Historical trends in movies according to IMDb

Helen Wang, Jason Yu and Wanhua Feng

April 10, 2023

1 Overview

The stakeholder in the filmmaking industry is always keen to know the performance of the movie prior to its release. However, there is not a clear formula for what makes a movie successful, popular, or highly rated by users. In this project, our aims are to identify the factors or constituents that contribute to the success and popularity of the movies and how well these factors can predict revenues, popularity, and ratings of the movies. To do this, we have built several independent predictive models using pre-release data such as genres, actors, directors, runtimes, and release years to determine the constitution of each case. More specifically, the rating that was predicted does not take commercial success into account.

We have investigated the distribution of revenues and the number of votes (a proxy for popularity) to appropriately categorise the movie. The Random Forest algorithm was implemented with tuned hyperparameters for prediction and feature extraction. High accuracy was obtained: 0.821 for predicting revenue and popularity and 0.623 for the rating. To determine the success of movies, we have built a predictive model for classification that can successfully categorise movies into 3 grades based on revenue and popularity.

Additionally, we have used an AB test to determine if the runtime of the movie has a significant impact on revenue. Also, we have analysed the scarcity of some types of films based on the relationship between distribution and rating by genre.

2 Introduction

Context and motivation The film industry is a major global economic sector worth $43 billion[1], but not every movie is successful; millions of pounds were invested in the movie yet it can still lead to a flop at the box office. In parallel with the rapid growth of the market for streaming services [2] such as Netflix and Disney+, these leading streaming services heavily rely on recommendation systems to predict the movie preferences of their users.

It is critical to have an in-depth understanding of the constituents of movies to build predictive models based on the pre-existing data before the release of a movie. This understanding can be extended to modelling user preferences. IMDB is a recognised source of movie-related information, and with 83 million users registered[3], it can show a representative trend and pattern for the years 2006 to 2016.

This data science study explores factors that impact the success or popularity of the movie and detects if there is a hidden pattern for such a reason. We are intending to find correlation between the different variables and to discover some unfound behaviours. Given the average cost of making a larger film is $100 million[4], it is beneficial for different stakeholder groups(filmmakers,producers ..ect) to make accurate predictions at an early stage in order to minimize financial risks and avoid losses. The unexpected behaviour may also inspire the stakeholder to come up with new ideas and stay on track with trends. Upstreaming services can further improve their recommendation systems from this. Additionally, we will be applying machine learning techniques and statistical inference for further proof of our assumptions.

Previous work From the previous studies, many researchers have implemented machine learning techniques such as Random Forest, Gradient Boost, KNN, and SVM as the foundation of their predictive models for pre-released movie data. However, the best-performing model varied between studies. One study by R. Dhir and A. Raj [5]found that correlation analysis was not effective for categorical data types, and Random Forest had the highest accuracy at 61%. Another study by Bristi [6] addressed imbalanced datasets by implementing the SMOTE technique and using weighted features classification, and Random Forest performed the best again with a result above 90%. Interestingly, some researchers such as Quader[7] did not include genre and series as features in their models. This study claimed that predicting sequels was a challenge because the success of previous movies in the series can lead to false positives. Other studies[8]used multi-regression models and concluded that the success of a film could be accurately predicted using classical features.

Objectives As the previous studies suggest, we can build a predictive model that could accurately predict success using the current classical data such as genres, directors, runtime, and year of release. However, the underlying reasoning behind the model still needs further explanation. We aim to analyse the roles of factors and the model and see if there's a specific pattern within them. We also aim to investigate if there have been any trends in moviemaking over the years and analyse the impact these have had on popularity and revenues. This will reveal a great overview of the dataset.

3 Data

Data provenance The dataset for our project was obtained from Iván González (Contributor) on the Kaggle website[9]. Although the direct source of the dataset was promptCloud. This dataset contains the top 1000 IMDB movies from 2006 to 2016, based on data from IMDB. The data was made available under the CC0 1.0 licence, which is a public domain dedication that allows anyone to use the data to the extent permitted by law without needing to ask for permission. Therefore, it is both legally and ethically permissible to use this dataset for our project.

