CMPT 353 —

Movie Success Data Analysis

Parmvir Johal, Raymond Eng

August 3, 2018

1 Problem

Producers, investors and organisations are highly interested in determining what makes a successful movie. What makes a successful movie? How should we determine success?

2 Data

We used multiple sources of data, the first being data to describe movies from WikiData. The wiki data was downloaded as a compressed JSON and was filtered using spark to get the meaningful data. The code was provided by Greg Baker but we made changes to include more data such as box office prices, movie cost and actor labels. Greg baker used a scraping algorithm to get data from Rotten Tomatoes. The rotten tomatoes data included audience and critics data, and could be joined with wikidata using the uniques imdb ID. Lastly we used the OMDb API to get movie plots based on the imdb ID. Most of the data was clean however, the occasional empty values were filtered out using pandas in our notebooks.

3 Analysis & Techniques

3.1 Success correlation

We suspected the ratings percentages from Rotten Tomatoes was correlated with success. We defined success as the amount of profit a movie generated. SciPy stats tools were used to analyze correlations and matplotlib to generate visuals.

- Matplotlib, SciPy stats tools

Additionally, we wanted to take a look at a general perspective to see if movies box office revenue is correlated to directors or actors.

- ANOVA test was used to test whether there is a difference in means on actors or directors in regaurds to box office Dataframe manipulation

3.2 Genre Prediction

We inquired if a model could predict the genre given a plot. Our pipeline consisted of:

- CountVectorizer, to convert the plot into a dictionary of features and removes stop words such as 'the'
- TfidfTransformer, which includes a term frequency (Tf) to transform occurences to frequencencies and a inverse document frequency to downscale weights on words that appear often in the corpus but not in the plot
- SGDClassifier with alpha=0.001 to classify our data efficiently

3.3 Genre Popularity based by year

We wanted to see if genres have stayed consistent over the years or if they have changed. For example have Action films just become popular or have they always been. We know that genres play a role in profit so we graphed out all the genres based on box office revenue to see the most popular genres (by year) for a successful movie.

- Used pandas dataframe algorithms to group the data by year and separate the multiple genres for each movie into separate rows.
- Numpy is used to calculate the means
- We visually displayed the data using seaborn to show all the genres revenue based on year

3.4 Predict audience rating

we tried to predict audience rating using box office revenue, cost, number of ratings, critic ratings, awards and genre fed into a machine learning model.

- Custom vectorized functions to extract nominations and wins
- StandardScaler to scale the data for the model
- SVC model with a c value of 1

4 Results

From the correlation tests we have determined that both average audience and critic reviews do not correlate well with "success" (r=0.23) and (r=0.19) respectively. We defined success to be profit because as a business that is the most important thing. Because of these findings we will move onto other data that might influence profit.

An ANOVA test to see if the means of each director's box offices revealed that directors made an significant difference to a movie's box office (p=7.94e-05). On the contary, we failed to reject that the actor's box officice means are different (p=0.33).

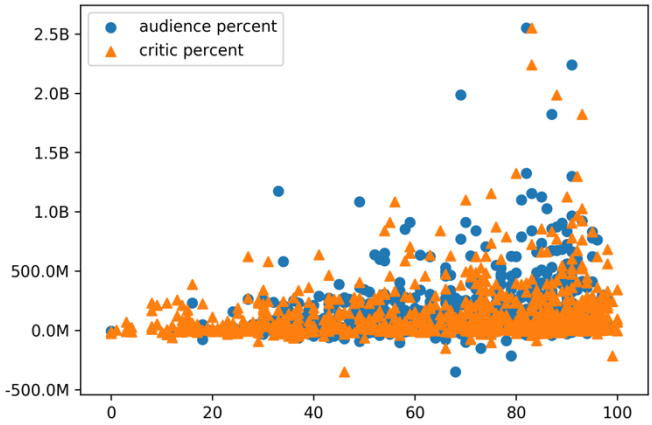

Figure 1

Next, we wanted to see if there was a correlation between the number of ratings (popularity) on rotten tomatoes and the box office revenue. We obtained an positive but weak correlation of (r=0.18). At first, (Figure 2) did not look right because we assumed more ratings would mean more ticket purchases. However the data we have does not include other sources of viewing the movies like netflix.

