Movie Research Project
A study in movie business

Hrushik Reddy Tadvai

170007110

I. Introduction (163 words)

The purpose of this report is to identify critical success factors for the popularity and revenue of a movie and build visualizations to build later predicting models to provide an inside into the movie industry.

In an era of digital movies and the following complementary data, relevant investors and production companies push a more significant push to establish clear connections between factors and popularity to appeal to a greater audience, likely resulting in a more substantial profit. According to industry statistics, the Movie industry is now worth over 100 billion dollars, but six out of ten movies are unprofitable; investing is risky. Given these inherent risks, Production companies must virtually gamble on a film for success.

Online Streaming services like Netflix recently started producing their original movies using the data they procured from their customers to construct appropriately trained machine learning to understand the business, make predictions, and set expectations. Netflix has used their models to an incredible extent to produce popular originals.

II. Analytical questions and data (296 words)

A. Data and variables

This study has been constructed by leveraging data from the Kaggle website containing over 50,000 movie information, including cast, genres, overview, Budget, Revenue, and popularity based on a vote. Popularity and revenue will be the two dependent variables to explore against all other factors available. TMBD has constructed the metric popularity to describe a movie's popularity by accounting for multiple factors, including audience ratings, how many people viewed it, and whether it is still popular long after its release date. This information will be an accurate source of data to compare movies' popularity over IMBD ratings, which does not account for the number of people who viewed the film.

The source included two separate CSV files, including films' details and the other providing information about the cast. There was a common constituent "id" in both sets, which was used as a pivot to merge the two sets as a singular data frame effectively.

B. Research Questions

The Hypothesis is that there are common genres that are more likely to be more popular over others. Furthermore, casting and the film's budget will play a considerable role in influencing a film's popularity. Films' popularity then is likely to play a direct role in increasing revenue. There is an excellent source of data in Kaggle that provides the necessary data. The analysis should then either prove or reject the theories mentioned above and identify trends and correlations within the dataset to help relevant bodies decide on film matters. Following questions have been derived to form the scope of the analysis:

- Which factors manipulate a film's popularity and revenue
- Are there clear distinctive aspects, genres or casting a Production company should consider when choreographing a successful movie
- Do geographical, chronological data have an impact?
- Do the Manufactured models predict a film's acclaim accurately.

III. Analysis and Discussion (952 words)

As expected, there was a significant standard deviation in budget and revenue; this is possibly linked to the type of studio and country in which the movies were produced. An average movie had a popularity of 4 and revenue of 23,700,000 dollars.

The figure below describes the correlations for the whole dataset. Surprisingly there were no clear indications for popularity discounting revenue and budget. This visualization was reciprocated for the top 500 movies, and the results were no different from that of earlier analysis.

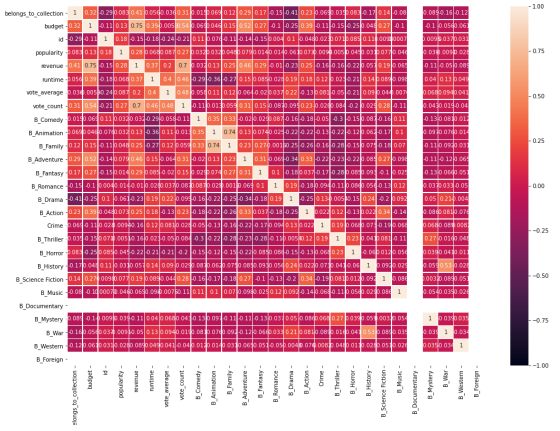

Figure 1. Correlation matrix of entire data set (*Heat map*)

A. Chronological Data

The number of movies produced has increased exponentially over time; from 2004 to 2013, the number of films released every year doubled. Technological advancements mean that movies are now both easier and cheaper to produce. There is an increase in revenue for the movie industry, which is also proliferating, as shown in figure 3.

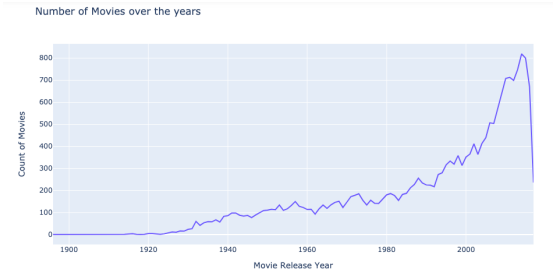

Figure 2.

