MAST 6251 - HW 2

2023-02-19

Group Members: Kinnera Puppala, Sumita Nair, Sahar Khan, Ganesh Nimmala, Varun Gokhale, Harsh Tandel

Introduction

The primary aim of this analysis was to identify and report important factors from the 'Full MovieLens' dataset that contribute to determine whether a given movie is considered 'good' or 'bad'. Given that the dataset had over 10 data points, it was important to identify the key data points from the cohort that have a higher influence on a movie's success, when compared to others. The first step in that process was data cleaning and manipulation which allowed us to have a better understanding of the data. Furthermore, identifying a key set of data points allowed us to compare them with user ratings data (acquired from the official GroupLens website) and helped set a narrative for this report.

The data points that we selected as a base factor in determining whether a movie is considered good or bad were user-ratings and revenue. Our definition of a 'good' movie is a movie that has a rating score of 6 or higher and a verdict set to 1 for such movies. Any movies with a rating score lower than 6 are 'bad' movies and a verdict is set to 0 for those. We also considered the effects of other data points in the dataset that contribute to a poor rating score and hence classify a movie as 'bad'. These data points include, but are not limited to, movie meta data such as the director, production house, runtime, and popularity. Also, including the associated data points and their effect on ratings score helped us in building a robust model for predicting whether a movie was 'good' or 'bad'. All the other data points are compared to our base data points in determining which factors have the most influence on a movie's rating score.

Executive Summary

- In order to improve the potential for higher ratings, it is recommended that movies falling under the genres of Animation, Comedy, Action, History, and Science Fiction, which are at least 106 minutes in length and have top directors, top cast members, a popularity rating of 9 or greater, a revenue of at least 7 million, and have been released after 2003, should be selected. It should be noted that the Family, Drama, Action, and Documentary genres have the highest impact on revenue based on the Revenue model. As a result, selecting action movies could provide a good tradeoff for better value for money and a higher success rate between revenue and ratings.
- Despite the notion that only popular movies earn good revenue, it is important to note that some lesser-known movies, including critically acclaimed works and foreign language films with a specific audience, have also achieved substantial financial success.
- It is worth noting that a movie's number of votes does not necessarily reflect its rating. In many cases, movies with fewer votes have achieved high ratings, suggesting that a smaller but more dedicated audience can sometimes be a better indicator of a movie's quality.
- While it is generally true that popular movies tend to receive high ratings due to their wide viewership, it is important to recognize that not all popular movies are highly regarded by audiences. As such, a high rating may not always be indicative of a movie's quality or appeal to individual viewers.

Assumptions and Modifications

In order to analyze the dataset, we had to make certain assumptions and modifications that would streamline the findings and allow us to make an objective conclusion. The following assumptions and modifications were made based on our exploratory data analysis:

Ratings score: The TMDB ratings scale had a range from 1 through 10, with 1 being the lowest rating possible and 10 being the highest rating possible. We rescaled the movie ratings into TMDB scale by calculating the weighted average rating for each movie. On the basis of these values, we set a ratings value of 6 as a threshold value. Any movie with a score of 6 and above was considered a 'good' movie and any with a lower score than 6 was a 'bad' movie.

Directors: We also classified the Top 22 directors according to TMBD as 'good' directors when compared to the others in that list, who were classified as 'bad' directors.