Data description The IMDB dataset contains 1000 movies spanning the year 2006 to 2016 with 12 variables:

Numeric Variables	Rank, Year, Runtimes (Minutes), Ratings, Revenues (Millions), Votes, Metascores
Non-numeric Variables	Title, Genre, Description, Directors, Actors

Table 1: IMDB variables

'Ratings' represent average of the movie on a scale of 0 to 10, based on user rating. 'Votes' contains the number of users who votes the rating. 'Revenue (Millions)' shows the domestic revenue in millions of dollars. Finally, 'Metascore' represents the which is a weighted average of critic reviews on a scale of 0 to 100. Revenues (Millions), Ratings, Votes and Metascores are post-release movie data. It was shown there is 128 missing values in Revenue (Millions) and 64 missing values in Metascores. On non-numerical variables, 'Genre' holds list of multiple genre(s) for each movie, separated by a comma, the same applies for 'Actors' as well. 'Title' contains the name of the movie; 'Description' provides short description of the movie.

Data processing We have dropped all the NA values(128 missing values in Revenue and 64 missing values in Metascores). This reduced the number of movies to 838. Merging data is not an option from other sources as inconsistency of the data occurs for different same metrics. For instance, some other datasets use global revenue. Filling data with the mean or median is also not an option since it cannot

provide an accurate representation of the movie's data. Based on the previous studies, we decided not to use genre as a feature to build our model. This left the options of Title, Description, Directors, Actors, Runtimes (minutes), and Year. Here, Actor and Directors are more useful. To transform the categorical data (director and actors) into a measurable metric, we have calculated the historical movies average revenues participated by the actor as an indication of the star power, then summed up all the actor values for the movies overall star power. Similarly, votes are an indication of each star's popularity. We have done the same to indicate director power and director popularity. Hence, new variables, act_avrg_rev, act_avrg_votes, drct_avrg_rev and drct_avrg_votes were created and used for the predictive model, along with year and runtime.

4 Exploration and analysis

Several factors can influence a movie's success, such as its director, cast, duration, and release year. To conduct an overall evaluation, we developed a random forest classifier model to forecast a movie's success. To prepare categorical variables like the director and actors for the model, we converted them into numerical variables by assigning weights based on their average revenue and voting scores.

4.1 Using Random Forest model to predict the commercial success and artistic rating

Predict the commercial success To begin with, we define a movie's commercial success by examining its box office earnings and level of popularity, as indicated by the number of votes it receives. To classify movies, we categorise them into three groups: A, B, and C. Our classification criteria is based on the top 30% quartile of revenue (95.297 million) and number of votes (236184.59). According to our definition, movies that exceed the benchmark in both revenue and number of votes are ranked as A. Those that exceed the benchmark in either revenue or number of votes, but not both, are classified as B. Movies that do not meet either benchmark are placed in category C. By utilising a random forest classifier, we can predict a movie's ranking based on its characteristics. After splitting the dataset into a training set of 70% and a test set of 30%, our classifier achieves an accuracy of 0.821. The precise confusion matrix, which illustrates the correlation between the predicted and true labels, is presented in Figure 1.

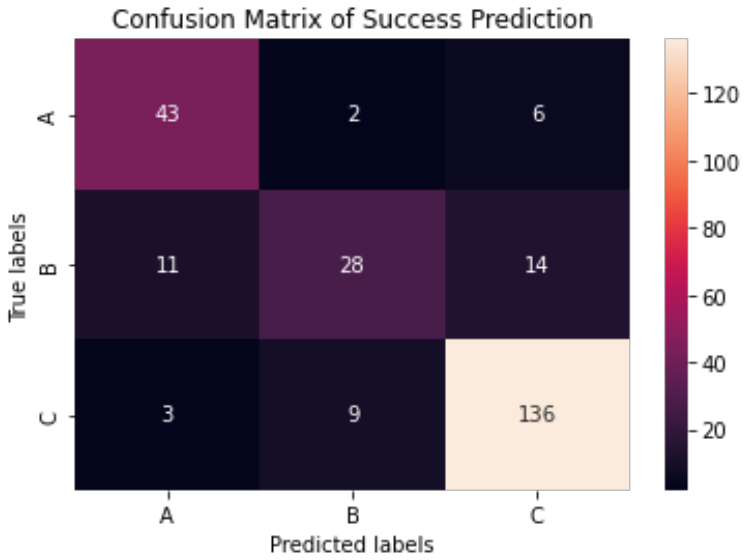

Figure 1: Prediction Confusion Matrix of Success

In addition, Figure 2 demonstrates the Random Forest model's overview of the features that contribute significantly to a movie's success.

