Predicting Movie Ratings

Movie Money Matters, jsfarrel, bkaravan, ystepane

Goal

A movie's IMDb rating is a quantitative measure of success, but will not always correlate to other benchmarks such as box-office gross. The use case for this project is to determine underlying factors that influence a movie's rating. Underlying factors include statistics such as the writer's name, runtime, budget, gross, genre, release date. Thus, we are most interested in the following: given a movie's statistics, predict the movie's average IMDb rating.

Data

Our dataset consists of various movie attributes and their average IMDb ratings. We combined existing IMDb datasets with web scraped box-office statistics from The Numbers. Our dataset consisted of 10,715 entries, reduced to 1,311 after cleaning. Cleaning included removing duplicate movie entries and entries with missing data points. Preprocessing included formatting dates and monetary values so they are uniform across the dataset and easy to analyze later on. Our final dataset had 14 attributes, including movie name, budget, domestic gross, worldwide gross, date (day, month, year), writers, directors, genres, runtime, number of votes, and average IMBd rating.

Model+Evaluation Setup

We trained Elastic Net and Random Forest models to predict the rating of a movie based on the attributes we have in our dataset. We chose to represent movie ratings as numerical continuous values, rather than categorical, therefore making this a regression problem. We believed a categorical representation may be an oversimplification of the problem. The target variable was IMDb average rating, and attributes included everything but the movie's title. During preprocessing, columns were transformed using a one-hot encoder for categorical values, ordinal encoder for ordinal values, and standard scaler for numerical values. We used a K-fold cross validation pipeline consisting of 5 splits to tune hyperparameters. The models were evaluated using mean squared error to penalize large errors and prevent overfitting of the data. We performed an 80%-20% train-test split of the dataset.

Results and Analysis

Claim #1: Both regression models outperform the baseline by a significant margin.
Support for Claim #1: Table 1 shows the MSE of Random Forest and Elastic Net models and the baseline on the test dataset. Both models MSE < 0.850, while the baseline > 1.100.

Model	**Mean Square Error (MSE)**
Random Forest	0.596
Elastic Net	0.808
Baseline	1.145

Table 1: MSE of different models on testing dataset

Claim #2: The relationship between our movie dataset attributes and movie rating is non-linear.

Support for Claim #2: Random Forest is a non-linear model, while Elastic Net is a linear regressor. The relationship between the attributes in our dataset and movie rating is non-linear, because as Table 1 (above) shows, our movie rating predictions are best with non-linear Random Forest compared to Elastic Net and the baseline.

Claim #3: Number of votes and runtime are the two most important factors for a movie's rating.
Support for Claim #3: After determining the best model from Table 1, Random Forest, we ranked features by permutation importance and global feature importance in Figures 1 and 2, respectively. The features that stand out as being most important globally and due to permutations include runtime in minutes and number of votes.

