Comparing Critics to the General Public: Who is a better predictor of a movie's success?

Jacob Inwald and Ollie Jones

1 Overview

This report analyses the impact of critics on the success of movies within the film industry. We carried out this analysis using data from IMDb which provided statistics on one thousand movies released between 2006 and 2016. A least squares multiple regression model was used to determine which factors of a movie can be used to predict their box office success, while another least squares multiple regression model was used to determine if critics were able to predict success based on a metric which combines the revenue and viewer rating of that movie. We found that a movie's revenue is correlated by the number of votes it has on IMDb, its runtime and the experience of actors in its leading roles. We further found that critics ratings cannot be used to predict the box office success of a movie, but can predict how the public will recieve it, even though they tend to be less lenient than the general public. Moreover, we discovered that actor experience is a better prediction of box office success than director experience.

2 Introduction

Context and motivation The global movie industry is immense, with a total of 629,807 titles listed on IMDb as of December 2022 [1], some dating as far back as 1894 [2]. With such a vast selection, it is impossible for all movies to become successful blockbusters. This raises the question: what factors contribute to a movie's success? There is an abundance of data available regarding movies, their cast, crew, and even their audiences. This data can be used to gain insight into the production, reception, and overall success of films. This report aims to explore how different metrics can predict the success of a movie before it gets released to the general public.

When it comes to movie ratings, there are often two different scores available: one from the general public and the other from professional movie critics. At times, scores may be nearly identical. For instance, "The Godfather" [3] received nearly identical scores from both groups. On the other hand, "The Last Jedi" [4] saw a significant difference between the two. The disparity between ratings for certain movies can make it challenging to determine whether they were successful.

Previous work Recent research has explored the various factors that contribute to the success or failure of films. Several studies and papers have been conducted in this field, examining a range of elements. Previous research has investigated the potential of utilizing IMDb data to predict the success of films [5]. Other studies have sought to identify the factors that contribute to the making of a blockbuster movie [6]. Moreover, further studies have attempted to develop their own mathematical models to predict the success of upcoming movies [7].

Objectives This report seeks to identify the most influential factors in determining a movie's success. With the multitude of factors that could potentially affect a movie's performance at the box office, it is not feasible to investigate them all. Therefore, we will focus our research on the director, actors, genre and runtime of a movie, and compare these with its viewer rating, critic score and box office revenue. We will be exploring the following question in depth: who is more accurate in predicting a movie's success, critics or the general public? Additionally, we will investigate the influence of a movie's director and actors on its success, as well as the genres that are most likely to produce successful films.

3 Data

Data provenance We utilized two movie datasets: an IMDb dataset [8] and TMD [9] (The Movie Dataset). The IMDb dataset contains data about movies made from 2006-2016, while the TMD dataset contains data about movies released on or before July 2017. Both datasets were obtained from kaggle.com and are shared under the CC0 1.0 Universal Public Domain Dedication. TMD is merged data from TMDb (The Movie DataBase) and grouplens.org, a movie ranking site. It is worth noting that the provenance of the IMDb dataset has been subject to controversy as it was scraped from IMDb.com.

Data description The IMDb movie dataset contains 12 columns with string or floating point values. The only column of note is the Genre column, which can only take on 12 distinct values, shown in Table 1. A summary of the columns is provided in Table 1.

Column Name	Description	Data Type
Rank	The rank the movie has in the IMDb database	Integer
Title	The name of the movie	String
Genre	The genres that apply to the movie, there can be anywhere from 1-3 genres. A genre can be any from: Action, Adventure, Sci-Fi, Thriller, Animation, Comedy, Family, Fantasy, Drama, Music, Romance, History, Crime, Western, War, Musical, Sport, Horror, Mystery, Biography.	Genre
Description	The description of the movie	String
Director	The person who directed the movie	String
Actors	The lead roles in the movie	String
Year	The year the movie was released	Integer
Runtime (Minutes)	The runtime in minutes of the movie	Integer
Rating	The mean rating of the movie, taken from IMDb.com	Float
Votes	The amount of users that voted on a movie to give it that rating	Integer
Revenue (Millions)	The gross income the movie made at the US box office	Float
Metascore	The rating of movie, determined using aggregated weighted Critics scores	Integer

Table 1: The different columns in the IMDb-Movie-Data.csv file.

