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 analyzes the relationship between different aspects of a movie and its gross income. These aspects include director list, actor list, runtime and so on. To investigate these relationships, we looked at a comprehensive dataset compiled by the movie ranking website IMDB.com. This dataset includes information about a movie like the director, the actors, the genre.

We found ..

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 study aims to explore how different metrics can predict the success of a movie before it gets released to the general public. Success in this context is defined as the amount of revenue generated by the movie.

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 study 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 used two movie datasets for our investigation: an IMDb dataset [8] and The Movie Dataset (TMD) [9]. The IMDb dataset contains data about movies released between 2006 and 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 a merged dataset from TMDb (The Movie Database) and grouplens.org, a movie ranking site. We only used the part of the dataset that was obtained from TMDb. It is worth noting that the provenance of the IMDb dataset has been subject to controversy as it was scraped from IMDb.com. However, for this investigation, we used only a sample of this dataset which is publicly available on Kaggle.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. However, 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 to 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, including all the people who made it.	.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. he 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 merged datasets columns is shown in Table 3.

Column Name	Description	Data Type
Rank	See table 1	Integer
Title	See table 1	String
Genre	See table 1	Genre
Director Exp.	The number of movies that the director of the movie has made	Float
Mean Lead Roles Exp.	The mean number of movies that the lead actors have been in	Float
Year	See table 1	Integer
Runtime (Minutes)	See table 1	Integer
Rating	See table 1	Float
Votes	See table 1	Integer
Revenue (Millions)	See table 1	Float
Metascore	See table 1	Integer

Table 3: The different columns in the merged data set

We checked that the data had been properly normalized by creating a histogram plot of all the numeric variables, as shown in Figure 1. As anticipated, some variables were not normally distributed, including Revenue (Millions), Votes, Runtime (Minutes), Director Exp., and Rating.

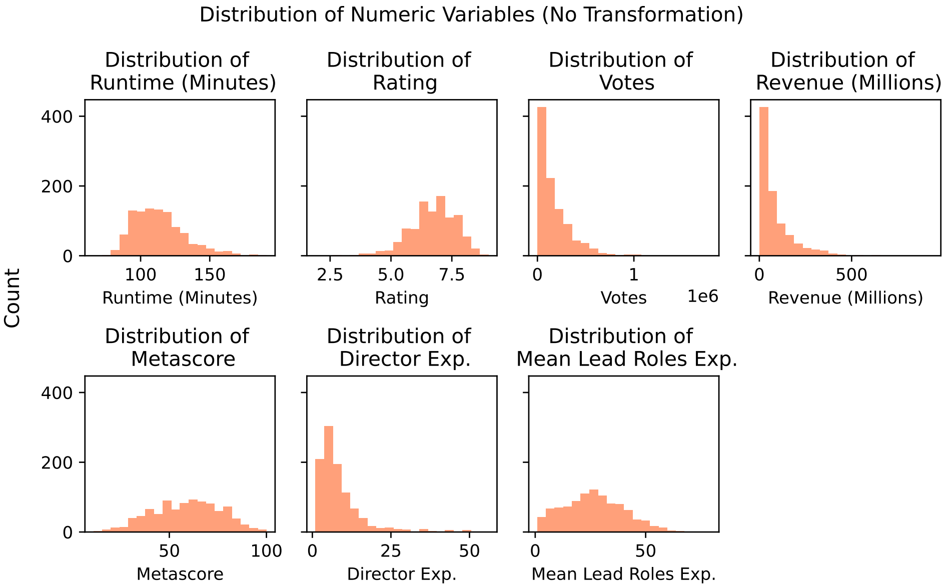

Figure 1: The distributions of the numeric variables in the merged dataset

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.