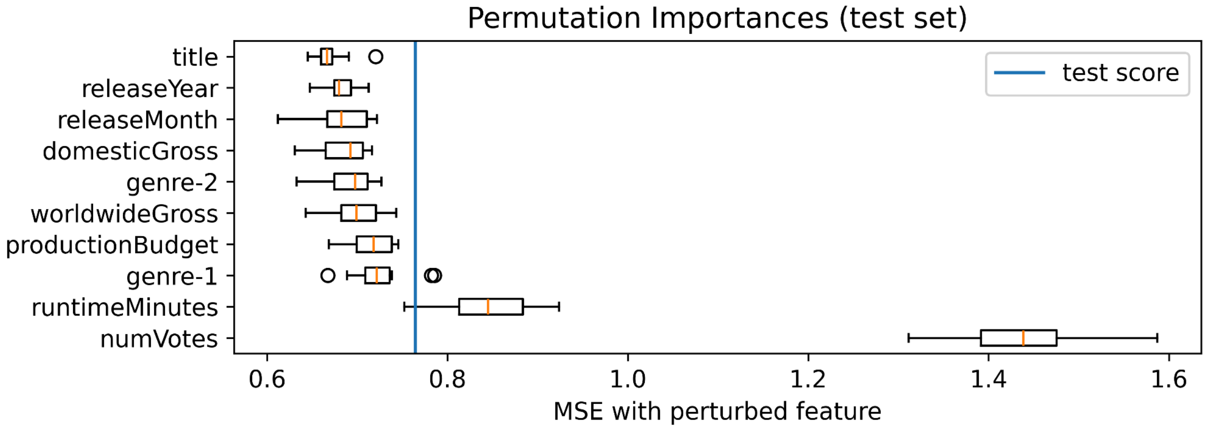

Figure 1: Permutation importance for Random Forest. Higher MSE means greater significance.

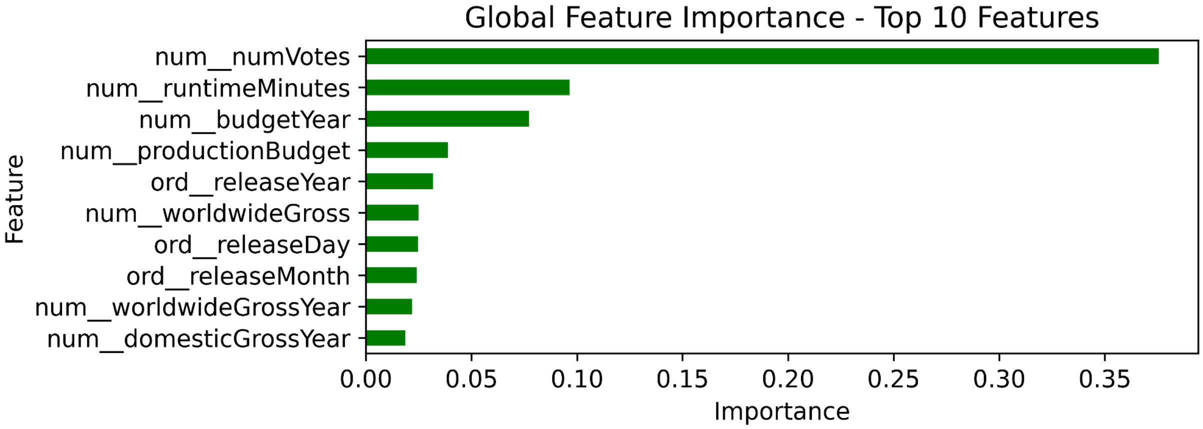

Figure 2: Random Forest global feature importance.

Recorded Video Presentation

Socio-historical Context and Impact Report

Socio-historical Context

- ○ Research the socio-historical context of your project to identify a few societal factors that could affect your data, prediction goal, and/or hypothesis.

When predicting a movie's rating, one additional factor that is very hard to track with data is the current societal trend. The trend can directly impact the popularity and social outreach of the movie, which, in turn, can define how people like it and rate it. Over the past decade, there has been a growing demand in properly representing different cultures and backgrounds. As our source states, diversity matters, and "[diversity] provides some of the best opportunities for unique storytelling, especially as things change" ("Entertainment Industry Trends"). We can draw a conclusion that movies that follow this trend have a better chance of reaching a better overall audience ranking.

- ○ Who are the major stakeholders in this project?

In the context of movie rating, there are two main groups of stakeholders: consumers and producers. Our project focused on the classical definition of a movie and consulted the IMDb database, where the first group, the consumers, can directly "assess" the work of the second group. However, we did not account for a newer popular method of entertainment: streaming services. Undoubtedly, they compete for a similar audience, and there are clear things that affect the movie-making industry. On the one hand, content creators can benefit from the steady revenue stream that gets provided from a giant like Netflix. On the other, people seem to lose their power of "monetary vote" – there is no need to go to the theatre and promote a movie with your dollars when you can watch all of them on a streaming platform. This effect can discourage producers from creating quality pictures that are aimed at a higher voting and gross average, but rather, quicker content to receive a cut from a streaming service ("The Impact of Streaming Services").