We worked exclusively with the Crew table in TMD, which is a small part of the entire database. This table contains three columns - Cast, Crew, and ID. The Cast and Crew columns are not composed of discrete data points, but rather JSON files representing the entire cast or crew list. As a result, we extracted the data using string parsing methods. A summary of the columns is provided in Table 2.

Column Name	Description	Data Type
cast	The cast list of the movie, including all actors who appeared in it.	.json file
crew	The entire crew list of the movie.	.json file
id	The movie id - used in the larger dataset to connect tables together	Integer

Table 2: The different columns in the credits.csv file.

Data processing We used a Python script to parse the TMD dataset and count the number of movies each director and actor has been involved in. The resulting datasets were saved as CSV files named actor_counts.csv and director_counts.csv. We then merged this data with the original IMDb dataset. This involved dropping the Description column and replacing the Director and Actor columns with Director Exp. and Mean Lead Roles Exp. A summary of the new columns is shown in Table 3.

Column Name	Description	Data Type
Director Exp.	Amount of movies director has made	Float
Mean Lead Roles Exp.	Mean amount of movies lead actors have made	Float

Table 3: The different columns in the merged data set, it also shares all the columns described in Table 1, except for Director and Actors.

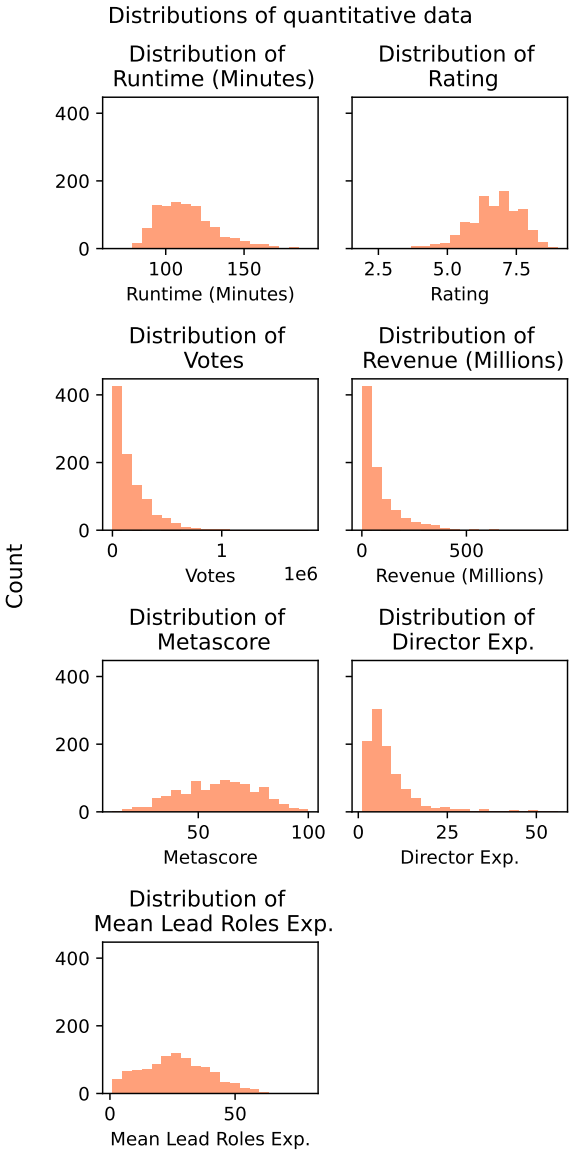

Figure 1: The distributions of the numeric variables in the merged dataset.

