Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/08BE38Ostojic

PREDIKCIJA ZARADE FILMOVA TOKOM VIKENDA

PREDICTING WEEKEND FILM BOX OFFICE

Miloš Ostojić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U radu je predstavljeno istraživanje o predviđanju zarade filmova tokom vikenda. Podaci su prikupljeni sa najrelevantijih sajtova koji se bave ocenjivanjem i analitikom zarade filmova i iz jednog naučnog rada. Zarada filma je predviđana na osnovu više atributa, jedan od njih je i koeficijent zainteresovanosti publike po polu za film na osnovu opisa zapleta filma. Za predviđanje koeficijenta zainteresovanosti publike korišćene su NLP tehnike, logistička regresija, SVM i k-NN. Problemu predikcije zarade filma se prišlo pomoću regresionih i klasifikacionih modela.*

Ključne reči: *Predviđanje zarade filma, regresija, klasifikacija*

Abstract – *This paper presents a study about predicting weekend film box office. The dataset that was made and described in the paper was taken from various relevant websites and from one scientific paper. Box office was predicted based on many attributes, one of which is an attribute that speaks of interest of public for a film based on their gender. That attribute was made when the synopsis of the film was run through NLP techniques, logistic regression, SVM and k-NN. The problem of predicting box office was approached with regression and classification models.*

Keywords: *Predicting film box office, regression, clasification*

1. 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 uzimaju najveći procenat od prodatih karti za filmove. 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. Rađeno je i predviđanje zainteresovanosti publike po polu za film na osnovu opisa zapleta filma.

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, način pripreme podataka za obučavanje modela i predstavljeni su modeli koji su korišćeni. U četvrtom poglavlju su prikazani i prokomentarisani rezultati primene modela. Poslednje poglavlje je sumiralo sadržaj ovog rada i u njemu su iznete za buduće unapređenje istraživanja teme rada.

2. PRETHODNA REŠENJA

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. Imali su 65% uspešnosti predviđanja klase zarade.

Hoang je u svom radu [4] 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 meru od 0.56 i 80.5% uspešnih pogađanja.

NAPOMENA:

Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, docent.

3. METODOLOGIJA I ALATI

3.1 Skup podataka

Inicijalni skup podataka je preuzet sa *Cinemascore*-ove internet stranice. On sadrži ocene filma od strane publike koja ga je gledala tokom prvog vikenda tog filma u bioskopima na teritoriji SAD-a. Potom je skup podataka obogaćen sa skupom podataka dobijenim pomoću OMDB API-a, koji preuzima podatke sa *rottentomatoes.com*, *imdb.com* i *metacritic.com* sajtova. Tom skupu podataka je potom dodat skup podataka preuzetih sa *boxofficemojo.com*, koji sadrži detaljne podatke o zaradi filmova. Preuzeti su podaci o filmovima koji su bili u bioskopima u periodu od 2000. do 2018. godine.

Konačni skup podataka je sadržao, posle izbacivanja podataka sa nepoznatim vrednostima, 15419 redova i sledeće podatke o filmovima: naziv filma, godina izlaska (eng. *wide release*, srp. pušten u većini bioskopa u SAD), tačan datum izlaska u SAD, opis zapleta filma (OMDB API ga preuzima sa *imdb.com*), *Cinemascore* ocena, ocena korisnika sajta *imdb.com*, *Metacritic* ocena, *Rotten Tomatoes* ocena kritike, žanrovi kojima film pripada, vreme trajanja filma u minutima, MPAA rejting film (sistem klasifikacije filmova u SAD na osnovu pogodnosti za uzrast), produkcijska kuća, režiser filma, glavni glumci u filmu, vikend u godini (u kom je prikazan film), godina u kojoj je posmatrani vikend za koji će se vršiti analiza, broj bioskopa u kojima je film bio prikazan za taj vikend, zarada filma za posmatrani vikend u američkim dolarima, ukupna zarada filma do posmatranog vikenda u američkim dolarima i produkcijski budžet u milionima američkih dolara.

