# Predviđanje filmskih rejtinga

Projekat u okviru kursa Mašinsko učenje

Nenad Perišić
perisicnenad96@gmail.com

31. avgust 2021.


Matematički fakultet
u Beogradu

# Sadržaj

# 1 Uvod

U ovom radu ću pokušati da dam odgovor na pitanje da li je moguće predvideti rejting filma i proceniti njegovu uspešnost pre nego što on bude pušten u bioskopima. Ovo ću uraditi analiziranjem skupova multimodalnih podataka sa raznim atributima (svojstvima) kao što su režiser tog filma, glumačka ekipa, produkcijska kuća, opis filma, žanr filma, poster filma, budžet, vreme trajanja filma itd. Fokus će biti predviđanje rejtinga na osnovu tekstualnih atributa i primene algoritama linearne regresije, grebene regresije, slučajnih šuma na njima.

# 2 Skup podataka

Koriste se dva skupa podataka, "The movie dataset"[1] i "Movie genre from its poster dataset"[2], oba dostupni na Kaggle [3] veb sajtu. Oni sadrže metapodatke za 45000 filmova koji su objavljeni pre jula 2017. godine. Podaci se sastoje od glumačke ekipe (cast), celokupne ekipe (crew), ključnih reči radnje, budžeta, prihoda, postera, datuma objavljivanja, jezika, produkcijskih kuća, država, broja glasova i njihovog proseka.

Skup podataka je organizovan na sledeći način:

- *movies_metadata.csv* - glavna datoteka metapodataka o filmovima
- *keywords.csv* - sadrži ključne reči filmskog zapleta dostupog u formi JSON objekta
- *credits.csv* - sastoji se od glumačke ekipe takođe u formi JSON objekta
- *links.csv* - sadrži TMDB i IMDB id-ove svih filmova
- *ratings_small.csv* - sadrži rejtinge od 100 000 rejtinga koji su prikupljeni od 700 korisnika na 9000 filmova
- *movie_genre.csv* - sadrži žanrove filmova koje ćemo spajati preko imdb_id-a, a koja takođe sadrži i link do postera filma

# 3 Preprocesiranje

Nakon spajanja svih navedenih tabela, uklonićemo kolone koje se ponavljaju, a to su: *Title*, *Original title*. Zatim uklanjamo još i kolone *Poster* - link do postera filma, *homepage* - link do stranice filma, *belongs_to_collection*.

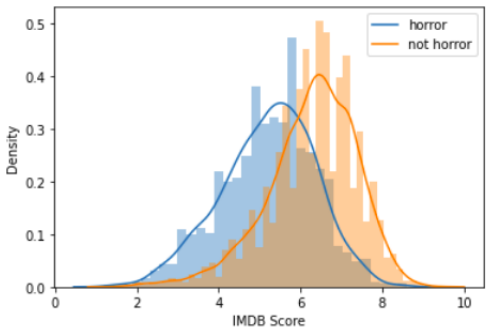
Za ulazne parametre posmatraćemo atribute cast, crew, synopses, genre i runtime, a IMDB Score ćemo uzeti za cilj predvidjanja. Iz kolona cast i crew ćemo da izdvojimo dva glavna glumca, režisera i prve dve produkcijske kuće za svaki film ali tako da su se pojavljivali u barem 5 filmova kako nam se ne bi narušilo predviđanje ukoliko je neko imao samo jedan film u svojoj karijeri. Takođe, zarad boljih rezultata možemo uzeti samo filmove koji su prvi put prikazani posle 1970. godine i čiji je originalni jezik engleski.

Nakon izdvajanja potrebnih atributa iz JSON formata uradićemo binarizaciju nad atributima *Genre*, *actors* i *companies* koje smo izdvojili. Što se tiče kolone *overview* (synopses) koristićemo biblioteku *spacy* [4] za tokenizaciju i ostavićemo samo reči koje se javljaju u bar 20 filmova.

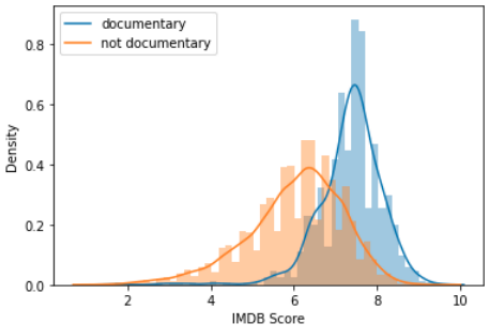
Podela podataka na trening i test skup je izvršena u odnosu 70%:30%.

