Analiza uticaja parametara serija sa Hulu platforme na ocenu gledalaca

Damjan Vinčić SV58/2022
Decembar, 2023

Opis problema

U današnjem vremenu, raznolikost serija na streaming platformama je ogromna, a gledaoci često biraju serije na osnovu ocena i preporuka. Naš cilj je da istražimo faktore koji doprinose visokim ili niskim ocenama serija na Hulu platformi. Kroz analizu, hoćemo da razumemo kako žanr, trajanje, broj sezona i ostali faktori utiču na ocene gledalaca.

Metodologija

Eksplorativnom analizom podataka istražićemo distribuciju serija po žanrovima, dužini trajanja i broju sezona kroz tabele, historgram i druge vizualizacije. Ciljna labela nam je ocena serije. Hoćemo da vidimo šta utiče na nju. Da li žanr utiče na ocenu? Ili možda broj sezona, trajanje epizode… Možemo napraviti više modela i uporediti ih, naći neku seriju koja nije u skupu podataka i proveriti koju bi ocenu imala. Podelićemo podatke na train i test setove u odnosu 80/20. Pre primene regresionih modela, sprovešćemo preprocesiranje podataka, što može uključivati:

- **Popunjavanje nedostajućih vrednosti** – Ukoliko postoje nedostajuće vrednosti, razmotrićemo strategije za njihovo popunjavanje, na primer upotrebom prosečnih vrednosti, splajna, ili nekom drugom tehnikom
- **Kodiranje kategoričkih varijabli** – Pretvaranje kategoričkih varijabli u numeričke oblike (One-Hot encoding, Label encoding…)

Za regresionu analizu, možemo primeniti različite modele kao što su:

- **Linearna regresija** – Pronalazi linearnu vezu između nezavisnih i zavisnih promenljivih.
- **Random Forest regresija** – Kombinuje više stabala odlučivanja kako bi poboljšala predikcije.
- **Gradient Boosting regresija** – Gradi niz slabih modela kako bi stvorila bolji model.

- Skup Podataka

Koristićemo set podataka sa sajta Kaggle koji sadrži potake o serijama sa Hulu platforme. Opis nekih kolona:

- **Show Name** – Ime serije
- **Genre(s)** – Žanr kome serija pripada
- **Run Time** – Dužina u minutima za svaku epizodu ili prosečno trajanje svih epizoda
- **Number of seasons** – Ukupan broj sezona
- **Rating** – Prosečna ocena serije, između 0 i 10

Link do seta podataka: https://www.kaggle.com/datasets/thedevastator/hulu-popular-shows-dataset

- Način evaluacije

Za procenu performansi modela regresije možemo koristiti neke od metoda:

- **R2 (R-squared)**
- **RMSE (Root Mean Squared Error)** – Meri prosečnu grešku između stvarnih i predviđenih vrednosti.
- **MAE (Mean Absolute Error)** – Meri prosečnu apsolutnu razliku između stvarnih i predviđenih vrednosti.

- Tehnologije

Koristićemo programski jezik Python, i biblioteke poput Pandas, Matplotlib, Numpy, Statsmodels.

- Literatura

- https://towardsdatascience.com/random-forest-regression-5f605132d19d
- https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502
- https://stats.oarc.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis/
- https://www.educative.io/blog/one-hot-encoding
- https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/