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 datasets to aid our investigation: an IMDB movie dataset, containing various details about movies made from 2006-2016; and a larger movie dataset, containing details about movies released on or before July 2017. The IMDB dataset has a bit of a contentious past as it was scraped off of the movie ranking website IMDB.com, and is actually only a sample dataset from a much larger dataset that has every movie made from 2006-2016 that is in the IMDB. The large movie dataset (TMD) was collated using data from TMDB (The Movie DataBase) and grouplens.org; but we only use the part that was obtained from TMDB as it is used to find Director and Actor experience.

Both datasets were downloaded from kaggle.com, a data science platform that enables users to access and share datasets. These datsets are shared underneath the CC0 1.0 Universal Public Domain Dedication, as such we will use them underneath fair use.

Data description The IMDB movie dataset has 12 columns, all either containing string values or floating point values. The only odd column is the Genre column which contains the different genres that can be applied to a particular movie. The genres that are in this dataset are the arbritary genres:

- Action
- Adventure
- Sci-Fi
- Mystery
- Horror
- Thriller
- Animation
- Comedy
- Family
- Fantasy
- Drama
- Music
- Biography
- Romance
- History
- Crime
- Western
- War
- Musical
- Sport

The column summary is shown in Figure 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 Category
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

Figure 1: The different columns in the IMDB-Movie-Data.csv file

The TMD dataset used had a bit of a more odd structure. We only used the crew.csv file that is a small part of the larger dataset, so we only explain the structure of crew.csv. This file was composed of three columns: cast, crew and id. The odd part of the dataset is that the cast and crew columns are composed of .json files, and as such we needed to parse out the useful data from these json files to actually get useable data. The column summary is shown in Figure 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. This is used in the larger dataset to connect data together	Integer

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

Data processing How you have processed the dataset, e.g., cleaning, removing missing values, joining tables. To cleanup the f

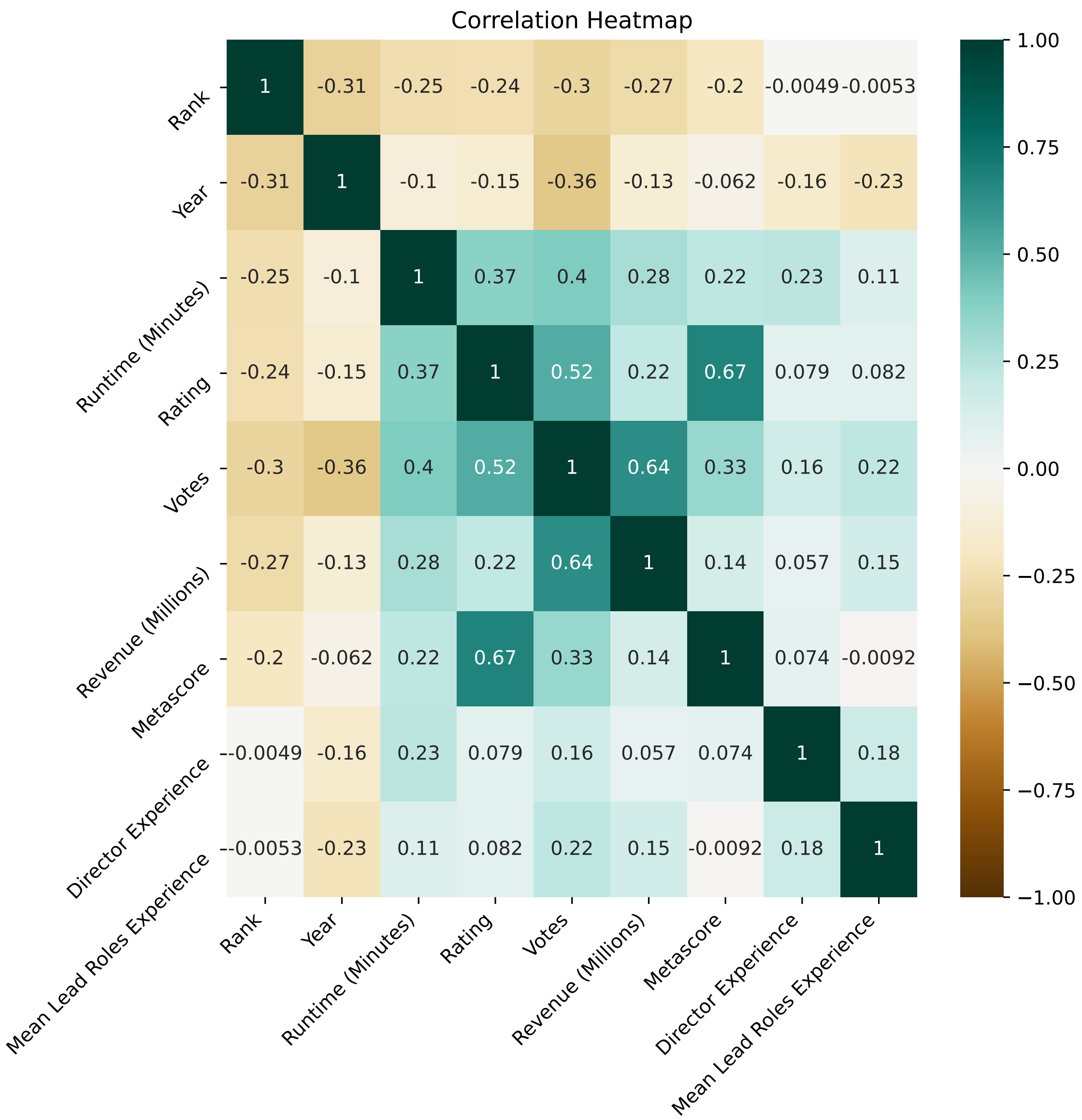

Figure 3: Demonstration figure. This caption explains more about the figure. Note that the font size of the labels in the plot is 9pt, which is obtained by the settings as shown in the Jupyter notebook.

4 Exploration and Analysis

A data science analysis of the paper, including:

- Visualisations (for example Figure 3) and tables (for example Table 1). 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:

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

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. ".

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`.

Table 1: Excerpt from Scottish Index of Multiple Deprivation, 2016 edition. `https://simd.scot`. You may put more information in the caption.

Location	Employment	Illness	Attainment	Drive Primary	Drive Secondary	Crime	…
Macduff	10	95	5.3	1.5	6.6	249	…
Kemnay	3	40	5.3	2.4	2.4	168	…
Hilton	0	10	6.3	2.2	3.0	144	…
Ruchill	8	130	4.9	1.7	5.6	318	…
Belmont	2	50	6.1	3.1	3.2	129	…
…	…	…	…	…	…	…	…

[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`.