MEDIAGENIX

MEDIAGENIX

MEDIAGENIX

DATA SCIENCE ASSIGNMENT

Data

IMDb is the most popular movie website and it combines movie plot description, Metascore ratings, critic and user ratings and reviews, release dates, and many more aspects. The website is well known for storing almost every movie that has ever been released (the oldest is from 1874 - "Passage de Venus") or just planned to be released (newest movie is from 2027 - "Avatar 5"). IMDb stores information related to more than 6 million titles (of which almost 500,000 are feature films) and it is owned by Amazon since 1998.

Datasets:

- The movies dataset includes 81,273 movies with attributes such as movie description, average rating, number of votes, genre, etc.
- The ratings dataset includes 81,273 rating details from demographic perspective.
- The names dataset includes 175,719 cast members with personal attributes such as birth details, death details, height, spouses, children, etc.
- The title principals dataset includes 377,848 cast members roles in movies with attributes such as IMDb title id, IMDb name id, order of importance in the movie, role, and characters played.

Data has been scraped from the publicly available website https://www.imdb.com.

All the movies with more than 100 votes have been scraped as of 17/11/2019.

Assignment

Build a predictive model for predicting whether or not a movie will achieve an average voting score > 7.5. Show which variables have the biggest impact on the model's decisions. Also show the predictive performance of your model on a random sample of 10% of the data.

Prepare a short and concise presentation (PowerPoint or Notebook) containing your approach and the choices you made, your main results and insights. Please also email us a zip file containing your code.

Note that the primary focus is not on the predictive performance of your model, but rather on your approach, choices and main insights. As such, it is optional to include datasets other than the movies dataset in your model.

We prefer the open source technologies Python (preferred) and R.