Predicting Total World Revenue for Movies

Mitchell Seiter

Abstract

The goal of this project was to build a scraper and web navigator to automate the collection of data from a movie based website, and also, to then be able to use that data to build linear regression models to try and predict how much revenue a movie would make based off of the features gathered. The data was gathered from boxofficemjojo.com and initially consisted of over 2000+ movies and ten key features of those movies. Some linear, lasso, and ridge regression models were built to test these features against a target to try and get different metrics such as R^2 and MAE.

Design

The goal of this project came from a client who was hoping to maximize the amount of revenue a movie could make based off of certain features. The data comes from boxofficemojo.com. If proper linear regression models are built we can determine whether or not total revenue can be determined off of certain features.

Data

The dataset consisted of 2000+ movies coming from the top grossing #1 movies of all time, as well as top grossing movies that never made it in the #1 spot. I initially scraped off eleven features, with three of them being categorical. Some of the most prominent features that were utilized were, budget, production company, and genres.

Algorithms

For the web scraping I was able to build an algorithm that would automatically scrape all of the data that was needed once the initial URL was put into place. From there, using Selenium and BeautifulSoup, the algorithm would click the movie title page, scrape the necessary info, click on any additional links that were needed, and then return to the movie list page to gather the next unit of data.

From there I cleaned up the DF to remove any NA values, removed outliers, and cleaned up the data types. Once that was done I was able to do some feature engineering to make the categorical features usable, and did a polynomial test on a couple features where appropriate.

Lastly, I was able to split the data, and build some different models that would best fit the data. Some models that were built included a LASSO regression, RIDGE regression, and simple linear model.

Unfortunately none of these models scored above 0.61 for R^2 and had a fairly high MAE value indicating a lot of variance in the model.

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting
- Selenium and Beautiful Soup for web scraping

Communication

This information will be presented to the client in a powerpoint format. We will go over the key metrics and what their values mean, how to interpret the results, and how to move forward with the information that was found.