Metis Week 2 - Project Luther Proposal

Steven Bierer, October 2, 2018

Overview

The movie industry in the United States is truly a big business, with studios and distributors always looking for the next blockbuster to pay the bills for their ongoing projects. But many much-loved films do not haul in loads of money when they appear in theaters. While these movies often have relatively small budgets, studios won't produce them if they continuously hurt their company's bottom line. For this reason, there is substantial interest in optimizing the release time of movies to maximize exposure to viewers and increase ticket sales.

My Week 2 Data Science project aims to tackle this problem by analyzing available databases of films released in the U.S. and developing a linear regression model to predicting how a number of variables - release date as well as budget, movie genre, language, etc - affect the amount of money brought in at theaters. The approach of the project fits the goals of the Project Luther learning experience by emphasizing data scraping, data exploration, and regression modeling. Additionally, movie databases are a popular topic for data scientists, especially in the class room, so I expect the data set to be reasonably easy to work and that solutions can be found online if I run into a problem. This will allow me to focus on learning new analytical and visualization tools rather than become burdened by a myriad set of hurdles.

Data Resources and Tools

I will analyze data from the Internet Movie Database (IMDb - www.imdb.com), a website compiling a trove of information on over 4 million movies and TV shows. Box Office Mojo (boxofficemojo.com) and other movie/TV production aggregate sites may be scoured as needed.

For collecting information from the database, I will mainly use the web scraping tool Selenium, but will employ Beautiful Soup as well when necessary. Analysis will be performed in the Python programming language using the modules pandas, numpy, and skikit-learn. The

modules matplotlib and seaborn will be used for graphical display.

Variables to Analyze

The proposed dependent variables will be box office receipts, both for the entire first run of the movie (total) and for the first week. A break-down to subsequent weeks may be evaluated if preliminary analysis suggests that's appropriate. Only U.S. receipts will be included, unadjusted for inflation as any year-by-year trends will be evaluated.

Following is a list of data elements I plan to initially include as variables in the regression model. These do not include variables that are derived from others (e.g. calendar week or holiday indicator derived from the release date). Other data may be added as the model evolves.

Variable Name	Description	Type
Total Box Office	All U.S. receipts for first run in theaters	Numeric (U.S. dollars)
First Week's Box Office	Only the first week, usually Friday to Friday	Numeric (U.S. dollars)
Budget	Including production and marketing	Numeric (U.S. dollars)
Release Date	Date when movie is in most targeted theaters	Numeric (date)
Opening Week Theaters	Number of theaters in the first week	Numeric (integer)
Movie Genre	Action/Adv., Comedy, Drama, Family, etc	Categorical (string)
Language	English or non-English	Categorical (yes/no)
Sequel	Will it be better than the first?	Categorical (yes/no)
Source Material	Is it based on a book or play, or original?	Categorical (string or yes/no)

MPAA Rating	G, PG, PG-13, R, Unrated	Categorical (string)

Preliminary Analysis

I will take the following preliminary steps as I explore the data:

1) Use Selenium to scrape movie information for the 10 year period from Jan 2008 - Dec 2017 (last full year). Tables with this data are obtainable from the IMDB website across multiple pages.
2) Obtain additional information using Selenium to iterate across html links to the individual movies. All data will be compiled into a pandas dataframe.
3) Look for expected trends in the data as a whole: a) Box office as function of release date, with expected peaks in early summer and December, especially around holidays; b) Larger receipts and budgets for Action/Adventure movies, substantially smaller for documentary and other genres; c) A significant correlation between box office and number of theaters playing the movie at release.
4) Look for correlations among the non-box-office (independent) variables in the above table, to anticipate collinearity issues in the regression models.

Regression Modeling

The sckit-learn statistics module will be used to generate an initial first-order linear regression model (and a zero-order intercept-only model for "monkey" comparison), with box office receipts (total or for opening day) as the dependent variable and the data elements in the table above as independent (predictor) variables. From a preliminary investigation of this model, variables with little impact on the residual errors, or those determined as being highly collinear with other variables, will be modified or removed. Interactions among select variables will be introduced in a progressive manner. Temporal variables like week-of-the year will be analyzed as cyclical where appropriate.

Since my objective is to predict the success of lower budget movies, I will focus subsequent analysis on movies with budgets in the lower quartile of

all movies, as calculated on a yearly basis. Thus, removing movies with high budgets and/or high box office receipts (e.g. the highest 10%) may improve the prediction power for low-budget movies. The rationale is that some movies (e.g. The Last Jedi) will simply gain many viewers no matter when they are released.

Some “feature engineering” may be necessary to improve the regression model. For example, the proximity in release date of one or more high-budget movies, or of several movies in the same genre, may reduce ticket sales to a particular movie. I anticipate generating a new composite variable, “threat score”, assigned to every movie that will reflect such competition.

Finally, a mixed-modeling approach may be prudent, to account for unobserved factors related to the known variables. So, for example, specifying year as a random factor, such that every year can be assigned its own intercept, may help account for unknown effects of monetary inflation and other influences of the U.S. economy that could impact ticket sales.

Hypotheses

I expect that low budget movies, perhaps especially those in less-popular genres like documentaries or those filmed in a foreign language, will be sensitive to both absolute release date (to the month level, and accounting for holidays) and temporal proximity to big-budget and same-genre films.

Potential Pitfalls

I expect many hurdles in the form of bad coding and missing data elements. Also, it’s possible that the model variables chosen won’t yield a significant prediction. I will add new variables from movie databases or other sources (e.g. user ratings, at risk of drowning out my hypothesis) and explore interactions among variables depending on the nature of the data.