Voter Rating Prediction System

Group Number 12

Project Topic: Movie Voter Rating Prediction System

Crew: **Anuva Sehgal,as18774@nyu.edu**
Ravan Buddha,rb5579@nyu.edu

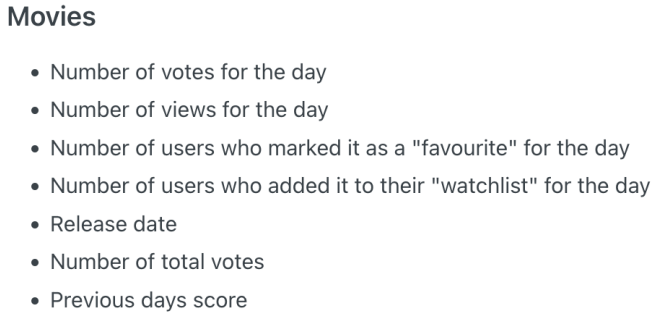
Objective:

The goal of this project is to build a machine-learning model that can predict how the audience would rate a movie given information like Genre, Budget, Overview etc.

Dataset:

The dataset will consist of movie metadata and summaries from the following dataset: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata. The columns that we will extract from this dataset are:

1. Original title
2. Budget (USD)
3. Genres (one movie can have multiple genres, for example: Avatar is categorized as both Sci-fi and Action, among other things)
4. Keywords + Overview (we will be breaking the “overview” into a bag of words and combining it with the already given column for keywords)
5. Popularity: This metric is based on a number of factors as described by TMDB itself (https://developer.themoviedb.org/docs/popularity-and-trending):



6. Production companies
7. Release date
8. Revenue
9. Runtime
10. Tagline

11. Vote count
12. Vote average (output)

Note: The dataset contains 4505 (out of 5000) English movies and the rest are other languages. For the sake of this project, we'll only consider English movies as it'll give us a better prediction and prevent overfitting for other languages(due to lack of information).

Steps:

Data preprocessing:

1. Filter movies with language: "en" (English)
2. Extract appropriate columns from the dataset (given above)
3. Convert the Overview into bag-of-words and combine it with the "keywords" column
4. Clean data: Remove movies with empty features
5. Create a numerical representation of features like genres, keywords, production companies etc. (Feature Engineering)
6. Normalize the data
7. Split the dataset for training and testing: we will split the data based on the release date (70% - 30%)

Implementation of prediction models:

1. We choose various regression models to explore results and increase model efficiency
2. We train the models and compare the R^2 score to conclude (accuracy measure for regression models)

References:

Dataset: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata/data

- https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/
- https://towardsdatascience.com/working-with-multiple-types-of-data-in-a-single-problem-in-machine-learning-31b667930179
- https://www.kaggle.com/code/bhsraman/predict-movie-ratings-via-machine-learning
- https://www.cs.cmu.edu/~ark/personas/
- https://machinelearningmastery.com/neural-network-models-for-combined-classification-and-regression/
- https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/