Predviđanje filmskih rejtinga

Projekat u okviru kursa Mašinsko učenje

Nenad Perišić
perisicnenad96@gmail.com

2. septembar 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*.

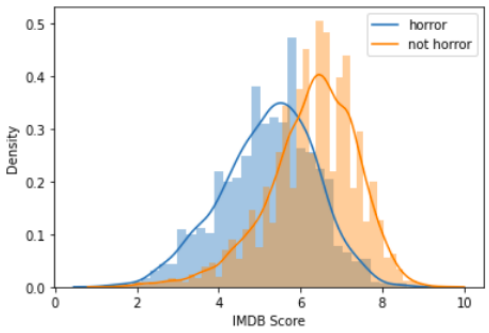
Za ulazne parametre posmatraćemo atribute cast, crew, synopses, genre i runtime, a IMDB Score ćemo uzeti za cilj predvidjanja. Iz kolona cast i crew ćemo da izdvojimo dva glavna glumca, režisera i prve dve produkcijske kuće za svaki film ali tako da su se pojavljivali u barem 5 filmova kako nam se ne 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.

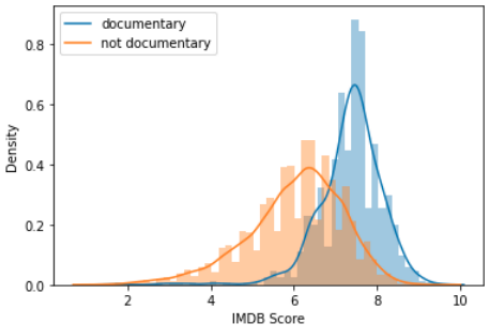
Podela podataka na trening i test skup je izvršena u odnosu 70%:30%.

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.

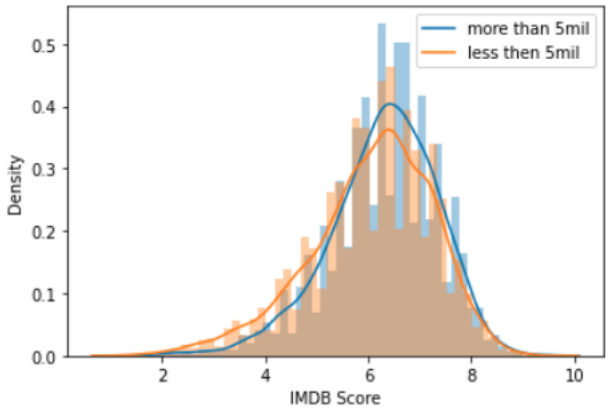


Slika 1: Genre: Horror



Slika 2: Genre: documentary

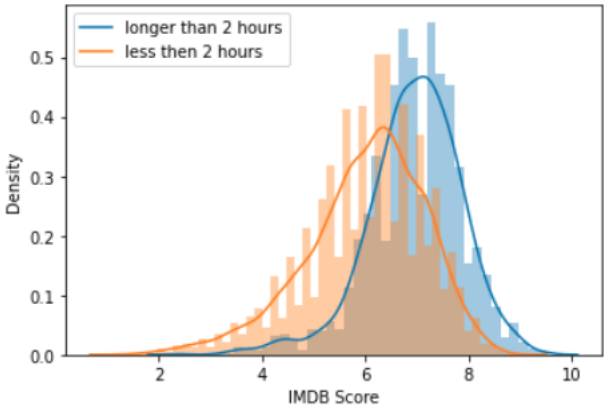
Pokušajmo sada isto sa budžetom. Na slici 3 se može zaključiti da ima nešto više filmova koji imaju budžet preko 5 miliona dolara, ali da njihov rejting nije nužno viši u odnosu na filmove sa manjim budžetom.



Slika 3: Budget

Na kraju možemo da pogledamo još i dužinu trajanja filma u odnosu na rejting. Na slici 7 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 regresije na 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

Pošto rezultati nisu savršeni, isprobaćemo više modela sa različitim parametrima da vidimo kako se menjaju vrednosti grešaka modela. da bismo dobili što bolje rezultate. Ovo ćemo uraditi korišćenjem GridSearchCV metoda[5]. Parametri koji su korišćeni su *fit_ intercept: True, False, normalize: True, False*, *scoring=r2* i za parametar cv je uzeta vrednost 5. Dobijene vrednosti su: *best_params: fit_ intercept: True, 'normalize': False*. Kada ponovo treniramo model sa novim parametrima, rezultati koje dobijamo se mogu videti u tabeli 2.

