Historical trends in movies according to IMDb

Jennifer Gao, Karen Zheng, Fiona Yu

April 11, 2023

1 Overview

This report analysed the features that can be used to describe and predict movie revenue, the Academy Award for Best Picture, and the distribution of movie ratings based on the movies released from 2006 to 2016. The data was cleaned, and we incorporated additional columns, including the budget column. This study begins by selecting the factors that would influence revenue. A multiple linear regression model was trained to predict the revenue, the model has an adjusted R-squared of 0.718, with the budget being the continuous variable having the highest coefficient of determination and a positive impact on revenue. A random forest model is used to identify the most significant variables for predicting whether a movie can be nominated for Best Picture, the model ended with scores of sensitivity of 0.98 on the training set and 0.6 on the test set. Finally, the chi-square test is applied to verify the difference in distributions of ratings between IMDb rating and Metascore, the reason for the difference was further explained based on the rating mechanism. Firms seeking to maximise their profitability and reputation can use our results to inform strategic decisions to optimise profitability and popularity by understanding the underlying mechanisms.

2 Introduction

Context and motivation The emergence of streaming platforms like Netflix, Hulu, Amazon, Disney+ and others producing original television and movies, movies are perceived to be underestimated compared to television in the entertainment industry. We may learn more about the elements that influence a movie's success by examining data from a variety of sources. The Academy Award for Best Picture is one of the most esteemed honours in the movie business. It is one of the main recognition to highlight how successful the movie is. As a result, it raises questions about the factors that underlie a movie's success. With the ever-growing demand for new movies and the diversity of movie genres, runtimes etc. can offer valuable insights into what factors can forecast the revenue of a movie and what features can make the movie a nominee for Best Picture.

Previous work There has been a growing interest in analysing the trends in the movie industry using machine learning techniques. Previous research has explored the factors that influence movie revenue and Oscar award nomination features. For instance, a paper from Krauss, Jonas; Nann, Stefan; Simon, Daniel; Fischbach, Kai [1] investigated the impact of the movie industry and predicting movie success and academic awards, also find the rating can also be used as an indicator to predict the success of a movie. Similarly, a study of Kaggle by Saba Tavoosi [18] used regression to predict box office revenue and found that factors such as genre, production budget, and movie duration had a significant impact on revenue.

As a study published in 2012 conducted a regression model to examine the relationship between the genres and the resulting revenue boost. The conclusion states that 'action' and 'horror' have negative coefficients with the overall grossing [20].

Objectives This study aims to offer valuable insights into the 'successes' of the movie industry. In particular, this study aims to address the following questions:

- Which variables could be used to predict the revenue of a movie?
- What are the reasons that we can not predict the revenue of a movie well based on the existing variables?
- Which features are important in predicting whether a movie could be nominated for Best Picture?
- How do the IMDb ratings and Metascore distributions differ, and why there is such a difference?

3 Data

Data provenance The main dataset we used was downloaded from Kaggle and was scraped by IVÁN GONZÁLEZ from IMDb [11].

An additional dataset(we refer to it as the TMDb dataset) we used was from Kaggle provided by RITAYAN DHARA [6], the data was scraped from themoviedb.org(known as TMDb), and its API is available and free for everyone to use for non-commercial purposes.

The second additional dataset(the Oscar Award dataset) on Kaggle includes information about all types of Oscar Awards and was provided by RAPHAEL FONTES [8].

The last additional dataset(the Best Picture dataset) from Kaggle including information about Oscar Best Picture Movies is provided by MARTIN MRAZ. [15].

All datasets above are available under the CC0(Creative Common Zero). As described by creativecommons.org, "You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission. "[4].

Data description The main dataset includes information about the 1000 movies(records) that were released between 2006 and 2016, it has the following columns :*Rank*, *Title*(movie title), *Description*, *Director*, *Actors*, *Genre*, *Year*, *Runtime (Minutes)*, *Rating*, *Votes*, *Revenue (Millions)*(in dollars) and *Metascore*. For *Rank*, the meaning was not provided on Kaggle, and we didn't use this column in our study.

The TMDb dataset includes information about 10000 popular movies(records), it has the same variables as the main dataset, the only additional variable we used for merging is *Budget*.

The Oscar Award dataset includes 10800 records about Academy Award winners and nominees from 1927 to 2023, the variables we used in this dataset are the following: *category*(the type of the award. e.g. Actor, Actress, or directing, etc), *film*(name of the film), *name*(the name of the nominee or winner), *year_film*(the release year of the film).

The Best Picture dataset includes 571 movies(records) that were nominated for Oscar Best Picture from 1928 to 2020. The only variables we used are *Film*(name of the film), and *Year of Release*.

Data processing In our main dataset, we split the column *Year* into 11 different year columns: *2006*, *2007*, *2008*...and *2016*. If the movie was released in 2006, then the value on column *2006* is 1, and 0 for the rest of the year columns.

Similarly, for column *Genre*, we get all genres from this column and create a new column for each genre. For example, the movie *Avatar* has genres of Action, Adventure and Fantasy, its values for the columns: *Action*, *Adventure* and *Fantasy* are all "1", the values for the rest of the genre columns are all "0".

