Initial Project Proposal: Predicting IMDb ratings

Members: Cindy Tran, Elliscope Mingzhe Fang

I. Problem we are planning to tackle

For our project, we want to solve the question of whether one can accurately predict IMDb ratings for movies using features like the budget of the movie, content rating, or popularity of the movie director, actors, and actress (with popularity determined by the number of facebook likes they have). This can have real world applications, as movie studios can predict whether their movie will have a good rating in the concept phase without relying on critics to first view the movie after it has been produced.

II. State of the Art/ Related Work

There has been some machine learning work relating to movies.In one of the data science post, Sun mentioned the application of multilinear regression and random forest regression model on IMDB prediction. He queried over 5000 movies info from IMDB and analyze their correlations. His came to the conclusion that random forest regression works better than multilinear one by generating a square residual result of 0.89.

III. Dataset that we will use

We will use a dataset found on Kaggle, titled IMDb 5000 dataset, that has 5000 movie data scraped from the IMDb website. This dataset has 28 variables including data such as the movie_title, duration, director_name, gross, genres, and budget. For this project, we will try to reduce the number of features to about 4 or 5 by using feature engineering methodology. We will also convert string value features to number based for the convenience of model application. We may also scrape from the IMDb website ourselves for some more recent movie data, as the above mentioned dataset is about 6 months old.

III. Proposed Approach

First, since there are some missing values for some movies in the Kaggle dataset, we will clean the dataset first before we apply any machine learning models on the data. This process includes excludes moves that have 0 in their data.

Secondly, we will conduct some feature engineering and dimensionality reduction to reduce the number of input features we are feeding our learning algorithms. We are so far planning on using the following input variables: duration, budget and director_facebook_likes, based on our expectations about the correlation between these variables and the IMDb rating. However, these feature selections may change as we start work on this project. Next, we need to search for an efficient method to evaluate the features we have right now. Since some features are described as strings while others are numbers, we need to convert the strings to values so that it can be applied with scikit-learn library directly.

Then, we can conduct the learning, by splitting our dataset into 5 folds, with 4 folds as training and 1 fold as a test set, using 4 cross-validation for the training set to tune our parameters and having the labels be the IMDb rating. In the end, we will try to compare the accuracy result between different machine learning algorithms and to determine which model is the best fit for our problem.

A. Machine Learning Algorithms

We will use supervised, regression machine learning algorithms to predict the value of the IMDb rating. Some machine learning algorithms we are planning to use are linear regression, Support Vector Regression, Neural Networks, and Regression Trees, validating the hyperparameters with our 4-fold cross-validation.

B. Metrics to Evaluate Our Approach

We will compare the mean absolute error, the mean squared error and R squared to determine how well our model generalizes the training set against the test set. From these metrics, we can see the average error that each model has in predicting the IMDb scores, and determine which algorithm and model is the best one for our data.

References:

1.Movie-Rating-Prediction
http://blog.nycdatascience.com/student-works/machine-learning/movie-rating-prediction/
2.Movie Kaggle Dataset
https://www.kaggle.com/reza2866/datamining