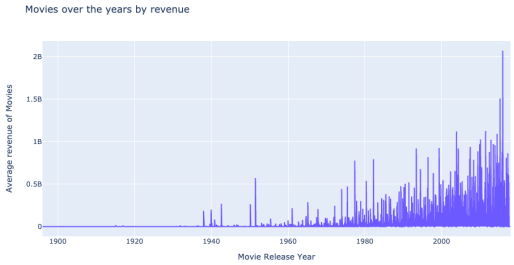

Figure 3.

An interesting observation from Figure 4 is that films released from May to July and November to December time-frames tend to generate more significant revenue. These periods coincide with the school holidays and summer and Christmas in many countries where the general public may have more free time.

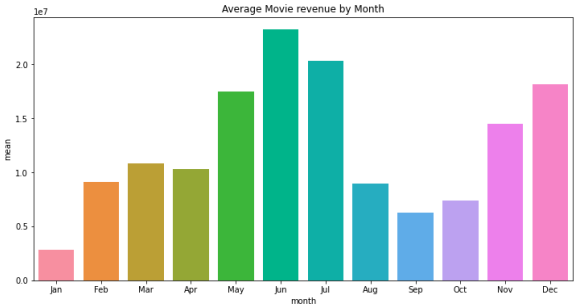

Figure 4.

B. Production Companies

Production Companies play a large role. It is clear that a well-recognized production company is expected to produce both a greater count of movies and expect a higher revenue; Warner Bros has produced 1244 movies by July 2017 with an average gross of 50 million dollars as displayed in Figure 5. A point to note is that companies focusing on animations or CGI appear to top the revenue charts. This suggests that these companies focusing on family and action genres, are likely to reap high rewards.

Figure 5 illustrates the top 50 production companies in terms of average revenue and the number of movies they produced.

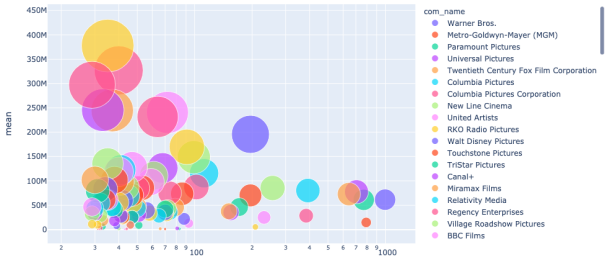

Figure 5.

C. Boxen plot between genres

The boxplot below indicates the spread of popularity of movies over genres. The Boxen feature is employed to highlight the distributions between the elements. Ultimately, Science Fiction has the best averages and maximum upper quartile; Science Fiction, Fantasy action movies have a tremendous potential to become a blockbuster. Most of the highest-grossing movies of all time fall under these categories.

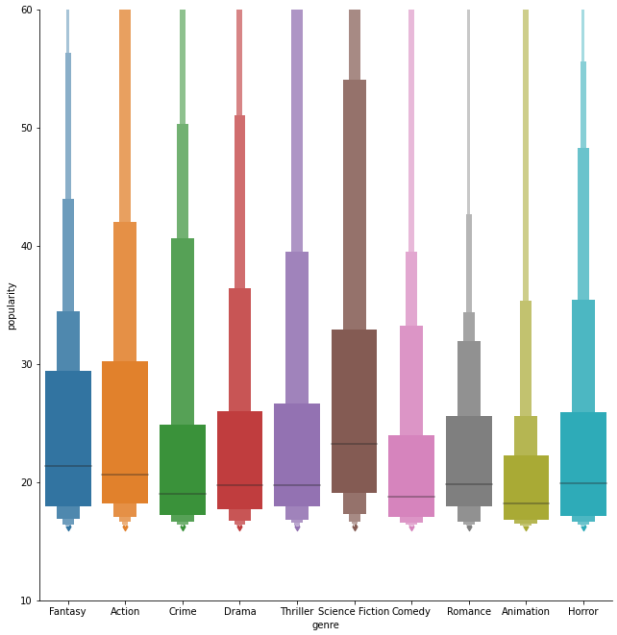

Figure 6.

D. Movies by country produced

Figure 7 shows the distributions of revenue by dollars for an average movie of the respective countries. There is a

clear trend that movies produced in western countries tend to produce more significant revenue. In general, the Anglosphere has a more excellent average than most, potentially suggesting that movies in English are likely to be watched more worldwide due to increasing English literacy in many countries. In Asia, China and Japan also lead the charts in terms of average movie revenue; China has the world's largest theatrical market.

