Movie Money Matters

Julia Stepanenko, Bohdan Karavan, John Farrell

Introduction

There are numerous factors that drive a film's success. Acknowledging the several agents at play, our project hones in on a particular metric: the correlation between movie metrics and their impact on the silver screen. 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. This project aims to determine how such factors influence a movie's rating.

Hypotheses

We decided to investigate what part of the movie contributes the most to its "success". Using machine learning, we could determine which weight was the most significant when predicting movie ratings.
Additionally, we considered the following hypotheses:

1. Release Timing Hypothesis
2. Directorial Influence Hypothesis
3. Budget Correlation Hypothesis

Data

Our data came from:

- The Numbers: Web scraped the box-office results and budgets of many movies.
- IMDb: The go-to database for movie ratings and statistics, offering a comprehensive view of a film's reception.

We curated a dataset of 10,715 entries, which was refined to 1,311 after cleaning and removing duplicates. The steps we took include:

- Identified and removed duplicate records.
- Checked for and addressed any missing data points.
- Combined data, resolving issues with movies having the same title and release year.
- Focused on movies with complete data sets, including director, box office, release dates, and ratings.

Methodology

After aggregating the data into a pandas dataframe, we decided to use the following libraries to help with our project:

- Hypothesis testing: scipy, matplotlib, statsmodels
- Machine Learning: sklearn, joblib, numpy, pandas

Analysis & Results

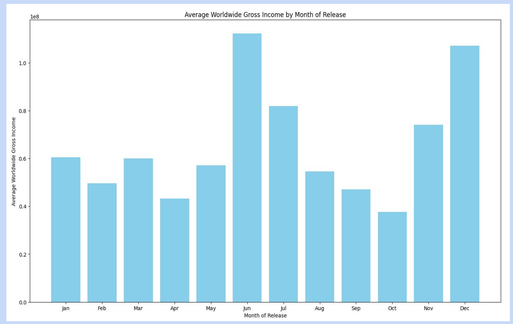

Figure 1: Month vs Worldwide Gross

December Result:
T-statistic: 1.84
P-value: 0.067
Not significant

June Result:
T-statistic: 2.65,
P-value: 0.009.
Significant

Hypothesis 1
To test our Release Timing and Worldwide Gross hypothesis, we used an independent t-tests for each month. The visualization Figure 1 helps visualize the potential outliers: June and December.

The **significant** results for June confirm that movies released during this summer month tend to enjoy **higher** worldwide gross, likely benefiting from school holidays and peak movie-going season.

Hypothesis 2
For the Director's Influence on Average Rating hypothesis, we created an additional "Top director" dataframe column that would mark directors with at least 3 entries in the dataframe. We then ran a Chi-squared test of independence for movies that were "High" or "Low" in rating compared to the mean of the dataframe.

Chi-squared: 18.43
P-value: 0.00001766
The significant association indicates that experienced directors are an important factor in achieving higher ratings, reflecting their ability to deliver quality films that resonate with audiences.

Hypothesis 3
To test our Budget's Correlation with Gross and Rating hypothesis, we used a Pearson's correlation coefficient.
Correlation coefficient: 0.60
P-value: < 0.001.
We can see a strong positive correlation between production budget and worldwide gross suggests that movies with larger budgets have a better chance of achieving higher earnings, possibly due to more extensive resources and marketing.

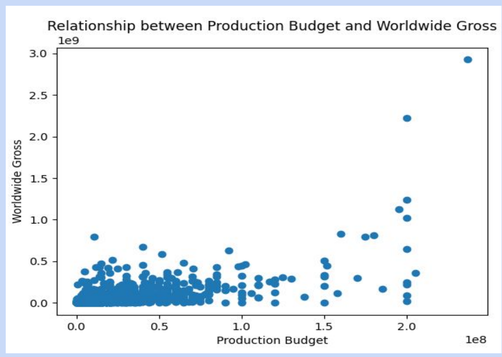

Figure 2: Product Budget vs Worldwide Gross

Machine Learning Analysis
We trained an Elastic Net and a Random Forest model on our movies dataset. Our target variable was movie rating, with other dataset values being attributes. We opted for regression over classification to capture the nuances of numerical ratings. The mean squared error (MSE) was our chosen metric, as it penalizes larger errors, helping to prevent overfitting.

The Random Forest model outperformed Elastic Net, with an **MSE of 0.596 compared to 0.808**, indicating a complex, non-linear relationship between features and ratings. This performance was significantly better than our baseline model's **MSE of 1.145.** We determined the most influential features by performing permutation importance of all features. The most influential features for predicting ratings were found to be **movie runtime, number of votes, and budget variations**.

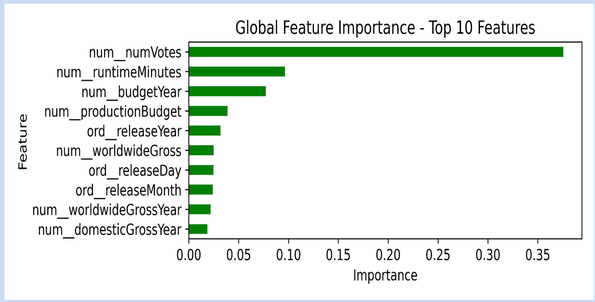

Figure 3: Random Forest global feature importance.

Challenges

Data Cleaning: Our dataset was merged from scraped data of The Numbers box office statistics and various IMDb datasets. The uncleaned dataset included many missing values and duplicate movies. Cleaning this dataset was challenging, due the high quantity of duplicates, but strategically removing duplicates and missing values allowed us to maintain the original IMDB average ratings proportions of our dataset, normally distributed.

Reverse Causality: Reverse causality influences directors to make certain marketing decisions based on anticipated outcomes. For example, directors may be influenced by holiday season masses to schedule big premiers during peak holiday months like June and December. However, in our movie rating analysis, distinguishing between cause and casualty is critical.

Conclusions

- Our findings align with our initial hypothesis that budget and popularity (number of votes) would influence ratings. The significance of runtime was unexpected but suggests that mainstream, feature-length films are rated more favorably.
- June Releases: Significantly higher worldwide gross, likely due to holidays.
- Budget and Success Correlation: Strong Positive Correlation: Higher production budgets are associated with greater worldwide gross earnings.
- Directorial Influence: Experienced directors are more likely to produce highly-rated movies.
- Random Forest's superior performance indicates a non-linear relationship between features and movie ratings.
- Key Predictive Features: Movie runtime, number of votes, and budget variations are crucial in predicting ratings.
- Findings support the initial belief that budget and popularity influence ratings.
- An interesting key point for future research could be movie's profit potential. It can expand our third hypothesis and check how much higher budget movies differ in their potential profit as a percentage of budget.
- As it is challenging to determine a "top" director, our second hypothesis might need cross checking with more data outside of our immediate dataset.