Jesse Reonardy

# Predicting The Profitability of Movies That Are About To Be Released

September 22nd, 2021

**Question**

I am a Data Scientist from Metis, and I was approached by a confidential client from a movie studio. The client would like to know if we can predict the profitability of movies that are about to be released by using web scraping and linear regression tools.

**Data Description**

To model the linear regression of the problem, I am planning to gather the following datasets from web scraping:

- 5000+ movies (data points) from boxofficemojo.com, imdb.com, and www.the-numbers.com
- 10+ features. Here are some of the potential features: Production Budget, Opening Weekend Gross, Domestic Gross, International Gross, Running Time, Number of Release Theaters, Ratings, Genre, Producer(s), Director(s), Actor(s).

I will build the model using the correct techniques of regularization and/or polynomial features. Also, I will carefully assess each of the feature correlation and drop/add features as appropriate. In addition, I will study the model selection and evaluation rigorously with proper validation and testing. Finally, I will study the model from total profitability (revenue - cost) and profit margin ((revenue-cost)/revenue) to see which model best represents our question.

**Tools**

- SQLAlchemy on Jupyter Notebook for preliminary database exploration
- Python modules of pandas and numpy for deeper data analysis and manipulation.
- Python module of Scikit-learn linear regression for modeling
- Web scraping tools of BeautifulSoup and Selenium for data collections
- Seaborn (or Bokeh/Plotly) for plotting and visualization

**MVP Goals**

MVP will include a two preliminary linear regression plot (total profit and profit margin) with truncated data sets of around 1000 data points and 3-5 features. These visualizations will include at least one short paragraph of the initial conclusion.