Tabela 2: Errors

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

Dakle, može se zaključiti da su rezultati isti kao sa modelom koji je treniran sa *default* parametrima.

5.2 Grebena regresija

Rezultati grebene regresije na trening i test skupu sa parametrima *alpha=10, normalize=True* se mogu videti u tabeli 3.

Tabela 3: Errors

	R^2	MAE	MSE
train	0.14921849413953459	0.8430023954492178	1.169638290656023
test	0.12596594545080564	0.8545487903574828	1.197038964388495

Rezultati nisu dobri, greške odstupaju puno od svojih idealnih vrednosti, zbog toga probajmo sada GridSearchCV metod. Uzećemo sledeće vrednosti parametara: *'alpha':[1,5,10,30], 'fit_ intercept':[True,False], 'normalize':[True,False]* i pustiti algoritam da radi. Najbolje dobijene vrednosti su: *'alpha': 1, 'fit_ intercept': True, 'normalize': True*. Kada ponovo treniramo model sa novim parametrima, rezultati koje dobijamo se mogu videti u tabeli 4.

Tabela 4: Errors

	R^2	MAE	MSE
train	0.4471618617605746	0.6677622829582377	0.7600313953296844
test	0.35788726526000325	0.7228114470349141	0.879409628278469

Vidimo da su rezultati dosta bolji sa parametrima koje je GridSearchCV pronašao, ali i dalje nedovoljno dobro. Nastavimo dalje.

5.3 Lasso regresija

Rezultati grebene regresije na trening i test skupu sa parametrom *alpha=0.01* se mogu videti u tabeli 5.

Sa GridSearchCV metodom za parametara: *'alpha':[0.1, 0.01, 0.5], 'fit_ interceptTrue,False], 'normalize':[True,False], 'selection':['cyclic', 'random']* se ne dobija nikakvo poboljšanje.

Tabela 5: Errors

	R^2	MAE	MSE
train	0.3067452704795627	0.7552564534961796	0.9530734639152751
test	0.304028552361731	0.752210485497123	0.953172175144332

5.4 Slučajne šume

Probajmo sa slučajnim šumama da li možemo da dobijemo bolje rezultate od dosadašnjih. Pokrenućemo algoritam za dubinu 32. Na osnovu rezultata u tabeli 6 zaključujemo da je algoritam slučajnih šuma do sada doe najbolje rezultate.

Tabela 6: Errors

	R^2	MAE	MSE
train	0.7741702821868292	0.4291284543398816	0.3104664162336058
test	0.3865944458760707	0.6995345386544246	0.8400935243162557

Nakon provere najboljih parametara sa GridSearchCV metodom, dobijeni rezultati su se neznatno poboljšali kao što se može videti u tabeli 7.

Tabela 7: Errors

	R^2	MAE	MSE
train	0.5893342000209669	0.5757413356809938	0.5645755590709117
test	0.4007266263235302	0.6939344630588781	0.8207387056346158

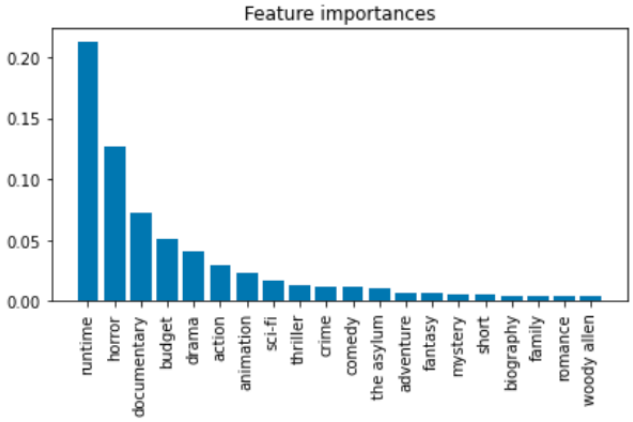
5.5 Extreme gradient boosting

Na kraju, probaćemo još i *gradient boosting* algoritam. Dobijeni rezultati u tabeli 8 su najbolji od svih modela koje smo do sada probali.

