DSA 210: Final Report

Project Title: The Relationship Between Awards, Box Office, and IMDb Ratings

Prepared by: Ozan Avşar

Student ID: 34138

Submission Date: 01/08/26

1. Motivation

The film industry is a high stakes environment where millions of dollars are invested with the hope of both commercial success and critical acclaim. However, a recurring question in the entertainment industry is whether financial investment correlates with quality. Does a $200 million budget guarantee audience satisfaction, or is "cinematic quality" driven by intangible factors?

My primary motive for this project was to determine if financial investment (Budget/Box Office) truly correlates with high audience ratings, or if artistic recognition (Awards) is a more reliable predictor. Additionally, I wanted to test my own cinematic taste against the "wisdom of the crowd" by comparing my personal ratings to global IMDb scores. This comparative analysis aims to quantify the "subjectivity gap" between an individual critic and the general public.

2. Datasets & Data Enrichment

The project utilizes a multi-source data pipeline to create a comprehensive dataset of 3,173 movies.

2.1 Data Sources

- **Box Office Mojo:** This served as the financial backbone of the project. It provided production budgets, domestic gross, and worldwide gross figures.
- **OMDb API (Open Movie Database):** To capture "quality" metrics, I enriched the financial data using the OMDb API. This provided metadata including:
    - **Awards:** Textual descriptions of wins and nominations (e.g., "Won 3 Oscars").
    - **Metascore:** A weighted average of critic reviews.
    - **IMDb Votes:** The volume of audience engagement.
- **Personal Ratings:** A curated list of my own movie ratings was merged to perform the bias analysis.

2.2 Data Cleaning & Enrichment Process

A critical challenge in this project was merging datasets with slightly different naming conventions. I employed a rigorous cleaning strategy:

- **Merging Strategy:** The primary join key was imdbID. For records lacking IDs, I used a composite key of Title + Year, cleaning punctuation and spacing to maximize matches.
- **Feature Engineering:**
    - **Log-Transformation:** Financial variables (Budget, Gross) were extremely right-skewed (power-law distribution). I applied a logarithmic transformation (np.log1p) to normalize these features for the Machine Learning model.

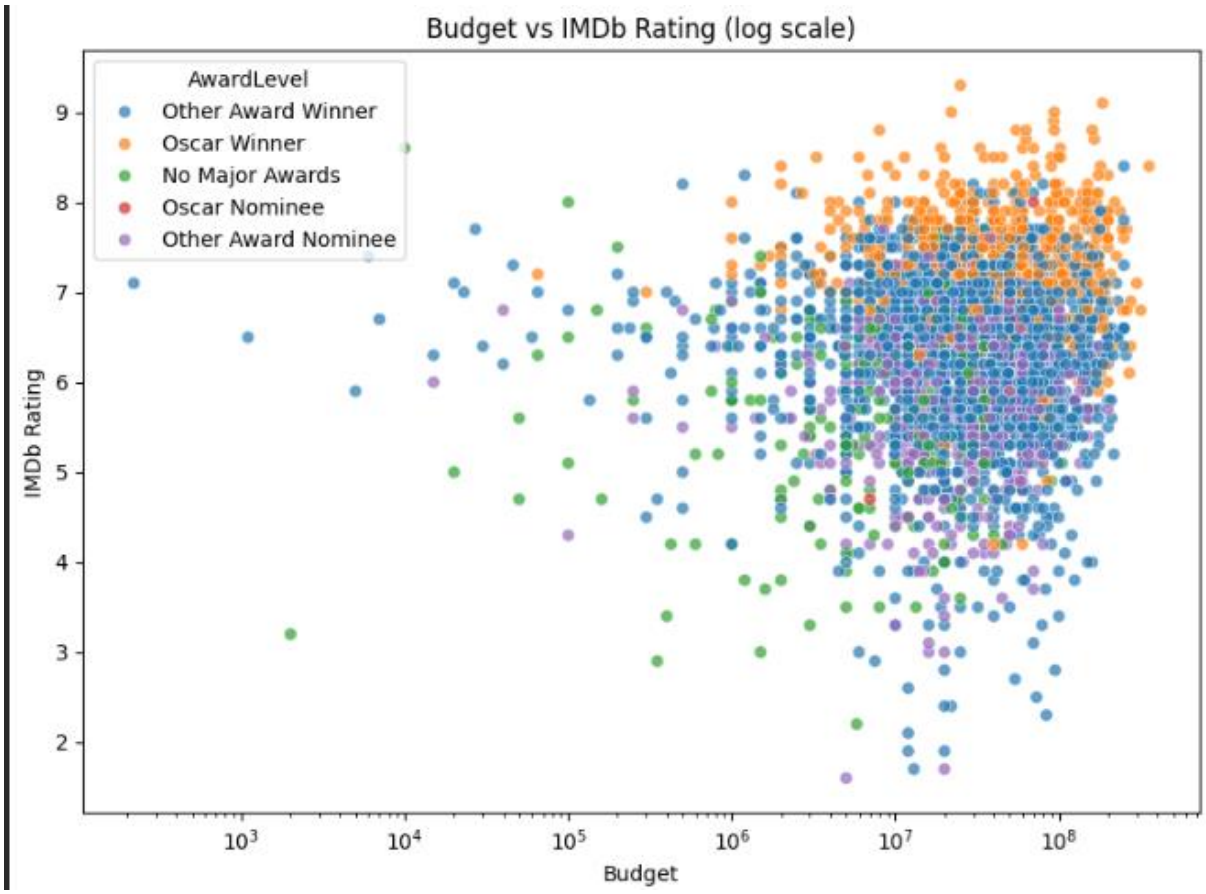
- **RatingGap:** I created a custom variable RatingGap = MyRating - IMDbRating to quantify personal bias.
- **Award Parsing:** Unstructured text from the API was parsed to create categorical levels: *Oscar Winner*, *Oscar Nominee*, *Other Awards*, and *No Awards*.

