What Makes a Movie Successful?

Roberta Pošiūnaitė

October 7, 2023

1 Overview

The movie industry generates billions of dollars in revenue each year and has a significant cultural, political, and social impact. Movies are often described as successful or unsuccessful, but how can the success of a movie be measured? Many factors can contribute to how well a movie is received, whether it garners acclaim or generates a sizable revenue. This report focuses on the definition of a "successful" movie and how the success of a movie can be predicted by looking at a dataset from the Internet Movie Database (IMDb) covering a thousand movies released between 2006 and 2016. Machine learning techniques such as random forest and k-nearest neighbors were used to analyze the data. The study found that there are no specific factors that can predict the revenue of a movie. Additionally, it was determined that viewer and critic ratings cannot accurately predict the revenue of a movie. This study suggests that the success of a movie is more complex than just the revenue and ratings, and may be determined by many immeasurable factors such as marketing, timing, and audience preferences.

2 Introduction

Context and motivation With an annual revenue of about 77 billion US dollars [3] and hundreds of movies released each year, the movie industry has a significant impact on global and pop cultures, entertainment, and the economy. Movies can help shape our worldviews and beliefs, inform us about social and political issues, and help create a sense of community. With so many movies released each year, not all will be considered good, successful, or enjoyable, but how can the success or quality of a movie be measured? If a movie is considered successful, does it necessarily mean it's good and vice-versa? Is there anything production companies can try to do to ensure that a movie will be successful, or is there any way for them to predict a movie's reception? It seems that describing the success of a film is just as complex as cinema itself. But it is this complexity that makes it interesting to explore data about movies.

The most common way to look at the success of a movie is to look at its box office revenue, critical and audience ratings, and cultural impact [6]. However, these can sometimes be contradictory to each other. IMDb, or the "Internet Movie Database," is an online database established in 1990, holding information on over 400 million (as of December 2022) movies, TV shows, podcasts, and many other forms of entertainment [5]. With a collection of data obtained from IMDb on 1000 movies released from 2006 to 2016, we can look at how different factors determine the success of a movie and analyse whether the success of a movie can be measured by revenue and ratings.

Previous work Factors Affecting the Success of Movies - A Case Study of Twin Movies [7] by Niharika Sood and Prof. Balamurugan J is a case study that looked at four pairs of twin movies released between 1998 and 2013. Twin films are those that deliver a very similar story or plot. The study found that factors such as good promotion, the reputation of the actors, and the quality of the script affected the reception of movies.

Exploring the key success factors of films: a survival analysis approach [1] by Ahyun Kim, Silvana Trimi, and Sang-Gun Lee is a study that formed five hypotheses relating to how different factors affect the

number of screening days of a movie. The study found that positive critic comments, genre, ownership nationality, and age restrictions have an impact on the screening days of a movie.

Objectives The main question I aim to answer with this project is whether the way we measure the success of a movie a good approach. I will answer this by looking at some sub-questions:

- Do factors such as runtime, director or genre have an affect on the revenue?
- Do critics and viewers have similar opinions on movies?
- Do high reviews from critics and or viewers affect the revenue of a movie? Can their reviews be used to predict the revenue of a movie?

3 Data

Data provenance The dataset was uploaded on Kaggle.com by a user of the website [4]. It is not made clear how the user had gathered the data, only that it was a summary of the full dataset on movies released from 2006 to 2016 obtained from IMDb, which is hidden behind a paywall. I obtained the dataset by downloading the CSV type file from the website. The user has set the license of the dataset to "CC0: Public Domain," meaning anyone can "can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission." [2]. IMDb allows limited non-commercial use of their data. Since this project is for non-commercial use and the licence type is of public domain, this dataset can be used.

