Predviđanje vikend zarada filmova

Miloš Ostojić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad, Republika Srbija
milos.ostojic@uns.ac.rs

Dejan Grubišić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad, Republika Srbija
grubisic.dejan@yahoo.com

Milorad Trninić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića
21000 Novi Sad, Republika Srbija
milbor16@gmail.com

***Apstrakt*—Filmska industrija zarađuje 40ak milijardi dolara svake godine. Glavno tržište za tu industriju su teritorije SAD i Kanade, jer tu ima najviše publike i najveći procenat zarade od naplaćenih karti. Ovaj rad se bavi predviđanjem zarada filmova tokom vikenda, pošto su vikendi delovi nedelje kada filmska publika najviše ide u bioskope. Podaci su prikupljeni sa najrelevantijih sajtova koji se bave ocenjivanjem i analitikom zarade filmova, kao i iz jednog naučnog rada. Predviđanje zarade se vrši pomoću ocena filma od strane kritike i publike, broja bioskopa u kojima se film prikazuje, budžeta filma i koeficijenta zainteresovanosti publike (gledano kroz pol publike) za film u odnosu na žanr i opis zapleta filma. Za predviđanje koeficijenta zainteresovanosti publike korišćena je logistička regresija, SVM i kNN. Modeli logaritamske regresije i SVM pokazali su slične performanse dok je model kNN bio značajno lošiji. U svakom modelu korišćenje tehnike *tf-idf* dovelo je do značajnog poboljšanja preciznosti. Za predviđanje zarada filma je korišćena linearna regresija, neuronska mreža, SVM, naivni Bajes i AdaBoost. Najbolje rezultate je pokazala neuronska mreža, sa između 76% i 91% tačnosti, u zavisnosti od skupa podataka.**

Ključne reči —predviđanje zarade filmova; regresija; analiza sentimenta; klasifikacija;

I. UVOD

Filmska industrija je tokom 2018. godine zaradila 41.5 milijardi američkih dolara. Najbitnije filmsko tržište na svetu za holivudske studije je tržište Sjedinjenih Američkih Država i Kanade. To tržište uzima najveći udeo ukupne zarade, 11.5 milijardi dolara. Kinesko filmsko tržište je u porastu (prošle godine zarada od 8.9 milijardi dolara) i preti da će u skorijoj budućnosti pretići po zaradi američko tržište. Holivudski studiji će i kada se to dogodi i dalje smatrati američko tržište najbitnijim, jer tu najveći procenat od prodatih karti za filmove uzimaju. Iz tog razloga najveći broj analitika i predviđanja zarada se vrši nad podacima vezanim za američko tržište.

U ovom radu biće predstavljeno jedno rešenje za predikciju vikend zarada filmova na teritoriji SAD. Rešenje je realizovano pomoću više modela za predikciju, na osnovu dostupnih podataka. Korišćena je i model za predikciju zainteresovanosti publike po polu u skladu sa žanrom i kratkim opisom zapleta filma. Za predviđanje su korišćeni podaci o prihvaćenosti filma od strane kritike i publike.

Postojalo je mnogo izazova pri realizaciji. Podatke je teško prikupiti. Mnoge organizacije nisu voljne da dele svoje podatke, što je otežalo ili onemogućilo dobijanje željenih podataka. Izazov je predstavljalo i spajanje podataka iz različitih izvora. Kod različitih organizacija se znalo desiti da je isti film zaveden pod različitim imenima ili pod različitim godinama izlaska. Opisi zapleta filma mogu biti previše kratki ili neodređeni, što je otežalo predikciju žanra i zainteresovanosti publike.

Detaljniji opis podataka i metodologija je izložen u ostatku rada. Drugo poglavlje se bavi srodnim istraživanjima na ovu temu. U trećem poglavlju je opisan skup podataka i način pripreme podataka za obučavanje modela. U četvrtom poglavlju je predstavljena metodologija korišćena za predviđanje vikend zarada filmova. U petom podatku su prikazani dobijeni rezultati. Poslednje poglavlje je zaključak rada.

II. SRODNA ISTRAŽIVANJA

Quader i drugi su u svom radu [1] predviđali uspeh filma (da li je profitabilan ili ne) koristeći podatke dobijene sa *Box Office Mojo*-a, *Metacritic*-a i drugih. Predviđali su uspešnost pomoću tehnike potpornog vektora (eng. *Support vector machine*, skr. SVM), neuronskih mreža i NLP-a (eng. *Natural language processing*). Došli su do zaključka da budžet, broj *IMDb* ocena i broj bioskopa u kojima se film prikazuje najviše utiču na finansijski uspeh filma.

Sharda i *Delen* u svom radu [2] su koristili neuronske mreže i logističku regresiju za predviđanje zarada filmova. Podelili su filmove na 9 različitih kategorija po zaradi i predviđali su pripadnost tim kategorijama.

Cook i drugi su napisali rad [3] koji predviđa uspeh na osnovu na osnovu da li je zaradio 110% svog budžeta. Od klasifikatora su koristili i naivni Bajes (eng. *Naive Bayes*). Imali su 65% uspešnosti.

Hoang je u svom radu [10] koristio naivni Bajes, *Word2Vec+XHBoost* i rekurentne neuronske mreže za tekstualnu klasifikaciju. Za označavanje žanra je koristio k-binarnu transformaciju, rang metodu i probabilističku klasifikaciju nad 250,000 filmova. Postigao je F-*score* od 0.56 i 80.5% pogodaka.

III. Opis skupa podataka

Skup podataka koji je korišćen u radu prikupljen je sa nekoliko najpoznatijih sajtova specijalizovanih za filmsku i televizijsku industriju. Ti sajtovi su *boxofficemojo.com*, *rottentomatoes.com*, *metacritic.com*, *imdb.com* i *cinemascore.com*. Inicijalna lista sa nazivima filmova I ocenama publike preuzeta je sa *Cinemascore*-ove internet stranice, na kojoj su ocene publike za film po izlasku iz bioskopa. U filmskoj industriji on služi kao indikator da li je film pogodio publiku. Na osnovu preuzetih naziva filmova su pretraženi drugi atributi na gore navedenim sajtovima uz pomoću njihovog javnog i OMDB API-ja. OMDB API je besplatan servis koji prikazuje podatke sa već navedenih sajtova, a njihovu pretragu omogućava po nazivu filma ili po *IMDB* identifikacionom broju.