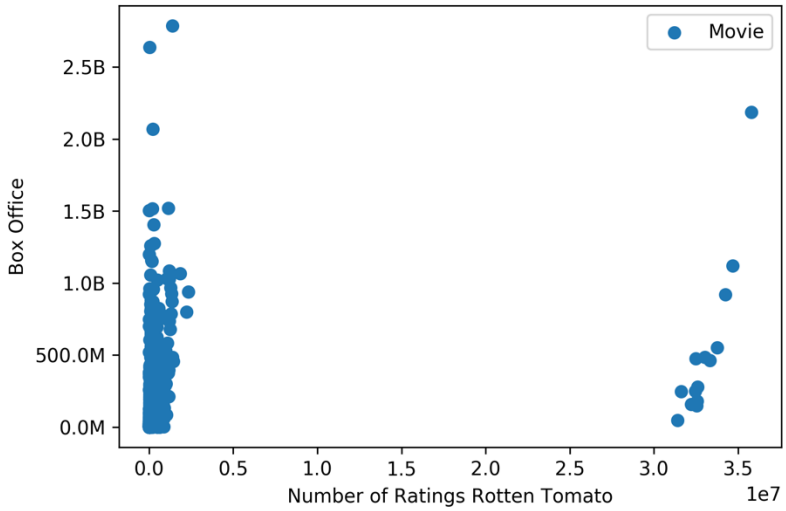

Figure 2

We used a Chi-squared test on genres with counts of the movies that made profit and those that did not. The test rejected the null hypothesis that the means between the genres were same with a result of (p=0.008). From (Figure 3). Over the years we can conclude from the data, the genre, superhero films were most likely to succeed (96.7%) from a sample of 31 movies.

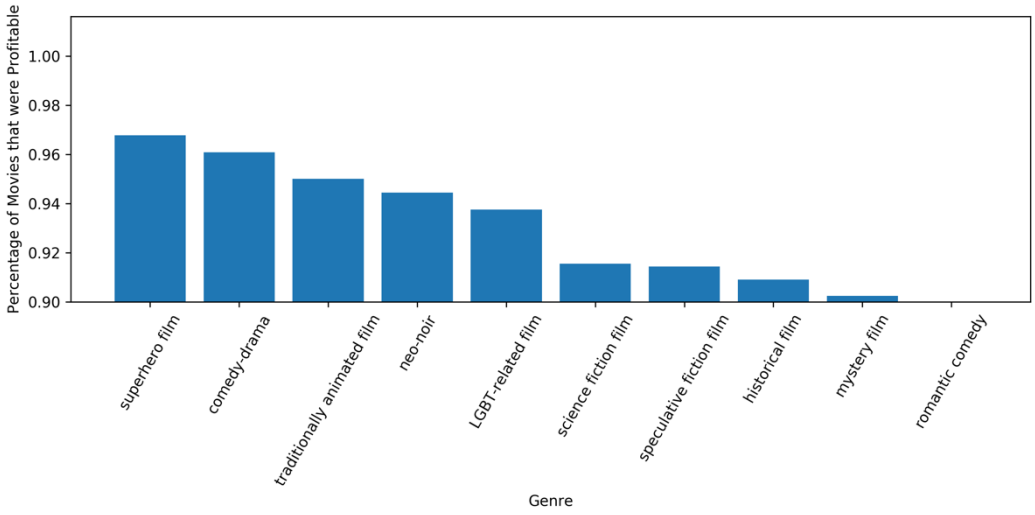

Figure 3

From (Figure 4) it may be hard to tell but from the data we found that most genres tend to change with inflation aka more movies are being made now compared to the past because of resources and overall people. We can easily find the most popular movies by year with this dataframe. Genres such as history tend to have spikes in box office over the years because of its limitation on ideas (You can not just invent a history film). One genre may be more popular than the other but it does not mean it will 100% successful it was popular in the past.