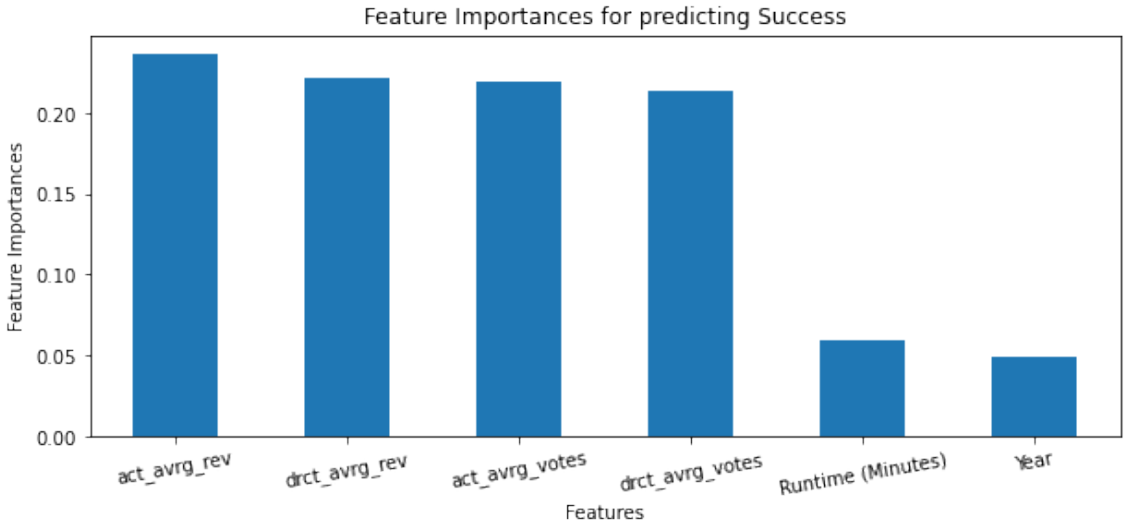

Figure 2: Feature importances for predicting the success. drct_avrg_rev standings for average revenue for director; drct_avrg_votes for average votes for director; act_avrg_rev for cumulative average revenue for all actors; act_avrg_votes for cumulative average votes for all actors;

Predict the artistic rating In addition to commercial success, it is also crucial to examine artistic success. For this purpose, we consider the audience rating and metascore, with the benchmark set at 69.0 for metascore and 7.3 for rating. We adopt the same methodology and criteria used for determining commercial success to evaluate artistic success. Our classifier achieves an accuracy of 0.623, with the corresponding confusion matrix shown in Figure 3 and the feature importances chart illustrated in 4.