Skup podataka preuzet pomoću OMDB API-ja sastoji se od 4077 filmova gde su za svaki prikupljene sledeće osobine:

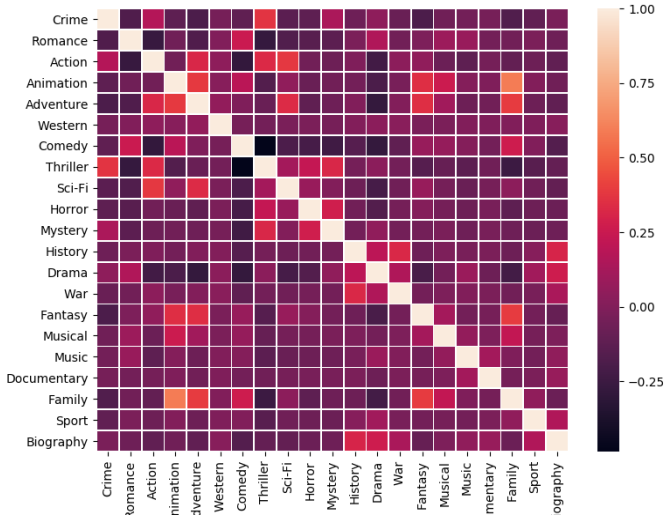
- Naziv filma
- Godina izlaska (*wide release*, pušten u većini bioskopa u SAD)
- Tačan datum izlaska u SAD
- Opis zapleta filma (IMDB)
- *Cinemascore* ocena
- IMDB ocena korisnika sajta
- *Metacritic* ocena
- *Rotten Tomatoes* ocena korisnika sajta
- Žanrovi kojima film pripada
- Vreme trajanja u minutima
- MPAA rejting (za koje godište je predviđen film)

U podatke su uključeni i *Metacritic* i *Rotten Tomatoes* ocene kritike, jer se razlikuju metrike kojom računaju ocenu filma. *Metacritic* računa prosečnu vrednost ocena kritičara (od 0 do 100 su ocene). *Rotten Tomatoes* ocenu pravi kao procenat pozitivno ocenjenih filmova.

Sa sajta *boxofficemojo.com* su preuzeti su podaci o zaradama filma i o budžetu filma. Podaci preuzimani godinu po godinu, počevši od 2000, zaključno sa 2018. godinom. Preuzeti su potom spojeni u jedinstvenu datoteku gde se nalaze podaci o svim zaradama tokom vikenda svakog filma koji je izašao u tom vremenskom periodu. Sledeće spajanje podataka je izvšeno sa bazičnim podacima o svakom filmu sa sajta *boxofficemojo.com*. Time je dobijen skup podataka sa 36450 redova. U spajanju sa podacima dobijenim preko OMDB API-ja, prilikom kog su izbacivani redovi sa nedostajućim i nevalidnim podacima, dobijena je datoteka sa 15419 redova. Podacima dobijenog preko OMDB API-ja pridodati su sledeći podaci:

- Vikend u godini (u kom je prikazan film)
- Godina u kojoj je posmatrani vikend za koji će se vršiti analiza
- Rang mesto zarade filma na listi vikenda
- Broj bioskopa u kojima je film bio prikazan taj vikend
- Zarada filma za posmatrani vikend u američkim dolarima
- Ukupna zarada filma do posmatranog vikenda u američkim dolarima
- Ukupna domaća zarada filma (SAD) u američkim dolarima
- Produkcijski budžet u milionima američkih dolara

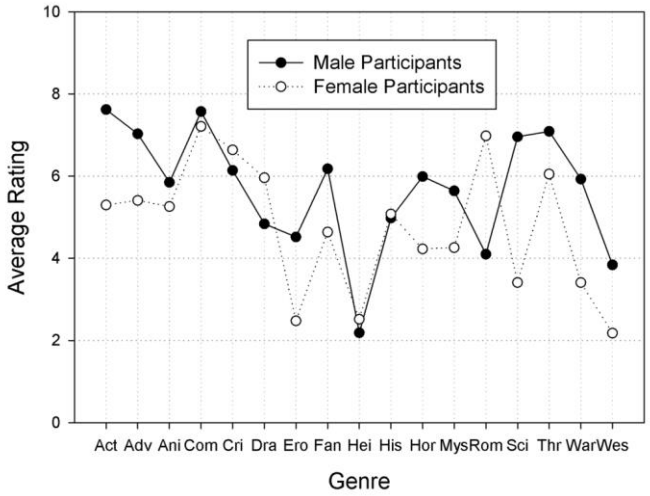
Od OMDB API podataka se određuje se lista svih žanrova u setu podataka i kreira se novi skup podataka za trening u kome svaki žanr ima kolonu. U opštem slučaju jedan film može imati više žanrova i analizom skupa podataka dobijeni su sledeći žanrovi i njihov udeo u celom skupu: *Action*(29.8%), *Adventure*(24.5%), *Animation*(6.6%), *Biography*(5.9%), *Comedy*(46%), *Crime*(21%), *Documentary*(0.8%), *Drama*(43.8%), *Family*(15.3%), *Fantasy*(15%), *History*(3.3%), *Horror*(9.6%) *Music*(4.5%), *Musical*(2.3%), *Mystery*(11.1%), *Romance*(22.7%), *Sci-Fi*(15%), *Sport*(4.8%), *Thriller*(31%), *War*(3.2%) i *Western*(1.1%). Između žanrova takođe postoji korelacija prikazana na slici 1, gde 1 predstavlja najveću pozitivnu, -1 najveću negativnu korelaciju, dok 0 znači da nema korelacije.



Slika 1. Korelacija žanrova

Rad [8] je izvršio ispitivanje o zainteresovanosti za neki žanr filma u odnosu na pol i dobijeni rezultati se porede sa stereotipima o muškim i ženskim filmovima.

U ispitivanju su učestvovali 80 muškaraca i žena, koji su ocenjivali 17 žanrova ocenom od 1 do 10 (1-najmanje, 10-najviše). Kao rezultat ispitivanja dobijena je tabela, na osnovu koje je pravnjena predikcija prihvaćenosti filma u odnosu na pol. Te ocene su ušle kao ocene žanra u podatke.



Slika 2. Ocene privlačnosti žanrova po polu [8]

Ova studija nije obuhvatila sve žanrove u setu podataka, pa su njihovi koeficienti određeni na osnovu ličnog iskustva. Koeficijenti i žanrovi koji su dobijeni na ovaj način su *Biography(6, 6), Documentary(5, 5), Family(5, 6.3), Music(4, 5.5), Musical(4, 5.5), Sport(5.1, 5).*

IV. METODOLOGIJA

Metodologija ovog rešenja može biti podeljena u pet celina.

A. *Analiza i generisanje podataka*

Podaci koji su dobijeni spajanjem više izvora prolaze kroz četi obrade.

U prvoj obradi im se pridružuje kolona koja govori koji vikend predstavlja po redosledu od izlaska filma, da bi se olakšalo dalje generisanje podataka. To je učinjeno prebrojavanjem koliko vikenda je film bio aktivan pre vikenda koji se obrađuje.