3. Exploratory Data Analysis (EDA)

The EDA phase focused on understanding the distributions and relationships between money, fame, and quality.

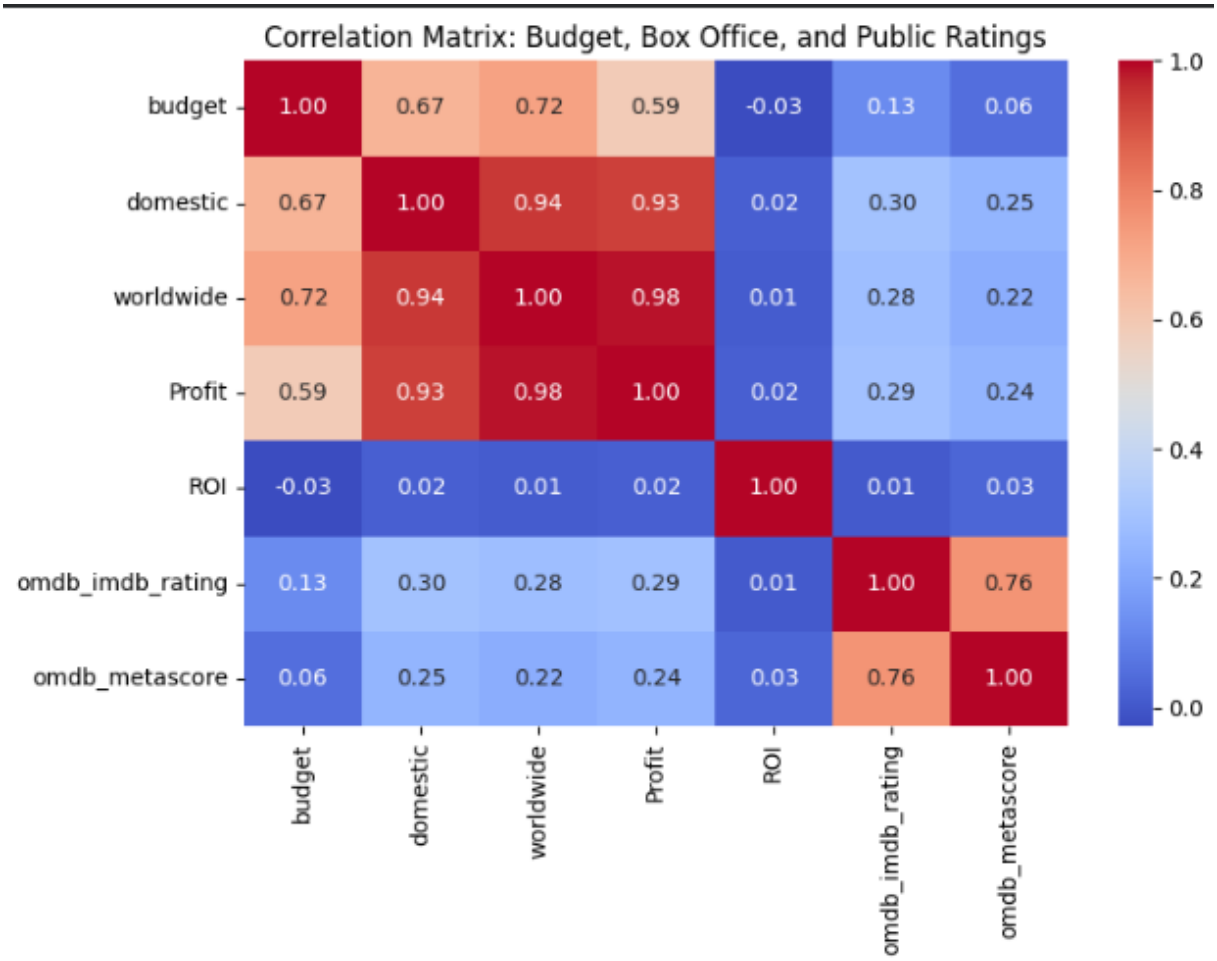
3.1 Financial Distributions

A small fraction of "Blockbusters" accounted for the majority of the total budget and revenue, necessitating the log-transformations used later.