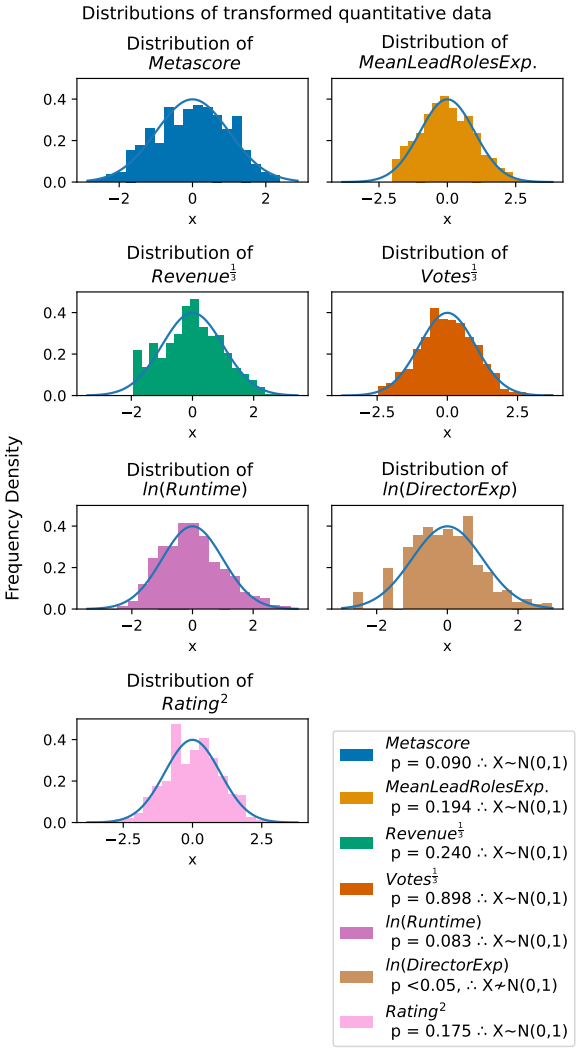

Figure 2: The distributions of the standardised and normalised numeric variables in the merged dataset, the legend has the p-values from testing with $H_0 : X \not\sim N(0,1)$. Also shown on the plots is the Gaussian Distribution with the columns mean and standard deviation.

To check this data was properly normalised we made a histogram plot of all the numeric variables, shown in Figure 1. As expected, a few variables did not appear to be normally distributed, namely Revenue (Millions), Votes, Runtime (Minutes), Director Exp., and Rating.

Figure 1 shows that Revenue (Millions) and Votes are severely right-skewed, implying an exponential distribution. Runtime (Minutes) and Director Exp. appear to be less severely right-skewed, implying a lognormal distribution. Rating seems to be left-skewed. The transformations that give the best approximations to a normal distribution are cube root transform for Revenue (Millions) and Votes, log transform for Runtime (Minutes) and Director Exp., and square transform for Rating.

Figure 2 shows the standarised and normalised numeric variables. The legend contains the p-values from the Kolmogorov-Smirnov normality test [10], which was used to test the data against the null hypothesis $H_0 : X \not\sim N(0,1)$. It is worth noting that the Director Exp. column failed the test. Despite this, the histogram follows a normal distribution curve and there appears to be missing data, providing convincing evidence that ln(Director Exp.) has a normal distribution. Ultimately, these transformations helped to normalize the dataset and improve its suitability for further analysis.

To make use of the Genre column, one-hot encoding was used. We added a column for each Genre and set each entry to 1 if the corresponding movie fit that genre and 0 otherwise. Genres with less than 100 movies made were excluded to ensure a large enough sample size was present, such that we could draw meaningful conclusions.