U drugoj obradi je pridružena kategorija kojoj vikend zarada filma pripada. Empirijski je određeno da se filmovi rasporede u osam kategorija. Kategorije su:

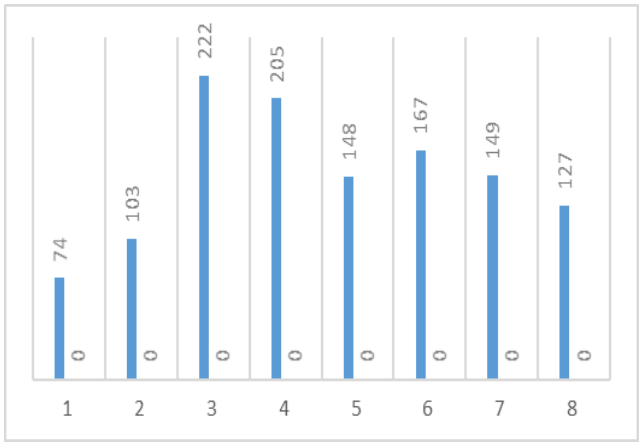
1. do milion dolara,
2. od milion do 5 miliona dolara,
3. od 5 miliona do 10 miliona dolara,
4. od 10 miliona do 15 miliona dolara,
5. od 15 miliona do 20 miliona dolara,
6. od 20 miliona do 30 miliona dolara,
7. od 30 miliona do 50 miliona dolara,
8. od 50 miliona dolara.

Kategorije su predstavljene numerički, brojevima od 1 do 8. Pri određivanje kategorija pomoglo je i što su izračunati kvartili.

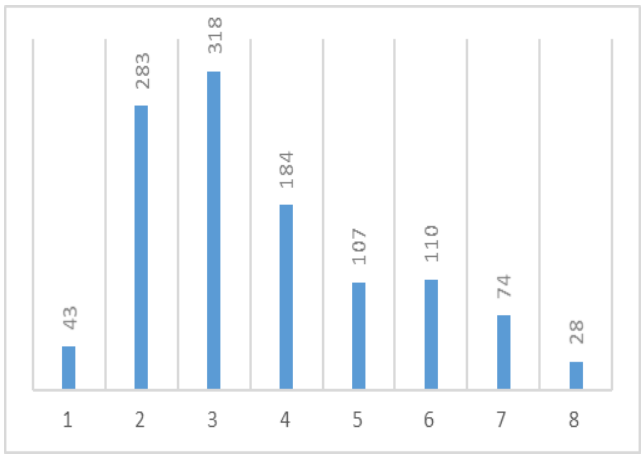
Prvi kvartil (25% vrednosti skupa je ispod vrednosti) za zaradu filma tokom prvog vikenda iznosi 7,588,084 dolara.

Drugi kvartil (50% vrednosti skupa je ispod vrednosti, medijana) iznosi 14,800,723 dolara.

Treći kvartil (75% vrednosti skupa je ispod vrednosti) iznosi 28,405,524 dolara.



Slika 3. Prikaz rasporeda filmova po kategorijama za prvi vikend



Slika 4. Prikaz rasporeda filmova po kategorijama za drugi vikend

U trećoj obradi upisana su generisana predviđanja o zainteresovanosti publike za film na osnovu opisa zapleta i žanra, više o tome u sledećem potpoglavlju.

U četvrtoj obradi su korišćeni rezultati druge obrade i na osnovu toga su generisana dva skupa podatka, u prvom su podaci o filmovima kojima je pridružena zarada tokom prvog vikenda, a u drugom su podaci o filmovima kojima je pridružena zarada tokom drugog vikenda.

Iz analize podataka sa boxofficemojo.com utvrđeno je da skoro pola zarade filmovi zarade kada se završi druga nedelja od kada je film izašao na tržište. Zato se u ovom radu predviđa zarada za prva dva vikenda od kada je film izašao.

B. *Predviđanje žanra filma u odnosu na opis zapleta filma*

Predviđanje žanra na osnovu opisa zapleta filma spada u probleme analize sentimenta. Ovakav problem obrađen u [9] i po uzoru na njega realizovane su sledeće tehnike za predviđanje žanra filma u odnosu na opis zapleta filma:

1. Logistička regresija
2. Tehnika k-najbližih komšija (eng. *k-NearestNeignbors*)
3. Tehnika potpornih vektora

Ove tehnike unapređene su još sa tehnikom relativne frekvencije izraza (eng. *term frequency–inverse document frequency, tf-idf*), prilikom obrade opisa filma, kako bi se dodatno poboljšale njihove performanse.

Prvi korak za svaku od navedenih tehnologija predstavlja pretprocesiranje opisa filma, koje je zasnovano na tehnici zbrajanja reči (eng. *bag of words*). Za svaki film se opis se svodi na mala slova, određuje se set zaustavnih reči, koje nemaju konteksnu vrednost, da bi se na kraju ostavile sve ostale reči koje nose značenje.

Problem određivanja liste žanrova može se odvojiti na predviđanje pojedinačnih žanrova na osnovu opisa radnje (eng. *one-versus-all*). Ovaj pristup je korišćen kod svih tehnologija. Na osnovu udela žanrova se može videti da su podaci nebalansirani. Ovo je ipak ostavljeno prilikom obuke, jer se na ovaj način mreža bi trebala da odbaci žanr, osim ako ne postoje velike indicije za njega. Kako u većini slučajeva postoji više od jednog žanra, slučaj kada greškom nije predviđen žanr (eng. false-negative) manje utiče na konačan rezultat, nego u slučaju pogrešno procenjenog žanra (eng. false-positive), što će biti opisano u nastavku. Pored ovoga postoji zavisnost između nekih žanrova, pa nedostatak jednog žanra može biti kompenzovan drugim žanrom.

Prvi model zasniva se na se na logističkoj regresiji. Ova tehnologija odabrana je jer omogućava da jednostavan model da ostvarimo binarnu klasifikaciju. Logistička regresija efikasnije radi u slučajevima kada se veća vrednost pridodaje ulazima od interesa. Kako je u ovom slučaju određivanje žanra zasnovano na opisu radnje filma i tehnici zbrajanja reči, model je unapređen davanjem veće težine rečima, čija je zastupljenost veća nego u ostalim opisima filmova (tf-idf). Ovaj koncept se često koristi u obradi teksta i nalaženju semantike.