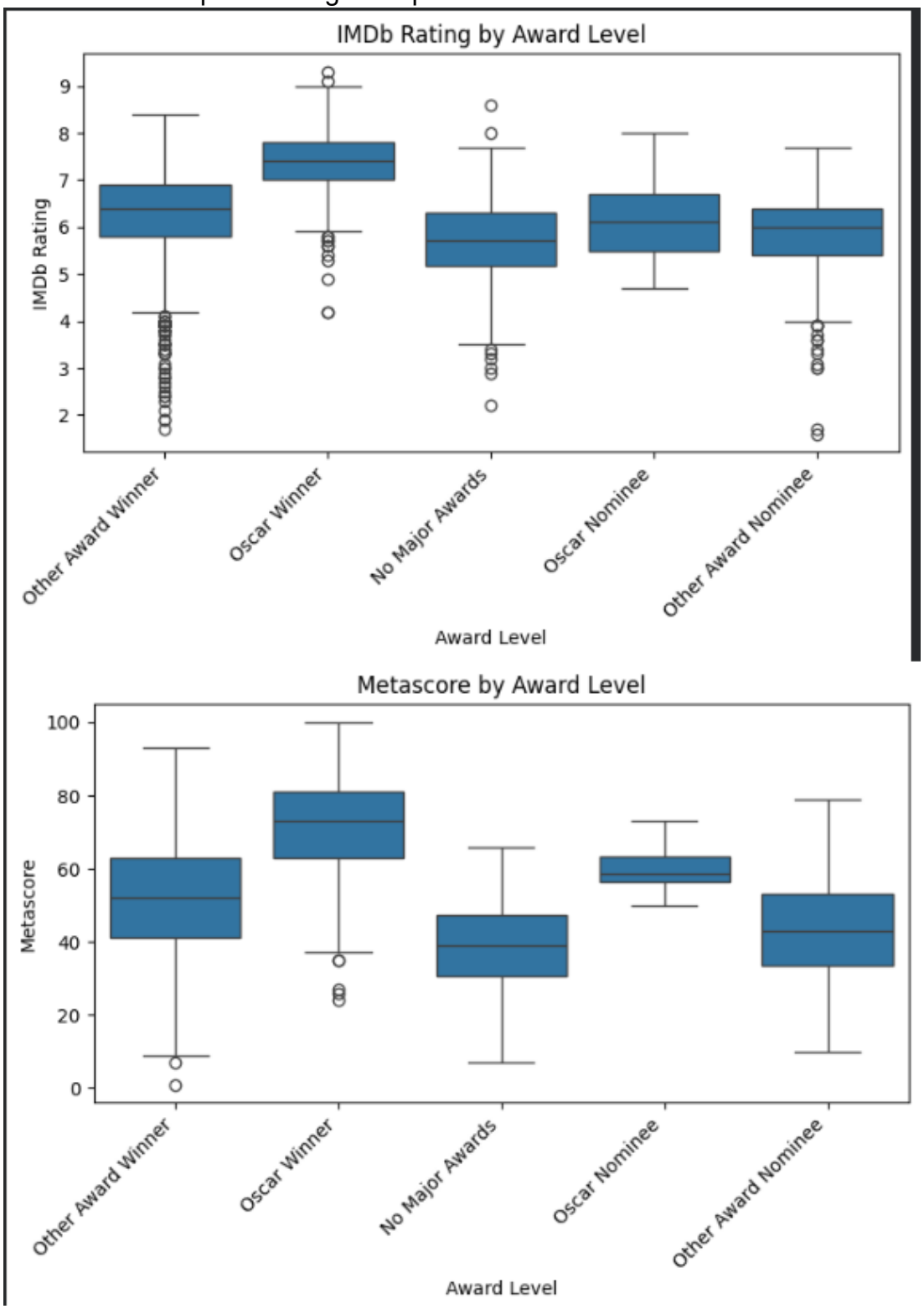
3.2 Correlations: Money vs. Quality

Visualizing the relationships revealed distinct patterns. While Budget and Box Office Gross were strongly correlated with each other ($r > 0.7$), their correlation with IMDb Rating was surprisingly weak ($r \approx 0.1$). This visually confirms that spending more money does not linearly increase audience satisfaction.



3.3 The "Award" Effect

Boxplots illustrating the relationship between Award Status and Ratings showed a clear hierarchy. Movies that won Oscars had a significantly higher median rating and a narrower interquartile range compared to non-awarded films.