We removed the duplicates from all additional datasets. We merged the main dataset with the TMDb dataset, the Oscar Award dataset, and the Best Picture dataset based on the release year and movie title. From the TMDb dataset, we got the budget for each movie and we created a new column *Budget (Millions)*.

We created a new column *hasOscarActorNomineeWinner* to indicate whether any of the main actors in this movie was a nominee or a winner for Oscar's Best Actor/Actress in a leading role before the release year of the movie. In this column, we use "1" to indicate True and "0" to indicate False.

We also created a new column *OscarDirectorNomineeWinner* to indicate whether the director of the movie was a nominee/winner before this movie was released.

From the Best Picture dataset, we created a new column *OscarPictureNomineeWinner* which contains only '0' and '1'. '0' means that the movie was not nominated for Best Picture, and '1' means that the movie was a nominee or a winner.

For missing values, whenever answering a particular question, we choose some of the columns that will be involved for this question, and then we drop the rows if a missing value appears in any of these columns.

4 Exploration and analysis

Firstly, we studied which factors could be used to predict the revenue of a movie before its release. We found that the *Revenue (Millions)* in the dataset corresponds to the domestic box office [22].

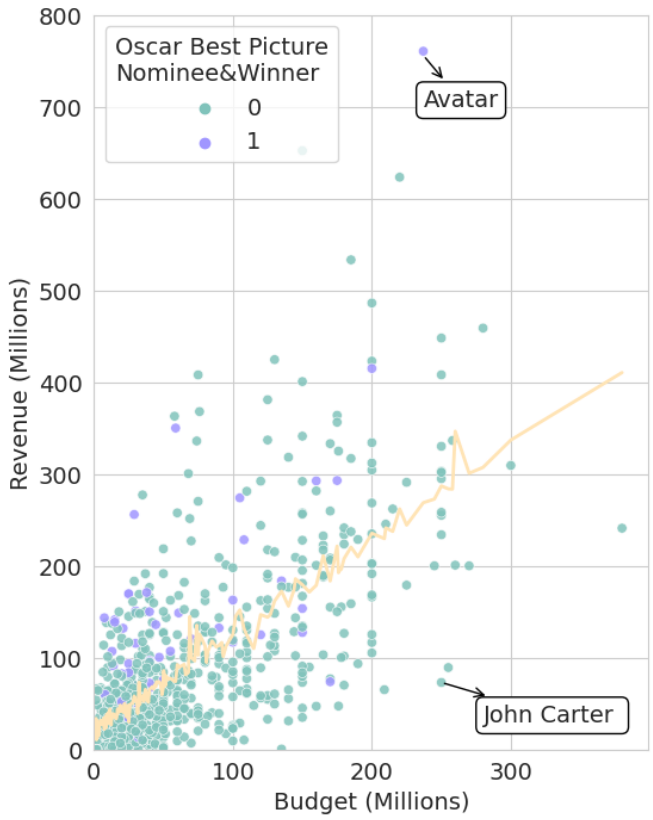

Figure 1: The golden line indicates the predicted revenue using the variables in Table 1. The purple dots are movies that were nominated for Best Picture in Oscar Awards. There are 2 outliers: *Avatar* and *John Carter* which deviate most from the predicted revenue.

The continuous variables in this dataset were *Budget (Millions)* and *Runtime (Minutes)*, so we calculated their correlation coefficients with revenue. We found that the revenue had the highest correlation coefficient with the budget, which was 0.678.

Figure 1 shows the relationship between revenue and budget. We can see from this graph that the general trend is that as the budget increases, the revenue also increases accordingly. However, as the budget gradually increases, the data begins to spread from the centre to both sides, indicating significant heteroscedasticity and suggesting the existence of luring variables.

Using the OLS(ordinary least squares) class provided by Statsmodels, we trained a multiple linear regression model to predict revenue. We used the following variables to train the model: *Budget*, *Runtime(Minutes)*, *OscarDirectorNominee*, *hasOscarActorNominee*, all year variables(we excluded a year column *2006* since we need to avoid dummy variable trap), and all genre variables. Each variable was raised to the first power. We also had a null hypothesis for each variable, which was that the variable had no effect on revenue, this means that we assume the coefficient of each variable was 0, and we set the significance level to 0.05. We used backward elimination for feature selection, therefore, when the p-value of a variable was ≥ 0.05, we would conclude that it had no significant effect on the revenue and remove the variable.

After the first training, we eliminated any variables with p-values ≥ 0.05, including the intercept. Then we trained the model again on the remaining variables, this time, every remaining variable has a

Dep. Variable:	Revenue (Millions)	**R-squared:**	0.720
Model:	OLS	**Adj. R-squared:**	0.718
Method:	Least Squares	**No. Observations:**	735

	coef	**std error**	**t**	**P > |t|**	**[0.025**	**0.975]**
Budget (Millions)	1.0483	0.061	17.322	0.000	0.929	1.167
Runtime (Minutes)	0.3143	0.053	5.954	0.000	0.211	0.418
Animation	59.6796	12.076	4.942	0.000	35.972	83.387
Adventure	-17.0571	7.626	-2.237	0.026	-32.028	-2.086
Drama	-15.7547	6.226	-2.530	0.012	-27.978	-3.531
Action	-13.5077	6.402	-2.110	0.035	-26.077	-0.939

Table 1: OLS summary results from statsmodels.

p-value ≤ 0.05. The results of the second training are summarised in Table 1, and we can only reject the null hypothesis of variables included in this table. The adjusted R-squared of our model for the remaining variables was 0.718. However, it is evident that the actual revenue gradually deviates from the golden line(predicted revenue) as the budget increases. Therefore, our model cannot perform well on high-budget movies. One piece of evidence is that our RMSE is 69.84 million, which is very huge. The main reason for this is that the RMSE was scaled up by the outliers.