Težine koje se dobijaju korišćenjem tehnike tf-idf su:

$$\text{tf-idf}(t, d) = \text{tf}(t, d) * \text{idf}(t)$$

$$\text{idf}(d, t) = \log [(1 + n) / (1 + \text{df}(t))] + 1.$$

tf(t) - broj reči u svim opisima filma,

n - broj opisa filmova, df(t)-broj opisa sa zadatom reči

Tehnika najbližih komšija izabarna je kao drugi model. Argument u ovom slučaju bio je da ako su dva filma istog žanra, i reči koje ih opisuju će da budu slične. Algoritam za pretraživanje k-najbližih komšija izabran je *brute-force*, a na osnovu empirijskih rezultata ustanovljeno je za broj komšija K=13, dobijaju najmanja odstupanja, dok je za realizaciju *sa tf-idf* tehnikom K=11. Kao metrika za razdaljinu je korišćena kosinusna metrika. Ova metrika računa udaljenost u prostoru na osnovu ugla izmedju dva vektora. U slučaju prostora reči, svaka reč predstavlja jednu dimenziju i na osnovu broja ponavljanja reči i njihovom linearnom kombinacijom dobija se vektor u tom prostoru. U ovakvom prostoru ova metrika je znatno pogodnija od euklidske udaljenosti, jer bi u slučaju euklidske metrike, razdaljina bila najvise uslovljena brojem ponavljanja reči u vektoru, umesto razlikom reči u vektorima. Drugim rečima vektori sa jednom reči bi bili svi međusobno blizu, dok bi se povećavanjem broja reči ta udaljenost povećavala.

Kao poslednji model uzet je model potpornih vektora. Ovaj pristup je odabran jer pruža dobro predviđanje kada je prostor visoko dimenzionalan u odnosu na broj podataka. U ovom slučaju za neke žanrove važi da imaju samo nekoliko instanci. Na osnovu rada [7] pokušano je da se poveća težina penala prilikom promašaja za žanrove sa manje od 15% podataka, ali dobijeni rezultati nisu značajno odstupali.

U svakom od predhodnih modela urađena je unakrsna validacija metodom *k-fold cross validation* kako bi se dobila predikcija za ceo skup podataka i kako bi parametri preciznosti bili pouzdaniji.

C. *Određivanje naklonosti publike u odnosu na žanr*

Predikcija zarade filmova može zavisiti od toga da li film privlači više mušku ili žensku publiku.

Labela prihvaćenosti filma u odnosu na pol nalazi se u opsegu od [-1,1], gde -1 označava veće interesovanje ženske publike dok 1 označava veće interesovanje muške publike. Kako su koeficijenti muške publike veći u proseku od koeficijenata ženske, određivanje labele dobijeno je na sledeći način. Za svaki žanr se računa razlika koeficijenata i deli se sa većim, kako bi se dobio zadati opseg. U koliko je razlika veća od nule dodaje se listi "muških", odnosno listi "ženskih" filmova u suprotnom. Dobijene vrednosti se unutar lista dele sa svojom pozicijom, počevši od 1, čime se smanjuje uticaj viših koeficienata muške publike. Na kraju se vrednosti dve liste sumiraju i zaokruživaju na -1 ili 1 ako je dobijena suma manja odnosno veća od te vrednosti, čime se dobija labela.

Kao model za obučavanje mreže korišćena je linearna regresija sa unakrsnom evaluacijom. Kao lista ulaznih parametara uzete su binarne vrednosti žanrova po filmu, dok je za očekivanu vrednost uzeta izračunata labela.

D. *Predviđanje žanra filma u odnosu na opis filma*

Testirani modeli za predikciju žanra na osnovu opisa su model logističke regresije, k-najbližih komšija i model potpornih vektora. Svaki od ovih modela ispitan je u osnovnom obliku i korišćenjem tehnike tf-idf, opisanoj u metodologiji. Modeli su obučavani nad 2336 instanci, dvostrukom unakrsnom validacijom.

E. *Predviđanje vikend zarada filma*

Korišćena je *Scikit-learn* biblioteka za mašinsko učenje u *Python*-u.

Iz skupa podataka izdvojeni su za potrebe predviđana sledeći podaci o filmovima:

- *cinemascore*
- ocena korisnika na sajtu imdb.com
- ocena filmskih kritičara na sajtu metacritic.com
- ocena filmskih kritičara na sajtu rottentomatoes.com
- broj bioskopa u kojima je prikazan film
- budžet produkcije za film.

Za predviđanje zarade filma tokom drugog vikenda je korišćen i podatak o zaradi prvog vikenda.

Za predikciju zarada filmova tokom vikenda korišćeni su regresioni modeli:

- linearna regresija
- linearna regresija sa *lasso* regulacijom
- linearna regresija sa *ridge* regulacijom
- *AdaBoost* konfiguracija

Svaki regresioni model je primenjen na oba skupa podataka. Mera koja je korišćena za evaluaciju performansi modela je R^2 i desetostruko presavijena unakrsna validacija (eng. *10-fold cross validation*).

Multivarijabilna linearna regresija je prvo izvršena. Ona spada u osnovne i najčešće korišćene metode prediktivne analize podataka. Generalno se preporučuje da se linearna regresija prvo izvrši, radi poređenja sa drugim modelima.

Laso (eng. *Lasso*, skraćeno od eng. *least absolute shrinkage and selection operator*) je korišćena za optimizaciju parametara i smanjenje *over-fitting*-a, videti TABELA I. Tačnost rezultata je bila slična onim dobijenim linearnom regresijom.

Ridge (dobilo ime od engleske reči za greben[4], u literature se može naći i pod imenom Tihonova regularizacija) regresija jeste tip linearne regresije koja pravi "škrt" model u slučajevima multikolinearnosti. Za razliku od proste linearne regresije u kojoj se koristi metoda najmanjih kvadrata ridge regresija pravi razliku između važnih i manje važnih prediktora (nezavisnih promenljivih). Zbog toga su manje šanse da dođe do *over-fitting*-a modela. Takođe mnogo bolje radi sa podacima u kojima postoji multikolinearnost. *Ridge* regresija izbegava ove probleme iz razloga što ne zahteva *unbiased* procene (precizan procenitelj, precizan u smislu da ne postoji precenjivanje ili potcenjivanje). *Ridge* regresija dodajte taman toliko *bias*-a da su mu procene razumno pouzdane. Tačnost rezultata je bila slična onim dobijenim linearnom regresijom.

AdaBoost ili prilagođeno ubrzanje (eng. *Adaptive Boosting*) se bavi na problem klasifikacije i ima za cilj da konvertuje skup slabih klasifikatora u jedan jak klasifikator. Koraci algoritma su:

- Treniranje slabih klasifikatora i odabir onog sa najmanjom greškom.
- Računanje težina za klasifikatore. Biraju se svi oni imaju tačnost različitu od 50%..
- Ažuriranje težina za svaku tačku.