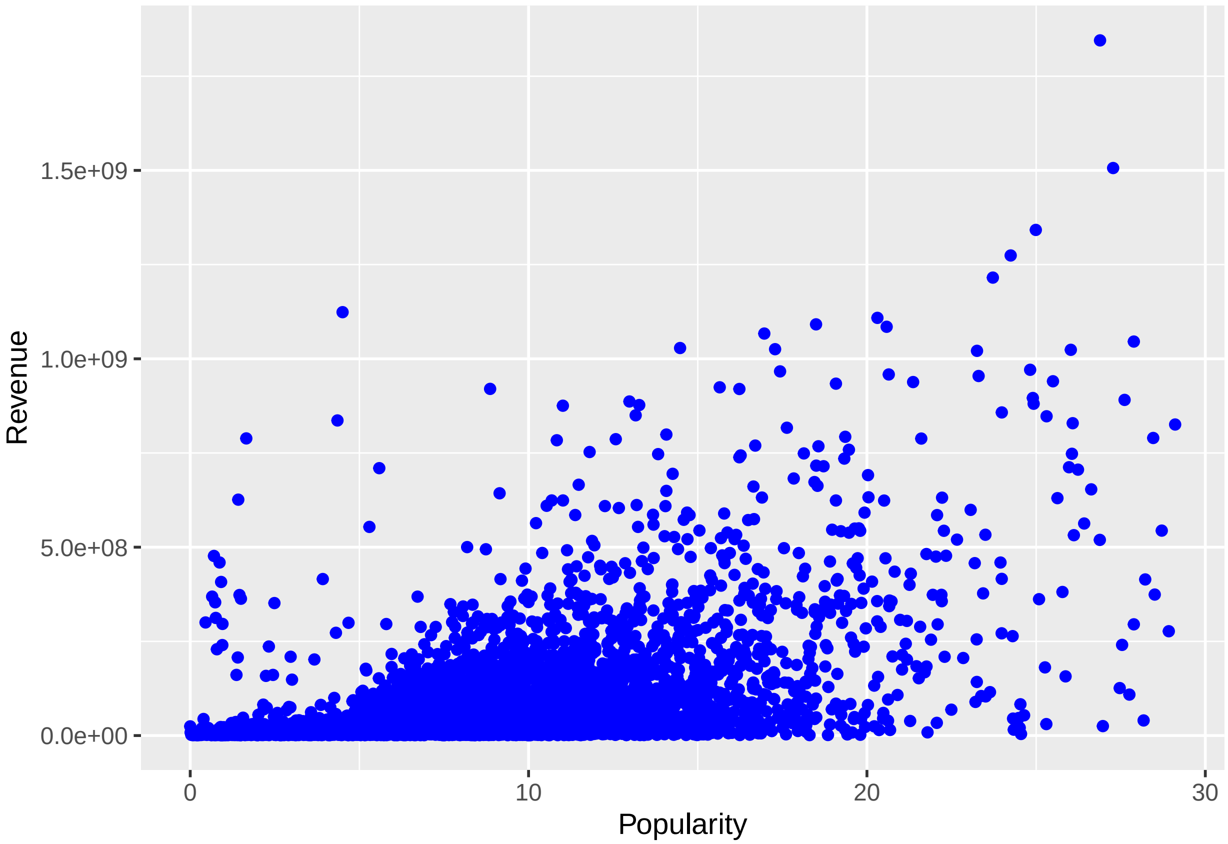
Actors: We also classified the Top 70 actors according to TMBD as 'good' actors when compared to the others in that list, who were classified as 'bad' actors.

Production Houses: We also classified the Top 16 production houses according to TMBD as 'good' production houses when compared to the others in that list, who were classified as 'bad' production houses.

