Atributi kategorij so manj pomembni kot pri logistični regresiji. Podobno velja tudi za metodo Extra trees, kjer je ciljna vsota šele 6 najbolj pomemben atribut. Najpomembnejše atribute te metode prikazuje slika 3.13.



Slika 3.11: Najpomembnejši atributi metode naključni gozd.



Slika 3.12: Najpomembnejši atributi metode Extra trees z označenim standardnim odklonom ocene pomembnosti.

Tabela 3.13: Najpomembnejši atributi glede na mesto uvrščenosti v logistični regresiji (LR), naključnem gozdu (RF) in metodi Extra trees (ET).

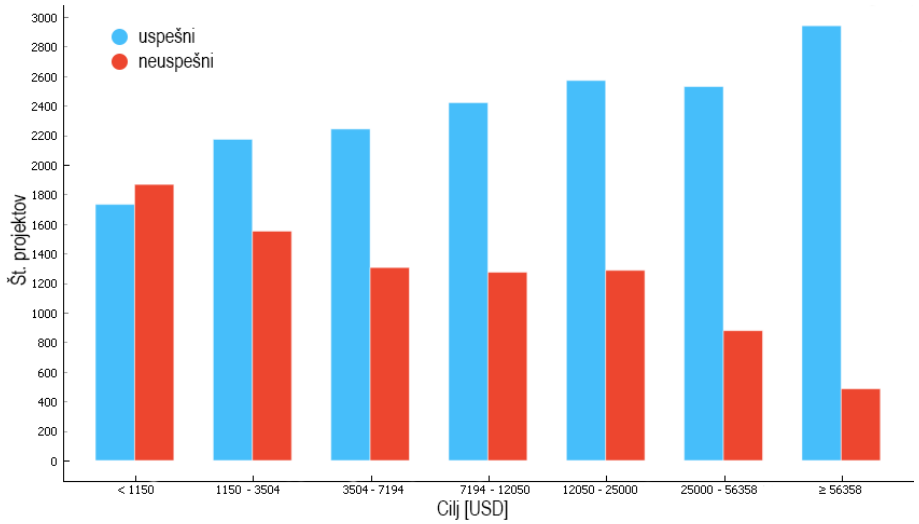
	Uvrščenost		
Atribut	LR	RF	ET
ciljna vsota	1.	2.	6.
podprti projekti avtorja	2.	1.	2.
št. slik v opisu	3.	3.	1.
št.ponujenih nagrad	5.	5.	4.
prisotnost videa	8.	13.	8.
št. znakov v opisu	10.	4.	5.
trajanje kampanije	11.	9.	11.
znesek največje investicije	17.	6.	7.

3.3.1 Pomembne značilke

Iz podatkov o pomembnosti atributov posameznih metod opazimo, da med najpomembnejše spadajo atributi, ki so prikazani v tabeli 3.13. Od vrednosti teh atributov je odvisen uspeh projekta na Kickstarter-ju.

Pridobljeni podatki projektov iz Kickstarter-ja imajo povprečno vrednost ciljne vsote uspešnih kampanij 15.638$, neuspešnih pa kar 94.048$. Povprečni vrednosti se precej razlikujeta, saj je med neuspešnimi projekti kar nekaj takšnih, ki so poskušali zbrati nekaj milijonov ameriških dolarjev. Podatki zbranih projektov kažejo, da so projekti, ki zbirajo manjše vsote denarja, tipično bolj uspešni. Število uspešnih (označeno z rdečo) in neuspešnih kampanij (označeno z modro) glede na ciljno vsoto denarja, ki jo projekt hoče pridobiti, prikazuje slika 3.13. Delež uspešnih kampanij, ki zbirajo manj kot 1.150$, denimo presega 50%. Pri tistih s ciljno vsoto višjo od 56.000$ pa je delež uspešnih kampanij manjši od 17%.