Data description The dataset had 12 columns: *Rank, Title, Genre, Description, Director, Actors, Year, Runtime (Minutes), Rating, Votes, Revenue (Millions),* and *Metascore*. Each column included a thousand records, except for *Revenue (Millions)* and *Metascore*, which included 872 and 936 records respectively. The *Genre* column consisted of one to three different genres, and the *Actors* column included four actors. The metascore is the weighted average of the scores assigned to a movie's reviews by a large group of world-renowned critics. It is calculated if at least four critics' reviews are collected. The higher the value, the more positive the reviews.

Data processing I removed the *Rank, Title, Description, Year*, and *Votes* columns since they are irrelevant to the questions I am answering, and *Rank* seems to just be an index. I also removed rows missing the *Revenue (Millions)* and or *Metascore* values. Since the *Genre* column held one to three values, I created a new column called *Main Genre* that only had the first genre mentioned and removed the *Genre* column. This way only the main genre of the movie would be considered. Since the *Actors* column included four actors, I followed the same steps and created a new column *Lead Role*. This way only the actor or actress playing the lead role would be considered.

4 Exploration and analysis

Interpretation of Results Figure 1) shows the importance of different features in improving the prediction of the random forest algorithm. The figure shows that the feature with the greatest importance is the runtime of a movie (in minutes). The action genre and the metascore and ratings, certain actors, such as Will Smith or Daniel Radcliffe, and directors such as David Yates and Rudley Scott, can also help predict the revenue of a movie more accurately.

Figure 2) shows the relationship between the metascore (the critics' rating of a movie), the rating (the viewers' rating of a movie), and the category of revenue that a movie falls into (categories being a range from very low to very high). The graph shows a clear correlation between critic and viewer rankings of a