Table 1 revealed that both budget and runtime have a positive impact on revenue, as movies with longer duration tend to have higher production costs and therefore attract more audiences and generate higher box office. We only retained 4 genres in our final model: Animation, Adventure, Action, and Drama. We found that only Animation has a higher impact on revenue, possibly due to its wider audience appeal. Parents often accompany their children to the cinema, which contributes to higher box office revenues, even though Animation is primarily targeted at children.

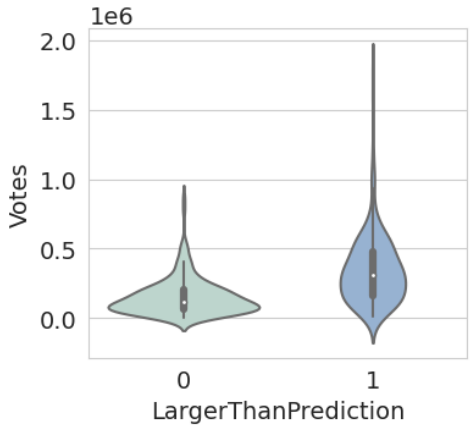

Figure 2: The distributions of votes on the group(left) that has real revenue < predicted revenue(left), and the group(right) that has real revenue > predicted revenue.

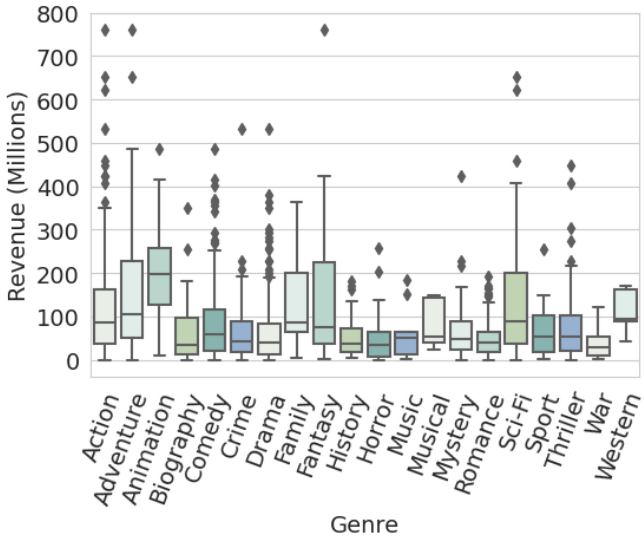

Figure 3: The range of revenue in each genre

From Figure 3, we can see that the range of revenue for Animation is generally higher, while Adventure has a broad range, but its minimum revenue is relatively low, and the ranges of revenue for Action and Drama are also lower. These are all reflected in the coefficients of the variables in the model, but in fact, there are no significant differences in the ranges between many genres. After we performed bootstrapping on the entire dataset and refit the linear model again, we found that the p-values for these genres changed significantly. One major reason for this is the limited quantity of our dataset. Besides, each

movie has more than one genre. Therefore, when we break down the original *Genre* into individual genre variables, we often encounter different genres sharing the same movies making it difficult to distinguish genres from each other and leading to the difficult prediction of revenue.

We created a new variable *Difference*, which is calculated by $Difference = real revenue - predicted revenue$. If the real revenue is larger than the predicted revenue, the difference is a positive number, and vice versa. We sorted the data based on *Difference* in ascending order and calculated the 75th percentile and the 25th percentile. Furthermore, we divided the data into 2 groups, one group has *Difference* $\geq$ the 75th percentile and the real revenue is larger than the prediction; the other group has *Difference* $\leq$ the 25th percentile and their revenue is all smaller than the prediction. The two groups are shown in Figure 2, the left group with real revenue smaller than prediction has generally lower vote counts compared to the right group with real revenue larger than prediction. Vote count can to some extent reflect the movie production company's promotion efforts, as well as the quality of the movie itself. This implies that the popularity of the movie can also affect its revenue.

To further analyse factors that make it difficult to predict the revenue of some movies, we specifically selected 2 movies(see Table 2), *Avatar* and *John Carter*, which have the largest positive and the largest negative differences, respectively.

The director of *Avatar* is James Cameron, a famous director with extensive experience in the movie industry [16]. In *Avatar*, he utilised innovative and new visual effect techniques, including an advanced 3-D camera for motion-capture. Additionally, the mature 3-D technology contributed to the higher ticket price [9]. The theme of *Avatar* is about the relationship between humans and the natural environment, which has been a popular topic in recent years, also increasing in popularity. *John Carter* is a science fiction movie about Civil War-to-Mars. Its failure at the box office has especially been criticised on its marketing: its trailer didn't convey its Civil War elements to attract young people [7]. In addition, after the movie was released, the rating was not excellent and therefore failed to win more audiences.