Podobno velja tudi za število projektov, ki jih je avtor projekta do sedaj podprl. V tem primeru so tipično bolj uspešne kampanije, kjer je število takšnih projektov večje. Kot kaže tabela 3.14, avtorji uspešnih projektov povprečno podprejo skoraj 11 projektov več kot avtorji neuspešnih. Odvisnost atributa od uspešnosti projekta prikazuje slika 3.14. Uspešnost 14.046



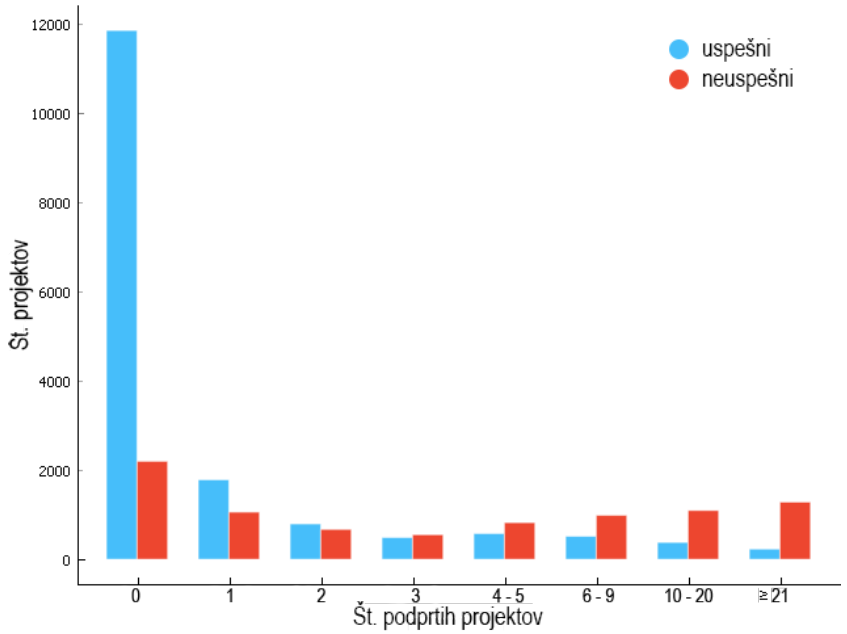
Slika 3.13: Število uspešnih in neuspešnih kampanij glede na želeno ciljno vsoto.

Tabela 3.14: Statistični podatki števila podprtih projektov avtorja glede na uspešnost.

	Povprečje	Varianca
uspešni	12.69	34.37
neuspešni	1.71	11.17

projektov, katerih avtorji niso podprli drugih projektov, znaša le 15,6%. Med zbranimi projekti je tudi 1.508 projektov, katerih avtorji so podprli več kot 20 drugih projektov. Uspešnost takšnih projektov znaša 84,9%.

Uspeh kampanije pa je odvisen tudi od podatkov v opisu projekta, ponujenih nagrad in trajanja kampanije. Uspešnost projektov, ki v opisu ne vsebujejo slike, je denimo le 13,3%, uspešnost projektov z več kot 25 slikami v opisu pa kar 70,5%. Porazdelitev uspešnih kampanij glede na število slik v opisu prikazuje slika 3.15. Uspeh kampanije je odvisen tudi od števila različnih donacij oz. števila nagrad in njihovih vsot. Kot prikazuje tabela 3.15, je število nagrad tipično večje pri uspešnih projektih. To velja tudi za razpon možnih donacij. Podatki v tabeli 3.15 kažejo, da so na platformi



Slika 3.14: Uspešnost kampanij glede na število podprtih kampanij ustvarjalca projekta.

uspešnejši projekti s krajšo dobo trajanja, z daljšim opisom projekta, avtorji uspešnih projektov pa imajo v povprečjo več izkušenj s projekti.

Med izdelavo naloge smo našli objavo na portalu Genius Games[6], v kateri poročajo, da je uspeh projekta na Kickstarter-ju odvisen od meseca, v katerem se odvija kampanija. Zato smo v nabor atributov vključili tudi mesec kampanije. Pri analizi pomembnosti atributov nismo opazili, da bi ta atribut ključno vplival na rezultat naših metod. Smo pa vseeno opazili manjši delež uspešnih kampanij v mesecu juliju in avgustu, vendar pa razlika ni tako očitna, kot je prikazano v omenjeni objavi. Porazdelitev uspešnih in neuspešnih projektov glede na mesec začetka kampanije prikazuje slika 3.16. V analizi projektov iz leta 2014, ko je omenjena raziskava nastala, smo opazili večjo odvisnost uspeha od meseca začetka kampanije. Rezultate iz leta 2014 prikazuje tabela 3.16.