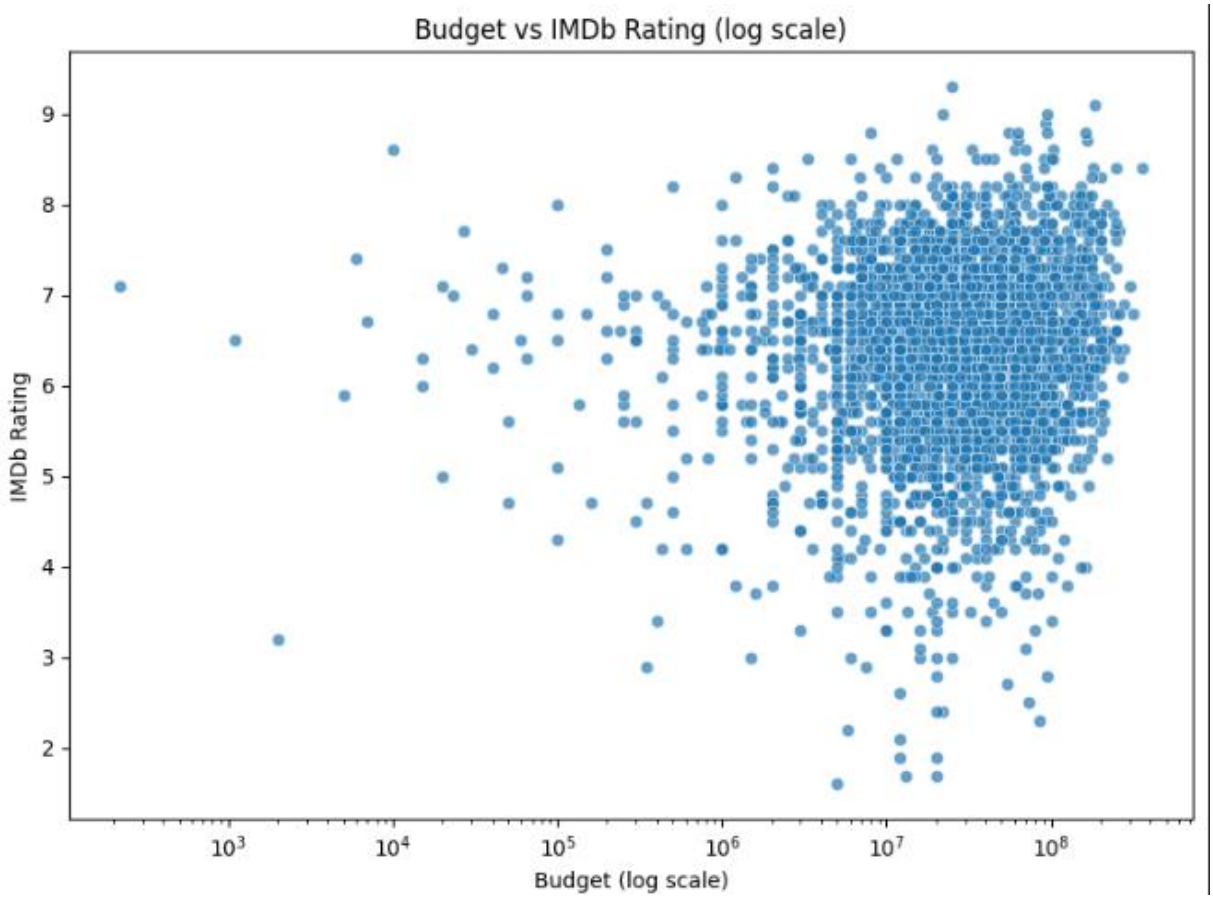


4. Hypothesis Testing

To statistically validate the visual findings, I performed the following tests using a significance level of α = 0.05.

4.1 Budget vs. Ratings

- Hypothesis (H_1): Higher budget films tend to receive higher IMDb ratings.
- **Test:** Spearman Rank Correlation (chosen due to non-normality of budgets).
- Result: The correlation (r) was very weak positive (approximately 0.13). While statistically significant due to the large sample size, the practical effect is negligible. We conclude that budget is a poor predictor of quality.



4.2 Personal vs. Public Ratings

- Hypothesis (H_1): My personal ratings differ systematically from the public average.
- **Test:** One-Sample T-Test on the RatingGap variable (H_0: = 0).
- **Result:** The test yielded a p-value of **0.0000** with a mean gap of **-0.19**.
- **Conclusion:** We reject the null hypothesis. The results prove a statistically significant "downward bias," indicating I am a tougher critic than the average audience member.

5. Machine Learning Modeling

In the final phase, I developed a predictive model to quantify the importance of different features in determining a movie's rating.