Taken together, the revenue of a movie depends on various factors, including the fame of the production crew and actors, promotion, budget, runtime, genre, rating etc. In particular, the ground-breaking special effects technologies used in movies have significantly boosted the box office, as they can create a captivating and innovative experience for audiences.

Title	Genre	Director	Year	Rating	Revenue (Millions)	Budget (Millions)
Avatar	Action,Adventure, Fantasy	James Cameron	2009	7.8	760.51	237.0
John Carter	Action,Adventure, Sci-Fi	Andrew Stanton	2012	6.6	73.06	250.0

Table 2: 2 movies which deviate most from our prediction.

Additionally, we are interested in which factors can help us identify a movie that is likely to be nominated for Best Picture. This award is a valuable indicator of popularity and success.

When observing the data presented in Figure 4, it becomes clear that certain movie genres are more likely to win an Oscar than others. The genre of Drama, which has won the most awards compared to other genres, has a higher probability of winning an Oscar. Conversely, horror movies are the least popular and less likely to win an Oscar.

From Figure 5, the distribution of the Metascore of nominated movies is prominently higher than the distribution of unnominated movies, this suggests that we could use Metascore as an index to help us recognise a nominee.

Additionally, Figure 1 shows that the distribution of revenue of nominees mainly concentrates between 50 million dollars and 200 million dollars.

In order to verify our observations, we used the Random Forest algorithm to predict which movies would be nominated for Best Picture. A "random forest" is composed of many decision trees. Whenever we ask the Random Forest classifier to predict whether a movie would be nominated, each decision tree

in the forest will classify the movie accordingly, and the class with the most votes becomes the model's final prediction [5].

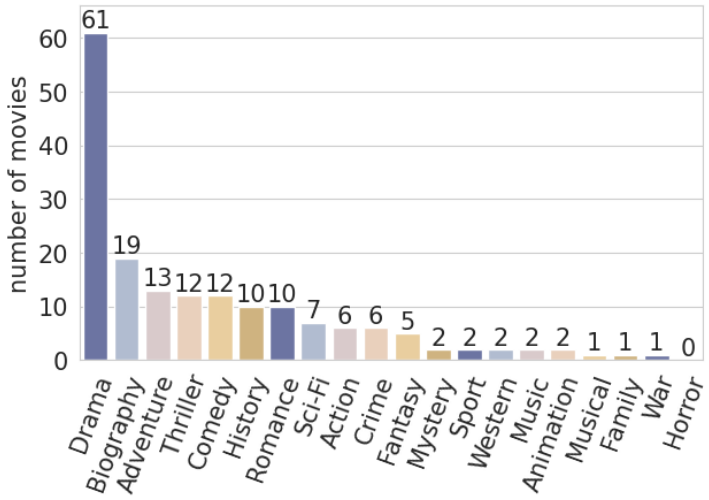

Figure 4: The number of movies that were nominated for Best Picture in each genre.

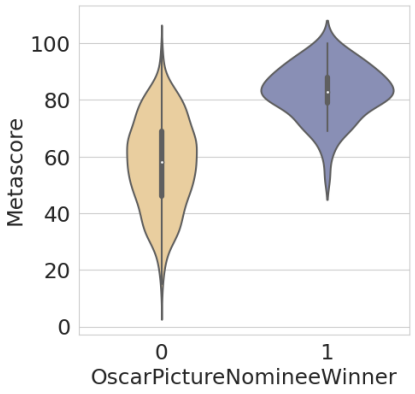

Figure 5: The distributions of Metascore of non-nominated movies(left), and nominated movies(right).

To build a decision tree, the algorithm starts from the entire dataset (the root node) and splits it into smaller subsets based on a specific feature. Each subset can be viewed as a node of a tree. Typically, the splitting process stops when the decision tree reaches an appropriate depth, and the bottommost nodes become the leaf nodes. While classifying a new data point using the tree, the data point is assigned to the class in the leaf node that has the largest portion of data points. We used the CART algorithm for building the decision tree, which stands for Classification and Regression Trees. We choose CART because it uses Gini impurity which can handle both continuous and categorical variables well. The Gini impurity of a node measures the probability of misclassifying a new data point and thus a low Gini impurity is preferred for correct classification. The Gini impurity is calculated for each node D as $Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$, where p_i is the probability of choosing a data point with class i [3]. For a split based on feature A, CART produces 2 subsets D_1 and D_2 with sizes n_1 and n_2 respectively, and the Gini impurity of the subsets is $Gini_A(D) = \frac{n_1}{n} Gini(D_1) + \frac{n_2}{n} Gini(D_2)$, where n is the size of D [3]. Therefore, the Gini Gain that feature A can produce is $\triangle Gini(A) = Gini(D) - Gini_A(D)$ [3]. Each time, CART chooses a feature that can maximise the Gini gain.