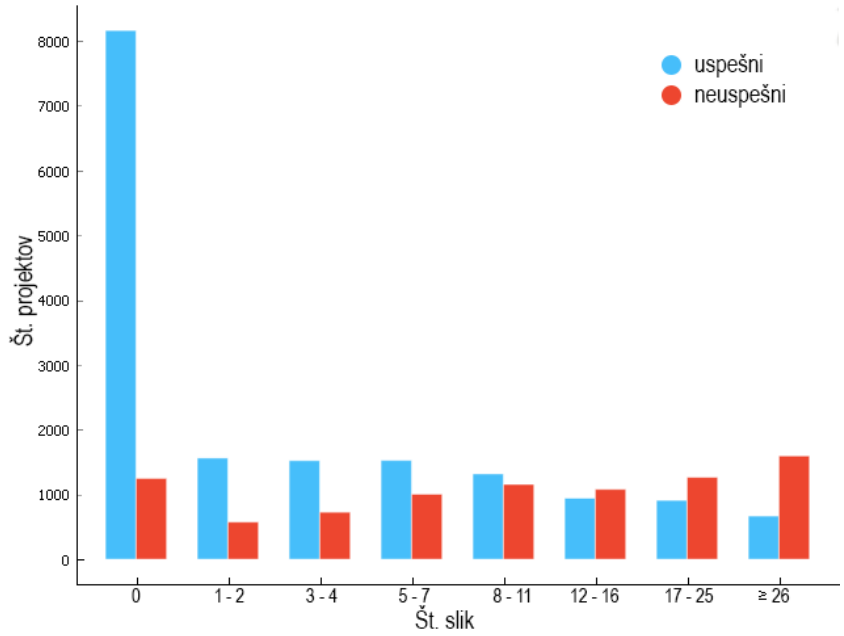
[6] `https://gotgeniusgames.com/kickstarter-stats-101-does`

Tabela 3.15: Podatki o uspešnih in neuspešnih projektih glede na vrednosti atributov.

Atribut		Povprečje	Standardni odklon
št. nagrad	neuspešni	6.06	4.73
	uspešni	10.22	6.66
prisotnost videa	neuspešni	0.59	0.49
	uspešni	0.86	0.35
št. znakov v opisu	neuspešni	3909.19	4125.12
	uspešni	6777.58	5163.55
trajanje	neuspešni	34.66	12.37
	uspešni	31.31	11.06
največja donacija	neuspešni	1789.46	2931.56
	uspešni	1917.36	2860.91
najmanjša donacija	neuspešni	73.69	593.37
	uspešni	10.91	49.86
št. ustvarjenih projektov	neuspešni	0.70	15.50
	uspešni	2.02	5.14

Tabela 3.16: Delež uspešnosti projektov glede na mesec začetka kampanije.

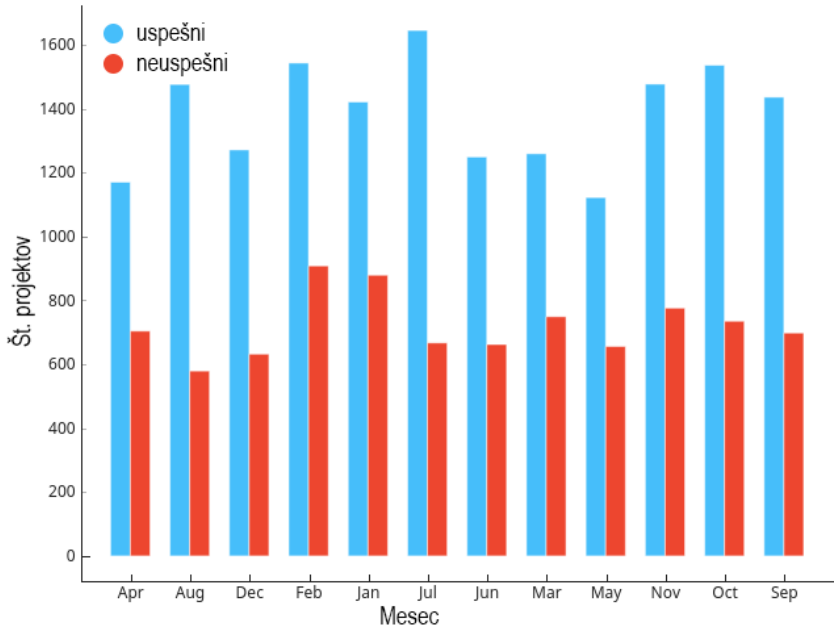
Mesec	Projekti iz leta 2014	Vsi projekti
januar	0.65	0.38
februar	0.65	0.37
marec	0.60	0.37
april	0.59	0.38
maj	0.54	0.37
junij	0.42	0.35
julij	0.27	0.29
avgust	0.29	0.28
september	0.33	0.33
oktober	0.30	0.32
november	0.38	0.34
december	0.57	0.33