Tabela 8: Errors

	R^2	MAE	MSE
train	0.5674401301819887	0.5913678678131937	0.5946751113596823
test	0.4355463606452872	0.6716326508373056	0.7730511144064957

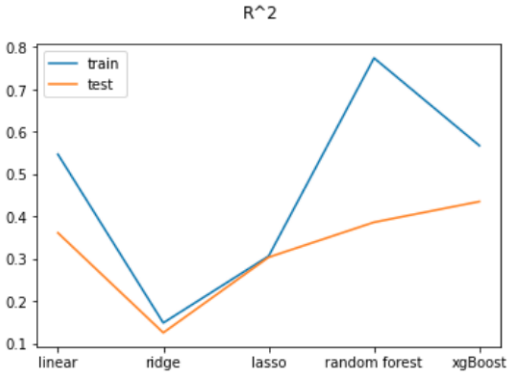
Kada smo već kod ovog algoritma, možemo napraviti grafik bitnosti atributa za predikciju rejtinga. Na slici 5 se može videti da dužina filma, i žanr igraju veliku ulogu u oceni tog filma.



Slika 5: Feature importance

5.6 Vizuelno poređenje modela

Sada ćemo vizuelno uporediti dobijene rezultate od svih modela, za svaku grešku posebno.



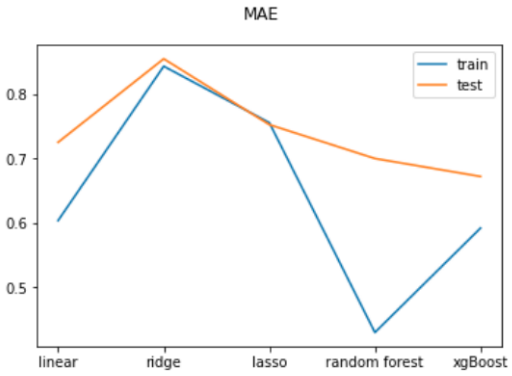
Slika 6: MAE errors

Vidimo da xgBoost daje najbolje rezultate, a da ridge i lasso regresije najbolje generalizuju podatke.

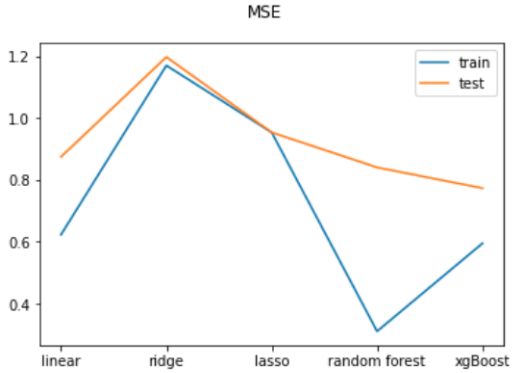
6 Ponovno preprocesiranje podataka

U cilju dobijanja koeficijenta determinacije što bliže vrednosti 1, a srednje absolutne i srednje kvadratne greske što bliže 0, pokušaćemo dodatno da obradimo ulazne podatke i da vidimo da li ćemo u tom slučaju dobiti bolje rezultate. Ideje za unapređenje su sledeće:

- iz svakog filma ćemo uzeti pvih 6 glumaca, potom od njih izdvojiti 200 najfrekventnijih glumaca - onih koji se javljaju u najvećem broju filmova, izbaciti filmove gde oni ne glume, a potom primeniti binarizaciju



Slika 7: MAE errors



Slika 8: MSE errors

- da se poboljša tokenizacija opisa filmova (*synopses*) tako što ćemo koristeći nltk biblioteku izbaciti stop reči, jako učestale reči i specijalne karaktere kako bismo smanjili broj kolona u podacima na kojima ćemo trenirati modele
- standardizacija atributa *runtime* i *budget*

7 Modeli sa poboljšanim preprocesiranjem

Nakon izvršenog ponovnog preprocesiranja, može se primetiti da se broj atributa značajno smanjio, sa 4993 na 1959. Istrenirajmo sada ponovo modele i pogledajmo da li će rezultati biti bolji ili ne.

7.1 Linearna regresija

Dobijeni rezultati linearne regresije se mogu videti u tabeli 9.
Nakon puštanja GridSearchCV algoritma, nisu dobijeni ništa bolji rezultati.

