Prediction for the Greatness of a Movie

Jin Huang Taihong Cao Hanzhao Wang

Computer Science and Engineering Department, New York University

New York City, USA

{jh5442, \*\*\*\*, tc2305}@nyu.edu

*Abstract*

***The growth of the World Wide Web has resulted in troves of reviews for products we wish to purchase, destinations we may want to travel to, and decisions we make on a day to day basis. Using machine learning techniques to infer the polarity of a product is of great importance in this age of information. Since there are too many movies released every year, we need a reliable way to judge whether a movie is worth to watch. The right combination of talent, timing, all with some good luck can have something to do with the popularity of a movie. So, we use a dataset contains the information of more than 5,000 movies all over the world, analyzing the 28 variables and try to find how different factors influence the success of a movie.***

Keywords—machine learning, prediction, regression

# Introduction

Online review has become crucial to film industry, because most people are relying on reviews and social media like Facebook to judge the quality of a film and to decide whether to watch it or not. The likes on Facebook, actors, director, IMDB review and other factors can influence the popularity of a movie. Thus, what makes a movie successful? Among all the factors, there are important ones and less important ones. The purpose of our project is to use Machine Learning methods to analyze the dataset and furthermore to predict the greatness of a movie and help people to choose good movies to watch.

This topic can be taken as a supervised learning problem, which requires analysis and prediction. Since there are over 30 variables in the dataset, we will first use the correlation matrix to find the most important variables and do analyze to the dataset. We plan to use regression model on the dataset and our main method will include linear regression and logistic regression. To be specific, we would like to predict the score (varies from 1 to 10) to each movie which can present how good it could be, which means the IMDB\_score is the label we want to predict. After finishing this, we will compare the result of the different methods and make a conclusion for which one is better for our dataset.

# Motivation

Feature films are a multibillion-dollar industry. Given plenty number of films produced as well as the level of scrutiny to which they are exposed, it may be possible to predict the success of a film based on publicly available data. As we may know, movie ratings are influenced by many factors, so accurate prediction of new movie ratings can be challenging. In recent years, various kinds of analysis techniques were successfully applied to analyzing user reviews, which in turn were applied to predict IMDB movie ratings (based on IMDB reviews, YouTube movie trailer reviews etc.).

What we are aiming at in the project is to predict the IMDB score of a movie. Why is the IMDB score what we concern much? This is because among the factors of a movie, the IMDB score may be the most intuitive factor for normal audience. Although we care about the director, the promotion PV or the main actor from time to time, it is more regular for us to check the score of a movie and judge whether it is good or not because the IMDB score is to some degree a reliable reference as a quantitative indicator.

We focus on the preprocessing and analyzing of the dataset by midway of the project. This includes the

# Related Work

(Add a reference to each paper you consulted. When you refer to reference #1 in your paper, for example, use this notation: [1])

# data

In this pro

# Algorithm(s) used

(In this section, can describe: Your experiment setup/issues with data/performance/etc.

Describe your baseline algorithm and leave the improvements for the final report)

# Code

*(Future: Include a link to your github repo with the code)*

# Future Work

(In this section describe how you will expand your data/algorithms etc, for the final report)

# Conclusion

(Future… One or two paragraphs about the accuracy/ROC/goodness of your analysis.)