Cecilia Bell 5.14.2021

Project Write-Up

Domestic Success of Foreign Language Films

Abstract

In 2014, IndieWire's Anthony Kaufman claimed that US viewership of foreign language films had fallen. But with Bong Joon-Ho's *Parasite* becoming the first foreign language feature to win Best Picture at the Oscars in 2019, it is clear that some foreign language films have the potential to capture wide US audiences. The goal of this project was to predict the domestic success of foreign language films, taking domestic success as total domestic profit. Using Box Office Mojo and IMDB data, various models were trained and cross-validated. A Ridge regression model with regularization strength of 0.01 was eventually selected and the R2 score for the model on the test dataset came out to 0.192. Some features hypothesized to have predictive power ultimately did not correlate with the target, while genre and domestic distributor turned out to be more significant.

Design

The project was designed to be utilized by US distributors who are interested in buying the licensing rights to foreign language films such as A24 and Neon, as well as film sales agents, streaming companies and international production companies. As such, it focused only on foreign language films that were not first released in the US. Since the goal of the project was to predict the domestic success of foreign language films, the target was taken to be total domestic profit rather than total domestic gross.

Data

Data on foreign language films was first scraped from Box Office Mojo. The original dataset consisted of 2500 data points and 20 features, including genre, international gross, earliest release location, etc. Data on films that were first released in the US were dropped. IMDB was then scraped to fill in missing values found in the BoxOfficeMojo dataset, most notably budget. Rows with budgets that appeared to be misreported (budgets under $10,000) were then dropped. The final dataset consisted of 550 data points.

Algorithms

The project began with web scraping. A scraping pipeline was created using BeautifulSoup and Selenium, allowing data from thousands of Box Office Mojo webpages to be acquired. Error handling was built into the pipeline, so that webpages would be repeatedly refreshed when the server was down. The data was then stored in a Pandas dictionary which was combined and converted into a Pandas dataframe. Cleaning followed, with missing values either imputed based on the variables distribution or dropped. Further scraping was performed to gather missing budget data from IMDB. Budget data in foreign currencies was then converted to USD with Forex-Python and all monetary data was adjusted for inflation with the CPI library. The target, total domestic profit, was engineered using total domestic gross revenue and a scaled portion of the budget (total domestic gross - budget(total domestic gross/total worldwide gross). Alongside cleaning, exploratory data analysis was performed with Pandas, Matplotlib and Seaborn, to scope out regression viability and look for multicollinear features. A baseline model was then constructed with numerical features. A cross-validation scheme was then set up to test different models with different features including newly engineered categorical dummy features. The categorical dummy features included certain genres and certain distributors that were found to be significant through boxplot visualizations. The model was expanded and refined and it was found that a Ridge regression model with scaled coefficients performed best. Regularization strength was then tested, and it was found an alpha of 0.01 performed best. Visualizations were then selected to showcase correlations and unexpected findings.

Tools

- BeautifulSoup and Selenium for scraping
- Numpy and Pandas for data manipulation
- Forex-Python and CPI library for currency conversion and inflation adjustment
- Scikit-learn for modeling and statistics
- Matplotlib and Seaborn for plotting

Communication

Regression model scores, as well as Matplotlib and Seaborn plots visualizing correlations and unexpected findings, are displayed in a PowerPoint, which will be presented to the Metis class and shared on GitHub.