Tabela 9: Errors

	R^2	MAE	MSE
train	0.6975373878682696	0.4414965886950981	0.325694684152396
test	0.08361906453769419	0.8171506801963895	1.0701402184640283

7.2 Grebena regresija

Rezultati grebene regresije se mogu videti u tabeli 10. Sa GridSearchCV metodom, rezultati su malo bolji (tabela 11).

Tabela 10: Errors

	R^2	MAE	MSE
train	0.19174609984964475	0.7217692891909088	0.8703356651888018
test	0.12165415806380964	0.7268126983311105	0.8674203342644023

Tabela 11: Errors

	R^2	MAE	MSE
train	0.5430554470969404	0.5358217404979934	0.4920423413129251
test	0.2672915613818133	0.6653916126978011	0.7235944754329561

7.3 Lasso regresija

Dobijeni rezultati lasso regresije se mogu videti u tabeli 12.

Tabela 12: Errors

	R^2	MAE	MSE
train	0.22539566709947767	0.7071151600038588	0.8341014837140812
test	0.23082228287647877	0.6769756376029746	0.7596101223924109

7.4 Slučajne šume

Dobijeni rezultati se mogu videti u tabeli 13.

Tabela 13: Errors

	R^2	MAE	MSE
train	0.8434910739949135	0.3320578907345683	0.16853033458580594
test	0.3480215435800472	0.6226180725231611	0.6438686717686649

Slični rezultati su dobijeni i sa GridSearchCV metodom.

7.5 Extreme gradient boosting

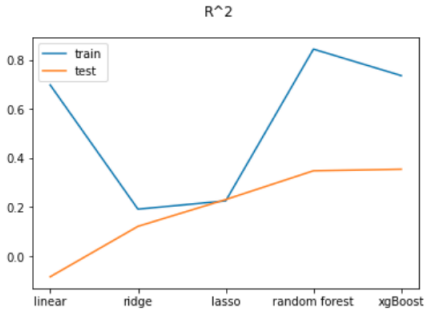
Dobijeni rezultati se mogu videti u tabeli 14.

Tabela 14: Errors

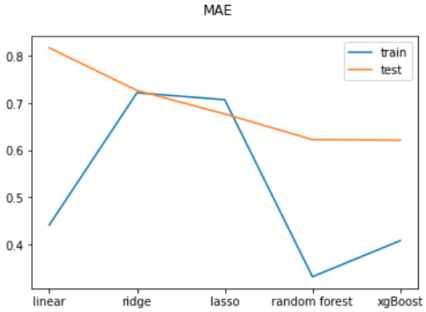
	R^2	MAE	MSE
train	0.7356594644033094	0.40847821777075716	0.2846444611552286
test	0.35436000629533615	0.6213208720406557	0.6376090514861819

7.6 Vizuelno poređenje modela

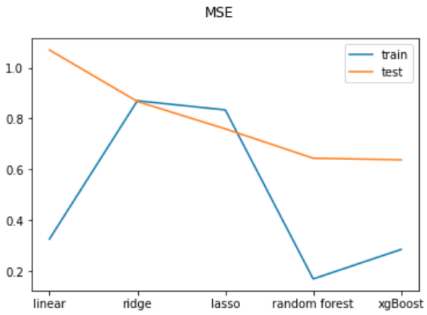
Sada ćemo vizuelno uporediti dobijene rezultate od svih modela, za svaku grešku posebno.



Slika 9: R^2 errors



Slika 10: MAE errors



Slika 11: MSE errors

Vidimo da xgBoost daje najbolje rezultate, a da ridge i lasso regresije najbolje generalizuju podatke.

8 Zaključak

Isprobano je dosta modela nad različitim skupom podataka. Može se zaključiti da je najbolje rezultate dao model dobijen xgBoost regresijom, a da najbolju generalizaciju vrši lasso. Što se tiče atributa možemo da izvedemo zaključak da će zanr imati dosta uticaja na samu ocenu filma, na primer vidimo da ukoliko je film horor, dokumentarac, drama ili akcija će imati veću ocenu od ostalih. Dužina trajanja filma se takođe pokazala kao bitna stavka, dok se budžet pokazao kao manje bitan u odnosu na ova dva parametra.

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.