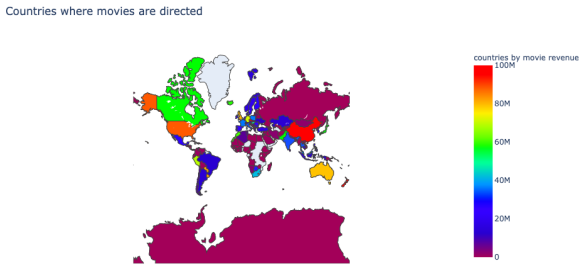

Figure 7.

E. Movie Collections

Through some exploratory data analysis, it is clear that there is indeed a link between movie collections and popularity; movies under collections have average popularity of 8.713 while movies without a collection have an average of just 4.012. Film directors realized that there is potential in making prequels, sequels, and replays as many fans for a particular movie will see the next continuation of the film. This is on display in figure 8, where series like Avengers and Harry potter are clear maximums.

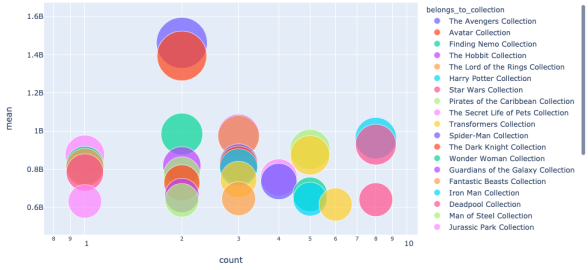

Figure 8.

F. Casting Influence

Casting drives the success of movies. There are clear distinctions between stars and their average movie revenue from the tables. Film studios who end up paying millions of dollars to stars over lesser-known actors seem to be driven by this idea. There are actors/actress like Emma Watson, who seems to do so consistently, as shown below. Therefore it can be vital to consider picking prominent artists.

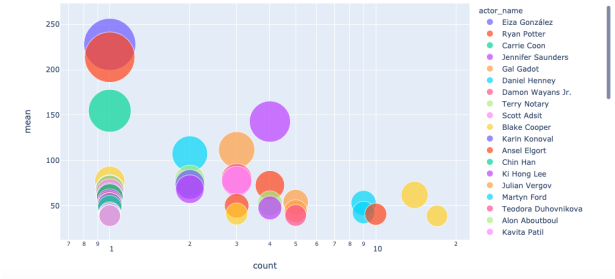

Figure 9.

G. Wordcloud

The Wordcloud generated highlights the keywords in the overview and visualizes the words that have been the most repeated. We will be using the top 500 most popular films for this model's scope only to identify key text for popular movies. This information can be used by relevant marketing teams when creating their overviews and taglines. World, find, life, and new are some of the most repeated words in the WordCloud below, implying Action or similar genres are likely more prevalent.

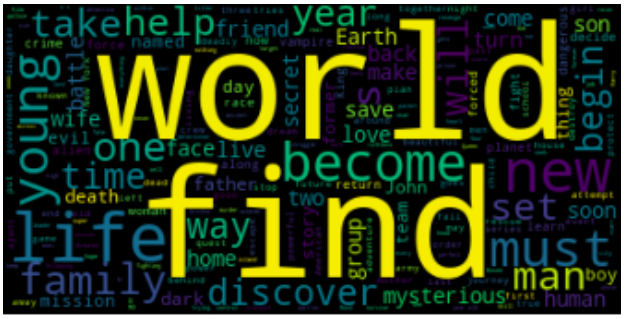

Figure 10.

H. Machine learning models

The principal metric that these models will predict is popularity. Two different machine learning models are used and compared: a Regressor model and a classifier model. For model fitting, the numeric data available from the dataset is interpreted. The string data of genres, production companies, cast, and the overview is evaluated to increase the models' accuracy and reliability. The text is Vectorised and later dimensionally reduced to allow for more efficient analysis.

I. Linear Regression

Linear Regression is a linear approach where a dependent variable is compared against one or multivariate other metrics, and a line that best fits the data is calculated. The Line is used to estimate y-values to the corresponding x values. Evidence from the correlation matrix contrived that there are indeed few correlation links between popularity and other values; therefore, it was a plausible idea to consider this approach.

The model had an accuracy score of 0.11, which indicates that 11% of test numbers were correctly predicted and an R2 score of 0.248. The results yield a low accuracy. A clear limitation is that the model can only consider linear relationships and they may not necessarily be causing each other to change.