4 Exploration and Analysis

4.1 What makes a movie successful at the box office?

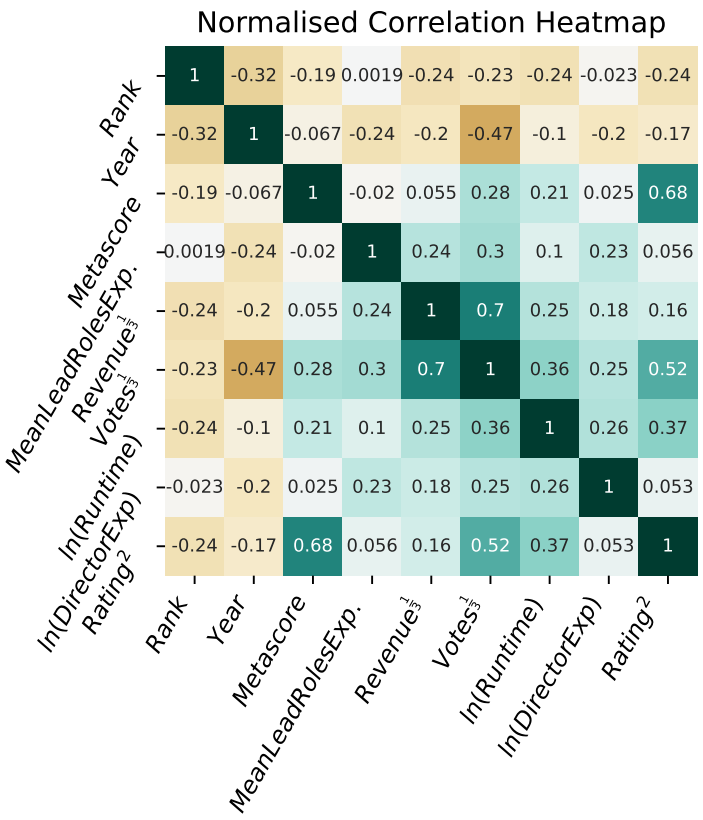

Figure 3: A heatmap illustrating the strength of correlations (using PMCC) between various normalised variables from the IMDb dataset [8] and the TMD data [9].

By examining the correlations between the movie's revenue and other factors, we can gain insight into which factors have the greatest impact on the success of a movie. These correlations are visualised in Figure 3.

Figure 3 reveals a strong positive correlation ($r = 0.7$) between a movie's revenue and the number of votes it receives on IMDb, suggesting that movies which have more votes may have sold more tickets, which shows a possible interaction or causal link. Additionally, the heatmap shows moderate positive correlations between the movie's revenue and its runtime ($r = 0.25$) and between the movie's revenue and the experience of the actors in lead roles ($r = 0.24$), indicating that longer runtimes and more experienced actors may be associated with higher revenues. In contrast, there is a moderate negative correlation between a movie's revenue and its rank on IMDb ($r = -0.24$), suggesting that higher ranks on IMDb do not necessarily translate to higher revenues. These four correlations are shown in more detail in Figure 4.

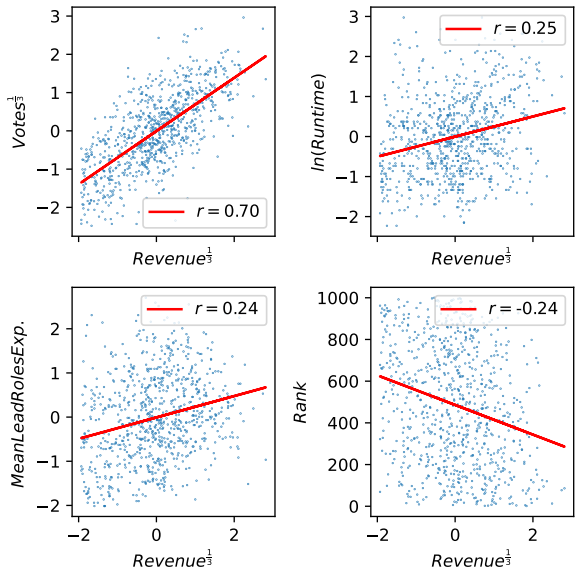

Figure 4: Scatter plots showing the relationships between a movie's revenue and the four factors with moderate or strong correlations.