5.1 Model Selection

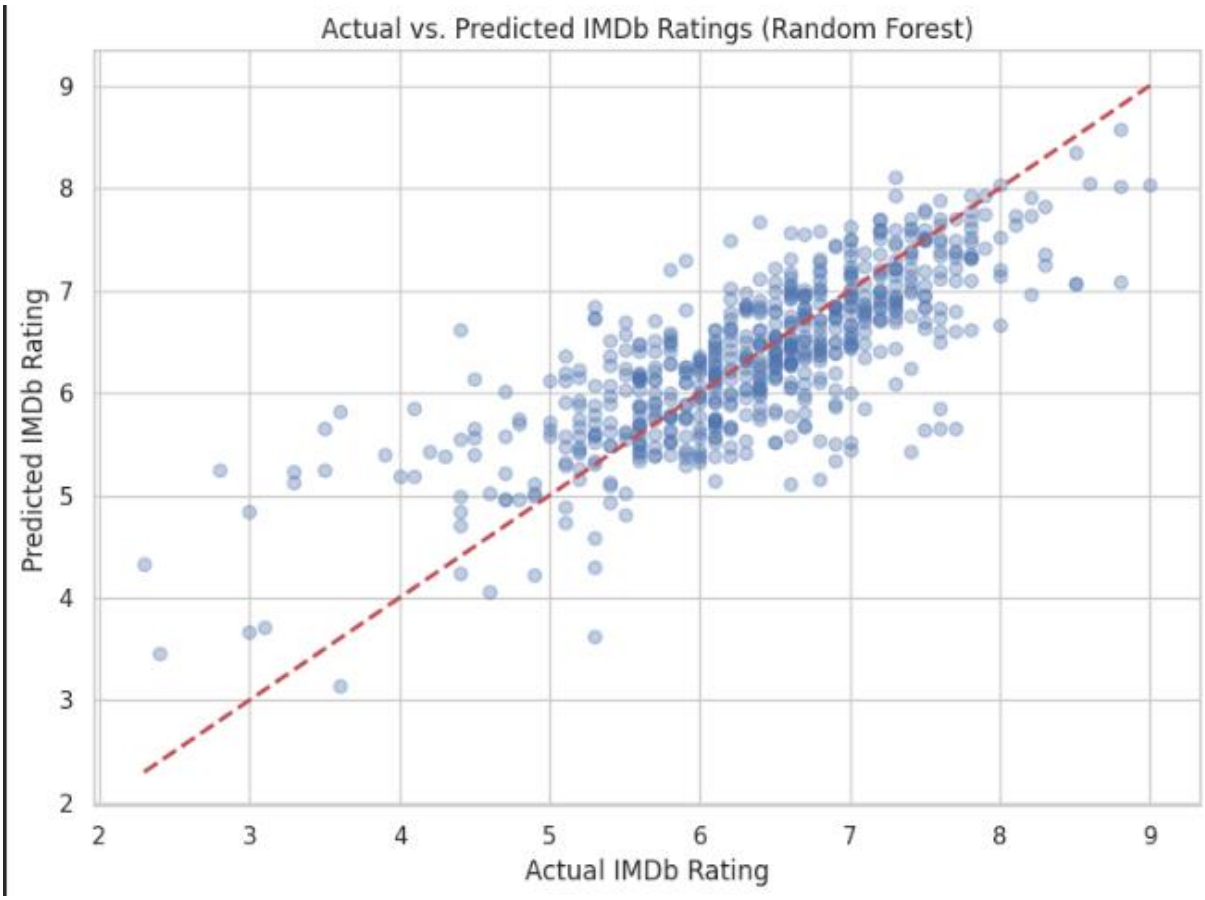
I chose a **Random Forest Regressor** (an ensemble of decision trees) for this task. Unlike linear regression, Random Forest captures non-linear interactions between features, for example, how a high budget might boost a rating *only* if the movie also has a high Metascore.