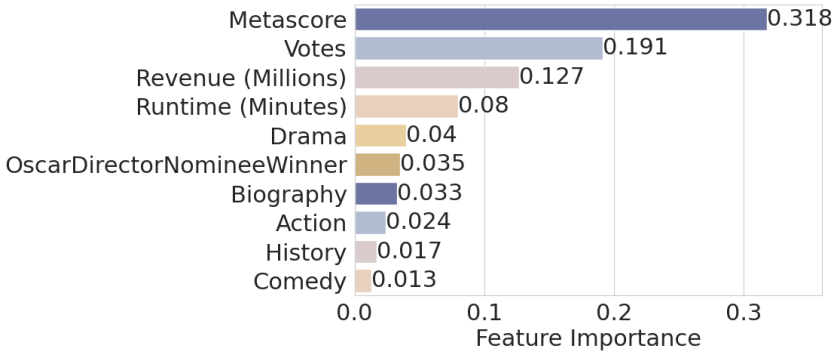

Figure 6: The variables with top 10 highest feature importance scores

We split our data into a 90% training set and a 10% test set. The training set consists of 754 data points and the test set has 84 data points. We used the following variables for training the random forest: *Revenue (Millions)*, *Votes*,*Metascore*, *Runtime (Minutes)*, *OscarDirectorNomineeWinner*, *hasOscarActorNomineeWinner* and all genre variables. We didn't use *Rating* since *Metascore* is provided by more professional critics. The number of decision trees in the random forest was set to 100, which means that we performed bootstrapping 100 times and trained a decision tree for each bootstrap sample. A 3-fold cross-validation was applied to find the maximum

depth of each tree that can help produce the highest accuracy score on the validation set, this also prevents overfitting. Finally, we ended up with a random forest model with 100 decision trees and a maximum depth of 11. The accuracy scores of the model on the training set and the test set are 0.999 and 0.976 respectively. In addition, the sensitivity on the training set is 0.98, and 0.6 on the test set, a sensitivity of 0.6 means that whenever the model tries to predict a movie that is actually a nominee, there is a possibility of 0.6 that the model predicts the movie as a nominee. One reason for a lower sensitivity on the test set could be that the number of nominated movies is small on both training and test sets, so the model tends to predict the majority class (non-nominated movies).

The feature importance of variables is shown in Figure 6. A variable with a higher feature importance signifies that it plays a more crucial role in classifying the data. The feature importance of a variable is calculated based on the Gini gain produced by this feature among all decision trees in the random forest, a higher Gini gain implies higher feature importance.

Both Drama and Biography have particularly higher feature importance than other genre variables, this aligns with the fact that they both have a large number of nominated movies displayed in Figure 4.

In addition to the genre, a movie's popularity and critical reception also play an important role in determining its success. Figure 6 highlights the importance of votes and Metascore as crucial elements in evaluating a movie, as they provide valuable insight into how a movie is perceived by the public and its potential success. Moreover, revenue and runtime are also important factors to consider. Revenue serves as an indicator of a movie's financial success, which is often linked to its popularity and quality. In addition, a movie's runtime influences its success by determining the amount of time viewers invest in the movie. Whether the director of the movie has been nominated for Best Director also appears to be an important feature since it is an indicator of the good quality of the movie.

Notice that, although Drama in Figure 3 has a very low range of revenue, it has a great impact on the chances of a movie being nominated for Best Picture.

Overall, the results suggest that the interaction between a movie's genre, critical reception, popularity and financial success are crucial in predicting the likelihood of a movie being nominated for Best Picture. Therefore, moviemakers and production companies should take these factors into account when making movies if they want to win an Oscar.

In the end, we analysed the differences between the score of reviews from the general public and the professional critics. This study can help movie production companies better position their movies to cater to specific groups.

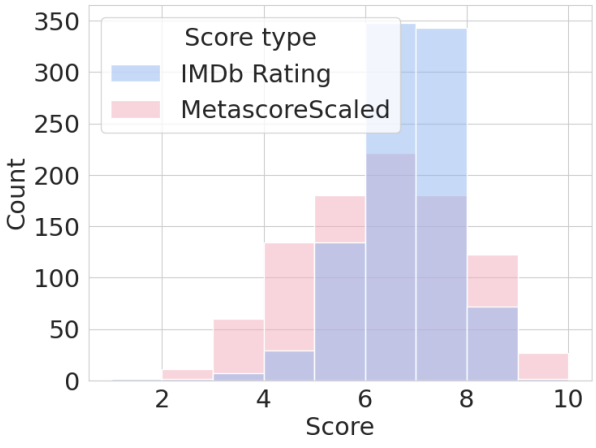

Figure 7: The distributions of IMDb ratings and scaled Metascore for all movies.

We investigated if there were any differences in the distribution of IMDb ratings and Metascore. The IMDb ratings range from 1 to 10, while the Metascore ranges from 0 to 100. To help with comparison, we created a new variable *MetascoreScaled* to make Metascore the same scale as IMDb rating, we have $MetascoreScaled = Metascore \times 0.09 + 1$. Figure 7 shows the difference of distribution between Rating and scaled Metascore. It is clear that the distribution of the scaled MetaScore is very close to the normal distribution, with a peak from 6 to 7, and the differences between consecutive bins are small. On the other hand, the distribution of the IMDb ratings changes dramatically, with the highest point from 6 to 8, but there is an obvious lack of ratings between 1 and 5.