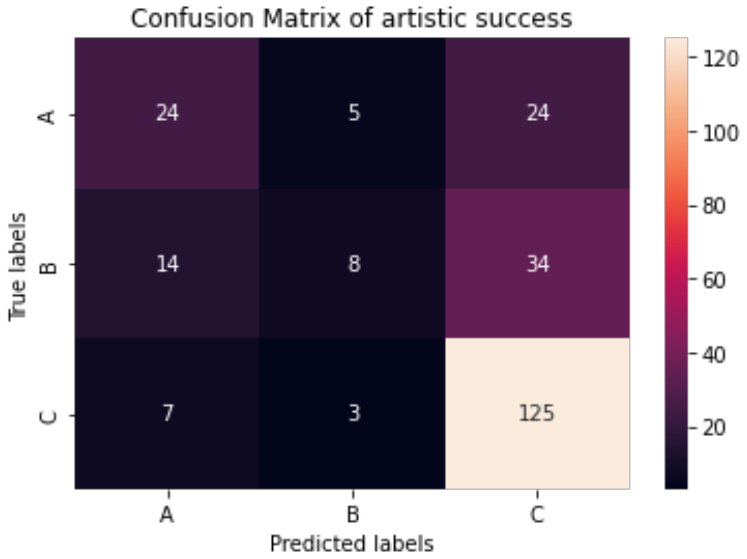

Figure 3: Prediction Confusion Matrix of Artistic Success

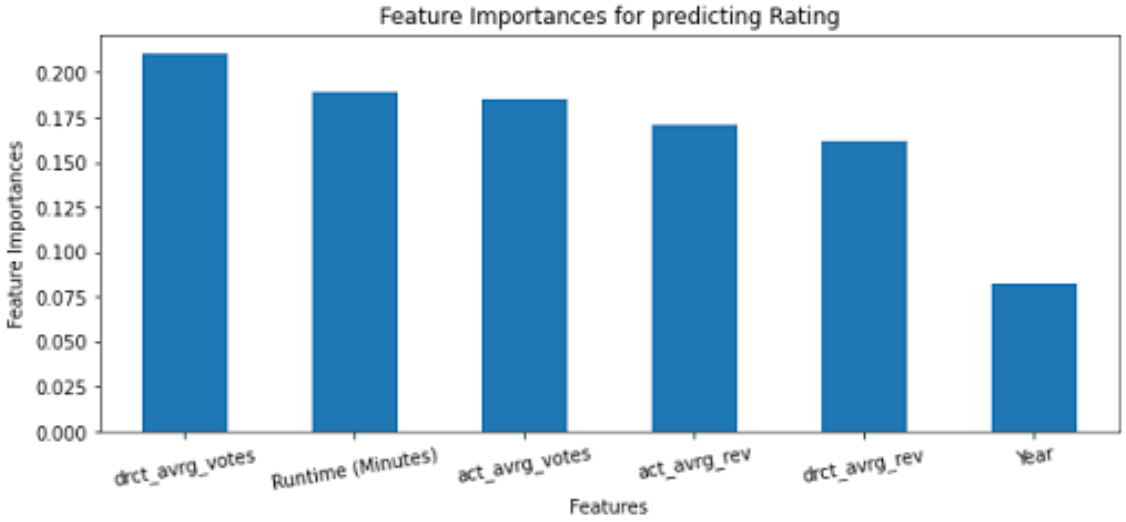

Figure 4: Feature importances for predicting the rating. drct_avrg_rev standings for average revenue for director; drct_avrg_votes for average votes for director; act_avrg_rev for cumulative average revenue for all actors; act_avrg_votes for cumulative average votes for all actors;

4.1.1 Interpretation of the result and the difference between commercial and artistic success

We noticed that directors and actors are consistently deemed highly important in both commercial success and rating prediction. This finding implies that movies with a talented director and cast may have a greater likelihood of success in the future. This is more pronounced in the case of predicting commercial success. In contrast, runtime appears to be a crucial predictor of ratings. This is because the runtime is a characteristic of a movie's artistic merit.
Additionally, Figure 5 illustrates a positive correlation between votes and ratings. This suggests that the presence of talented directors and actors is likely to be a common factor contributing to a movie's high popularity and rating.