Korišćen je model stabala odlučivanja kao slabi klasifikator. Za skup podataka sa zaradama filmova tokom prvog vikenda utvrđeno je da ima najbolju preciznost kada se podesi da je maksimalna dubina stabla odlučivanja 13, a broj *estimator*-a (procenitelja) 200. Za skup podataka sa zaradama filmova tokom drugog vikenda utvrđeno je da ima najbolju preciznost kada se podesi da je maksimalna dubina stabla odlučivanja 14, a broj *estimator*-a 150.

Za klasifikaciju na određene klase su korišćeni:

- SVM
- Naivni Bajes
- neuronska mreža
- linearna regresija
- *AdaBoost* konfiguracija

Za kernel SVM-a je korišćen *rbf* kernel (eng. *Radial basis kernel*, Gausova funkcija) jer je dao najbolje rezultate u [1] i jer se pokazao kao najučinkovitiji u radu sa prikupljenim podacima. Gama parametar definiše koliki je uticaj jednog trening primera, male vrednosti daju veliki uticaj, a velike vrednosti daju mali uticaj na okruženje. Za gama parametar odabrana je vrednost *scale*, i time je postignuto da je uticaj jednog trening primera veći. Za parametar *decision function shape* odabrana je vrednost *ono* (skraćenica od eng. *one-vs-one*), zbog preporuke u dokumentaciji biblioteke i jer je donosio bolje rezultate u odnosu na *ovr* (skraćenica od eng. *one-vs-rest*).

Naivni Bajes (eng. *Naive Bayes*) je korišćen jer se pokazao uspešnim u [3]. Korišćen je Gausov naivni Bajes.

Model neuronske mreže je baziran na neuronskoj mreži [5] koja je prvobitno bila namenjena predviđanju predviđanju na berzi. Model je modifikovan da radi sa podacima potrebnim za ovaj rad i dodat je još jedan sloj neurona, jer se pokazalo da se dobijaju bolji rezultati u tom slučaju. Upotrebljena je *Python*-ova biblioteka *tensorflow*. Za trening je korišćen *AdamOptimizer* sa parametrom 0.001.

Linearna regresija je korišćena za predviđanje klasa tako što je zarada za vikend koju predvidi kategorisana. Potom je dobijena kategorija upoređena sa kategorijom iz podataka za test.

V. REZULTATI

Prikaz rezultata je podeljen u dve celine. Prva celina se odnosi na prepoznavanje žanra i zainteresovanosti publike za film na osnovu njegovog opisa zapleta. Druga celina se bavi rezultatima predviđanja zarada filma tokom vikenda.

A. *Rezultati prepoznavanja žanra i zainteresovanosti publike za film na osnovu njegovog opisa zapleta*

Određivanje naklonosti publike u odnosu na žanrove ima R^2 meru 0.93. Prosečno odstupanje dobijenog rezultata od labele je 0.065, dok je medijan odstupanja 0.08.

Za svaku instance u podacima dobijena je prognoza za svaki žanr, kako bi se dobili sumirani podaci po žanru, prikazani u TABELA I.

Kao metrika korišćeni su preciznost(eng. *precission*) i odziv (eng. *recall*) , kako bi se prikazala realna slika klasifikatora, imajući u vidu veliku ne balansiranost seta podataka. Tačnost u ovom slučaju ima veliku vrednost za svaki žanr, ali ne oslikava stvarnu sliku, zato je većina tačnih predviđanja *true-false*, čime se predviđa da žanr nije prisutan. Kako bi bilo jasniji odnos prisustva nekog žanra u filmu, pored svakog žanra dat je procenat prisustva u celom setu podataka.

TABELA I: Prikaz rezultata predviđanja žanra na osnovu opisa filma

Predviđanje zanra na osnovu opisa filma

	Action (29.8%)			Adventure (24.5%)			Animation (6.6%)			Biography (5.9%)			Comedy (46%)			Crime (21%)			Documentary (0.8%)		
	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f
Log. Reg.	0.58	0.58	0.58	0.59	0.49	0.53	0.62	0.15	0.24	0.60	0.12	0.20	0.67	0.65	0.66	0.59	0.44	0.50	1.00	0.09	0.17
Log. Reg. tf-idf	0.71	0.60	0.65	0.71	0.41	0.52	0.90	0.10	0.19	0.83	0.07	0.12	0.71	0.65	0.68	0.72	0.42	0.53	1.00	0.09	0.17
KNN	0.79	0.36	0.50	0.76	0.20	0.31	0.00	0.00	0.00	1.00	0.03	0.05	0.63	0.57	0.60	0.68	0.12	0.20	0.00	0.00	0.00
KNN tf-idf	0.73	0.43	0.54	0.72	0.27	0.40	1.00	0.01	0.02	0.00	0.00	0.00	0.66	0.58	0.62	0.68	0.20	0.31	0.00	0.00	0.00
SVM	0.59	0.57	0.58	0.58	0.46	0.51	0.54	0.16	0.25	0.45	0.13	0.20	0.69	0.66	0.67	0.59	0.44	0.51	1.00	0.09	0.17
SVM tf-idf	0.74	0.57	0.64	0.76	0.34	0.47	1.00	0.03	0.07	0.00	0.00	0.00	0.73	0.67	0.70	0.71	0.35	0.47	1.00	0.09	0.17
	Drama (43.8%)			Family (15.3%)			Fantasy (15%)			History (3.3%)			Horror (9.6%)			Music (4.5%)			Musical (2.3%)		
	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f
Log. Reg.	0.58	0.57	0.58	0.53	0.24	0.33	0.59	0.37	0.45	0.33	0.03	0.05	0.50	0.27	0.35	0.79	0.21	0.33	1.00	0.07	0.14
Log. Reg. tf-idf	0.64	0.57	0.60	0.74	0.17	0.28	0.70	0.29	0.41	0.00	0.00	0.00	0.85	0.21	0.34	0.89	0.15	0.26	1.00	0.07	0.14
KNN	0.57	0.48	0.52	0.50	0.01	0.02	0.67	0.03	0.06	0.00	0.00	0.00	0.71	0.05	0.09	0.00	0.00	0.00	0.00	0.00	0.00
KNN tf-idf	0.62	0.52	0.56	0.64	0.05	0.09	0.74	0.14	0.24	0.00	0.00	0.00	1.00	0.08	0.14	1.00	0.06	0.11	0.00	0.00	0.00
SVM	0.59	0.57	0.58	0.50	0.25	0.33	0.56	0.37	0.44	0.22	0.05	0.08	0.49	0.31	0.38	0.70	0.27	0.39	1.00	0.07	0.14
SVM tf-idf	0.67	0.57	0.61	0.76	0.10	0.17	0.71	0.21	0.32	0.00	0.00	0.00	0.86	0.11	0.20	0.86	0.12	0.20	1.00	0.07	0.14
	Mystery (11.1%)			Romance (22.7%)			Sci-Fi (15%)			Sport (4.8%)			Thriller (31%)			War (3.2%)			Western (1.1%)		
	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f
Log. Reg.	0.57	0.28	0.37	0.47	0.35	0.40	0.69	0.46	0.55	0.78	0.25	0.38	0.57	0.49	0.53	0.67	0.29	0.40	0.00	0.00	0.00
Log. Reg. tf-idf	0.73	0.17	0.27	0.62	0.36	0.46	0.89	0.38	0.54	1.00	0.25	0.40	0.70	0.50	0.58	1.00	0.17	0.29	0.00	0.00	0.00
KNN	0.67	0.02	0.03	0.68	0.15	0.25	0.82	0.22	0.34	1.00	0.05	0.10	0.68	0.22	0.33	0.00	0.00	0.00	0.00	0.00	0.00
KNN tf-idf	1.00	0.02	0.03	0.63	0.19	0.29	0.88	0.26	0.41	0.78	0.12	0.22	0.67	0.31	0.42	1.00	0.03	0.06	0.00	0.00	0.00
SVM	0.53	0.29	0.37	0.48	0.37	0.41	0.66	0.48	0.55	0.74	0.25	0.37	0.58	0.51	0.54	0.71	0.29	0.41	0.00	0.00	0.00
SVM tf-idf	0.82	0.07	0.12	0.72	0.31	0.43	0.95	0.32	0.48	1.00	0.09	0.16	0.74	0.44	0.55	1.00	0.06	0.11	0.00	0.00	0.00

