Movie Earnings Executive Summary

Introduction

For our DATA-151 project, we researched what factors are most important in determining how much money a movie will earn. This topic is important because movies are a major part of modern culture. By looking into what leads to a movie earning more money, we can get a better understanding of what modern culture values, what stories appeal to people, and what people want out of entertainment. It is also possible that the results of this research may be used to find the most desirable movie to produce next, with the desire that such a movie earn as much money as possible (assuming all of the factors could reasonably be assembled).

Processes

We found and used a dataset containing the top 1000 ranked movies by IMDB rating on Kaggle. This dataset contains information about movies spanning from 1925 to 2019. It also includes a number of characteristics about each of these movies such as the release year, runtime, genres, director, and domestic earnings. We believed that this dataset gave a good overview of the film industry and provided sufficient information to determine what factors are most important in predicting how much money a movie will earn. To analyze and model this data, we used Python in the Google Colaboratory IDE.

We thoroughly cleaned the data, removing unnecessary columns, as well as data points that did not have enough information to properly use. We also calculated the inflation-adjusted domestic earnings for each remaining movie in our dataset indexed to 1983 dollars to provide a more common ground of analysis. We conducted exploratory data analysis to examine the data and the relationships between our explanatory and response variables. Finally, we developed a series of machine learning models, using both regression algorithms such as linear and lasso regression, and classification models such as decision tree and random forest classification, to try to predict earnings for a given movie.

Results

The coefficient of determination of the explanatory variables and the response variables for each explanatory variable revealed that many of the predictor variables did not seem to have a large impact on domestic earnings individually. The only variable with an R^2 value greater than .10 was the number of votes on a film's IMDB rating. There did seem to be a significant difference in earnings depending on whether a film was directed by a famous director. The average inflation-adjusted gross of a film made by a famous director was $69,344,856.78, while the average inflation-adjusted gross of a film made by a non-famous director was $52,474,195.54. This difference was significant at the 5% level. Furthermore, we found large differences in film earnings based on a film's genre. Despite these findings, our best performing regression model only explained about 19% of the variance in movie earnings. The classification models were a bit more successful in predicting blockbuster films.