J. KNN Classifier

KNN Classifier works by separating the data into classes and then using the predictor function to predict the new sample point classification. It assumes that similar data exist in close proximity. The model had an accuracy score of 0.45 and an R2 score of 0.179. The results yield a greater accuracy.

Further testing was carried out regarding dimensionality Reduction. Principal component analysis and neighborhood component analysis yielded relatively high accuracy shows visualized below. An increased number of classes from 2 to 5 also improved the accuracy of the model.

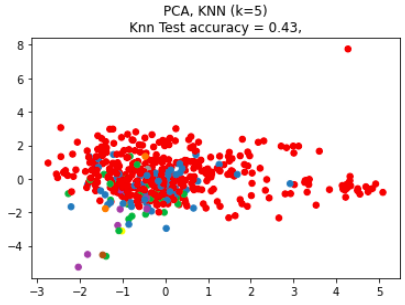

Figure 11.

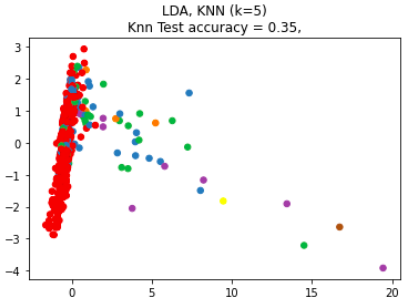

Figure 12.

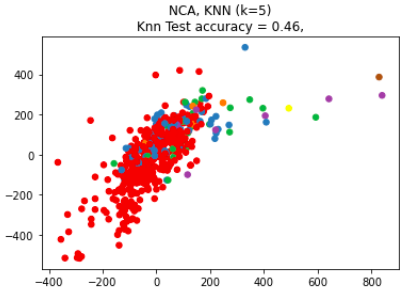

Figure 13.

IV. Findings and Reflection (472 words)

A. Future of the movie industry: analysis into Netflix

Another interesting point to touch upon is the ever-changing movie industry. It has evolved from cinemas and DVDs to online streaming services such as Netflix, which started to release their original movies. They started producing their content to keep subscription retention rates high, and it indicates a shift in terms of revenue from cinemas to an online platform. It becomes harder to predict how much a movie makes on these platforms as they profit from subscriber fees.
What is surprising is the sheer number of popular movies Netflix has produced over the years, such as Extraction and Irishman. According to an interesting article from Forbes, they have created accurate machine learning models from the data they collected from customers to predict movies' demands and identify the probabilities of success. This has given Netflix a clear edge in this industry. The simple truth is that The simple truth is that a directors inspiration does not make a Netflix film, but because a data model says it will work

B. Evaluation and Discussion

This analysis has provided vital statistics and visualizations has provided critical insights into the movie industry. It is an ever-increasing industry in terms of revenue, even if the count of movies produced is rapidly falling. Western-based popular production companies have the highest probability of producing a popular and profitable movie with genres like science fiction and adventure likely to reap the highest rewards. This is further evidenced by the fact that three companies produced the top 10 highest-grossing films globally in 2019 and all were adventure based with 6 being science fiction. Another interesting fact was that all these movies were released in the April-July or Nov-Dec period, further indicating the potential profitability in these periods. The analysis provides a deeper understanding about the movie industry and identify the major trends mentioned above.

Casting also has a significant influence, with crucial members consistently performing in a successful movie. This is evidenced by the fact that key actors and actresses appear more in popular movies than others. Hollywood recently is hamstring by a business model that relies on prequels and sequels being produced as they have a greater expectation to succeed as it is following a proven collection of movies. The statistics indicate that a movie in a collection is likely to be more popular than a stand-alone movie.

The machine learning models did yield a relative predictability success rate; The accuracy rate is just about 50%. There are clear limitations in terms of the volatility of the trends and the size of training set limitation, but it could still prove to be a valuable tool to determine at the start of the movie production whether or not the movie has the characteristics of being successful.

C. References

- https://towardsdatascience.com/how-to-vectorize-text-in-dataframes-for-nlp-tasks-3-simple-techniques-82925a5600db
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html)
- https://www.kaggle.com/priteshm/eda-movies
- https://www.forbes.com/sites/enriquedans/2018/05/27/how-analytics-has-given-netflix-the-edge-over-hollywood/?sh=56e2743f6b23.