U ovom slučaju komedije i drame imaju daleko veći udeo i balansiranost nego vestern i dokumentarni filmovi, zbog čega rezultati njihove klasifikacije daju mnogo bolje rezultate. Ovo se takođe oslikava i u rezultatima odziva. Mali odziv je posledica mnogo većeg broja negativnih odgovora u setu podataka, čime se model obučava da odbacuje žanrove sa većom verovatnoćom. Sa druge strane preciznost u nekim slučajevima ima veliku vrednost usled nedostatka podtataka i nebalansiranosti podataka.

Na osnovu prikazanih rezultata može se zaključiti da modeli logaritamske regresije i modela potpornih vektora podjednako dobro klasifikuju, dok model k-najbližih komšija daje značajnije lošije rezultate. U žanrovima sa većom zastupljenošću Thnika *tf-idf* u velikom broju slučajeva pozitivno deluje na predviđanje, što je očekivano.

F-mera dostiže i do 0.7 za slučaj komedije modelom potpornih vektora sa *tf-idf*. Ovaj rezultat predstavlja dobro predviđanje u poređenju sa modelima za klasifikaciju u ovoj oblasti, date u radu [7].

Pored nebalansiranog seta podataka koji utiče na predviđanje, problem u nekim slučajevima je sam opis filma i dodeljen žanr. Na primer film *Scary Movie 3* ima sledeći opis (prevedeno na srpski):

"Misteriozna video traka - ubica ide okolo. Jedan pogled na ovu traku i imate još sedam dana za život. Reporter *Cindy Campbell* (*Faris*) svedoči o ovoj video vrpci i pokušava pronaći način da spreči njenu smrt. Ali to nije jedina misterija koja se pojavljuje. Krugovi žitarica pojavljuju se u lokalnoj farmi *Tom* (*Sheen*) i *Georgeu* (*Rex*). Uz pomoć *Shaneequa* (*Latifah*), *Cindy* sumnja da su vanzemaljci možda povezani s trakom ubice i sada moraju razraditi obe tajne pre nego što bude kraj sveta."

Ovaj film prepoznat je kao triler i drama, dok je on obeležen kao komedija. Problem sa labelom u ovom slučaju je što je traženi film parodija koja ima radnju preduzetu iz filma drugog žanra, zbog ovoga se drastično razlikuju opis i dodeljen žanr. U ovakvim situacijama, teško je odrediti zahtevani žanr. Greške prilikom opisa mogu još biti i ako opis filma ne sadrži dovoljno indormacija.

B. *Rezultati predviđanja zarada filma tokom vikenda*

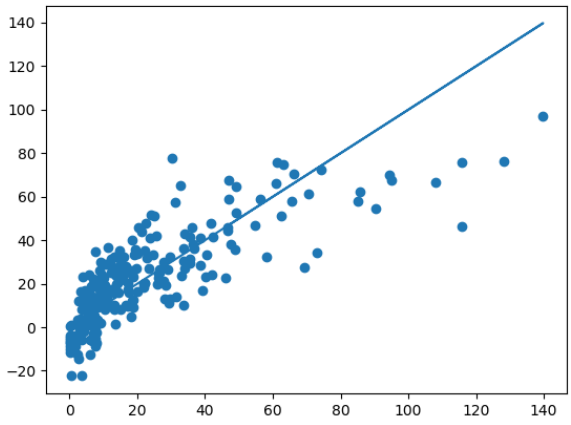
Regresioni modeli, SVM, naivni Bajes i neuronska mreža su testirani na skupu podataka opisanom u III i IV potpoglavlju E.

U TABELA II su prikazani rezultati dobijeni linearnom regresijom, metod za računanje je R^2.

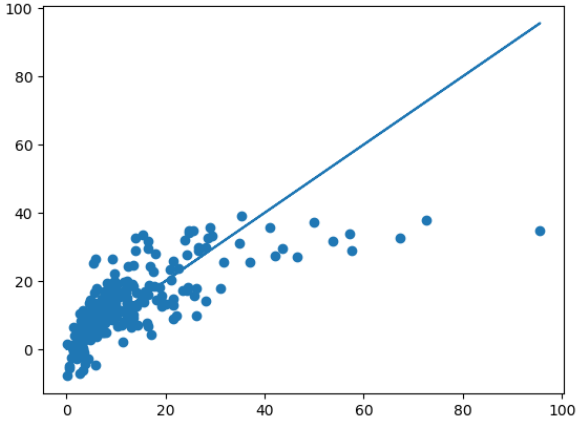
TABELA II: Rezultati linearne regresije

Linearna regresija sa skupom podataka	Rezultati modela linearne regresije	
	R^2 kada je skup podataka podeljen u odnosu 80% za trening, 20% za test	10-*fold cross validation*
Linearna regresija nad skupom podataka zarada filma za prvi vikend	0.63	0.57 (+/- 0.14)
Linearna regresija nad skupom podataka zarada filma za drugi vikend bez parametra sa zaradom filma za prvi vikend	0.52	0.58 (+/- 0.13)
Linearna regresija nad skupom podataka zarada filma za prvi vikend sa parametrom sa zaradom filma za prvi vikend	0.75	0.83 (+/- 0.13)

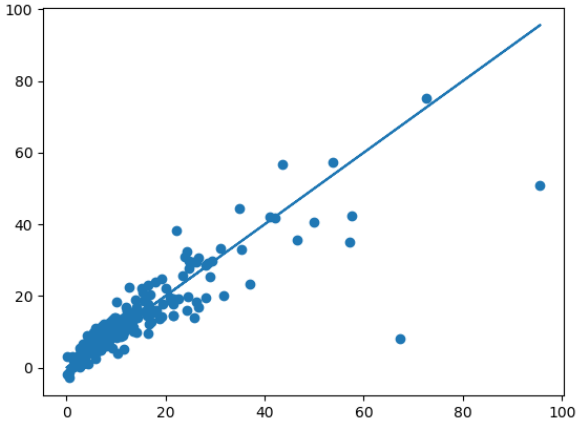
Na slikama 3, 4 i 5. može se grafički videti kako linearna regresija predviđa zaradu filma. Na graficima X-osa predstavlja originalnu vrednost zarade prvog vikenda, Y-osa predstavlja predviđenu vrednost zarade prvog vikenda.