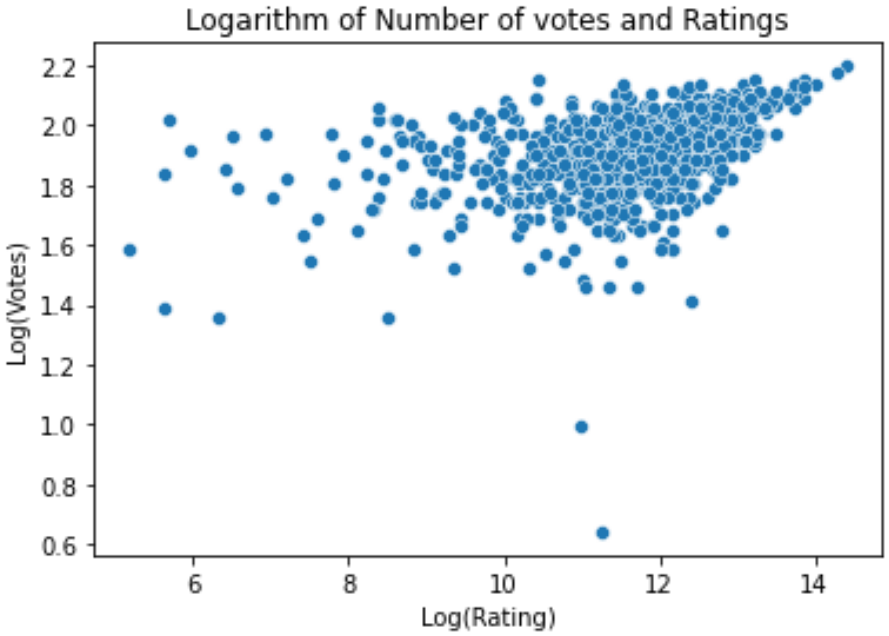

Figure 5: Relationship between Votes and Ratings after Log Transformation

4.2 Using A/B testing to investigate the relationship between runtime and success

As previously mentioned, runtime is a crucial factor in the success of the movie industry. In recent times, it is believed that feature-length movies have an advantage in terms of achieving higher box office earnings and ratings. Therefore, we aim to investigate whether feature movies perform better using A/B testing. However, the definition of a feature movie's length is not standardised. For our analysis, we will use 120 minutes as a reference point, and any movie that has a runtime exceeding 120 minutes will be considered a feature movie. We will also use the same benchmark as the random forest model: movies with box office earnings exceeding 95.297 million dollars will be classified as high revenue movies, while movies with ratings higher than 7.3 will be classified as high rating movies. The remaining movies will be categorised as either low revenue or low rating movies, respectively. Our null hypothesis is that "there will be no significant difference in performance between feature movies and non-feature movies," while our alternative hypothesis is that "feature movies will perform significantly better than non-feature movies." We will set a 95% confidence interval for our analysis.

Commercial success The High Revenue proportion is shown by Figure 6. As a result of A/B testing, we get a z score of - 6.84, a p-value < 0.001, a 95% confidence interval for control group (Non Feature Movies): [0.189, 0.259] and a 95% confidence interval for the treatment group (Feature Movies): [0.395, 0.512].

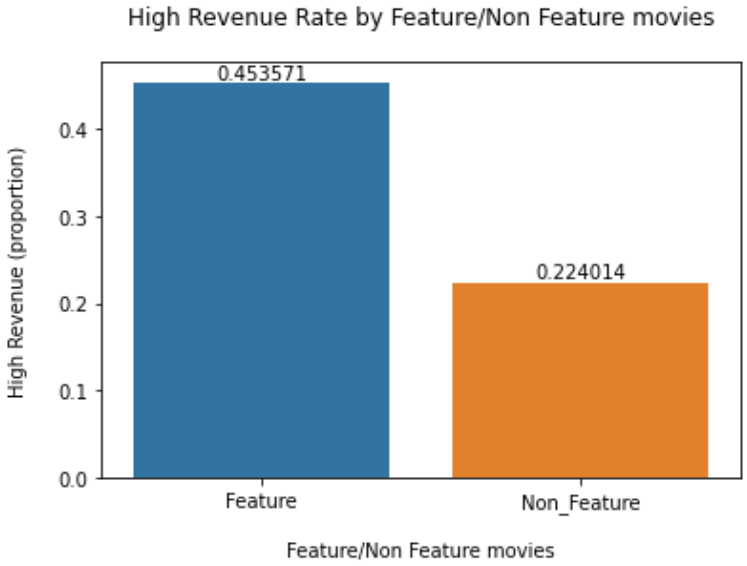

Figure 6: High Revenue Rate by Feature and Non Feature Movies (in proportion)

Artistic success Figure 7 shows the High Rating rate among all the movies. As a result of this A/B testing, we get a z score of -7.95, a p-value < 0.001, a 95% confidence interval for control group (Non Feature Movies): [0.205, 0.276] and a 95% confidence interval for the treatment group (Feature Movies): [0.456, 0.573]

4.2.1 Interpretation of the result

From Figure 6 and Figure 7, it is evident that there is a significant difference between the performance of Feature and Non-Feature movies, in terms of both box office and rating. This observation is supported by the statistical analysis, where the p-value for both cases is less than 0.001, which is well below our set threshold of α =0.05. Hence, we can reject the null hypothesis and conclude that Feature movies perform significantly better than Non-Feature movies. Moreover, the confidence intervals show a substantial difference of 20%, which further reinforces the significance of this conclusion.

