Can the success of a movie be predicted?

Dorna Hamed Barghi

March 2023

1 Overview

Many factors affect the success of a movie. Considerations such as budget, duration, and genre will impact the perception of a movie and the revenue it generates. In this report, the aim is to examine the extent to which the aforementioned factors affect the success of a movie and determine whether or not it is possible to accurately predict the success of a movie before it is even released.
In order to answer these questions, a range of statistical techniques were applied including linear regression, Random Forest, and k-Nearest Neighbours, and visualisations of their results were created. The results of these have led to the conclusion that while it is not possible to accurately predict the success of a movie with the small data set used, and only taking into account revenue when looking at success, it would be possible to make such predictions with a more comprehensive data set and defining success by a combination of variables.

2 Introduction

Context and Motivation A lot of leading streaming services such as Amazon Prime Video, Disney+, and Netflix make use of recommendation systems to enhance the experiences of their users. Nearly every major streaming platform uses its own unique algorithm that combines human knowledge and machine learning to create a decision-making process that will guide the viewer. All of these streaming platforms are competing against each other for subscribers, hours of view time, and notoriety, as well as aiming to reduce their churn rate (the number of subscribers who cancel the service within a certain amount of time) as much as possible [1].
The Internet Movie Database (IMDB) has made available its data on movies released between 2006 - 2016 detailing: ranking, title, genre, director, actors, year of release, runtime, rating, votes, Metascore (scoring by critics) and revenue.
In this study, I will be using this data set to examine the impact of these factors on a movie's success, which I have defined as the movie's revenue, in order to answer the question, can a movie's success be predicted?

Related Work There has been a lot of work in this area, particularly from a marketing perspective. Ericson and Grodman's paper[2] has aims that are particularly similar to this study. They make use of SVM and locally weighted linear regression to examine and predict a movie's success. Both this paper and Joseph's work [3] conclude that while a small study cannot be used to accurately predict a movie's success, we can provide some interesting insight into the factors that impact revenue.

Objectives The objectives of this project are to determine the extent to which other factors in our data set affect the movie's revenue and whether or not we can accurately predict the success of a movie using machine learning and other statistical techniques.

3 Data

Data Provenance The data was obtained from Kaggle [4]. It is a cut-down version of the original dataset IMDB released containing information on 1000 movies. The data was downloaded to a local system for initial exploration and processed using Anaconda and DataSpell. The second dataset was downloaded after some exploration of the original IMDB dataset, also from Kaggle[5], in order to gain information on the budgets of the movies in the original dataset. Both of these datasets have a CC0: Public Domain license meaning anyone can copy, modify, and distribute the work, all without asking permission [6].

Data Description The IMDB data was formatted as a .csv file, containing records for 1000 movies released between 2006 – 2016. The following variables were used in the analysis:

1. Title
2. Genre - all genres the movie falls into
3. Runtime (Minutes)
4. Rating
5. Votes
6. Revenue (Millions)
7. Metascore - IMDB critics' rating

The remaining variables in the data, including actors and directors who worked on the movies, were not used in the analysis. The Kaggle site [4] has more information regarding these.
The second dataset encompasses a much larger amount of information on the movies, and lists about 1000 movies however the majority of the variables were not useful for analysis, and the only variable that was used in the end was the budget of the movies. Further information on the rest of the dataset is available on Kaggle [5].

Data Processing In order to clean the data, all rows with null values in any of the columns were dropped. Additionally, the 'Description' column was dropped as there was no intent of using it as part of my statistical analysis.
From the second dataset, the 'Budget' column was selected and an inner join was performed with the original IMDB data frame. The 'Budget' column was also changed to conform to the same format as the revenue column where the values are in millions.
Finally, due to the categorical nature of the 'Genre' column, the data was encoded using one-hot encoding. Note, only the first genre value in the list of genres per film was kept in order to simplify encoding the data.

4 Exploration and Analysis

Visualisations Visualisations are on pages 3 and 4.

Interpretation of the results From Figure 1 we can see that the majority of factors are not strongly correlated with each other. The most highly correlated factors are budget with revenue and Metascore with rating. The latter means that critics and the general public, to an extent, agree on how good a movie is overall.