Napravljen je atribut koji je govorio o rednom broju vikenda filma u bioskopima. Na osnovu navedenih atributa napravljeni je još nekoliko atributa koji su korišćeni za predikciju zarade filmova:

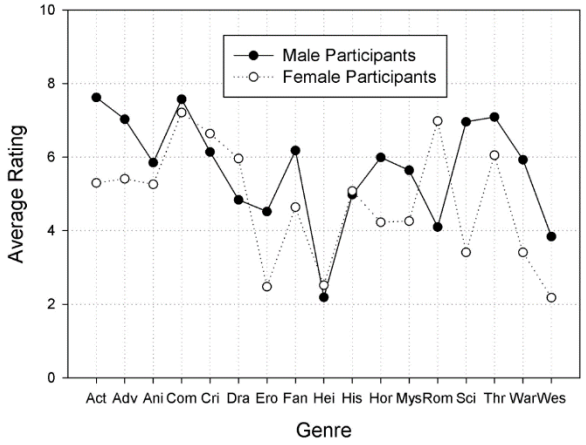
- Zarada filmskog studija: atribut je napravljen sabirajući zarade filmskih studija na filmovima u skupu podataka na teritoriji SAD.
- Zarada režisera: atribut je napravljen sabirajući zarade filmova koje je režirao režiser.
- Zarada glumaca: atribut je napravljen tako što su prvo sabirane zarade filmova koje je glumio glumac. Potom su za određeni film sabirane zarade glumaca koji su navedeni kao zvezde filma.
- Interesovanje publike za film: atribut je napravljen kombinacijom informacija o predviđenom žanru na osnovu kratkog opisa zapleta (detaljnije u potpoglavlju 3.2).
- Težina vikenda: atribut je dobijen na osnovu postavljenih upita o vikendu u kom se film prikazuje u bioskopima. Ako film u vikendu ima bar jedan film koji ima zaradu veću od 60 miliona dolara i sličnu zainteresovanost publike po polu i ako ima bar jedan film sa zaradom većom do 20 miliona dolara i sličnom zainterosavonšću po polu onda je težina vikenda kategorisana kao teška (vrednost 2). Ako ima više filmova koji su imali zaradu veću od 20 miliona dolara i sličnu zainteresovanost publike onda je težina vikenda kategorisana kao srednja (vrednost 1). U suprotnom je težina vikenda klasifikovana kao laka (vrednost 0).

Iz napravljenog skupa podataka izdvojeni su podaci koji govore o zaradi filmova tokom prvog vikenda u bioskopima. Pripremajući podatke za klasifikaciju dodato je jedno obeležje koja govore o klasi zarade. To obeležje predstavlja klasifikaciju na 8 klasa. Opsezi za klase po zaradi 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 i 8. od 50 miliona dolara.

3.2 Metode određivanja zainteresovanosti publike za film na osnovu opisa zapleta i žanra filma

Rad [6] 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 1. Ocene privlačnosti žanrova po polu [6]

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* (za muškarce 6, za žene 6), *Documentary* (5, 5), *Family* (5, 6.3), *Music* (4, 5.5), *Musical* (4, 5.5) i *Sport* (5.1, 5).

Predviđanje žanra na osnovu opisa zapleta filma spada u probleme analize sentimenta. Korišćeni su sledeći modeli:

- Logistička regresija
- *k*-NN
- SVM

Performanse modela su unapređene još sa tehnikom relativne frekvencije izraza (eng. *term frequency–inverse document frequency*, tf-idf), prilikom obrade opisa zapleta filma.

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 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 svesti na predviđanje pojedinačnih žanrova na osnovu opisa radnje (eng. *one-versus-all*). Ovaj pristup je korišćen kod svih modela.

Na osnovu udela žanrova se može videti da su podaci nebalansirani. Ovo je ipak ostavljeno prilikom obuke, jer bi na ovaj način mreža trebala da odbaci žanr, osim ako ne postoje velike indicije za njega..

U svakom od prethodnih 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.

Labela prihvaćenosti filma u odnosu na pol nalazi se u opsegu od -1 do 1, gde -1 označava veće interesovanje ženske publike dok 1 označava veće interesovanje muške publike.

3.3 Metode predviđanja zarada filmova

Postojala su 2 pristupa prilikom predikcije zarade filmova. Prvi pristup je bio da se problemu predikcije pristupi pomoću regresionih modela. Korišćeni su regresioni modeli linearne regresije i *Random forest*-a. Drugi pristup je bio da se problem predikcije reši pomoću klasifikacionih modela. Korišćeni klasifikacioni modeli u ovom radu su bili modeli SVM-a, naivnog Bajesa, veštačke neuronske mreže, *AdaBoost*-a, *Random forest*-a, *Gradient boosting*-a, *k*-NN-a i linearne regresije, čiji su rezultati regresionog modela iskorišćeni za klasifikaciju. Korišćena je *Scikit* biblioteka u *Python*-u za primenu modela.