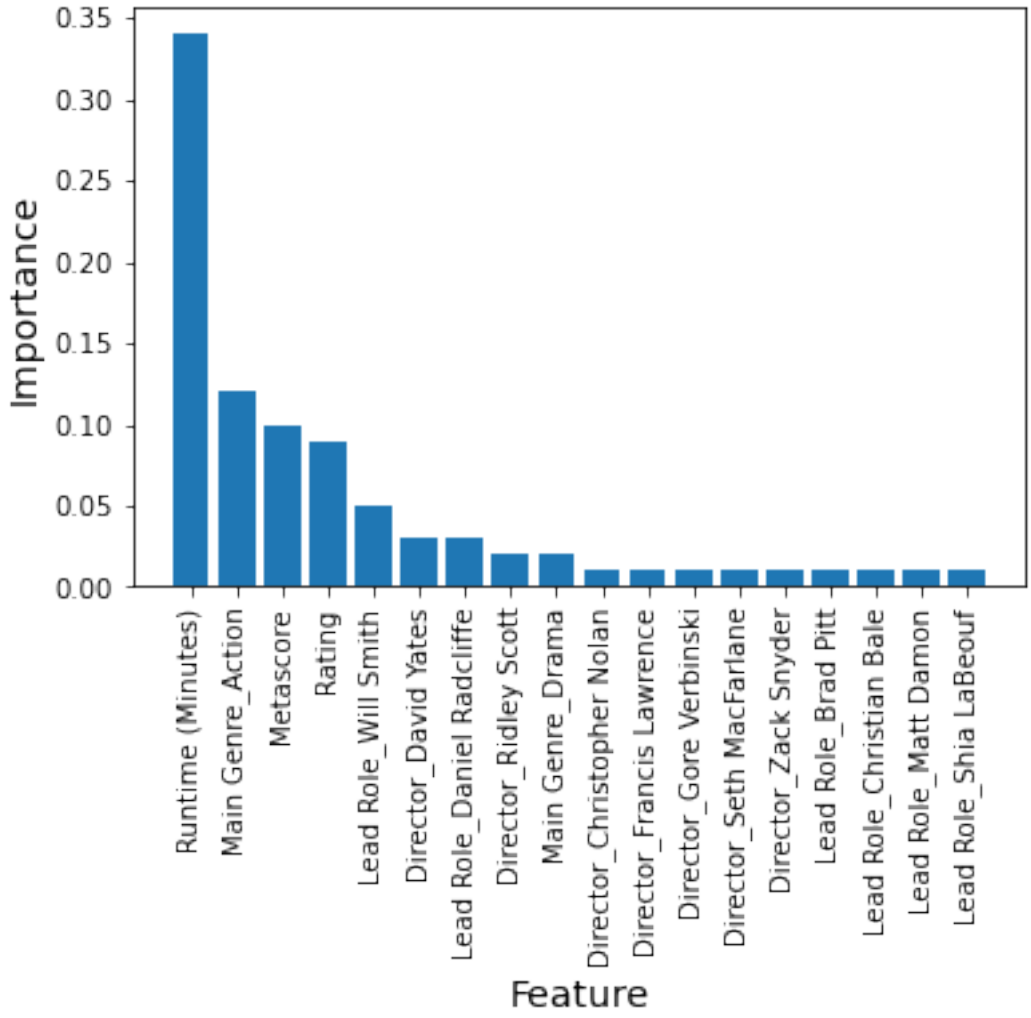

Figure 1: Importance of a feature, shows how much the inclusion of a feature can improve the predictions calculated by the random forest algorithm

movie. However, the category of revenue seems to be scattered, and not necessarily influenced by the ratings.

Figure 3 shows the relationship between the number of votes placed by users of the IMDb website (in thousands of votes) and the rating of a movie. There seems to be trend, that the more votes are submitted the higher the rating for a movie. The relationship is not linear, but seems to be more exponential.

Methods Used To determine how different features affect the revenue of a movie, I employed the random forest algorithm. I started by editing the data a little further. Since a director only having one movie in the dataset would not show a trend, I removed any rows of data with directors included less than three times in the dataset. I followed the same approach for actors in lead roles. This created a dataset with 155 rows. I then used one-hot encoding to encode the categorical data in the *Director, Main Genre* and *Lead Role* columns. Once the data was prepared, I separated the dataset into labels (revenue) and features (all other columns) for the algorithm to use. After running the random forest regressor, I found that it had a mean absolute error of 54.75 million dollars and an accuracy of 0.19%. To determine the features most improving the prediction of the model, I created figure 1.

To determine whether the revenue can be predicted by viewer and critic ratings, I used the k-nearest neighbors algorithm. For this algorithm, I removed all the columns from the dataset, besides the *Revenue (Millions), Ratings* and *Metascore* columns. I then created a new column *Revenue Category*, which held