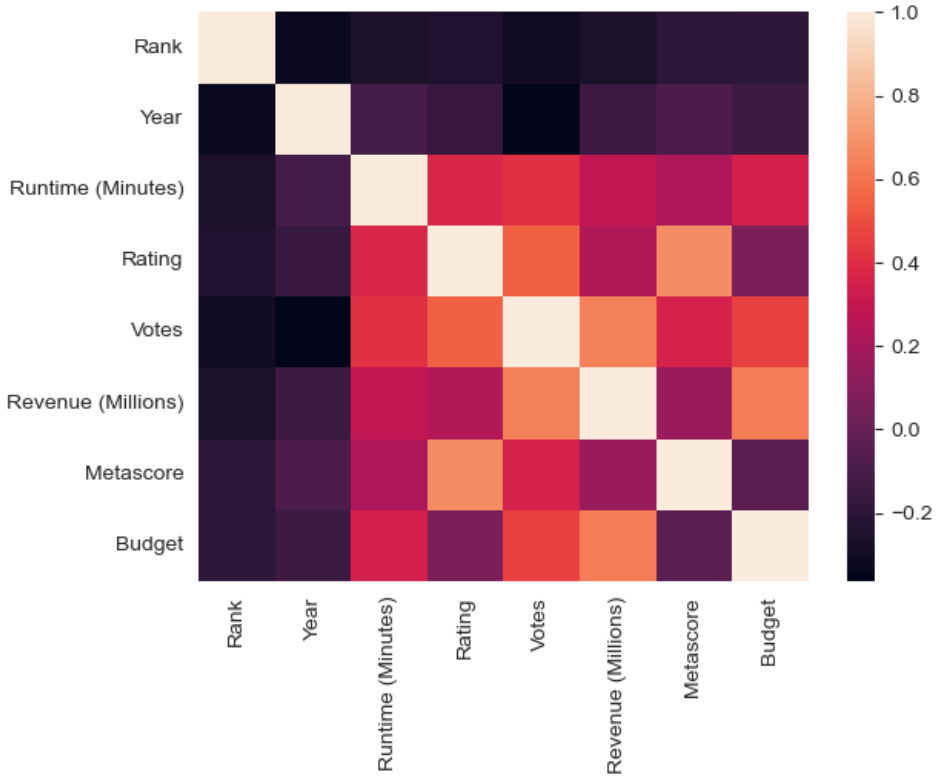

Figure 1: Heatmap showing correlation of factors. Made using .corr()

