Linear Regression Project Write-up

Movie Profit Analysis

By: Erica Stevenson

Abstract

The cinema industry is one of the oldest industries to exist globally. While having incredibly high costs to produce, market, and release movies, the industry has continuously seen a huge amount of profitability over time despite a rise in production value. Interestingly enough, it's also difficult to predict what will make a movie successful. For this analysis, I will be searching for a relationship amongst this data set that can help find a correlation in what makes a movie successful.

Data

- Rotten Tomatoes https://www.rottentomatoes.com/ There are over 10,000 movies with a critics score (Tomatometer) that is below 70%. I scraped these movie names as well as their Tomatometer score, year realeased, and number of ratings.
- IMDB https://www.imdb.com/ This site provided many additional metrics for the movie titles that I have: title, MPAA rating, run time, genres, director(s), writer(s), actor(s), movie studio. I used the total world gross as the target variable, and all other observations acted as features in the model.
- Box Office Mojo https://www.boxofficemojo.com/ This website provided the rank and title of each movie, as well as the worldwide, domestic, and foreign lifetime grosses.

Algorithms

I used Selenium and BeautifulSoup to scrape several movie database sites. I performed a thorough Exploratory Data Analysis of the data; cleaned, explored, aggregated, and visualized the data using Python and Pandas. Simple linear regression with StatsModels and an ElasticNet fit.

Tools

- Selenium is needed for scraping Rotten Tomatoes
- Requests and BeautifulSoup for web scraping IMDB and Box Office Mojo
- Pandas and Numpy for data analysis and exploration

- Scikit-learn
- Matplotlib and Seaborn for plotting and visualizations

Findings and Results

The target that was used for measurement was Worldwide Gross Revenue. Using the Rotten Tomatoes database, I found the movie title, release date, critic score, number of critic scores as key features. I merged this data with a couple of TSV files that I found through IMDB. Then I used IMDB's unique identifier to scrape Box Office Mojo for more information about movie revenue, MPAA ratings, etc. While the results of the models showed that it was difficult to predict exactly what can directly make a movie successful, as well as how to measure that success, there was some correlation with most of the features that I looked at. I was surprised to see that several factors that I anticipated to be highly correlated with success were not as impactful in the model as I had hypothesized (i.e. critic and audience ratings). Similarly, the analysis of the data confirmed that a movie's success can be measured in several different manners. I got a decent fit of a ElasticNet Regression on about 4500 movies worth of data.

Communication

All slides have been uploaded to my personal Github account at

https://github.com/ericajstevenson/METIS_projects under "Movie Profit Analysis"