5.2 Model Performance

The model was trained on an 80/20 split. The performance metrics on the test set were:

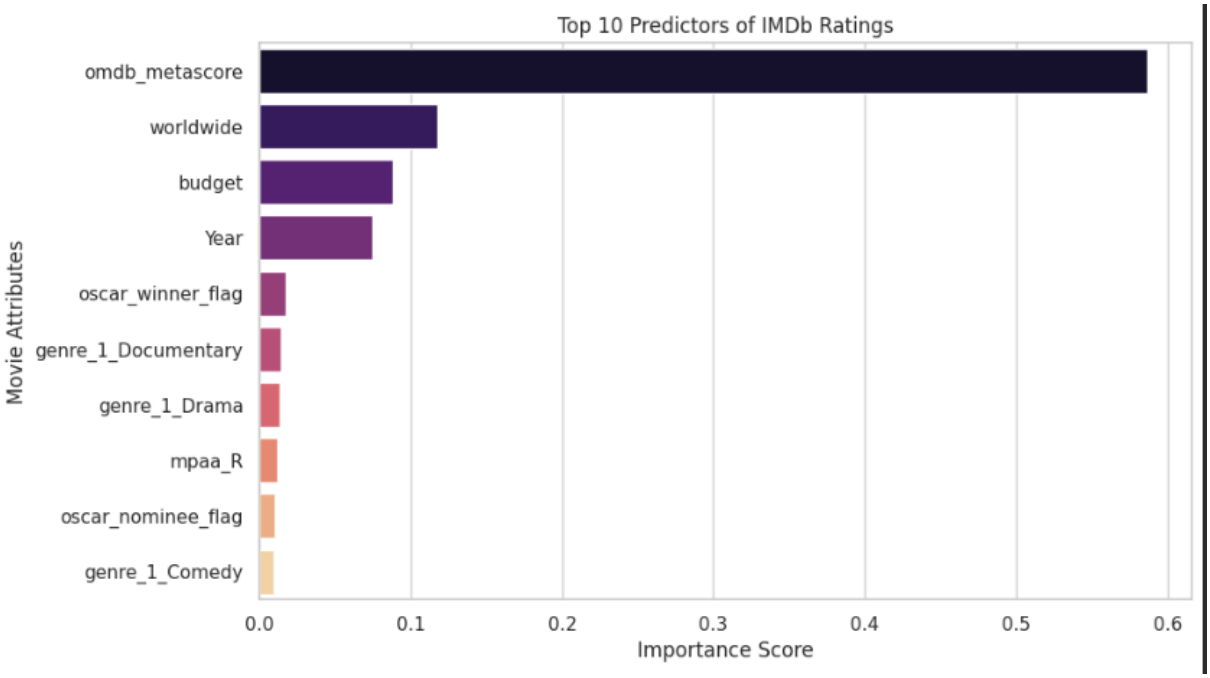
- **Mean Squared Error (MSE):** 0.3937
- **Root Mean Squared Error (RMSE):** 0.6275
- R-Squared (R^2): 0.5932