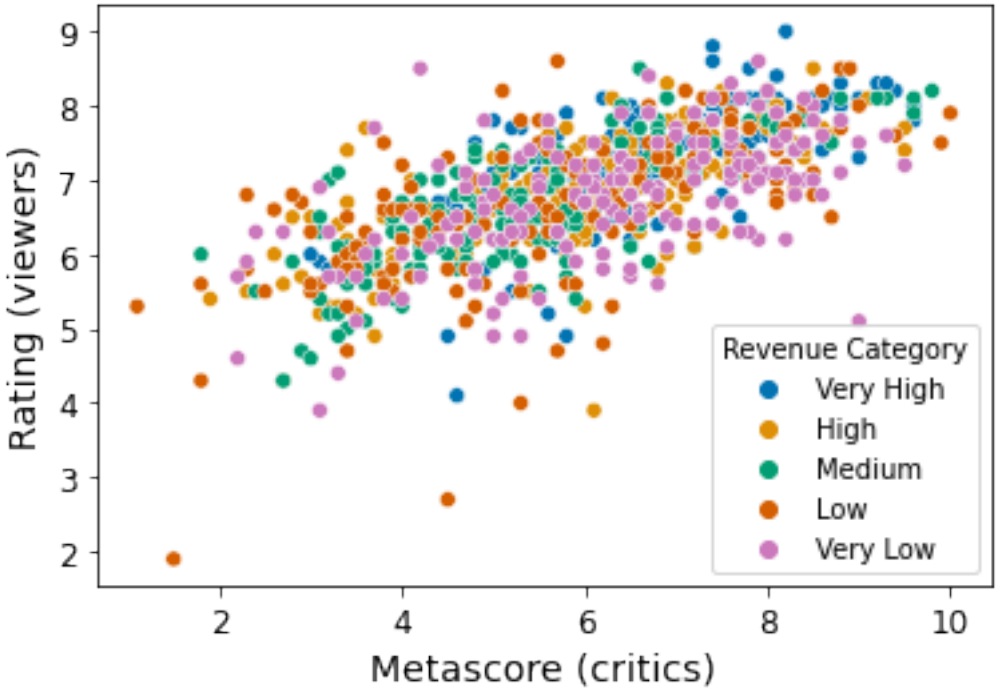

Figure 2: Relationship between metascore (the critics rating) and rating (the viewer rating) with the categories of revenue (from very low to very high) as the colours

the category of the revenue. I assigned the category by sorting the revenue in descending order, separating all the values into five sections, and assigning a category to each section. I then ran the k-nearest neighbors algorithm on the data. Every time it was run, it had an accuracy of about 0.25. I found that using particular values of k, the accuracy could be improved only to about 0.32 at best as shown by figure 4

Interpretation of Findings The inaccuracy of the random forest algorithm shows that the features provided in the dataset, are not sufficient enough to predict the revenue of a movie. Meaning that there may not be a large enough affect of any factor on how much revenue a movie generates.This is not too surprising, as renowned actors/actresses will have poor performances, different genres may be more popular during periods of time due to general trends in the media, and a big-name director or actor can bring in a lot of revenue, even if the movie itself may not be deemed good by critics or viewers. The low accuracy (about 25%) of the k-nearest neighbors algorithm shows that by using viewer and critic reviews we cannot accurately predict the revenue of a movie. If reviews cannot be used to predict revenue, these factors must not affect each other much. This means that looking at the success of a movie by looking at its ratings (both critic and viewer) and revenue is not a good approach. These factors can be contradictory, movies with high revenues may not be rated highly, and critics and viewers may not agree on how good a movie is. However, the trends seen in figure 3 show a trend that movies with more votes tend to have higher reviews. A larger number of votes means a wider variety of ratings. Thus, an average rating of a movie can be seen, with fewer outliers or individual opinions influencing the rating. However, movies that receive fewer ratings may show more individual opinions rather than the more generalized view. This trend may be affecting the found by the machine learning algorithms. Overall, these results show that the success of a movie is not something simply defined by revenue or ratings, and it cannot be predetermined by looking at measurable factors. Rather, the success of a movie is a complex combination of different variables that is tremendously elusive and nuanced.