- ○ Summarize the most relevant technical or non-technical research that has already been conducted about your project topic.

There have been a couple of attempts to figure out the relationship between a movie's components and its rating. One study was particularly close to our results, as they determined that Vote, Gross, Recency, and Duration all play a role in a movie's average rating (consistent with our results), while language, country, and genre might not be as responsible for the rating (“The Most Influential Factor of IMDb Movie Rating: Part II ”). With that said, there was another study arguing that higher-rated IMDb movies do not necessarily get more worldwide gross ("What Makes a High-Grossing Movie."). What we might agree on is that movies really depend on its audience, both in financial value and its average rating, so increasing the movie's potential outreach might be the best potential strategy to increase its statistics.

Ethical Considerations:

- What kind of underlying historical or societal biases might your data contain? How can this bias be mitigated?

Our data may reflect historical biases present in the film industry. Such biases can include underrepresentation of certain demographics in leading roles or directorial positions. This could skew the analysis towards the success metrics of a non-diverse set of movies. To mitigate this bias, we could incorporate diversity indices or control for demographic variables in the analysis, ensuring a more inclusive understanding of movie success.

The systems used to collect data, such as those we used – IMDb and The Numbers, may have inherent biases based on user demographics or industry reporting practices. For example, user ratings on IMDb might not represent the global audience if certain groups are less likely to use the platform. It is definitely very difficult to achieve for every movie as some are more popular in certain regions, or are only available in those regions.

- What biases might exist in your interpretation of the data?

Our biases might include:

- Confirmation Bias: There's a risk of favoring information that confirms pre-existing beliefs or hypotheses. For instance, if one believes that big-budget films always succeed, they might overlook data showing successful low-budget films.
- Selection Bias: The data analyzed might not represent the entire population of films, especially if it's limited to certain databases like IMDb or The Numbers, which may not fully capture global or independent cinema.

- Is data being used in a manner agreed to by the individuals who provided the data?

In our context, using the data from IMDb and The Numbers allows for analysis and research. Additionally, if we used any reviews that include personal data, it should have been anonymized for privacy reasons.

- What are possible misinterpretations or misuses of your project results and what can be done to prevent them?

Some of the misinterpretations of the project might include:

- Overgeneralization: The findings may not apply universally to all films or film industries globally. To prevent this, it's important to communicate the scope and limitations of the analysis clearly.
- Causation vs. Correlation: There's a risk that correlations found in the data (e.g., between the release month and the box office) might be interpreted as causal relationships. Emphasizing the difference between correlation and causation in the presentation of results can help mitigate this.

To prevent misinterpretations or misuses, it's important to provide comprehensive documentation of the methodologies, assumptions, and limitations of the analysis. We believe this is done well in our ReadMe and the poster. Additionally, engaging with a diverse set of stakeholders for feedback and perspective can help ensure the results are understood and applied appropriately.

Works Cited

"Entertainment Industry Trends." Pepperdine University, https://bschool.pepperdine.edu/blog/posts/entertainment-industry-trends-2022.htm.

"The Impact of Streaming Services on the Movie Industry." New York Film Academy, https://motionpicture.edu/socialcinema/2023/06/19/the-impact-of-streaming-services-on-the-movie-industry/.

Dai, Yuri. "The Most Influential Factor of IMDb Movie Rating: Part II - Data Analysis and Statistical Modeling." Medium, https://medium.com/@yd334/the-most-influential-factor-of-imdb-movie-rating-part-ii-data-analysis-and-statistical-modeling-c6300b8d7d4d.

Dr Turner, Shruti. "What Makes a High-Grossing Movie." Towards Data Science, https://towardsdatascience.com/what-makes-a-high-grossing-movie-41ce3b2d0a6f