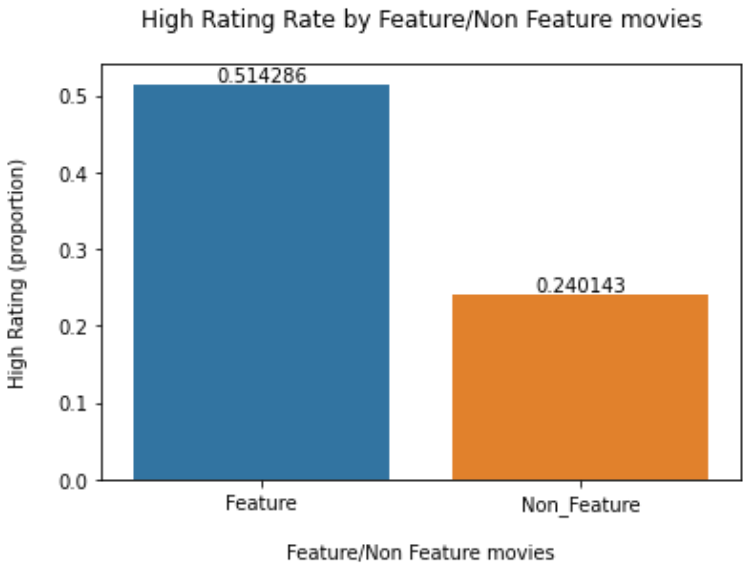

Figure 7: High Revenue Rate by Feature and Non Feature Movies (in proportion)

4.3 Understanding Genre differences based on their statistical data

In our previous analysis, we discovered a positive relationship between votes and ratings for movies. While some movie genres are more popular than others, not all movies in the popular genres are guaranteed to be successful. Typically, only movies with relatively higher ratings are well-received by audiences. Hence, it appears that other factors come into play, such as movie scarcity. When a new emerging genre is introduced and there are only a few movies within that genre, the competition within the genre is reduced. Consequently, movies belonging to that genre may have a higher chance of success.