# 4 Vizuelizacija podataka

U ovoj sekciji će na grafički način biti prikazana zavisnost budžeta, vremena trajanja filma i raznih žanrova od rejtinga filma. Vizuelizacija je urađena nakon preprocesiranja i izdvajanja određenih atributa iz JSON formata. Na slici 1 se može videti da filmovi čiji je žanr *horror* imaju nešto manji rejting od onih koji imaju drugi žanr. Na slici 2 možemo videti da filmovi čiji je žanr *documentary* u znatno većem broju imaju viši rejting. Dakle kolonu *genre* treba uzeti u razmatranje prilikom treniranja modela.

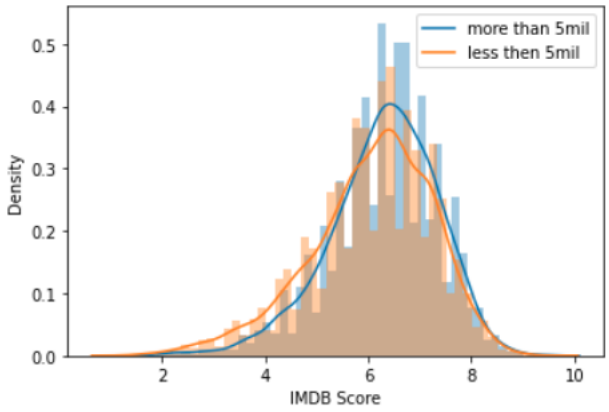



Slika 1: Genre: Horror




Slika 2: Genre: documentary

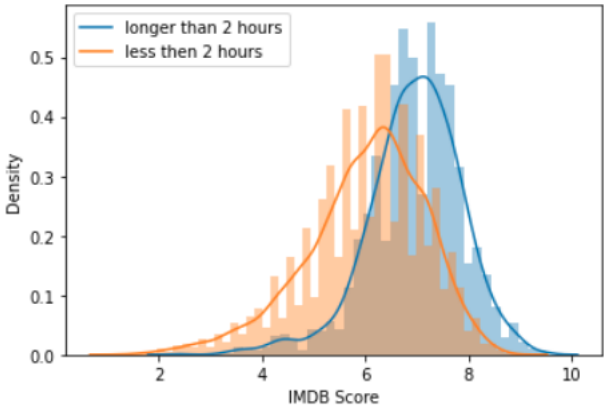
Pokušajmo sada isto sa budžetom. Na slici 3 se može zaključiti da ima nešto više filmova koji imaju budžet preko 5 miliona dolara, ali da njihov rejting nije nužno viši u odnosu na filmove sa manjim budžetom.




Slika 3: Budget

Na kraju možemo da pogledamo još i dužinu trajanja filma u odnosu na rejting. Na slici 4 možemo videti da filmovi koji traju duže od 2 sata imaju bolji rejting u odnosu na one čija je

dužina kraća od 2 sata pa ćemo i tu kolonu kasnije uzeti u razmatranje za treniranje modela.




Slika 4: Runtime

# 5 Modeli

U ovoj sekciji ćemo kreirati regresione modele nad preprocesiranim podacima i evaluirati pomoću kvadratne greške ($R^2$), srednje apsolutne (**MAE**) i srednje kvadratne greške (**MSE**). Modeli koji ce biti trenirani nad podacima su:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Random Forest Regression
- Extreme Gradient Boosting

Kao što smo mogli da vidimo na graficima bitni atributi koji će biti od značaja za treniranje modela su: *Genre, actors, director, companies, overview*, a ciljna promenljiva je *IMDB Score*. Nakon završenog preprocesiranja i podele skupa na trening i test možemo da krenemo sa treniranjem modela nad tim podacima.

## 5.1 Linearna regresija

Rezultati linearne regresijena trening i test skupu sa *default* parametrima se mogu videti u tabeli 1.

Tabela 1: Errors

| | $R^2$ | MAE | MSE |
|---|---|---|---|
| train | 0.5469532634174328 | 0.6031675024259086 | 0.6228400675303715 |
| test | 0.3615194589066031 | 0.7248295216204385 | 0.874435134094227 |

Nakon toga, da bismo dobili što bolje vrednosti za greške, napravićemo nekoliko modela sa različitim parametrima korišćenjem GridSearchCV metoda[5].

## 5.2 Grebena regresija

## 5.3 Lasso regresija

## 5.4 Slučajne šume

## 5.5 Extreme gradient boosting

# 6 Zaključak

zakljucak

## Literatura


[1] "Movie dataset." on line at: https://www.kaggle.com/rounakbanik/the-movies-dataset.

[2] "Movie genres." on line at: https://www.kaggle.com/neha1703/movie-genre-from-its-poster.

[3] "Kaggle." on line at: https://www.kaggle.com/.

[4] "Spacy." on line at: https://spacy.io/.

[5] "Gridsearchcv." on line at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.