4.2 The relationship between a movie's ranked position and its number of votes

IMDb's system for ranking movies is connected to the votes cast by viewers. The statistical distributions of the ranks and the votes are show in Figure 5. Plotting the normalised variables against each other produced a PMCC of $r = -0.23$. This moderate negative correlation suggests that having more votes on IMDb does not necessarily relate to higher revenues.

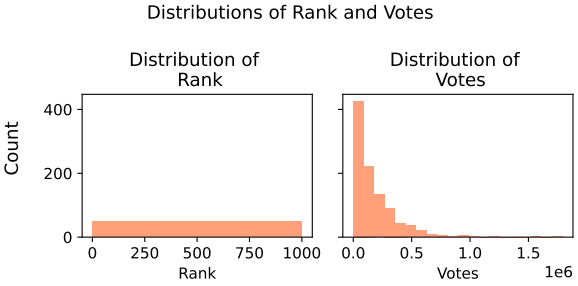

Figure 5

4.3 What factors can predict box office success?

In order to further analyze the correlations between the various movie details and the revenue they generated, we created a multiple regression model using the normalized data we collected. The target variable was the normalised revenue, with the rest of the dataset as the predictor variables, excluding Genre and Title. $Votes^{\frac{1}{3}}$ was excluded as it may be causally linked to Revenue (see sec 4.1). The model uses Ordinary Least Squares regression and a constant column has been added to show y-intercept. The summary results are provided in table 4.

This model performs respectably, explaining about 40% of the variance in box office success. It is also statistically significant ($R^2 = 0.404, F(15,810) = 36.6, p < 0.05$). The results indicate a well performing model, and as such we can postulate that at least some of these factors are impactful on the box office success a movie has.

The results suggest that box office sales can be predicted by lead actor experience ($\beta = 0.113, \sigma = 0.030, p < 0.05$), but cannot be predicted by director experience ($\beta = 0.051, \sigma = 0.030, p > 0.05$). Previous research has shown that movie advertising tends to be focussed around the actors present more than the director of the movie[11]. Examples of this are present in adverts like movie posters - actor faces feature front and center while directors are listed in name only. Actors who have been in many movies are more recognisable, meaning potential consumers may be more likely to invest money in seeing a movie they are part of.

The results also suggest that box office sales can't be predicted by critics (Metascore) ($\beta = 0.042, \sigma = 0.038, p > 0.05$), but can be predicted by user rating. ($\beta = 0.082, \sigma = 0.041, p < 0.05$). It is reasonable to assume both would have similar quality as predictor variables - they correlate strongly (see fig 3). As such, the difference in quality is surprising. This difference could be due to movies appearing in the box office for only a short period. Revenue is only calculated during this period, a period of time where consumers may rely on friends rather than a critics review, as those often come out later. With this in mind, it is reasonable to say that critics do not appear to be able to predict the likelihood of a movie becoming a blockbuster.

Finally, there are some interesting notes about the impact of genre has on the box office success of a movie. Adventure seems to have the biggest positive impact on box office success ($\beta = 0.477, \sigma = 0.074, p < 0.05$). Dramas appear to have the largest negative impact on box office success ($\beta = -0.558, \sigma = 0.071, p < 0.05$).

