Predviđanje filmskih rejtinga

Projekat u okviru kursa Mašinsko učenje

Nenad Perišić
perisicnenad96@gmail.com

26. avgust 2021.

Matematički fakultet
u Beogradu

Sadržaj

1 Uvod

U ovom radu ćemo pokušati da damo odgovor na pitanje da li je moguće predvideti rejting filma i proceniti njegovu uspešnost pre nego što on bude pušten u bioskopima. Ovo ćemo uraditi analiziranjem skupova multimodalnih podataka sa raznim atributima (svojstvima) kao što su režiser tog filma, glumačka ekipa, kompanija koja je radila produkciju filma, opis filma, žanr filma, poster filma itd. Mi ćemo se osvrnuti na predviđanje rejtinga na osnovu tekstualnih atributa i na njima primenjivati algoritme linearne regresije, grebene regresije, slučajne šume i neuronske mreže.

2 Skup podataka

Koristićemo dva skupa podataka, "The movie dataset"[?] i "Movie genre from it's poster dataset"[?], oba dostupni na Kaggle [?] 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

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

Podelu podataka na trening i test skup ćemo izvršiti u odnosu 70%:30%.

4 Modeli

5 Zaključak

zakljucak