Slika 3.15: Uspešnost kampanij glede na število slik v opisu projekta.

3.3.2 Primerjava pomembnih značilk s sorodnimi deli

Raziskave na področju napovedovanja uspešnosti projektov na Kickstarter-ju tipično poleg podatkov na strani projekta uporabljajo tudi podatke socialnih omrežij. Osredotočeni so predvsem na podatke na Twitterju, saj so javno dostopni. Vendar pa so v raziskavi KickPredict iz Kalifornije [1] ugotovili, da so podatki iz predstavitve projektov veliko bolj pomembni. V poročilu poročajo, da so najpomembnejši atributi: število podprtih projektov avtorja, število dosedanjih projektov avtorja, prisotnost videa in ciljna vsota. Te atribute smo tudi sami uvrstili med pomembne atribute, ki jih prikazuje tabela 3.13. O pomembnosti videa na strani projekta poročajo tudi v raziskavi iz Tehnološkega inštituta iz George [12]. Kot pomembne atribute navajajo še število ponujenih nagrad, povezanost avtorja s Facebook-om in število objav o dejavnosti projekta, ki ga sami nismo uporabili, saj smo ga v razdelku 3.2.3 označili kot neprimernega. Nekateri atributi podkategorij in trajanje kampanije imajo pri njihovem klasifikacijskem modelu neničelno utež.



Slika 3.16: Histogram uspešnosti projektov glede na mesec začetka kampanije.

Nabor atributov se med posameznimi raziskavami precej razlikuje. Na projektu univerze Stanford [2] in na projektu KickPredict iz Kalifornije [1] so dinamično napovedovali že med potekom projekta na podlagi trenutno zbranega denarja. V naši diplomski nalogi smo uvrstili vse atribute, ki so jih uporabile opisane raziskave, in dodali nove. Tako smo dodali število znakov v opisu projekta in naslovu, atribute meseca in leta ter število videoposnetkov v opisu. Izkazalo se je, da sta atributa dolžine opisa in naslova projekta bolj pomembna kot atribut meseca in leta.

3.3.3 Porazdelitev verjetnosti uspešnosti kampanij

Pri vrednotenju s prečnim preverjanjem je metoda stacking napačno negativnim primerom napovedala različne napovedi, katerih porazdelitev prikazuje slika 3.6. Opazimo, da je porazdelitev precej enakomerna. Pričakovali smo večji delež napovedanih verjetnosti bližje meji 0.5, ki pa jih pri napačno nega-

tivnih primerih ni. Boljše rezultate dobimo pri napačno pozitivnih primerih, kjer število primerov v splošnem pada z večanjem verjetnosti pripadnosti pozitivnemu razredu, kot prikazuje slika 3.7. Tako smo pri metodi stacking dobili manjše napake pri napačno pozitivnih primerih, saj imajo manjše število napovedanih verjetnosti, ki določajo močno pripadnost napačnemu razredu.

Zgoščene napovedne verjetnosti ob meji med razredoma opazimo tudi pri porazdelitvi napačno pozitivnih in negativnih primerov pri vrednotenju na novih projektih. Porazdelitve verjetnosti napačno negativnih in pozitivnih primerov prikazujeta sliki 3.9 in 3.10. Uspešne porazdelitve verjetnosti pripadnosti napačnemu razredu se izražajo v visoki meri AUC.