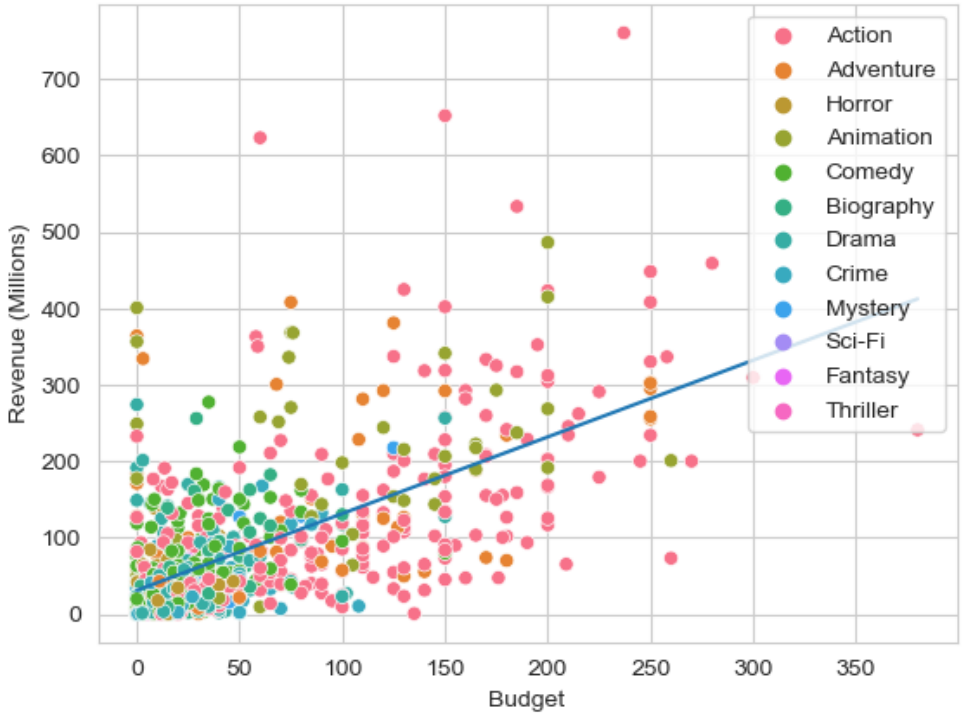

Figure 2: Linear regression to predict a movie's revenue using the budget. Mean Squared Error: 5898.989139257581, Root Mean Squared Error: 76.80487705385369, Adj. R-squared: 0.398

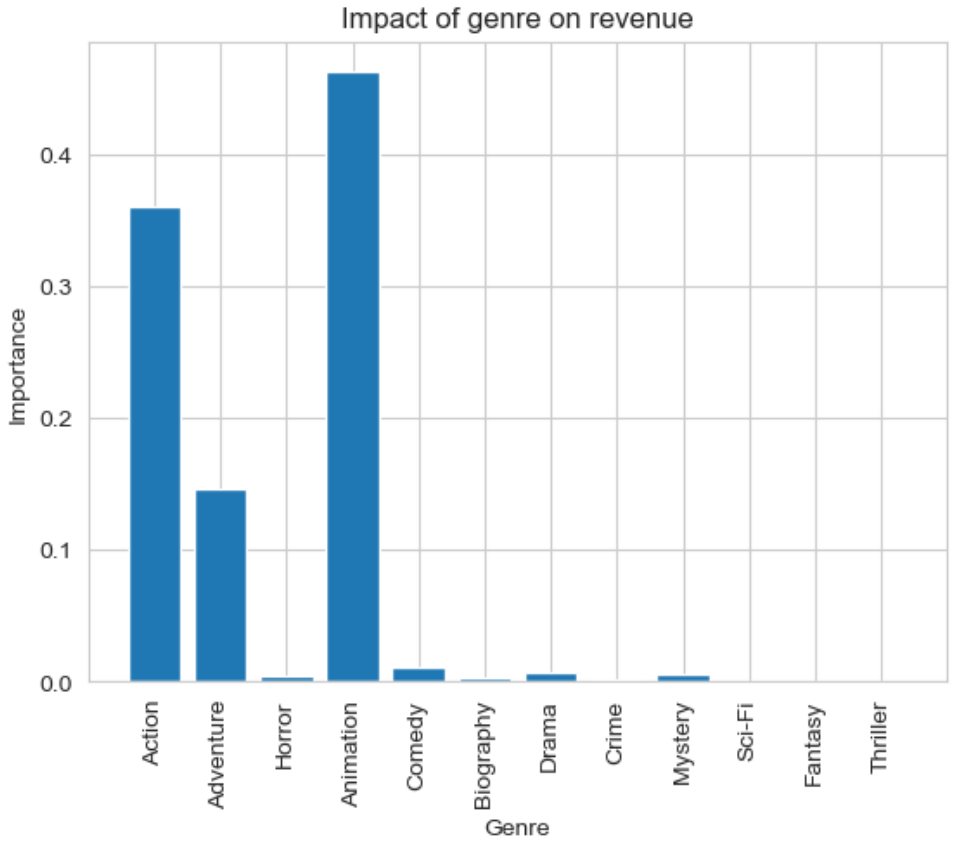

Figure 3: The importance of genre relating to revenue as calculated by applying Random Forest.