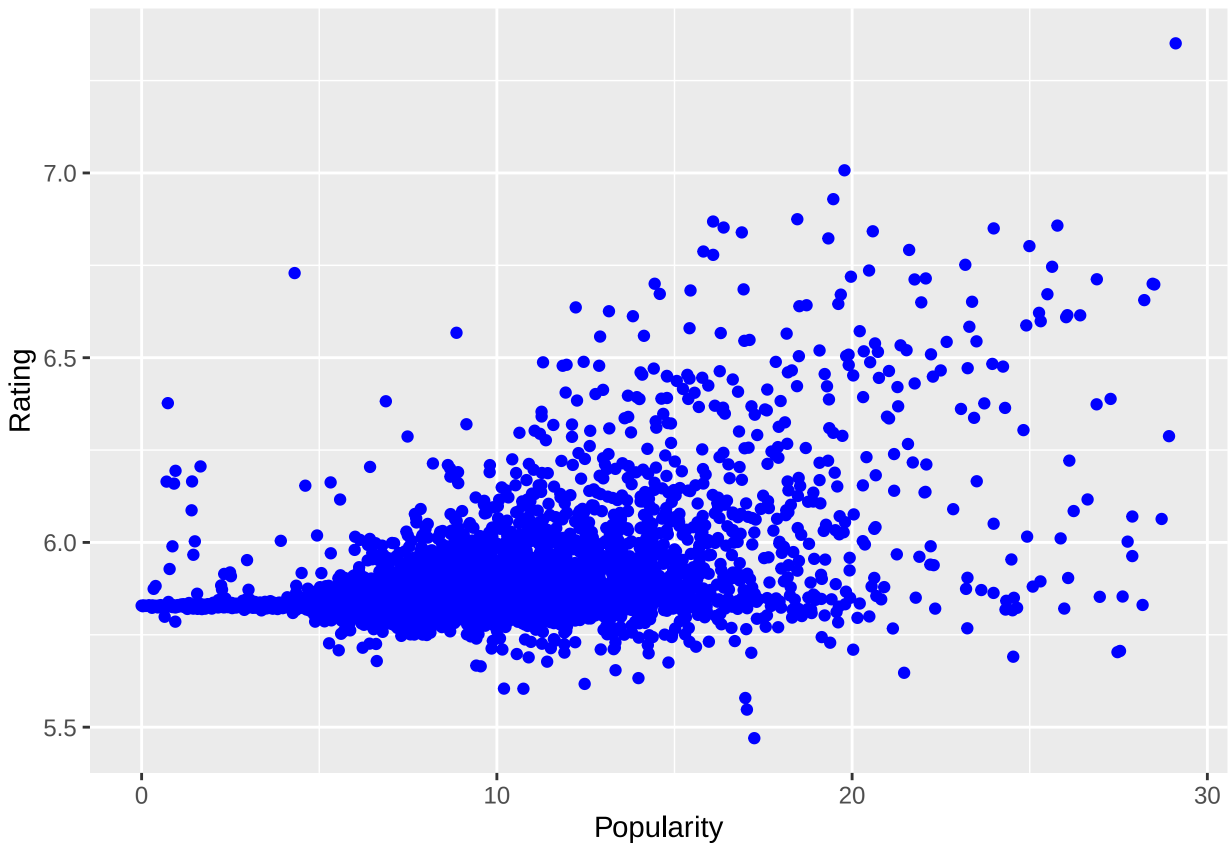
Movie Length: Another assumption we made was based on the runtime of a movie. We assumed that any movie above 106 minutes in length was considered a 'long' movie and anything below that was a 'short' movie.

Analysis/Outputs

1. Scatter plot Popularity VS Revenue: While most popular movies have earned good revenue, there are some less popular movies that have also earned good revenue (these may be movies that were critically acclaimed, foreign language movies with a niche audience)



2. Scatter plot Popularity VS Rating: Generally, popular movies tend to be highly rated. As most people watch these popular movies, they tend to receive a favorable rating



Our prediction model results in an accuracy of about 88%, allows us to generate predictions without overfitting it.

Insights from analysis that influence movie ratings:

- Most of the movies that received high ratings (above 6) were 'long' movies, which means they were greater than or equal to 106 minutes in length
- Most highly rated movies were directed by top 22 directors from the TMDB director's list
- Most highly rated movies had a revenue greater than the average revenue of $9.56 million

Recommendations

Our goal was to build a model that could predict the success of a movie based on various factors. We found that there are several important factors that can greatly influence a movie's success. In comparison with cast and director that certainly add value, they both often come at a high cost. In our model, we found that having a great director had a stronger impact than the cast and production, which could be a balanced trade-off between expense and success factors.

In addition, there are other good factors that contribute to the success of the movie:

1) Run time between 75-150 minutes
2) Genres: Revenue model - Animation, Comedy, Action, History and Science Fiction Rating model - Family, drama, action, and documentary Good tradeoff - Action movies
3) Popularity: Popular movies made it to the list of high rated movies. Thus, movies should be very well marketed or advertised