Putting the “Feature” in Films

Metis Linear Regression Project

Aathira Chennat, February 2022

Abstract

The goal of this project is to help aspiring film producers understand what make a film a box office hit. I used linear regression to build a model that predicts a film’s gross domestic box office based on its characteristics, including running time, distribution company, and genre. Using 2017-2019 film data scraped from The Numbers, I designed and cross-validated a model that was able to predict gross earnings at the box office with an R-Squared of 49% and a Mean Absolute Error of $23 million dollars. Presence of outliers and lack of features lead to overfitting and multicollinearity issues.

Design

This model predicts inflation-adjusted domestic box office earnings based on its characteristics, including running time, distribution company, and genre. The final model was selected via K-fold cross-validation, and the scoring metric used was mean absolute error (MAE).

Data & Algorithms

The dataset contains 1,448 observations, or films. The target was inflation-adjusted domestic box office earnings. The features included in the regression included running time and a host of categorical variables, denoting the production type (animated, live action, etc.), distribution company, genre, source (original screenplay, adaptation, etc), and (release date)

- Top line EDA showed an extremely right-skewed distribution of gross box office earnings. Upon further exploration, all outliers were identified as part of a major franchise from a major distribution company, i.e. Marvel Franchise (Disney), DC Franchise (Warner Brothers).
    - ✳ This finding lead to engineering interaction variables for Disney Franchise, Superhero Franchise (non-Disney), and Animated Children’s Films (non-Disney)

- A basic OLS regression rendered many variables insignificant, and a shortened features list was again optimized after running through a lasso regression, who's alpha has been selected through k-fold cross validation
- Assuming there are diminishing returns to an increase in running time on doc office, I squared the running time variable
    - ✳ Running time likely introduces multicollinearity, as larger action and superhero franchise films tend to have longer running times. Because running time was the only continous feature I had, I included it in the regression. In future iterations of this model, I would scrape films for larger timeframe, and gather more films that had data for the production budget. For my dataset, only 550 films had this feature, so I had to remove it from my features.
- I performed K-fold cross validation on three models, Linear Regression with regular running time, Linear Regression with squared running time, and a lasso regression. Using Mean Absolute Error as my scoring metric, I found that the squared regression worked best.
- Once I fit my model onto my test dataset, the scoring metric increased, suggesting an overfit model.
    - ✳ I believe this is mostly due to the outliers in the dataset, and how they were distributed across the Training and Testing datasets.
    - ✳ In future iterations, I would remove larger distribution companies from my analysis and hone in on the relationships for independent films, as this level of production is likely more accessible for aspiring producers.

Tools

- Beautiful Soup for Data Scraping
- Numpy & Pandas for Data Cleaning & EDA
- Stats Models and Scikit Learn for evaluating and validating Linear Regression
- Seaborne and MatPlotLib for visualizations