Dep. Variable:	$Revenue^{\frac{1}{3}}$	**R-squared:**	0.404
Model:	OLS	**Adj. R-squared:**	0.393
Method:	Least Squares	**F-statistic:**	36.60
No. Observations:	826	**Prob (F-statistic):**	1.35e-80
Df Residuals:	810	**Df Model:**	15

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	114.0116	19.712	5.784	0.000	75.319	152.705
Rank	-0.0006	0.000	-5.654	0.000	-0.001	-0.000
Year	-0.0565	0.010	-5.770	0.000	-0.076	-0.037
Metascore	0.0416	0.038	1.093	0.275	-0.033	0.116
MeanLeadRolesExp.	0.1125	0.030	3.793	0.000	0.054	0.171
$ln(Runtime)$	0.1705	0.033	5.203	0.000	0.106	0.235
$ln(DirectorExp)$	0.0507	0.029	1.756	0.079	-0.006	0.107
$Rating^2$	0.0816	0.041	2.013	0.044	0.002	0.161
Action	0.1824	0.070	2.598	0.010	0.045	0.320
Adventure	0.4765	0.074	6.452	0.000	0.332	0.621
$Sci-Fi$	0.0318	0.087	0.364	0.716	-0.140	0.203
Thriller	-0.0395	0.078	-0.504	0.615	-0.193	0.114
Comedy	0.1145	0.072	1.599	0.110	-0.026	0.255
Drama	-0.5579	0.071	-7.877	0.000	-0.697	-0.419
Romance	-0.0243	0.084	-0.288	0.773	-0.190	0.141
Crime	-0.0545	0.082	-0.664	0.507	-0.216	0.107

Table 4: Multiple regression results with $Revenue^{\frac{1}{3}}$ as the dependent variable and the normalised data as the independent variable. The OLS approach was used to find the best fit. The residuals for this model are shown in fig 6.

4.4 Can critics truly predict movie success?

To analyze the effect critics had on predicting general success of a movie, we define a metric of success, $Success = \frac{Revenue^{\frac{1}{3}} + Rating^2}{2}$. We chose this metric as Revenue and Rating are both measures of movie success, and Figure 3 shows that they are not strongly correlated. This means they describe independent aspects of success, and therefore the mean of both will better encapsulate the true success of a movie. We refer to this metric as success from now on.

To investigate the ability for critics to predict movie success, we present a multiple regression model, with the target variable being the success metric. We used the rest of the normalised dataset for predictor variables, excluding the $Revenue^{\frac{1}{3}}$, $Rating^2$, Genre and Title columns. Genre and Title were excluded as they are not quantative data and $Revenue^{\frac{1}{3}}$ and $Rating^2$ were excluded as they are causally linked to the success metric. We use the same approach as in table 4, with a constant column and using OLS. The summary results are provided in table 5.

Dep. Variable:	Success	**R-squared:**	0.758
Model:	OLS	**Adj. R-squared:**	0.753
Method:	Least Squares	**F-statistic:**	169.1
No. Observations:	826	**Prob (F-statistic):**	2.67e-237
Df Residuals:	810	**Df Model:**	14

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	-49.4262	10.871	-4.547	0.000	-70.765	-28.087
Rank	-0.0001	5.42e-05	-2.173	0.030	-0.000	-1.14e-05
Year	0.0246	0.005	4.560	0.000	0.014	0.035
Metascore	0.1858	0.015	12.006	0.000	0.155	0.216
MeanLeadRolesExp.	-0.0003	0.015	-0.022	0.983	-0.029	0.028
$Votes^{\frac{1}{3}}$	0.5669	0.020	28.674	0.000	0.528	0.606
$ln(Runtime)$	0.0820	0.016	5.185	0.000	0.051	0.113
$ln(DirectorExp)$	-0.0318	0.014	-2.286	0.023	-0.059	-0.004
Action	-0.0532	0.034	-1.561	0.119	-0.120	0.014
Adventure	0.1279	0.036	3.554	0.000	0.057	0.199
$Sci-Fi$	-0.2271	0.043	-5.300	0.000	-0.311	-0.143
Thriller	-0.0610	0.038	-1.610	0.108	-0.135	0.013
Comedy	0.0376	0.035	1.087	0.277	-0.030	0.106
Drama	-0.0391	0.035	-1.133	0.258	-0.107	0.029
Romance	-0.0858	0.041	-2.103	0.036	-0.166	-0.006
Crime	-0.0795	0.040	-2.003	0.046	-0.157	-0.002

Table 5: Multiple regression results with *Success* as the dependent variable and the normalised data as the independent variable. The OLS approach was used to find the best fit. The residuals for this model are shown in fig 7.

