IMDB Data Mining Project

D'Azevedo, Gloria
gad87@cornell.edu

Enriquez, Erik
eee37@cornell.edu

March 31, 2017

The problem that we will investigate is finding out which characteristics of a movie are significant to predicting the "success" of a movie. Since "success" can be hard to measure, we will predict the gross box office dollars as a proxy for success. In a preliminary analysis, we note that there are over 800 movies that have missing gross amount, so we will either get rid of these movies from our dataset or search for another dataset we can join to fill these missing values.

The data set that we will use is the Kaggle IMDB 5000 Movie data set which includes 28 features for over 5000 movies that was scraped from IMDB. The data includes names of the actors and directors, the number of likes on Facebook that they have, the budget for the movie, the score, the gross amount that the movie made, and other information about the movie. (https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset)

We hypothesize that certain factors will influence the success of a movie such as having A-list actors and directors while other factors such as duration or aspect ratio may not be as significant to the model. We also note that some standardization or classification may be needed for certain monetary factors including budget and gross dollars since the currency is reported in the movie's local dollars. It would be unwise to compare United States dollars with the Chinese RMB during any year since the exchange rates can wildly fluctuate. In addition to trying to predict the exact gross amount that a movie makes, we can divide up the range of values into bins and use classification techniques such as k-means clustering and one vs all logistic regression.

Some of the characteristics included in the dataset are breakdowns of Facebook likes for various movie aspects (for the actor or actress individually, for all the actors and actresses, the director, and the movie itself). These numbers are reflected at the time the data was scraped so there could be some bias if the actor or actress had some other box office hits between that movie and now or some movies may be exposed to audiences who cannot use Facebook. In addition we have movie-specific metrics such as the year it was released, the content rating, the IMDB rating, the genre, and the duration. If significant, they can play a role in planning and editing new movies. For example, if the content rating of a movie was lowered from a rating of R to a rating of PG-13, then more people are able to watch it which could increase the overall amount of money that it makes. Similarly, if a movie length was decreased from 150 minutes to 120 minutes, then additional showings of the movie could be scheduled, thus raising ticket revenues.

In addition to data processesing techniques, we plan to test the effectiveness of a wide variety of statistical methods including, but not limited to, multivariate linear regression, regularized regression, K-Nearest-Neighbors and classification while developing a model for "success" prediction of a movie using this data set. We might include the summary of our models to display the significance of each predictor on the gross amount. To test our model we will use cross validation on our data set and we can also scrape information for new or recent movies that are not currently in this dataset to assess the accuracy of our model.