Predicting Movie Profitability with Machine Learning

ITCS 3156 Final Project Report

Austin Profenius

Introduction

Movie studios spend millions on production but not every film makes its money back. In this project I use machine learning to predict whether a movie will be profitable or not based only on basic metadata such as budget, ratings, popularity, runtime, release year, and genre.

The goal is to build and compare two models—Logistic Regression and Random Forest—and see how well they can separate profitable movies from non-profitable ones. I also want to see which features matter most and what this says about which types of movies tend to make strong financial returns.

Data

The data comes from the Movies Metadata dataset on Kaggle. After cleaning (removing rows with missing values and budget or revenue ≤ 0), I ended up with 5,357 movies.

I defined profitability as:

- profitable = 1 if revenue ≥ 1.5 × budget
- profitable = 0 otherwise

This gives about 59.5% profitable and 40.5% not profitable films (Figure 1).

Basic visual analysis shows:

- Numeric features such as budget_log, vote_count_log, and runtime are right-skewed (Figure 2).
- Drama, Comedy, and Action are the most common genres, but Horror, Animation, and Science Fiction have the highest profitability rates (Figure 3 & 7).
- The correlation heatmap (Figure 4) shows vote_count_log and vote_average are most positively related to profitability, while release_year has a small negative correlation.
- Boxplots by class (Figure 5) indicate profitable movies tend to have higher popularity, more votes, and slightly higher ratings.
- The budget vs. revenue plot (Figure 6) clearly separates profitable vs. not-profitable films around the 1.5× budget line.