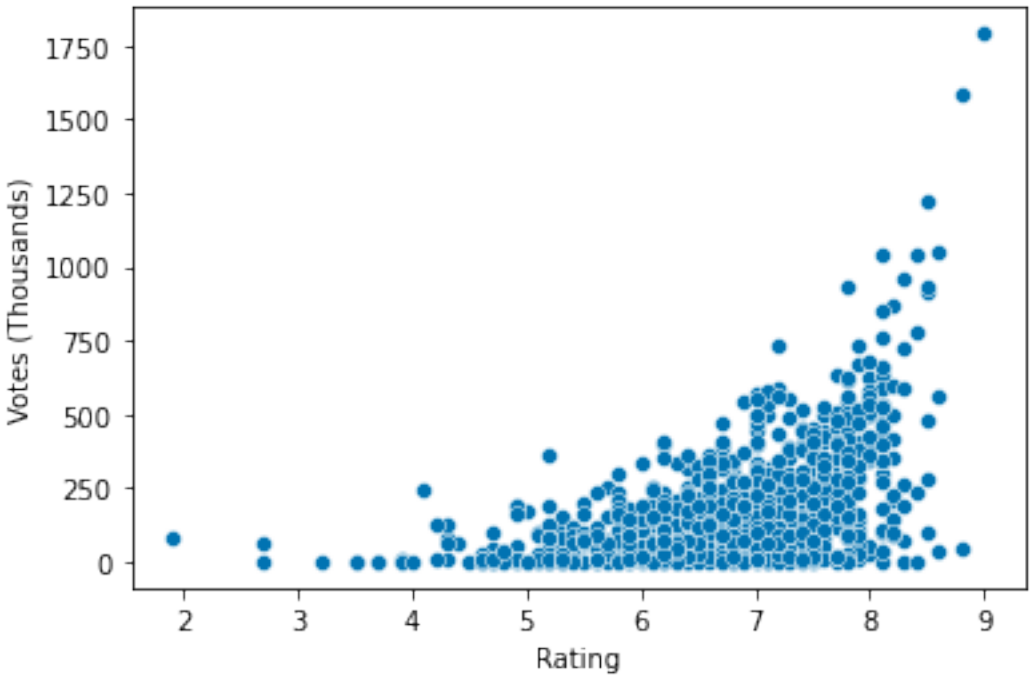

Figure 3: Relationship between rating (viewers) and the number of votes in thousands of votes

5 Discussion and conclusions

Summary of findings Using machine learning algorithms on the IMDb dataset covering movies released between 2006 and 2016, I have found that the way we determine the success of a movie is not a good practice. The random forest algorithm had a very low accuracy of only 0.19%, showing that the factors provided in the dataset were not sufficient in predicting the revenue of a movie. The k-nearest neighbor algorithm's accuracy of about 25% showed that the revenue of a movie cannot be accurately predicted using viewer and critic ratings, further showcasing how ratings may not determine the revenue of a movie. If both critic and viewer reviews and revenue do not have much of an effect on each other, then measuring the success of a movie using these factors is not a good practice.

Evaluation of own work: strengths and limitations The study presents clear visualisations that further illustrate values found by the machine learning algorithms and help support and demonstrate the conclusions drawn from the analysis. An interesting, multifaceted story that may not present desired outcomes, but highlights an interesting area within a subject that plays a large role in day to day life, is told. The limitations of study mainly lie in the lack of data, with a larger dataset, including more movies released in a larger time span, the data analysis could be significantly more accurate and interesting. We are unaware of exactly how the thousand movies for the dataset were chosen from the larger dataset that included all movies released in that time, and this could have presented selection bias in the dataset. Only looking at viewer rating on the IMDb website may be presenting bias too, since the dataset only considers the ratings of users of the website.

Comparison with any other related work Unlike both the studies carried out by Niharika Sood and Prof. Balamurugan J [7], and Ahyun Kim, Silvana Trimi, and Sang-Gun Lee [1], I found that the factors analyzed did not affect the revenue and, therefore, the success of a movie. This is not what was found in both studies. Niharika Sood and Prof. Balamurugan J factors such as a good cast, a well-planned budget, and good timing of release affected the success of a movie [7]. Besides the cast, these factors were not ones considered in my analysis. Ahyun Kim, Silvana Trimi, and Sang-Gun Lee found that the use of

