Cecilia Bell
5.5.2021

Linear Regression Project Proposal

Domestic Viewership of Foreign Language Films

The Question

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 framing question of my analysis, then, is: can I predict the US success of a foreign language film? I will take US success to be gross domestic box office profit adjusted for inflation, where gross domestic box office profit is gross domestic box office revenue minus a scaled proportion of the budget. 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, benefit from exploring this question. The question can be explored from a variety of perspectives/production stages (pre-production, post-production, post-early-release, etc.). To maximize the predictive power of my model, I will begin by exploring the question from the perspective of post-early-release. I will then try to maximize the utility of my model, exploring the question from the perspectives of post and pre-production.

The Data

I plan to use data on foreign language films (from at least the last twenty years) from Box Office Mojo, which I will scrape with Beautiful Soup. My ultimate target will be the gross domestic box office profit, which I will engineer using gross domestic box office revenue, gross worldwide box office revenue and budget (all adjusted for inflation). My features currently include language, budget, opening international box office profit, gross international box office profit, opening release number of markets, total release number of markets, genre, release date, MPAA, running time, number of festival screenings, number of awards, IMDB rating, critic ratings, distributor, and production company. Other features that require significant engineering include star power (combined actor rankings), director ranking and writer ranking. As such, an individual unit of analysis would be the data on all films for one feature (but could also be the data on all features for one film). I predict that the gross domestic box office profit will correlate most strongly with the gross international box office profit.

The Tools

As mentioned, I will use Beautiful Soup to scrape data from Box Office Mojo. I may also use Selenium if I find that my model would benefit from features that are unavailable on Mojo. I will use Pandas to create a list of dictionaries of my scraped data, which I will merge and put in a

dataframe. To explore the data and find features that seem to correlate with my target, as well as potential multicollinear features, I will visualize the data with Matplotlib or Seaborn and perform statistical analysis with Statsmodel. Finally, I will use SKLearn to build my linear regression model. I may also use Blokeh or Plotly for presentation visualizations.

The MVP Goal

My MVP goal is to build a linear regression model to determine the gross domestic box office revenue of foreign language films with one or two features. In order to do this, I will first need to scrape the required data and perform some exploratory data analysis and data cleaning.