Project Proposal

Authors: Jessie Wang (yw696), Eric Dai(emd88), Anil Vadali(arv42)

Question

Are there some important features of a film that determine how well it scores on the IMDB rating system? Can we predict how much a certain movie will earn in the box office based on its background?

Datasets

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

Why is this Project Important?

In the United States alone, there are over 600 movies made each year, however only a select few of these movies are very successful and do very well in the box office. Moreover, among these select few, an even fewer number number of films are scored very well by the infamous IMDB rating system and make it onto the IMDB's Top 250 movie list. With the emergence of new technology companies such as Google and Apple starting to make movies, and even Amazon Studios and Netflix winning Oscars last year, it has become extremely important to understand what makes a good movie to survive in this newly competitive film industry. Determining which movies to make and not to make, casting the optimal actors in certain roles, and predicting what consumers and critics will enjoy have been crucial decisions in creating a new movie and has become somewhat a science in recent years. As a result, this project hopes to accomplish at least one of the two following goals: predict how much a certain movie will earn in the box office based on its background, and determine the IMDB rating score of a certain film based on certain features and aspects.

Useful Features of the Datasets

There are numerous features of this dataset that will help us accomplish the goal of this project. One very interesting aspect of this dataset is that it provides social media statistics based on actors, directors and movies themselves. Analyzing social opinion of a movie is extremely important in understanding whether it will be successful and popular in the box office, and utilizing the number of Facebook likes of different aspects of a film will be crucial in accomplishing this. Additionally, features such as the time period a certain film was released is also key in determining whether a movie will earn a certain amount in the box office; this dataset only provides the year that a movie was released, but additionally we could perform data scraping and manipulate the dataset to include month and date as well. Moreover, this dataset includes information on the directors and the actors for each movie as well, which could play a significant role determining the IMDB rating of a film since acting and directing determine the success of a film from an artistic viewpoint. The data set also includes other very important features such as genres, key plot words, and poster analysis which we could also utilize. Overall, we were thinking of utilizing regression to determine the predictions based on these features. Additionally, other algorithms such as Naive Bayes may help us predict the probability of how likely a movie will achieve a certain rating or generate a certain gross box office value.