This model performs very well, explaining about 75% of the variance in user ratings. It is also statistically significant ($R^2 = 0.758, F(14,810) = 169, p < 0.05$). The results indicates a well performing model, and as such we can postulate that at least some of these factors are impactful on the success a movie has.

The results show that movie success can be well predicted by Metascore ($\beta = 0.186, \sigma = 0.015, p < 0.05$). This compares to how it is not a good predictor of Revenue (see table 4). This difference shows that while critics can't predict the box office success, they can predict the overall success of the movie. Metascore is one of the key predictors in this model - with the third largest absolute coefficient - demonstrating that critics are important when it comes to predicting the success of a movie. This idea is supported by the PMCC that Metascore has with the Success metric ($r = 0.48$). This is a relatively strong correlation, suggesting that critics can discern potentially successful movies.

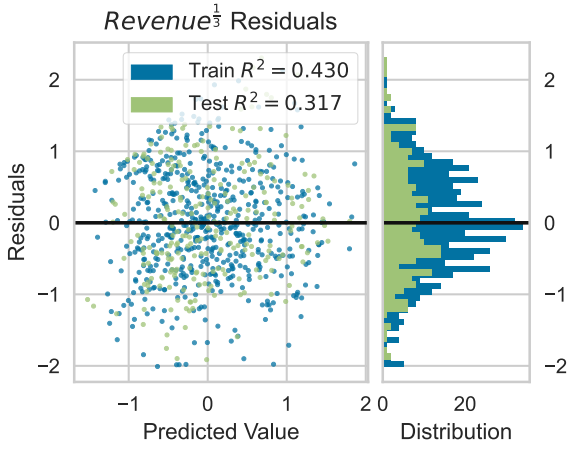

Figure 6: The residuals for the box office success revenue model. The R^2 it achieves on train or test data is shown in the legend of the plot.

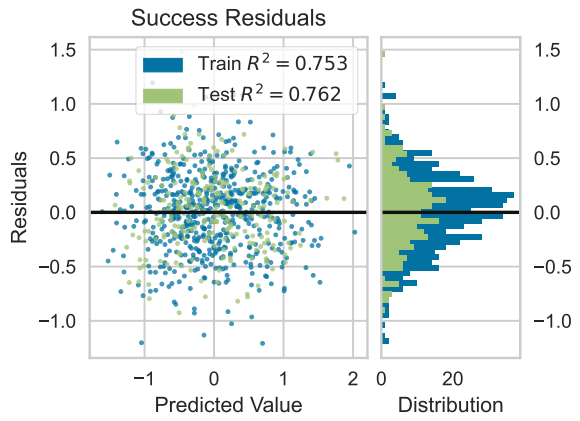

Figure 7: The residuals for the success revenue model. The R^2 it achieves on train or test data is shown in the legend of the plot.

Another observation is the constants coefficient is negative ($\beta = -49.4, \sigma = 10.9, p < 0.05$). This means that the average movie tends to not be very successful, which is inline with research that most movies do not succeed, with only a few actually garnering actual success [12].

Finally, there are some interesting notes about the impact of genre has on the box office success of a movie. Adventure seems to have the biggest positive impact on success ($\beta = 0.128, \sigma = 0.036, p < 0.05$). Sci-Fi appears to have the largest negative impact on success ($\beta = -0.223, \sigma = 0.043, p < 0.05$).