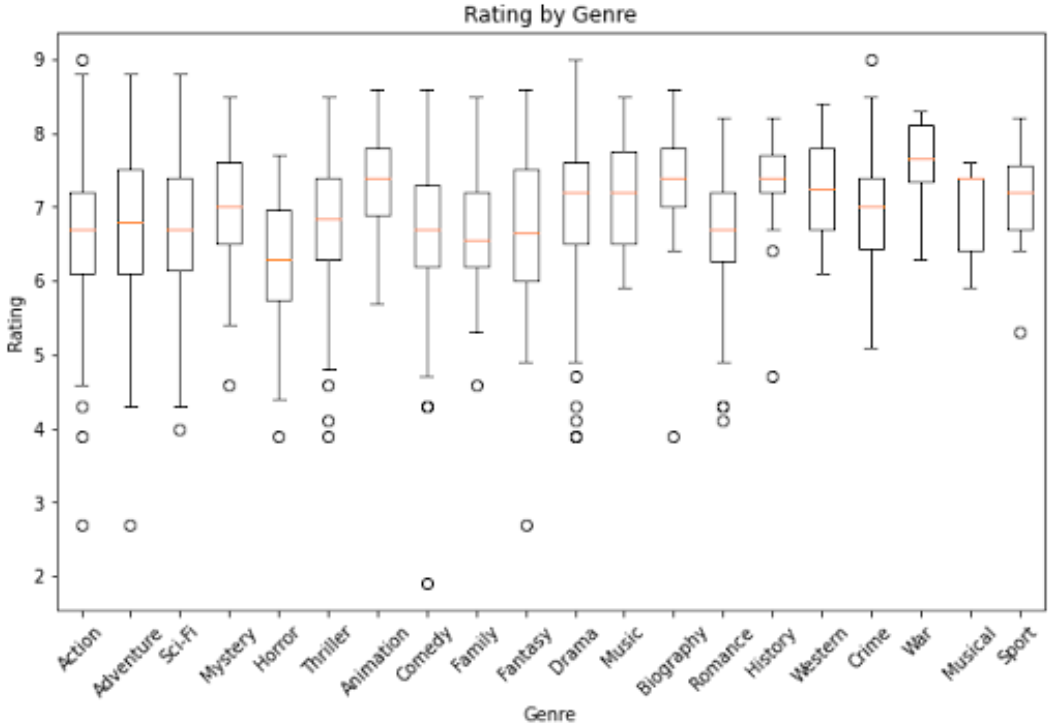

Figure 8: Distribution of Ratings for each Genre

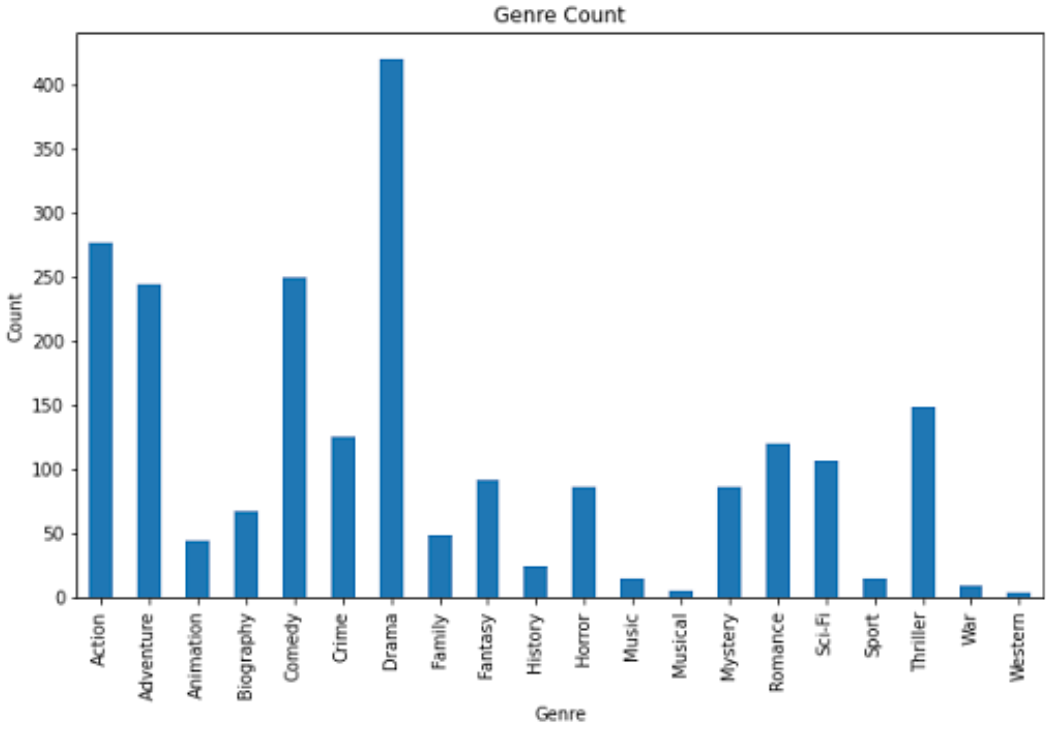

Figure 9: Movie count for each genre

Interpretation of the result An evident observation of Figure 8 is that the rating of Horror and Romance movies is relatively concentrated in a lower rating range in comparison to other genres. Meanwhile, the rating of other popular genres such as Adventure, Drama, and Actions are concentrated in a higher rating range. In addition, the movie count for Horror and Romance movies are all relatively low, while the count for Adventure, Drama and Action movies are relatively high.

This suggests that Horror and Romance movies face less competition and have lower quality counterparts. As a result, a new release in the Horror genre may only need to have a rating of 7 or 8 to surpass other movies in the same genre. When audiences would like to select a horror movie, it may be primarily selected due to the scarcity of Horror movies and its excellent performance in comparison with others. Therefore, there may be an advantage in popularity. In contrast, a new release in the Drama genre with a rating of 7 to 8 may not stand out as much because there are many other movies in this genre. This makes it harder for the audience to notice it when looking for a Drama movie.

However, we cannot ignore the large box office for the popular genres. Therefore, choosing the genre to achieve success depends on personal interpretation of 'success'.

5 Discussion and conclusions

Summary of findings In this study, we have identified several key findings. Firstly, we examined various factors that could affect a movie's popularity, commercial success, and ratings. We developed several random forest classifiers that yielded impressive accuracies of 0.823 and 0.621 in predicting a movie's commercial and artistic success. These classifiers showed that the director, actors, and runtime were important contributors to a movie's success. In addition, we used A/B testing to specifically investigate the impact of runtime on a movie's success, and found that Feature movies (movies longer than 120 minutes) tended to be more profitable and popular.

Furthermore, we used box plot charts and bar charts to explore the distribution of different genre ratings and movie competition. Our analysis revealed that movie genres with low competition (i.e., fewer movies and thus higher scarcity) and a low concentration of ratings in the lower quartile tended to have a higher chance of success in the future. Overall, these findings provide valuable insights into the factors that contribute to a movie's success and could inform future strategies for filmmakers and producers.