To confirm the differences between the two scores, we divided Rating and MetascoreScaled into 9 categories: $[1,2), [2,3), [3,4), [4,5), [5,6), [6,7), [7,8), [8,9), [9,10]$ respectively, and calculated the number of movies in each category. We then performed a chi-squared

test on them. The number of movies in each category of Rating is the expected result, and the number of movies in each category of MetascoreScaled is the observed result. Our null hypothesis is that the distribution MetascoreScaled doesn't deviate from the IMDb rating. The test had a significant level at 0.05, 8 degrees of freedom(since we have 9 categories, so 9-1=8), and a critical value of 15.507. Our result is $\chi^2 = 1731.78$, which exceeds the critical value. Therefore, we can reject the null hypothesis and conclude that the distribution of the scaled Metascore deviates from the distribution of the IMDb ratings.

From the information provided by imdb.com, it is clear that every registered user can provide a rating on a movie, and registration is open to everyone for free [2]. However, imdb.com also mentions that "we accept and consider all votes received by users, not all votes have the same impact (or 'weight') on the final rating... we don't disclose the exact method used to generate the rating". According to the information provided by the source website of Metascore, metacritic.com [13]. Metascore is not directly generated by professional critics. Instead, the scores of the reviews are assigned by metacritic.com itself, and the weighted average of the scores presents the final score. However, the exact weights are kept confidential by metacritic.com.

The above facts indicate that although both websites do not disclose the details of their rating mechanisms, the ratings on IMDb are still more dominated toward general audiences. This can be seen in its prominent peak from 6 to 8, where scores in this range are usually considered "good" or at least "mediocre". Ordinary reviewers tend to watch popular movies, and on the other hand, popular movies usually have decent quality. Additionally, ordinary reviewers could be more influenced by other people, leading to a biased number of votes with high ratings. In contrast, professional critics have more detailed analysis of the movie, and their judgments are more independent, which means they tend to be relatively conservative in terms of high ratings. However, in the area of extraordinarily high scores(from 8 to 10), there is more scaled Metascore compared to IMDb ratings, this could due to the small number of critics selected by metacritic.com, which means that it is less possible for these high scores to be scaled down by some low scores (According to metacritic.com, a score of 61 or above is considered good, and for scaled Metascore, it is 6.49 or above [13]).

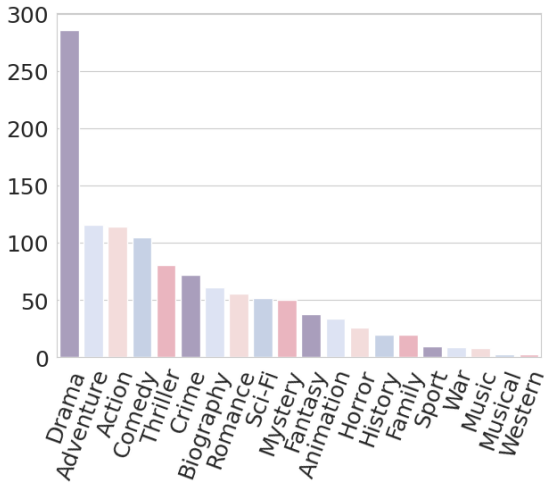

Figure 8: The number of movies in each genre for IMDb rating ≥ 7.

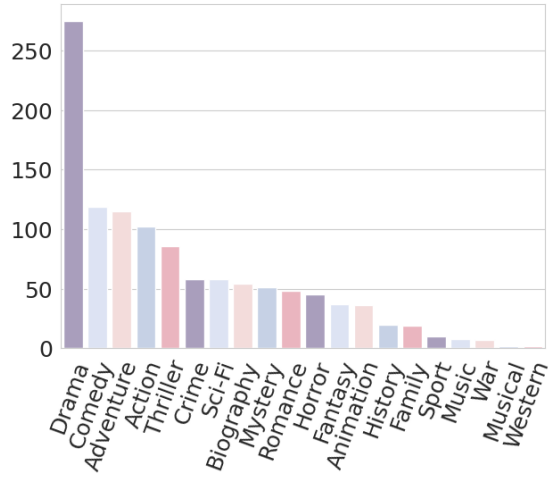

Figure 9: The number of movies in each genre for scaled Metascore ≥ 6.49.

To investigate whether there are any differences in the high-rated genres between the two types of ratings, Figure 8 shows the number of movies in each genre with a rating of 7 or higher, while Figure 9 shows the number of movies in each genre with a scaled Metascore of 6.49 or higher. From the two charts, it can be seen that although Adventure is the second most popular genre for high IMDb ratings, and Comedy is the second most popular for high *MetascoreScaled*, there is no significant variations in the overall ranking of genres. This indicates that there is not much difference in the preference of genre between professional critics and the general public.