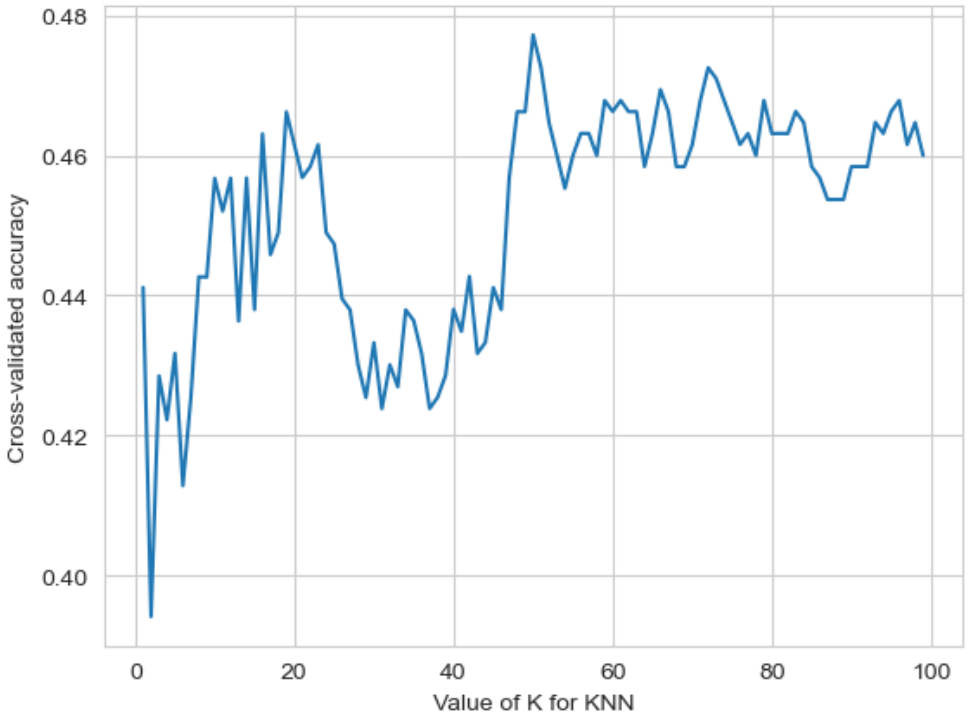

Figure 4: Cross-validation accuracy for values of k when performing k-NN

We look at the relationship between budget and revenue in Figure 2. From Figure 2 and the RMSE from the linear regression, we can see that budget cannot be used to accurately predict revenue - at least not on its own. However, it is also evident that the strength of the relationship between the two variables increases if we're looking only at certain genres, namely action films. The higher-grossing action films from cinematic universes such as the Marvel Cinematic Universe (MCU) typically have very large budgets, and intuitively we can say that films from franchises are guaranteed a level of success. If we then also take into context the time period this data set encompasses, where Marvel in particular was gaining popularity, we see that genre is an important factor in determining expected revenue.

Moreover, Figure 3 shows that different film genres have different amounts of impact on revenue. Animated films seem to be another genre that generates higher revenues. This might be because they typically cater to a wide audience encompassing a range of ages and interests. The same logic we used to explain the success of action films applies here. Animated films are often part of franchises and studios such as Disney have a guaranteed audience for a majority of their animated films - most likely due to their reputation and marketing.

In order to answer the question 'Can we predict a movie's success?', I made use of k-Nearest Neighbours (k-NN). Figure 4 shows the cross-validated accuracy of k values ranging from 1-100. We see that the peak accuracy is just under 0.48, showing that k-NN may not be a sufficiently accurate approach to predicting revenue. Other papers have made use of SVM [2] which seems to give better results.

Application of statistical method(s) Firstly, Ordinary Least Squares regression was used to examine how well the budget of a movie could predict revenue. This was done because it would be useful from a financial stakeholder's perspective to know if the budget can be a strong predictor of a high-grossing movie. Looking at the RMSE, I concluded that while as seen before, the two are highly correlated, budget on its own is not a reliable enough predictor of revenue.

k-Nearest Neighbours is one of the more basic classification algorithms in supervised machine learning. In this particular context, I wanted to see whether or not k-NN could accurately predict revenue using the budget, duration, and other numerical variables. I applied k-NN to the data and varied the k value (ranging from 1 to 100) in order to plot their accuracy (Figure 4). From this, we saw that optimal accuracy is 0.48 which happens when $k \approx 50$.

Random Forest, which was used in this report for analysing the importance of genre, is a commonly-used ensemble algorithm in machine learning. It combines the output of multiple decision trees to reach a single result[7]. The algorithm handles both classification and regression problems and I implemented it using Koehrsen's article [8]. Although Random Forest is not in the FDS course, I found it can produce some really interesting results and I would've found examining and visualising the impact of genres quite difficult otherwise. From my calculations, Random Forest has an accuracy of 0.72 when predicting revenue which is far better than k-NN. Looking at other articles I read while researching, it also seems to be the method that gives the most accurate results in comparison to other statistical techniques such as SVM.