Evaluation of own work: strengths and limitations The study's most significant advantage is the application of a Random Forest classifier that outperforms regression ML methods in terms of prediction accuracy. Regression techniques rely on the number of features to achieve precision to some extent. Meanwhile, the classification system can encompass more information in a single degree of information. For instance, the initial random forest model integrates box office data and number of votes, among other factors. Classification benefits are extended to the A/B testing method, which provides a clear answer about choosing 'Feature' or 'Non-Feature' movies.

However, the study is subject to certain constraints. For instance, converting lists of directors and actors into weighted revenue and weighted votes relies on a primary assumption and may influence the result of prediction. In real life, actors and directors have many more properties and labels, and evaluating the popularity of different actors requires more research. Additionally, the classification system could be more refined, as we used arbitrary 30% quartiles as a benchmark. This is partly due to the limited database. Furthermore, the number of instances in the dataset is relatively low to conduct a reliable A/B test. Finally, the genre analysis is primarily based on graph observations, which might result in relatively inaccurate conclusions.

Comparison with any other related work Our dataset is split relatively evenly between AB testing and reasonable movie categories. There is no need to apply an extra method like Bristi [6] does. The model using Random Forest performs quietly well with 0.83; this aligns with their results, which are all over 60% [6][5], and we have verified that accurate predictions can be made. Though it has been suggested by [8]that incorporating factors like social media can further improve our model. It is widely acknowledged that star power and director power have an impact on a movie's success[10], but not many studies unravel the trends bewteen the factors. The findings on runtime can be valuable for filmmakers as they develop their films.

Improvements and extensions In order to improve and expand upon the current study, there are several areas of focus for future research. One of the most important improvements is to expand the database beyond its current limited scope of mostly categorical features. If additional databases could be acquired for categorical variables such as actors, directors, and genre, the classification system could become more meaningful and representative. Alternatively, methods such as dummy variables could be used to deal with categorical variables. Additionally, other features that could potentially influence a movie's success, such as the country of release, exact release date, and production cost, should be incorporated into the analysis for a more complete conclusion.
In terms of machine learning and statistical methods, although the study has already attempted to find optimal hyperparameters for the random forest classifier, a more thorough analysis of these parameters could prove helpful. For example, running the search for hyperparameters more times could yield a more efficient classifier. Moreover, including a larger number of instances by expanding the database would improve the reliability of the A/B testing results. Finally, to increase the reliability of the study's final analysis for the genre, additional accuracy checks could be performed to arrive at a more precise conclusion.

References

[1] *25+ striking U.S. film industry statistics [2023]: Facts about the video production industry in the U.S.* Mar. 2023. URL: http://www.zippia.com/advice/us-film-industry-statistics.

[2] Caitlin Huston. *Don't expect streaming revenue to keep up its rapid growth rate*. June 2022. URL: https://www.hollywoodreporter.com/business/digital/streaming-revenue-to-keep-up-its-rapid-growth-rate-1235169023/.

[3] *Press room*. URL: https://www.imdb.com/pressroom/stats/.

[4] Annie Mueller. *Why movies cost so much to make*. Aug. 2022. URL: https://www.investopedia.com/financial-edge/0611/why-movies-cost-so-much-to-make.aspx.

[5] Rijul Dhir and Anand Raj. "Movie Success Prediction using Machine Learning Algorithms and their Comparison". In: *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*. 2018, pp. 385–390. DOI: 10.1109/ICSCCC.2018.8703320.

[6] Warda Ruheen Bristi, Zakia Zaman, and Nishat Sultana. "Predicting IMDb Rating of Movies by Machine Learning Techniques". In: *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2019, pp. 1–5. DOI: 10.1109/ICCCNT45670.2019.8944604.

[7] Nahid Quader et al. "A machine learning approach to predict movie box-office success". In: *2017 20th International Conference of Computer and Information Technology (ICCIT)*. 2017, pp. 1–7. DOI: 10.1109/ICCITECHN.2017.8281839.

[8] Anand Bhave et al. "Role of different factors in predicting movie success". In: Jan. 2015, pp. 1–4. DOI: 10.1109/PERVASIVE.2015.7087152.

[9] Iván González. *1000 IMDB movies (2006-2016)*. Jan. 2023. URL: https://www.kaggle.com/datasets/gan2gan/1000-imdb-movies-20062016.

[10] Anita Elberse and Assistant Professor. "The Power of Stars: Do Star Actors Drive the Success of Movies?" In: *Journal of Marketing - J MARKETING* 71 (Jan. 2007). DOI: 10.1509/jmkg.71.4.102.