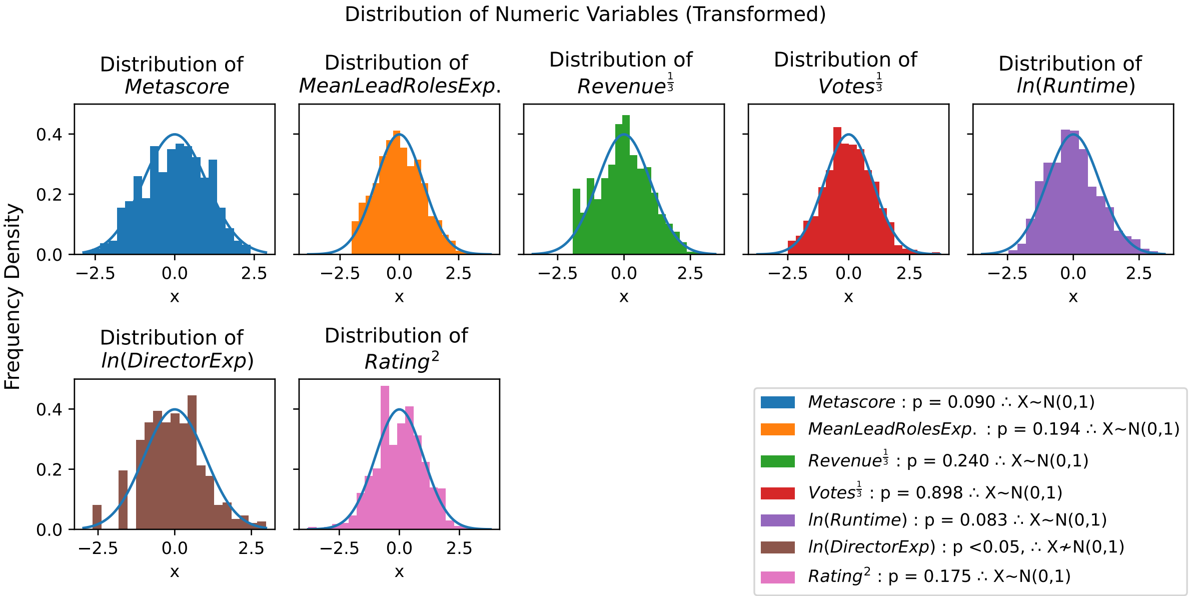

Figure 2: The distributions of the standardised and normalised numeric variables in the merged dataset, the legend has the p-values from testing with $H_0 : X \nsim N(0,1)$. Also shown on the plots is the Gaussian Distribution with the columns mean and standard deviation

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 $H_0 : X \nsim N(0,1)$. It is worth noting that the Director Exp. column failed the test, however, there is missing data present in the histogram. Despite this, the histogram follows a normal distribution curve, 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.

4 Exploration and Analysis

A data science analysis of the paper, including:

- Visualisations (for example Figure **??**) and tables (for example Table **??**). Please make sure that all figures and tables are referred to in the text, as demonstrated in this bullet point.
- Interpretation of the results
- Description of how you have applied one ore more of the statistical and ML methods learned in the FDS to the data
- Interpretation of the findings

You can use equations like this:

$$\bar{x} = \sum_{i=1}^{n} x_i \tag{1}$$

or maths inline: $E = mc^2$. However, you do not need to reexplain techniques that you have learned in the course – assume the reader understands linear regression, logistic regression K-nearest neighbours etc. Remember to explain any symbols use, e.g. "n is the number of data points and x_i is the value of the ith data point. ".

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 is causally linked to Revenue; the more votes, the more people bought movie tickets. The model uses Ordinary Least Squares regression and a constant column has been added to show y-intercept. The summary results are provided below.

Dep. Variable:	$Revenue^{\frac{1}{3}}$	**R-squared:**	0.209
Model:	OLS	**Adj. R-squared:**	0.200
Method:	Least Squares	**F-statistic:**	24.59
Date:	Thu, 30 Mar 2023	**Prob (F-statistic):**	8.23e-30
Time:	14:14:56	**Log-Likelihood:**	-851.28
No. Observations:	660	**AIC:**	1719.
Df Residuals:	652	**BIC:**	1754.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
$const$	146.6239	25.235	5.810	0.000	97.072	196.176
$Rank$	-0.0009	0.000	-6.886	0.000	-0.001	-0.001
$Year$	-0.0726	0.013	-5.799	0.000	-0.097	-0.048
$Metascore$	-0.0428	0.046	-0.927	0.354	-0.134	0.048
$MeanLeadRolesExp.$	0.1774	0.037	4.779	0.000	0.105	0.250
$ln(Runtime)$	0.1631	0.039	4.160	0.000	0.086	0.240
$ln(DirectorExp)$	0.0369	0.037	0.993	0.321	-0.036	0.110
$Rating^2$	0.0192	0.050	0.381	0.703	-0.080	0.118

Omnibus:	5.244	**Durbin-Watson:**	1.900
Prob(Omnibus):	0.073	**Jarque-Bera (JB):**	4.821
Skew:	0.154	**Prob(JB):**	0.0898
Kurtosis:	2.717	**Cond. No.**	1.52e+06

Table 4: Multiple Regression results summary

While this model does only explain 20% of the variance in the revenue, it is statistically significant, getting a very small $P(F-statistic)$. This indicates that the revenue of a movie can not be accurately predicted from data present in the dataset. However, despite the poor accuracy, there are a few interesting insights it provides.