An R^2 of ~0.60 indicates that our model explains 60% of the variance in movie ratings. The RMSE of 0.63 suggests that, on average, the model's prediction is within roughly half a star of the actual rating.



5.3 Feature Importance

The most critical insight from the ML phase came from analyzing Feature Importance. The model revealed that **Metascore** (critical acclaim) and **Award Counts** were the dominant predictors of a high IMDb rating. **Budget** and **Box Office** revenue were consistently ranked lower in importance. This aligns with the "Art over Commerce" hypothesis.



6. Key Findings

1. Money ≠ Quality: There is almost no linear relationship between production budget and audience rating. A $200M film is just as likely to get a 6.0 rating as a $20M film.
2. **The "Prestige" Factor:** Critical acclaim (Metascore) and Awards are the strongest indicators of audience appreciation. The "Oscar Bump" is real and measurable in the data.
3. **Subjectivity is Measurable:** Through statistical testing, I confirmed that my personal taste is distinct and stricter than the general public's, highlighting the value of personalized recommendation systems over generic averages.

7. Limitations & Future Work

- **Survivor Bias:** The dataset relies on movies listed in Box Office Mojo. This excludes many independent, direct-to-streaming, or foreign films that might have high ratings but no "Budget" data.
- **Imbalanced Classes:** There are fewer "Masterpieces" (Ratings > 9.0) than average movies. The model struggles to predict these extreme outliers accurately.
- **Future Work:**

- **Sentiment Analysis:** Scraping user reviews to perform sentiment analysis could provide a more granular "quality" metric than a simple 1-10 star rating.
- **Mitigating Survivor Bias:** The current financial model relies on Box Office Mojo, which inherently suffers from "survivor bias", it largely tracks films that successfully secured a theatrical release and reported a budget. This excludes many low budget independent films that "failed" or went straight to streaming. Future work should incorporate data from independent film festivals or direct to video databases to capture these missing "failures," providing a more realistic view of how low budgets correlate with movie quality.

8. Conclusion

This project successfully demonstrated that while financial investment ensures high production value, it does not guarantee audience satisfaction. By combining rigorous statistical testing with Machine Learning, we established that "Artistic Merit" (Awards/Critics) is a far superior predictor of success than "Financial Capital." Furthermore, the analysis of personal ratings emphasized the importance of individual taste in a landscape dominated by aggregated scores.

9. Technology Stack

- **Languages:** Python 3.10
- **Libraries:** Pandas, NumPy, Scikit-Learn, SciPy, Matplotlib, Seaborn
- **Tools:** Google Colab, OMDb API, Gemini(usage logs are documented in the ML MODEL.ipynb notebook)

10. Project Timeline

1 - Data Scraping & Cleaning (Box Office Mojo + OMDb)

2 - Exploratory Data Analysis & Visualization

3 - Hypothesis Testing (Spearman & T-tests)

4 - Machine Learning Model Development

5 - Final Reporting & Documentation