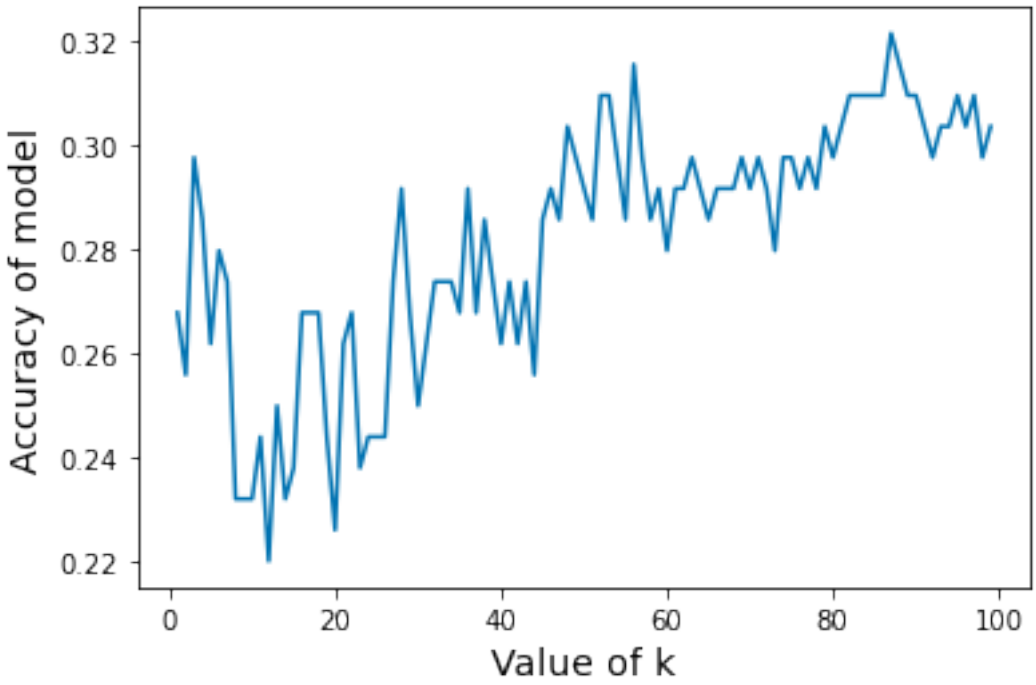

Figure 4: The accuracy of the k-nearest neighbours machine learning model for different values of k in the range of 1 to 100

specific language in reviews, genre, country of origin, and age restriction all impacted the number of screening days for a movie [1]. Besides genre, these are not factors I considered in my study. Ahyun Kim, Silvana Trimi, and Sang-Gun Lee found that movies that fell under the drama genre were likely to have more screening days than action or comedy movies [1]. I found that the inclusion of the genre of a movie, more specifically drama or action, helped more accurately predict the revenue of a movie.

Improvements and extensions I would have looked for a larger dataset, with information on more movies spanning a larger time frame. As well as more factors such as budget, length of screening, and awards won. It would also be interesting to look at a more international dataset, the movies in this dataset were mainly American or British movies, but looking at different markets and whether the measure of "success" fits those markets more could be an intriguing perspective.

References

[1] Sang–Gun Lee Ahyun Kim Silvana Trimi. "Exploring the key success factors of films: a survival analysis approach". In: *Service Business* 15.4 (2021), pp. 613–638.

[2] Creative Commons. *CC0 1.0 Universal (CC0 1.0) Public Domain Dedication*. URL: `https://creativecommons.org/publicdomain/zero/1.0/` (visited on 03/17/2023).

[3] Statista Research Department. *Film production worldwide – statistics & facts*. 2023. URL: `https://www.statista.com/topics/5431/film-production-worldwide/#topicHeader__wrapper` (visited on 03/17/2023).

[4] Iván González. *1000 IMDB movies (2006-2016)*. 2023. URL: `https://www.kaggle.com/datasets/gan2gan/1000-imdb-movies-20062016` (visited on 03/17/2023).

[5] IMDb. *IMDb Statistics*. 2022. URL: https://www.imdb.com/pressroom/stats/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=42d502de-837d-42ba-8d0a-7af777cde9b1&pf_rd_r=W9Q4VSTQ0EW55R5K18E7&pf_rd_s=right-2&pf_rd_t=60601&pf_rd_i=pressroom&ref_=fea_pr_pr_lk7 (visited on 03/17/2023).

[6] Adrian Johansen. *How to Measure a Film's Success*. 2022. URL: https://raindance.org/how-to-measure-a-films-success/ (visited on 03/17/2023).

[7] Prof. Balamurugan J. Niharika Sood. "Factors Affecting the Success of Movies – A Case Study of Twin Movies". In: *International Journal of Innovative Science and Research Technology* 2.11 (2017).