4. REZULTATI

4.1 Rezultati prepoznavanja žanra i zainteresovanosti publike za film na osnovu njegovog opisa zapleta

Određivanje naklonosti publike u odnosu na žanrove ima R2 meru 0.93. Prosečno odstupanje dobijenog rezultata od labele je 0.065, dok je medijan odstupanja 0.08.

Kao metrika uspešnosti predviđanja korišćeni su preciznost (eng. *precision*), opoziv (eng. *recall*) i F1 mera, kako bi se prikazala realna slika klasifikatora, imajući u vidu veliku nebalansiranost skupa podataka. Tačnost u ovom slučaju je velika 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 skupu podataka.

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

Na osnovu rezultata može se zaključiti da modeli logističke regresije i SVM-a podjednako dobro klasifikuju, dok model k-NN-a daje značajnije lošije rezultate. U žanrovima sa većom zastupljenošću tehnika *tf-idf* u velikom broju slučajeva pozitivno deluje na predviđanje, što je bilo očekivano.

4.2 Rezultati predikcije zarade filmova pomoću regresionih modela

Korišćeni su modeli linearne regresije i *Random forest*-a. Skup podataka je podeljen u razmeri 4:1 na skup podataka za trening i test. Za određivanje uspešnosti predikcije korišćena je R^2 metrika. Takođe je korišćena i unakrsna validacija, da bi se videlo da li postoji određeno odstupanje ako bi se drugačije napravio skup trening i test podataka.

Tabela 1. Rezutati regresionih modela

Model	R^2	10-*fold cross validation*
Linearna regresija	0.59	0.58 (+/- 0.17)
Random forest	0.76	0.72 (+/- 0.27)

Na osnovu tabele se vidi da *Random forest* pravi bolju predikciju zarade filmova od linearne regresije. Rad [5] koji se bavio predviđanjem ukupne zarade filmova za južnokorejsko tržište je dobio sličnu vrednost za R^2 za model linearne regresije, 0.588

4.3 Rezultati predikcije zarade filmova pomoću klasifikacionih modela

Rezultati klasifikatora su prikazani u tabeli 2. Korišćena je 10-*fold cross validation* i ukupan procenat pogođene klase. Takođe su posmatrani preciznost, opoziv, F1 mera i konfuzione matrice, radi boljeg uvida u rezultate klasifikacije.

Tabela 2. Rezultati klasifikacionih modela

Modeli	Procenat pogođenih klasa	10-*fold cross validation*
SVM	40.00%	0.39 (+/- 0.10)
Naivni Bajes	38.26%	0.38 (+/- 0.06)
k-NN	33.04%	0.35 (+/- 0.10)
Linearna regresija	33.91%	0.58 (+/- 0.17)
AdaBoost	40.87%	0.43 (+/- 0.09)
Random forest	42.61%	0.45 (+/- 0.10)
Gradient boosting	40.43%	0.42 (+/- 0.07)
Veštačka neuronska mreža	44.35%	0.42 (+/- 0.09)

Najuspešnijim od svih klasifikatora po procentu pogođenosti pripadnosti klasi zarade za dati skup podataka je model veštačke neuronske mreže, sa 44.35% uspešnosti pogađanja klasa za skup podataka. Model veštačke neuronske mreže se pokazao boljim od sledećeg najuspešnijeg modela, *Random forest* modela, za 1.74%. Najmanje uspešni modeli su bili *k*-NN sa 33.04% pogođenih klasa i linearna regresija sa 32.61% pogođenih klasa.

Poredivši F1 mere za svaku pojedinačnu klasu predviđanja najuspešnijih klasifikatorskih modela za skup podataka za prvi vikend može se zaključiti da model veštačke neuronske mreže uspešnije prepoznaje dve „ekstremne" (1. i 8. klasa), kao i u slučaju najbrojnije, 3.

klase. Predviđajući 3. klasu model veštačke neuronske mreže je oslabio F1 mere za predviđanje 4. i 5. klase. Model *Random forest* je konstantniji u predviđanju klasa, tj. nema izuzetno nisku F1 meru prilikom predviđanja klasa (za razliku od veštačke neuronske mreže). Zanimljivo da je kod oba modela 5. klasa (zarade od 15 miliona dolara do 20 miliona dolara) imala najmanju F1 meru, kod veštačke neuronske mreže je imala vrednost 0.08, kod *Random forest*-a je imala 0.23. Kod većine modela predviđanje 5. klase rezultovalo najmanjom F1 merom, sa izuzetkom naivnog Bajesa, kome je predviđanje 4. klase bilo najveći problem, i SVM-a, koji je imao vrednost 0 za F1 meru za 2. i 5. klasu, svrstajaću ih u 3. i 4. klasu, respektivno.