Slika 3. Grafički prikaz rezultata predviđanja linearne regresije nad skupom podataka zarada filma prvog vikenda.



Slika 4. Grafički prikaz rezultata predviđanja linearne regresije nad skupom podataka zarada filma drugog vikenda bez zarade prvog vikenda.



Slika 5. Grafički prikaz rezultata predviđanja linearne regresije nad skupom podataka zarada filma drugog vikenda sa zaradom prvog vikenda.

Sa grafika možemo videti da se najuspešnije predviđaju zarade filmova koji su ostvarili zaradu do 40 miliona, što ima smisla, budući da se najveći broj zarada nalazi u tom opsegu. Među *outlier*-ima (izuzecima) su često filmovi koji su imali ograničeno puštanje u promet (koji su imali na eng. *limited release*). To je taktika da se u prvom nedelji film pusti u ograničenom, malom broju bioskopa, da bi se u sledećoj nedelji povećao broj bioskopa, pa samim tim i zarada. Filmovi koji imaju velike zarade su ovim modelom potcenjeni, jer ih nema mnogo u skupu podatka, a faktori ocena publike i kritičara utiču u manjoj meri na njih. Podatak o zaradi filma tokom prvog vikenda značajno utiče na zaradu tokom drugog vikenda, što je verovatno posledica *word of mouth* (ljudi koji su videli film su preneli svoje utiske drugim ljudima usmeno) fenomena i preciznijeg znanja o tome koliko film može da zaradi.

Rezultati laso i *ridge* regresije su bili slični običnoj regresiji. Korišćena je podela skupa podataka na 80% za trening, 20% za test. Korišćena je R^2 metrika. Korišćeni je više vrednosti parametra α (0.01, 0.001 i 100). Nisu dobijene značajnije razlike u R^2 metrici. Najveća razlika je bila u *ridge* regresiji, gde je razlika u odnosu na linearnu regresiju u R^2 metrici bila 0.01.

Rezultati klasifikatora su prikazani u TABELA III. Korišćena je *10-fold cross validation* i ukupan procenat pogođene klase.

TABELA III: Rezultati klasifikatora

Model sa podacima	Rezultati klasifikatora	
	Procenat pogođenih klasa	10-*fold cross validation*
SVM sa prvim skupom podataka	35.56%	0.29 (+/- 0.06)
SVM sa drugim skupom podataka bez zarade filma tokom prvog vikenda	32.17%	0.34 (+/- 0.07)
SVM sa drugim skupom podataka sa zaradom filma tokom prvog vikenda	59.57%	0.62 (+/- 0.09)
Naivni Bajes sa prvim skupom podataka	42.68%	0.41 (+/- 0.08)
Naivni Bajes sa drugim skupom podataka bez zarade filma tokom prvog vikenda	38.26%	0.43 (+/- 0.06)
Naivni Bajes sa drugim skupom podataka sa zaradom filma tokom prvog vikenda	57.83%	0.60 (+/- 0.06)
AdaBoost sa prvim skupom podataka	40.59%	0.41 (+/- 0.06)
AdaBoost sa drugim skupom podataka bez zarade filma tokom prvog vikenda	41.74%	0.40 (+/- 0.07)
AdaBoost sa drugim skupom podataka bez zarade filma tokom prvog vikenda	67.39%	0.63 (+/- 0.08)
Linearna regresija sa skupom prvim podataka	24.27%	0.57 (+/- 0.14)
Linearna regresija sa drugim skupom podataka bez zarade filma tokom prvog vikenda	29.13%	0.58 (+/- 0.13)
Linearna regresija sa drugim skupom podataka sa zaradom filma tokom prvog vikenda	51.74%	0.83 (+/- 0.13)
Neuronska mreža sa prvim skupom podataka	78.66%	Nije urađeno
Neuronska mreža sa drugim skupom podataka bez zarade filma tokom prvog vikenda	86.08%	Nije urađeno
Neuronska mreža sa drugim skupom podataka sa zarade filma tokom prvog vikenda	91.74%	Nije urađeno

Najbolje rezultate je davala neuronska mreža. Najslabije rezultate je davala linearna regresija. *AdaBoost* je davala dosta dobre rezultate, s obzirom da je stablo odlučivanja davalo slabije rezultate od SVM, pa nije uključeno u pregled rezultata. Naivni Bajes se pokazao boljim od SVM-a, što je interesantno, jer to govori o relativnoj međunezavisnosti podataka, pošto Naivni Bajes pretpostavlja nezavisnost podataka dok SVM traži nekakvu međuzavisnost.

U radu [1] posmatran je i procenat uspešnosti predviđanja kategorije i ako se dozvoli da se pogreši kategorija za 1 stepen. To bi imalo smisla dozvoliti i u ovom radu, jer su razlike između nekih kategorija relativno male. Procenti predviđanja sa dozvolom promašaja kategorije za 1 stepen su dati u TABELA IV.

TABELA IV: Procenat pogođenih kategorija sa dozvolom da se pogreši kategorija za 1 stepen.

Model sa skupom podataka	Procenat pogođenih klasa sa dozvolom da se pogreši kategorija za 1 stepen
SVM sa prvim skupom podataka	66.95%
SVM sa drugim skupom podataka bez zarade filma	71.74%
SVM sa drugim skupom podataka sa zaradom filma tokom prvog vikenda	95.22%
Naivni Bajes sa prvim skupom podataka	81.59%
Naivni Bajes sa drugim skupom podataka bez zarade filma tokom prvog vikenda	81.30%
Naivni Bajes sa drugim skupom podataka sa zaradom filma tokom prvog vikenda	93.48%
AdaBoost sa prvim skupom podataka	74.48%
AdaBoost sa drugim skupom podataka bez zarade filma tokom prvog vikenda	82.61%
AdaBoost sa drugim skupom podataka bez zarade filma tokom prvog vikenda	96.09%
Linearna regresija sa skupom prvim podataka	61.51%
Linearna regresija sa drugim skupom podataka bez zarade filma tokom prvog vikenda	81.30 %
Linearna regresija sa drugim skupom podataka sa zaradom filma tokom prvog vikenda	95.22%

Dobijeni procenti slični su sa dobijenim procentima u [1], gde je procenjivana uspešnost filma.

