Predictions for Movie Performance

Bingxin Wu (bw383), Yiwen Jiang (yj76)

Project Description

During the past two decades, the movie industry has drascially evolved. The movies with highest grossing mostly happened in the past ten years,as the 3D and CGI techniques for films proceeded to maturity, which yielded high ticket prices and more public interests. Although the gross returns for movies have been increasing as a general trend, the cost of making these films have also increased due to many factors, for example, the technical costs for making high-quality pictures and the hiring costs for attractive actors. Therefore, the risk of movie-making has also grown higher. Hence, it is important for producers and investors to have a brief idea of how the outcomes of their movies might be beforehand.

For this project, we are interested in three fundamental criteria of movie performance: numbers of awards won, profitability, and public rating, as well as how different factors influence them.

Datasets and Methods

We will use two datasets from Kaggle for this project - "IMDB 5000 Movie Dataset" and "The Academy Awards, 1927-2015". The IMBD dataset contains the major factors of movies for us to analyze. We can use the gross returns and budget data to calculate profitability, and it also gives the IMDB user ratings of the movies. Other informations like director, actors, genre. etc. are useful for classification and performance prediction. "The Academy Awards, 1927-2015" dataset will be used as an add-on factor in our analysis: 1).what kind of movies are more likely to win awards 2). to see if there exists a relationship between profitability and rating of a movie and whether it won the academy awards or not.

We pre-ran the data and observed significant correlations between a lot of different factors in the dataset and our dependent variables, such as a postive correlation between gross returns and number of voted users or IMDB scores. Although we have not yet implemented specific tests or built specific models, our first glance into the data gave us confidence in creating a desired model. Hence, we believe that the datasets we plan to use will allow us to give some important insights in future movie performance prediction.