One such observation is that the experience of lead actors has a statistically significant effect on movie ticket sales (p<0.05), while the experience of directors does not (p>0.05). This raises the question of whether the director has as much of an impact as the actors when it comes to selling tickets. A possible explanation is the fact that promotional posters for movies often focus on the actors[**label**], with their faces being the first thing a consumer sees. Actors who have been in many movies tend to be more recognisable, making potential consumers more likely to see a movie they are in and thus improving ticket sales. In contrast, the director's name is usually the only thing featured on the poster, drawing little attention and thus being less impactful. This idea is inline with previous research[**label**].

Another insight is that user rating and critic scores are not reliable predictors of a film's revenue, with p-values greater than 0.05. This raises the question of how poorly rated movies can still make money in the box office. One explanation could be that ratings require someone to watch the movie, and since box office revenue is only generated while the movie is in cinemas, the rating and metascore don't have enough time to significantly impact the movie's revenue. A good example of this is Star Wars: The Rise of Skywalker, which made $1.074 billion despite having a low rating and metascore of 6.5 and 53%, respectively [**label**].

The same approach was used to further analyse the relationship between a movies rating and the other data collected about it. In this case, the target variable was the normalised user rating, with the rest of the dataset as the predictor variables, excluding Genre and Title. $Votes^{\frac{1}{3}}$ was not excluded here as there is no reason to believe it is causally related; the more votes a movie receives doesn't necessitate high ratings, i.e. Star Wars: The Rise of Skywalker. The model uses Ordinary Least Squares regression and a constant column has been added to show y-intercept. The summary results are provided below.

Dep. Variable:	$Rating^2$	**R-squared:**	0.421
Model:	OLS	**Adj. R-squared:**	0.416
Method:	Least Squares	**F-statistic:**	84.93
Date:	Thu, 30 Mar 2023	**Prob (F-statistic):**	1.11e-92
Time:	17:02:17	**Log-Likelihood:**	-926.48
No. Observations:	826	**AIC:**	1869.
Df Residuals:	818	**BIC:**	1907.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
$const$	-48.4648	21.934	-2.210	0.027	-91.518	-5.412
$Rank$	-0.0003	0.000	-2.971	0.003	-0.001	-0.000
$Year$	0.0242	0.011	2.219	0.027	0.003	0.046
$MeanLeadRolesExp.$	-0.0588	0.029	-2.059	0.040	-0.115	-0.003
$Revenue^{\frac{1}{3}}$	-0.4241	0.038	-11.234	0.000	-0.498	-0.350
$Votes^{\frac{1}{3}}$	0.7941	0.045	17.464	0.000	0.705	0.883
$ln(Runtime)$	0.1987	0.030	6.732	0.000	0.141	0.257
$ln(DirectorExp)$	-0.0877	0.028	-3.139	0.002	-0.143	-0.033

Omnibus:	8.164	**Durbin-Watson:**	1.960
Prob(Omnibus):	0.017	**Jarque-Bera (JB):**	8.964
Skew:	-0.170	**Prob(JB):**	0.0113
Kurtosis:	3.380	**Cond. No.**	1.75e+06

Table 5: Multiple Regression results summary

This model performs respectably, explaining about 41% of the variance in user ratings. It is also statistically significant, getting a very small $P(F-statistic)$. This indicates that the user rating of a movie can be predicted from the data present in the dataset, albeit not extremely well. However, the relatively good R^2 value does indicate a well performing model, as such we can postulate that users do use similar data when deciding the quality of a movie. Another good thing about this model is that each predictor has a low p-value, p<0.05. This indicates that each of these movie features have a statistically significant impact on user rating.

5 Discussion and Conclusions

Summary of findings

Evaluation of own work: strengths and limitations

Comparison with any other related work E.g. “Anscombe has also demonstrated that many patterns of data can have the same correlation coefficient”.

Wikipedia can also be cited but it is better if you find the original reference it for a particular claim in the list of references on the Wikipedia page, read it, and cite it.

The golden rule is always to cite information that has come from other sources, to avoid plagiarism.

Improvements and extensions

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.