5 Discussion and Conclusions

Summary of findings Critics play a major role in the movie industry. They are seen as wielding enormous power, with the ability to render years of work wasted with a single review. We aimed to investigate if critics truly have this power. We found that critics can't predict box office success, bringing into question their power over movie revenue. However, we did find that critics can predict the general success and reception of a movie. Their ratings have a high correlation with user ratings, and our multiple regression model used Metascore as a predictor for success. These findings suggest that while critics can predict the success of a movie, they don't appear to impact the revenue the movie will bring in. Moreover, we found that the general public is more lenient than critics. The critic scores follow a unskewed normal distribution, whereas user ratings is left-skewed (see fig. 1). This means that on average the general public rates movies higher than critics will. These findings suggest that critics provide a more unbiased review of movies than a consumer does. We also found that actor experience is a better predictor of box office success than the director experience. This was reinforced by its moderate correlation to box office success, which is inline with research about "star power", the idea that stars drive movie success [11].

Evaluation of own work As a way to factor for non-linearity in the data, we normalised and standardised the data. We tested our models by checking that they maintained similar R^2 values for train and test data and that the residuals verified linearity of data. One limitation of our approach was the small amount of data we worked with. The main dataset we used - the IMDb dataset - was just a sample of a larger database and had only 1000 entries. Another limitation was the lack of some would be useful columns from the dataset. One such example is budget, which was used in Ahmad et al. [7] to produce a model for predicting revenue of a movie.

Comparison with any other related work Elberse [11] has also demonstrated that actors influence the box office success of movies, also attributing this to lead actor experience. Dhir and Raj [5] also created a model to predict success, defining it in this case as user rating. Ahmad et al. [7] also created a model to predict success, defining it in this case as revenue generated.

Improvements and extensions An improvement we could make would be to purchase the full dataset and run our analysis on that dataset. Another improvement would be to find a dataset that combines various other aspects of movie creation, for example movie budget or production managers.

References

[1] IMDb. *IMDb Statistics*. 2022. URL: https://www.imdb.com/pressroom/stats/.

[2] Alexander Black. *Miss Jerry*. 1894. URL: https://www.imdb.com/title/tt0000009/?ref_=adv_li_tt.

[3] Francis Ford Coppola. *The Godfather*. 1972. URL: https://www.rottentomatoes.com/m/the_godfather.

[4] Rian Johnson. *Star Wars: The Last Jedi*. 2017. URL: https://www.rottentomatoes.com/m/star_wars_the_last_jedi.

[5] Rijul Dhir and Anand Raj. "Movie Success Prediction using Machine Learning Algorithms and their Comparison". In: *International Conference on Secure Cyber Computing and Communication (ICSCCC)* 1 (2018). URL: https://ieeexplore.ieee.org/abstract/document/8703320.

[6] Martin C. Snell Alan Collins Chris Hand. "What makes a blockbuster? Economic analysis of film success in the United Kingdom". In: *Managerial and Decision Economics* 23 (2002). URL: https://doi.org/10.1002/mde.1069.

[7] Amr Yousef Javaria Ahmad Prakash Duraisamy and Bill Buckles. "Movie success prediction using data mining". In: *International Conference on Computing, Communication and Networking Technologies (ICCCNT)* 8 (2017). URL: https://ieeexplore.ieee.org/abstract/document/8204173.

[8] Ivan Gonzalez. *1000 IMDB movies (2006-2016)*. Scraped from https://IMDB.com. 2023. URL: https://www.kaggle.com/datasets/gan2gan/1000-imdb-movies-20062016.

[9] Rounak Banik. *The Movies Dataset*. 2018. URL: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?resource=download.

[10] Laha Chakravarti and Roy. *Handbook of Methods of Applied Statistics*. Vol. 1. John Wiley and Sons, 1967, pp. 392–394.

[11] Anita Elberse. "The power of stars: Do star actors drive the success of movies?" In: *Journal of marketing* 71.4 (2007), pp. 102–120.

[12] W David Walls. "Modelling heavy tails and skewness in film returns". In: *Applied Financial Economics* 15.17 (2005), pp. 1181–1188.