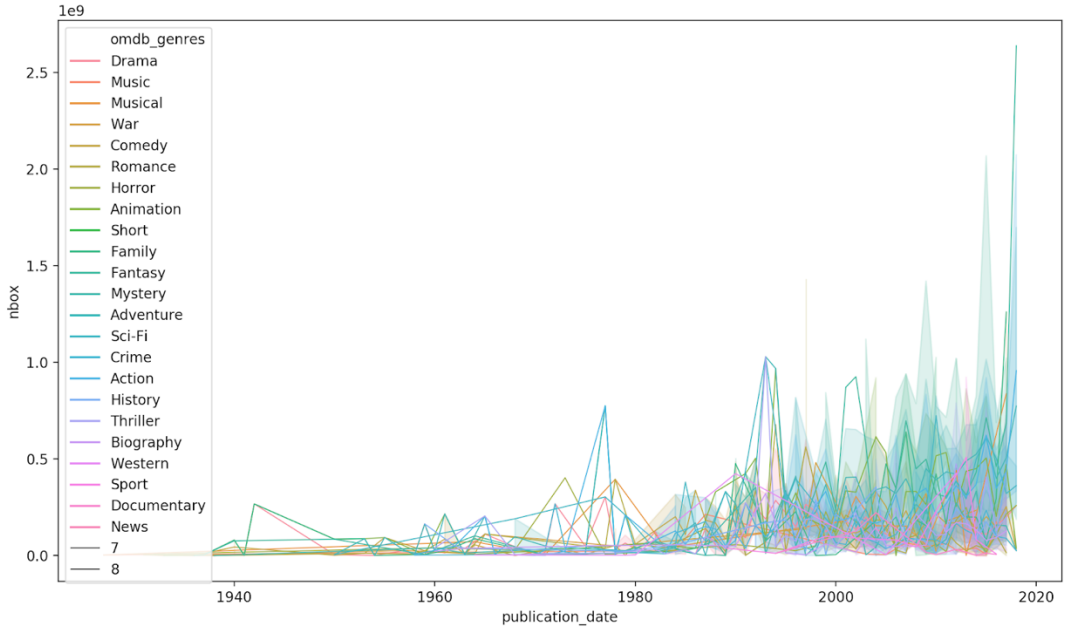

Figure 4

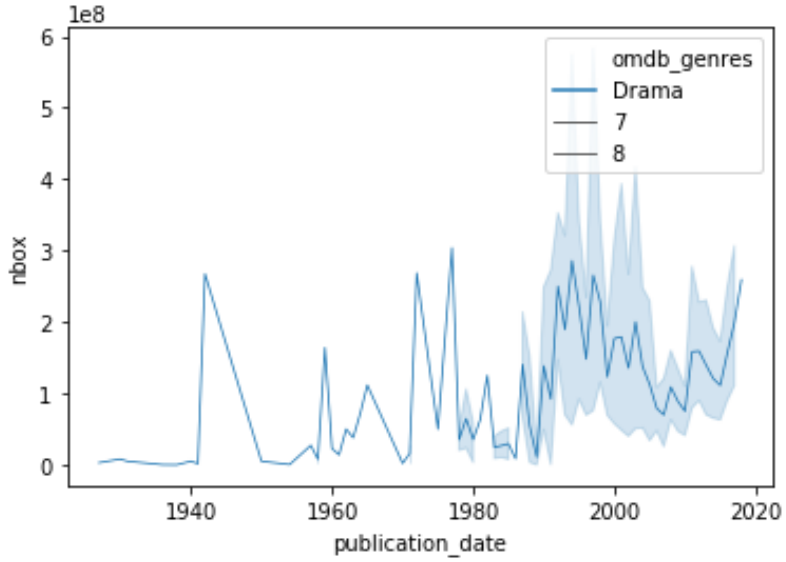

Figure 5

With our NLP model we can predict genres with a score of 49% given a plot summary. By using this model we can use it to determine success because success is associated with genre.

We were able to predict audience ratings using the SVC model and the StandardScaler. The results were not that great, we got a 50% score. We noticed that the critic review had the largest impact in accuracy, the score dropped roughly 20% when it was excluded. This result shows us that the other parameters did not make a huge impact on the audiences review.

5 Limitations

The data provided by Greg was good enough to get this project done but it in certain scenarios the number of data points was limited. For example when we took a look at movies with box office revenue we noticed that there were a lot of them missing this value. If we were to have scraped some data ourselves instead of relying on an API we could have gotten much better data catered to our needs.

We were limited on the type of data as well. When we took a look at the number of rotten tomatoes ratings per movie we noticed it did not correlate with the box office revenues. We believe that this is because we did not consider the other forms of application to view the movie. For example netflix would have been a great place to get data because it is so popular and because it has replaced the need for people to have to pay to watch a movie in the theatres.

6 Project Experience Summaries

Parmvir's Accomplishment Statement:

- Created a machine learning model that was able to predict audience ratings for movies
- Understanding the importance of data and how you should not just have a subset and take into account all scenarios
- Cleaning and filtering large data sets
- Working on multi data set project

Raymond's Accomplishment Statement:

- Extracted, transformed, and loaded data to perform data analysis
- Created data visualizations
- Leveraged machine learning models to perdict movie genres