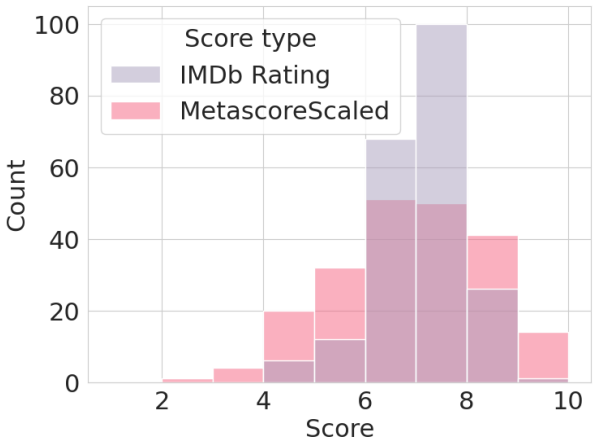

Figure 10: The distributions of IMDb ratings and scaled Metascore for commercially successful movies.

We categorised movies with revenue higher than twice their budget as commercially successful, Figure 10 shows the IMDb ratings and scaled Metascore of these movies, it can be seen that the distribution of ratings of commercially successful movies is generally similar to all movies: the distribution in the scaled Metascore is more smooth compared to IMDb ratings. Instead, most IMDb ratings are concentrated at 7-8. The proportion of scaled Metascore at 8-9 has increased, and the overall distribution is more concentrated in the high-scoring area than before. Therefore, we can conclude that if moviemakers want their movies to achieve commercial success, they need to improve their ratings with both the general public and professional critics.

Again, we noticed something contradictory when discussing about revenue: although Drama, Adventure, Action have a negative influence in terms of their negative coefficients for predicting revenue, they are the top 3 popular genres in movies that are highly rated on IMDb.

5 Discussion and conclusions

Summary of findings In conclusion, our study revealed that the budget and runtime have the most significant impact on revenue. Indicating that these two variables play a crucial role in determining a movie's commercial success. Our analysis has also demonstrated that animation has an obvious and positive impact on revenue with the smallest p-value among all genre variables. Our multiple regression was able to predict revenue with an adjusted R-squared of 0.718, however, due to the existence of luring variables, the performance of the model decreased for high-budgets movies.

Our further exploration of the factors that could influence the revenue demonstrates that movies with especially high revenue would have made a lot of effort in promotion. With a detailed analysis of 2 movies *Avatar* and *John Carter*, we conclude that the other important factors could be the theme of the story and the technology used in the movies.

Our random forest classifier demonstrates some important features of movies that were nominated for Best Picture, these include Metascore, vote, revenue, runtime, and Drama, with Metascore being the most important feature for classifying. In addition, whether a director has been nominated for Oscar Best Directing before also helps recognise the movies with high quality, or in other words, the movies that could be nominated for Best Picture.

We analysed the distribution of IMDb ratings and Metascore to investigate differences between public and professional reviews. Our study found that the distribution of scaled Metascore is closer to a normal distribution than IMDb ratings. IMDb ratings are more influenced by the opinion of the general public, while professional critics tend to be more conservative with high ratings. The commercially successful movies had a similar distribution of these 2 types of ratings, but with more concentration in the high-scoring area for both ratings. Therefore, it is crucial for moviemakers to satisfy both professional critics and the general public.

In the end, we found an interesting fact that the genres: Drama, Action, and Adventure with negative influence on revenue instead have the most movies in the high rating area.

Evaluation of own work: strengths and limitations The main limitation in our study is the small number of records in our main dataset. Additionally, the numbers of movies released each year are not evenly distributed in our main dataset, 29.7% of the movies were released in 2016, which makes it

difficult to analyse the trend. The provider of the dataset didn't explain whether the movies were randomly sampled, and there is a column *Rank* in the dataset without any description provided, which could imply that the movies were selected according to some criteria instead of being randomly sampled.

Our multiple linear regression model for predicting revenue has a robust R-squared of 0.718, but it has a large RMSE due to the outliers, indicating that our model can not perform well, especially on movies with a high budget. Since we used backward elimination for choosing variables, we removed many variables due to their high p-values. However, the p-values may change based on the technique of feature selection, so the variables in Table 1 may not accurately reflect the influence of each genre on the revenue.

Our random forest model for classifying nominated/non-nominated movies reflects some of the characteristics of nominated movies. However, the model has a much better sensitivity on the training set than the test set, but since there is only a small quantity of nominated movies, our model can not acquire enough information for nominated movies. Furthermore, our model was trained on movies released from 2006 to 2016, so it is restricted to this period especially when new trends appear in preferences of nominated movies.

Our investigation of the difference between IMDb ratings and Metascore reveals some characteristics of the general public and professional critics, but, it is worth noting that although we make Metascore on the same scale with IMDb ratings, we can not conclude that a scaled Metascore of 7 and an IMDb rating of 7 are exactly the same, since the original range of Metascore(0-100) can also influence the score given by metacritic.com.

Comparison with any other related work In the summary it draws the conclusion that there is a strong correlation between budget, runtime to the revenue. Similarly, the previous paper [12] also proves our conclusion that with an increase in budget, there is also an increase in revenue. Our study displayed consistent results with Litman (1983), Ravid (1999), and Terry et al. (2005) [14] which also shows the most significant contributor to revenue was found to be production cost, worth notoing that their model has a coefficient of determination of 0.647. Their model implies that a higher production budget would lead to higher 'global' box office revenue.

In addition, Prag and Casavant (1994) showed a negative relationship between drama and revenue, which is consistent with our result, however, this is in contrast with Terr et al. (2005) who demonstrated a positive relationship between the action genre and revenue [17]. Movie market predictions [10] show that animation (2021-2030) revenues are constantly increasing, confirming our conclusion that animation has the highest revenue.