Interpretation of Findings From my findings, I don't think the information in the IMDB dataset is sufficient for providing an accurate prediction of a movie's success. There is evidence that genre can be a strong indicator of expected revenue, but that is related to a typically higher budget than other genres, and the fact that certain genres typically belong to a franchise or are dominated by two or three well-reputed companies means these movies will most likely have a guaranteed amount of success.

5 Discussion and conclusions

Summary of findings In this report, I have found that a combination of factors results in a successful movie and that the success of a movie might also need to be a combination of indicators rather than just revenue. We have found that genre and the implications of a given genre - such as whether or not the movies typically belong to a franchise - play a significant role in predicting success. Looking at Netflix's algorithm [9], we see that genre is one of the main factors they look at when recommending a movie to their subscribers. Additionally, Netflix is known to have 'secret' categorization codes which subscribers can use in the Netflix search bar. No doubt, their algorithm also involves more specific categorizations, and users' activity and preferences are used to recommend them a movie.

Finally, looking at the main question I set out to answer, it is difficult to accurately and precisely predict the success of a movie with the given data. However, looking at the results of the Random Forest algorithm in particular, it could be possible to make such predictions given a larger data set, and taking into account marketing, the notoriety of the companies and people involved in the production of the movie (especially on social media), as well as other important factors.

Evaluation of own work: strengths and limitations In the study, I applied a variety of techniques and used a lot of the data available, particularly the numerical variables. I also looked at contextualising my findings by looking at the information available online, particularly when looking at the movie genre's impact. However, the data used is limited and outdated. It also lacks key information on marketing, franchises, etc. Additionally, k-NN is one of the more basic machine learning algorithms, and using another technique such as SVM might yield better and more accurate results.

Improvements and extensions As mentioned above, making use of a bigger, more comprehensive data set would enable a larger scope of analysis. Now more than ever, the production company and their reputation heavily affect the revenue of a movie. I also believe we should find a way to take into account the notoriety of actors and directors involved in a movie which no doubt will affect the initial buzz surrounding the release of a movie. A good example of this is Don't Worry Darling [10]. I also think more complex machine learning algorithms would be more effective and accurate. Finally, a big improvement would be to find a way to combine different indicators of success into one score, rather than looking at them individually, or choosing to only look at revenue.

References

[1] Devyn Hinkle. "How streaming services use algorithms". In: *Arts Management Technology Laboratory* (2021). Retrieved on 25 March 2023. URL: https://amt-lab.org/blog/2021/8/algorithms-in-streaming-services.

[2] Jeffrey Ericson Jesse Grodman. "A Predictor for Movie Success". In: (2013), p. 5.

[3] Sarah E. Joseph. "What Makes a Movie Successful: Using Analytics to Study Box Office Hits". In: (2019), p. 32.

[4] Ivan Gonzalez. "1000 IMDB movies (2006-2016)". In: (2022). Retrieved on 26 March 2023. URL: https://www.kaggle.com/datasets/gan2gan/1000-imdb-movies-20062016.

[5] Rounak Banik. "The Movies Dataset". In: (2018). Retrieved on 26 March 2023. URL: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset.

[6] "CC0 1.0 Universal (CC0 1.0) Public Domain Dedication". In: (). Retrieved on 26 March 2023. URL: https://creativecommons.org/publicdomain/zero/1.0/.

[7] Cloud Education IMDB. "Random Forest". In: (2022). Retrieved on 26 March 2023. URL: https://www.ibm.com/uk-en/topics/random-forest.

[8] Will Koehrsen. "Random Forest in Python". In: (2017). Retrieved on 30 March 2023. URL: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0.

[9] "How Netflix's Recommendations System Works". In: (). Retrieved on 26 March 2023. URL: https://help.netflix.com/en/node/100639.

[10] Harvey Austin. "Why Don't Worry Darling's Box Office Beat All The Drama Bad Reviews". In: (2022). Retrieved on 26 March 2023. URL: https://screenrant.com/dont-worry-darling-movie-box-office-success-explained/.