Poglavje 4

Zaključek

V diplomskem delu smo napovedovali uspešnost projektov na Kickstarter-ju z uporabo različnih metod klasifikacije. Z rezultati vrednotenja klasifikacijskih modelov smo bili zelo zadovoljni, saj smo dosegli mero AUC višjo od 0,9 in boljšo klasifikacijsko točnost ob začetku kampanije kot ostale raziskave na tem področju, katerih rezultati so opisani v razdelku 1.1.

Med izdelavo diplomske naloge smo se naučili uporabe knjižnic za pridobivanje podatkov iz spleta, obdelave podatkov, vizualizacije in strojnega učenja v programskem jeziku Python. Spoznali smo, da je pri napovednih problemih ključna izbira atributov, od katerih je odvisen razred, ki ga napovedujemo.

Napovedovali smo na podlagi podatkov zaključenih projektov, ki so dostopni že pred začetkom kampanije. Naš napovedni model bi lahko izboljšali z vključitvijo zanimivih jezikovnih značilk s pomočjo dostopnih programov za analizo jezika. Pri raziskavi iz George [12] so s programom LIWC analizirali devet milijonov fraz, zbrali kot atribute najpomembnejše in tako za več kot 14% izboljšali klasifikacijsko točnost njihovega napovednega modela. Atributom jezika bi lahko dodali tudi atribut števila napak v opisu projekta. Z dodajanjem atributov pa bi rasla časovna zahtevnost gradnje in vrednotenja modelov, ki nam je že sedaj pri ansambelskih metodah povzročala manjše težave.

V diplomsko nalogo bi lahko vključili tudi podatke o gibanju denarja skozi čas, kar bi nam omogočilo napoved uspeha že med samo kampanijo. V tem primeru bi lahko kot nekatere dosedanje raziskave na tem področju vključili tudi podatke o objavah na Twitterju. Tako bi dobili nove atribute o številu čivkov, njihovih odgovorih in številu sledilcev.

Rezultat napovednih modelov bi lahko izboljšali z razširitvijo učne množice z dodatnimi atributi. Na strani projekta na Kickstarter-ju se morda skriva značilka, ki je nismo opazili, vendar pa vpliva na uspešnost kampanije. Menimo, da bi točnost napovednih modelov lahko izboljšali z učno množico, ki bi vsebovala večje število projektov.

Literatura

[1] Kevin Chen, Brock Jones, Isaac Kim, and Brooklyn Schlamp. Kickpredict: Predicting kickstarter success. Technical report, California Institute of Technology, 2013.

[2] Nihit Desai, Raghav Gupta, and Karen Truong. Plead or Pitch? The Role of Language in Kickstarter Project Success. Technical report, Stanford University, 2015.

[3] Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple classifier systems*, pages 1–15. Springer, 2000.

[4] Saso Džeroski and Bernard Ženko. Is combining classifiers with stacking better than selecting the best one? *Machine learning*, 54(3):255–273, 2004.

[5] Vincent Etter, Matthias Grossglauser, and Patrick Thiran. Launch hard or go home!: predicting the success of kickstarter campaigns. In *Proceedings of the first ACM conference on Online social networks*, pages 177–182. ACM, 2013.

[6] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

[7] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[8] John Hartigan and Manchek Anthony Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[9] Hamid R Jamali and Mahsa Nikzad. Article title type and its relation with the number of downloads and citations. *Scientometrics*, 88(2):653–661, 2011.

[10] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, volume 14 of *IJCAI'95*, pages 1137–1145, 1995.

[11] Kai Ming and Witten Ian. Stacked generalization: when does it work? Technical report, Working Paper 97/3, Department of Computer Science, University of Waikato, 1997.

[12] Tanushree Mitra and Eric Gilbert. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 49–61. ACM, 2014.

[13] Gunnar Rätsch, Takashi Onoda, and K-R Müller. Soft margins for adaboost. *Machine learning*, 42(3):287–320, 2001.