5. ZAKLJUČAK

U ovom radu prikazani su modeli za predviđanje zarade filma u bioskopima na teritoriji SAD tokom prvog vikenda. Predviđana zarada su vršena na osnovu više obeležja. Obeležja koja nisu korišćena ili su drugačije konstruisana u radovi su obeležja zarade glumca, režisera i produkcijske kuće, težina vikenda i zainteresovanost publike po polu u odnosu na žanr filma. Podaci su prikupljeni sa vodećih sajtova za ocenjivanje i analitiku filmova. Prvo je rađena predikcija žanra u odnosu na opis zapleta filma. Ideja je bila da se zaključi zainteresovanost publike po polu za određeni film na osnovu par rečenica o filmu. Potom su primenjeni regresioni i klasifikatorski modeli za predikciju zarade filma.

Za predikciju žanra filma u odnosu na opis zapleta filma korišćeni su modeli logističke regresije, *k*-NN i SVM. Modeli logističke regresije i SVM-a pokazali su slične performanse dok je model *k*-NN bio značajno lošiji. U svakom modelu korišćenje tehnike *tf-idf* dovelo je do značajnog poboljšanja preciznosti. Kako postoji korelacija između žanrova, odsustvo prepoznavanja žanra u nekim slučajevima nije značajno uticalo na predikciju preferencija filma u odnosu na pol.

Od regresionih modela korišćena je linearna regresija i *Random forest* ansambl. Model linearne regresije je nad skupom podataka za prvi vikend imao R^2 vrednost od 0.59. Model *Random forest*-a je ostvario veću R^2 vrednost modela linearne regresije sa vrednošću od 0.76 za prvi vikend.

Od klasifikatorskih modela korišćeni su SVM, naivni Bajes, veštačka neuronska mreža, linearna regresija, *AdaBoost*, *Random forest*, *Gradient boosting* i k-NN. Zarade filmova su klasifikovane na 8 klasa..

Za skup podataka za prvi vikend kada je rađena klasifikacija na 8 klasa najuspešnijim u predviđanju se pokazao model veštačke neuronske mreže sa 44.35% tačnih predviđanja klasa. Najslabiji klasifikatori prilikom klasifikacije na 8 klasa su bili modeli k-NN-a i linearne regresije.

Ovaj rad bi se mogao unaprediti dodavanjem atributa koji bi govorili o aktivnosti na društvenim mrežama, npr. broj pregleda trejlera filma na Youtube-u u prvih 24 sata, ili aktivnost na Twitter-u u vreme objave trejlera i u vreme objave kritika na internetu. Time bi se dobili podatke koji bi govorili o zainteresovanosti publike za film, i moglo bi se pratiti koliko se ti komentari preneli na zaradu u bioskopu. Takođe bi se mogao pratiti uticaj pozitivnih i negativnih komentara na performans filma u bioskopima. Veliko proširenje ovom radu bi bilo kada bi bili dostupni podaci MPAA-a. Ta organizacija je jedan od glavnih autoriteta za podatke o zaradama filmova, ali nema ih dostupne za javnost. Svakog vikenda puštaju deo podataka koji su prikupili tokom njega, u kome imaju podatke o publici koja je videla film, počevši od broja prodatih karata, do podataka o rasnom, starosnom i polnom sastavu publike.

6. LITERATURA

[1] Quader, N., Gani, M. O., Chaki, D., & Ali, M. H. (2017, December). A machine learning approach to predict movie box-office success. In *2017 20th International Conference of Computer and Information Technology (ICCIT)* (pp. 1-7). IEEE.

[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. Predicting Blockbuster Success.

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

[5] Song, J., & Han, S. (2013). Predicting gross box office revenue for domestic films. *Communications for Statistical Applications and Methods*, *20*(4), 301-309.

[6] 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.

Kratka biografija:

Miloš Ostojić rođen je u Novom Sadu 1995. god. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva – Računarstvo i automatika odbranio je 2019. godine.
kontakt: milos.ostojic@uns.ac.rs