VI. Zaključak

U ovom radu prikazani su modeli za predviđanje zarade filma tokom vikenda koji uzima u obzir pored budžeta, ocena kritike i publike i koeficijent koji govori o tome koliko je koji film privlačan publici uzevši u obzir žanr gledajući pol publike.

Podaci su prikupljeni sa vodećih sajtova za ocenjivanje i analitiku filmova (videti poglavlje III).

Za predikciju žanra filma u odnosu na opis korišćene su tehnike logaritamske regresije, k-najbližih komšija i tehnike potpornih vektora. Modeli logaritamske regresije i potpornih vektora pokazali su slične performanse dok je model k-najbližih komšija bio značajno lošiji. U svakom modelu korišćenje tehnike *tf-idf* dovelo je do značajnog poboljšanja preciznosti. Na osnovu analize grešaka po filmu, dobija se u proseku ne više od 3 greške u izboru žanra za svaki model. Kako postoji korelacija između žanrova, odsustvo prepoznavanja žanra u nekim slučajevima neće značajno uticati na predikciju preferiranog pola.

Predviđanje sklonosti prema žanru u odnosu na pol realizovano je linearnom regresijom i ustanovljene su korelacije između žanrova. Srednja kvadratna greška bila je 0.93, dok je srednja vrednost odstupanje svih filmova iznosila 0.06 u opsegu od -1 do 1.

Kada se vrši predikcija sklonosti prema žanru i uzmu procenjene vrednosti žanrova, konačno apsolutno srednje odstupanje svih filmova kreće se od 0.29 za logističku regresiju sa tehnikom *tf-idf* do 0.34 za tehniku k-najbližih komšija u opsegu od -1 do 1(-1 ženski, - muški film).

Za određivanje mere kvaliteta predikcija regresijom su određene mere R^2 i 10-*fold cross validation*. Za određivanje mere kod klasifikatora korišćen je takođe 10-*fold cross validation*, kao i procenat pogođene kategorije zarade filma. Po uzoru na [1], gledan je i procenat pogođene kategorije zarade filma ako se dozvoli da model pogreši za 1 stepen kategorije.

Linearnom regresijom nad parametrima je za prvi vikend imala R^2 meru od 0.63, a za drugi vikend u najboljem slučaju je imala R^2 meru od 0.75.

Za prvi skup podataka u klasifikaciji najbolje se pokazala neuronska mreža, koja je imala 78.66% uspešnosti prepoznavanja klase. Najlošiji rezultat, ako ne uzmemo u obzir rezultat klasifikacije posle linearne regresije, je imao SVM, sa 35.56% uspešnosti.

Za drugi skup podataka, ako izuzmemo podatak o zaradi filma prvog vikenda, takođe se najuspešnijom pokazala neuronska mreža, sa 86.08% uspešnosti. Interesantno da je imala veću uspešnost u predikciji zarade drugog vikenda u odnosu na prvi vikend. Najlošiji rezultat, opet izuzimajuću klasifikaciju posle linearne regresije, imao SVM sa 32.17% uspešnosti. Samo su SVM i naivni Bajes imali pad uspešnosti u odnosu na prvi skup podataka.

Za drugi skup podataka, ako uključimo podatak o zaradi filma prvog vikenda, opet se najuspešnijim pokazala neuronska mreža sa 91.74%. Svi modeli su imali značajan porast uspešnosti prepoznavanja kategorije. Najlošiji rezultat je ostvario naivni Bajes, sa 57.83% uspešnosti.

Sa dozvolom da model greši za red veličine 1 stepen po kategoriji uspešnost prepoznavanja je znatno porastao. Sve metrike su imale preko 60% uspešnosti za prvi skup podataka, *adaBoost* i naivni Bajes su imali preko 80%. Za drugi skup podataka su svi modeli imali preko 80% uspešnosti prepoznavanja osim SVM-a, koji je imao 71.74% uspešnosti prepoznavanja. Za drugi skup podataka sa uključenim podatkom o zaradi prvog vikenda svi modeli imaju uspešnost preko 90%.

Modeli najviše greše kod bliskih kategorija, kao i u slučajevima velike zarade filmova, gde verovatno drugi faktori takođe utiču na zaradu, koji nisu obrađeni u ovom radu.

Planovi za dalje poboljšanje ovog rešenja obuhvataju:

- Dobavljanje dodatnih podataka, kao što su broj pregleda trejlera na *youtube.com* u prvih 24 sata, čime bismo mogli da merimo uticaj marketinga filma na zainteresovanost publike, broj taraba (#, eng. *hashtag*) na tviteru u trenutku objave trejlera iz istog razloga
- Pravljenjem metrike koja ocenjuje popularnost glumaca, režisera, produkcijske kuće koja pravi film.
- Pravljenjem metrike koja ocenjuje vrednost koliko je neki vikend potencijalno lukrativan za određeni žanr.

LITERATURA

[1] Quader, Nahid & Gani, Md & Chaki, Dipankar & Ali, Md. (2018). A Machine Learning Approach to Predict Movie Box-Office Success. 10.1109/ICCITECHN.2017.8281839.

[2] Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, *30*(2), 243-254..

[3] Cook, C., Cunningham, B., Reading, E., Sedgewick, M., & Tilcock, K. (2016). Predicting Blockbuster Success. *University of Victoria*..

[4] Hoerl, A.E. (1959). Optimum solution of many variables equations. *Chemical Engineering Progress*, **55** (11) 69-78.

[5] Model neuronske mreže [Online] link: https://medium.com/@rajatgupta310198/getting-started-with-neural-network-for-regression-and-tensorflow-58ad3bd75223 [Accessed: 3-May-2019].

[6] Tf-idf težine [Online] link: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html#sklearn.feature_extraction.text.TfidfTransformer [Accessed 3-May-2019].

[7] Ho, K. W. (2011). Movies' Genres Classification by Synopsis.

[8] Wühr, P., Lange, B. P., & Schwarz, S. (2017). Tears or fears? Comparing gender stereotypes about movie preferences to actual preferences. *Frontiers in psychology*, *8*, 428.

[9] Projekat sa temom predviđanja žanrova na osnovu opisa filma. [Online] link:https://github.com/ishmeetkohli/imdbGenreClassification [Accessed 3-May-2019].

[10] Hoang, Q. (2018). Predicting movie genres based on plot summaries. *arXiv preprint arXiv:1801.04813*.