Although some studies, such as those conducted by Smith & Smith (1986) and Eliashberg & Shugman (1997) [19], reported contradictory findings when using the Academy Awards as a metric for measuring a movie's success, other studies discovered that receiving an Oscar nomination or award, particularly for Best Picture/Actor/Actress, generally increased the likelihood of a movie's success [21].

Improvements and extensions We would like to find more movies released from 2006 to 2016. In addition, for predicting revenue, we would acquire more variables, such as release month, content rating, the salary of the main actors, and the country of the movie. In order to evaluate the effort the moviemakers make in marketing, we would find the number of views of trailers on YouTube, as well as the ratios of likes and dislikes of the trailer.

Regarding the revenue prediction model, we would like to try RandomForestRegressor provided by scikit-learn, and cross-validation could be used to find the best combination of the number of decision trees and the maximum depth of each tree. For predicting whether a movie could be nominated for Best Picture, we would also analyse the trend of the nominated movies in terms of genres, therefore we would need to collect all nominated movies that were released from 2006 to 2016.

References

[1] Krauss J. et al. “Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis”. In: *AIS Electronic Library* (2008). Retrieved on 10 April 2023. URL: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1185&context=ecis2008.

[2] Dennis Brunning. *IMDB.com/IMDB pro*. Oct. 2015. URL: https://www.ingentaconnect.com/content/charleston/chadv/2015/00000017/00000002/art00007.

[3] Fatih Karabiber Ph.D. in Computer Engineering and Fatih Karabiber Ph.D. in Computer Engineering. *Gini impurity*. URL: https://www.learndatasci.com/glossary/gini-impurity/.

[4] *Creative Commons License Deed*. URL: https://creativecommons.org/publicdomain/zero/1.0/.

[5] *Decision tree tutorial*. URL: https://www.projectpro.io/data-science-in-r-programming-tutorial/decision-tree-tutorial.

[6] Ritayan Dhara. *TMDb datasets*. Apr. 2020. URL: https://www.kaggle.com/datasets/ritayandhara/tmdb-dataset.

[7] Kevin Fallon. *Why John Carter flopped: 6 theories*. Jan. 2015. URL: https://theweek.com/articles/477390/why-john-carter-flopped-6-theories.

[8] Raphael Fontes. *The Oscar Award, 1927 - 2023*. Mar. 2023. URL: https://www.kaggle.com/datasets/unanimad/the-oscar-award.

[9] Trevin Geier. *Avatar is still the highest-grossing movie of all time, and here’s why*. Sept. 2022. URL: https://movieweb.com/avatar-highest-grossing-movie/.

[10] “Global Animated Films Market”. In: *animated films market* (2022). URL: https://www.sphericalinsights.com/reports/animated-films-market.

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

[12] Devansh Hingad. “Analysis of relation between budgets and revenues from movies - IJISRT”. In: *International Journal of Innovative Science and Research Technology* (2020). URL: https://ijisrt.com/assets/upload/files/IJISRT20MAY061_(1).pdf.

[13] *How we create the metascore magic*. URL: https://www.metacritic.com/about-metascores.

[14] Starling David Hunter III, Susan Smith, and Saba Singh. “Predicting box office from the screenplay: A text analytical approach”. In: *Journal of Screenwriting* (June 2016). URL: https://intellectdiscover.com/content/journals/10.1386/josc.7.2.135_1.

[15] Martin Mraz. *Oscar Best Picture movies*. Aug. 2021. URL: https://www.kaggle.com/datasets/martinmraz07/oscar-movies.

[16] Jennifer P. Nesbitt. “Deactivating feminism: Sigourney Weaver, James Cameron, and avatar”. In: *Film amp; History: An Interdisciplinary Journal* (July 2016). URL: https://muse.jhu.edu/pub/39/article/626142/summary.

[17] N.A Pangarker. “The determinants of box office performance in the Film Industry Revisited”. In: *ResearchGate* (Sept. 2013). URL: https://www.researchgate.net/publication/281730174_The_determinants_of_box_office_performance_in_the_film_industry_revisited.

[18] Tavoosi. Saba. “Predicting box office revenue with Random Forest”. In: *Kaggle* (2019). Retrieved on 10 April 2023. URL: https://www.kaggle.com/code/tavoosi/predicting-box-office-revenue-with-random-forest.

[19] Geethanjali Selvaretnam and Jen-Yuan Yang. *Factors affecting the financial success of Motion Pictures : What is the role of Star Power?* Jan. 2015. URL: https://research-repository.st-andrews.ac.uk/handle/10023/6786.

[20] Jeffrey S. Simonoff and Ilana R. Sparrow. “Predicting movie grosses: Winners and losers, blockbusters and sleepers”. In: *SSRN* (Oct. 2008). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1290219.

[21] Dean Keith Simonton. “Cinematic success criteria and their predictors: The art and business of the film industry”. In: *Psychology and Marketing* 26.5 (2009), pp. 400–420. DOI: 10.1002/mar.20280.

[22] *Where data and the movie business meet*. URL: https://www.the-numbers.com/movie/.