CS 593 PROJECT SUMMARY

TOPIC: IMDB Movie Rating Prediction

Team:

Vicky Rana 10416500

Vivek Ganjave 10411755

Advait Gupte 10415613

Bipin Pandey 10411830

- Dataset Source: https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

Our project problem statement is predicting movie rating based upon variables before it releases or the IMDB releases their rating chart.

It's hard to predict whether a movie is good or bad before its release in theatre. On top of that we can't rely on critic's reviews or TV shows every time. So, in order to find out the best possible solution for this, we try to predict movie ratings based on the models that we build. We will be using different factors from the dataset such as actor Facebook likes, director Facebook likes, number of user reviews etc. for predicting our target variable 'rating'. We performed cleaning over the dataset by removing missing values and dropping unwanted variables and converting target variable from continuous to categorical. For example: ratings less than 6.6 equal to 1, & ratings greater than or equal to 6.6 to 0. We build our model using Logistic regression and remove the collinearity as well as reduce the number of variables using PCA. Using the model that we built we trained our model and applied over the test dataset and got correct prediction of around 65 – 70 percent. We can use this predictive model on any new dataset having the same variables for optimal solution. We can work more on the accuracy rate of the model. We can predict movie ratings even deeper based on other factors, example: History of production house having it as a categorical factor etc.