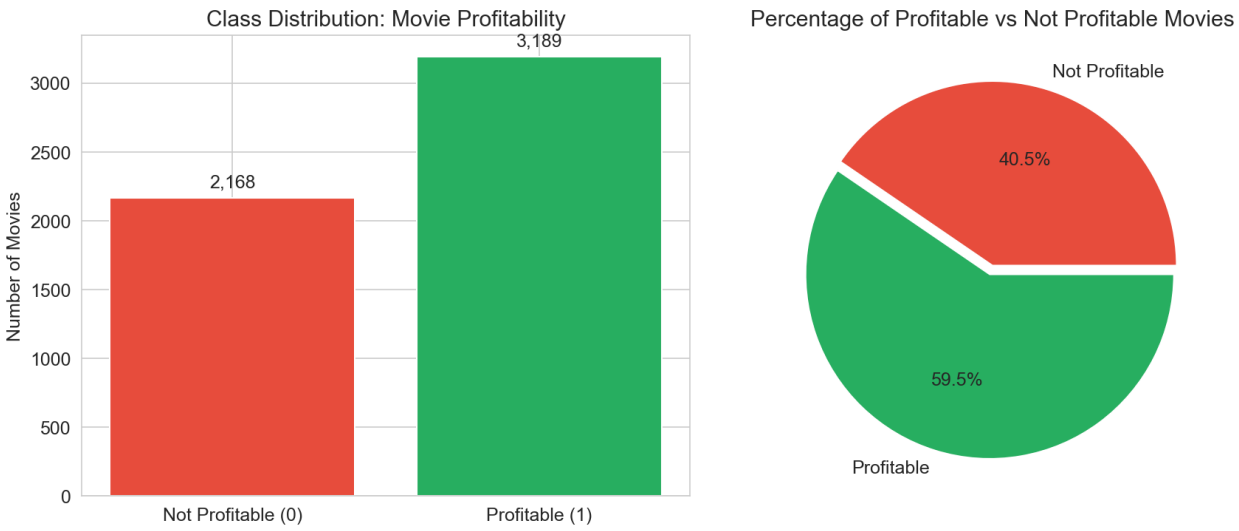

Figure 1

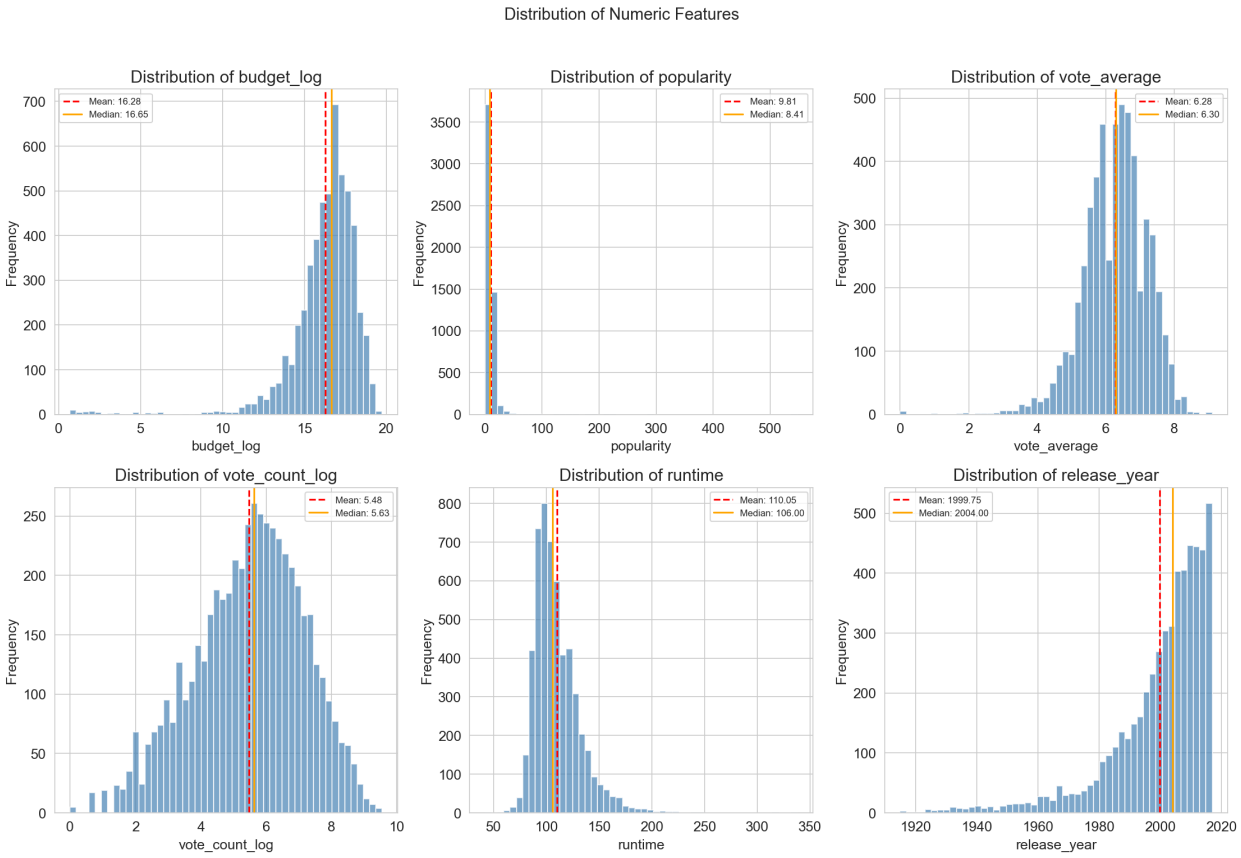

Figure 2

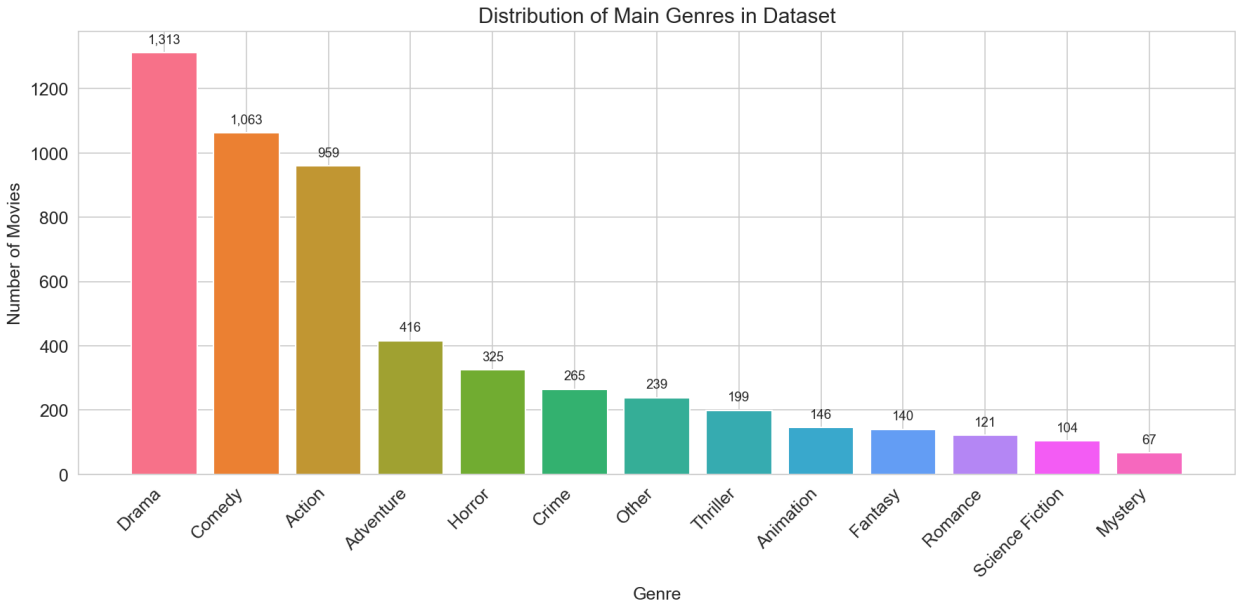

Figure 3

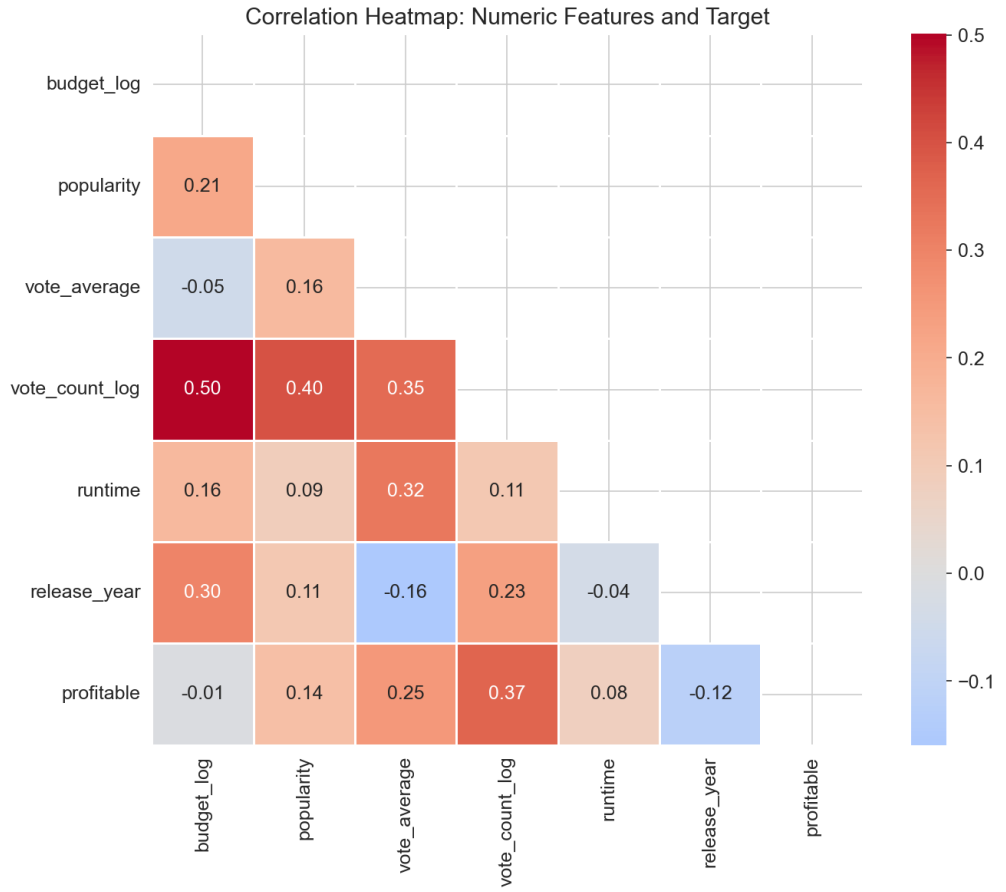

Figure 4

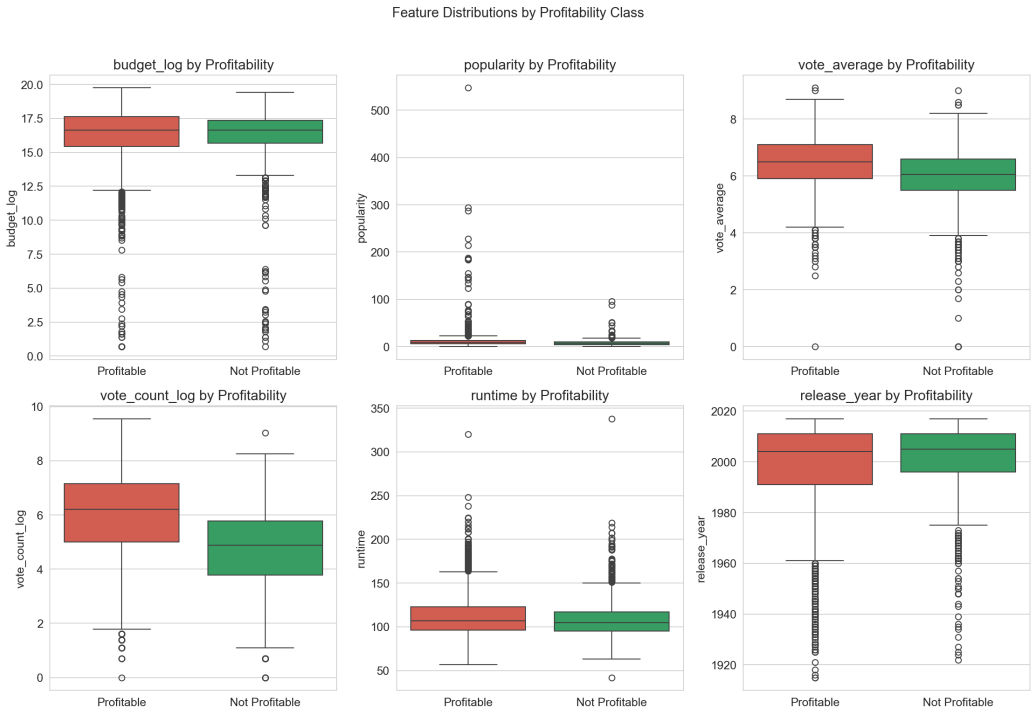

Figure 5

Figure 6

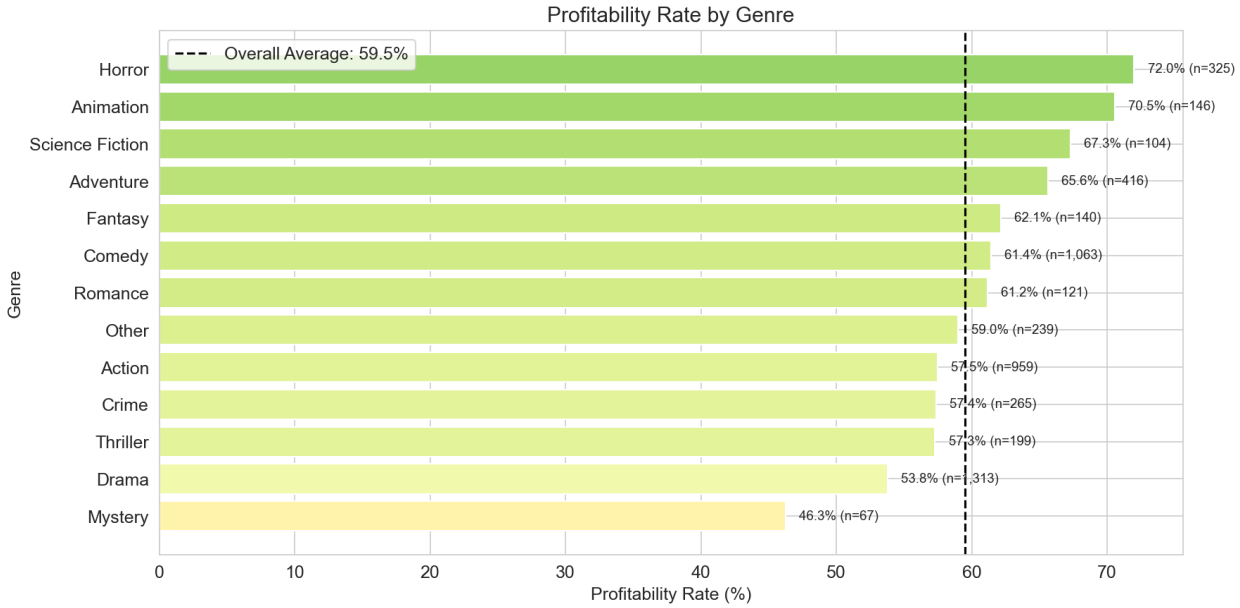

Figure 7

Preprocessing

Main preprocessing steps:

- Dropped rows with missing values in budget, revenue, popularity, vote_average, vote_count, runtime, release_date, or genres.
- Removed movies with non-positive budget or revenue.
- Engineered:
    - budget_log = log1p(budget)
    - vote_count_log = log1p(vote_count)
    - release_year from release_date
    - Binary profitable label as described above
- Parsed the JSON-like genres field and kept the first genre as main_genre.
- One-hot encoded the most frequent genres and grouped the rest into “Other”.
- Standardized numeric features for Logistic Regression using StandardScaler.
- Split data into 60% train / 20% validation / 20% test, stratified by the profitability label.

Methods

I trained and compared two supervised learning models:

1. **Logistic Regression**
    - Linear model that outputs the probability a movie is profitable.
    - Uses L2 regularization and class_weight="balanced" to handle the mild class imbalance.
    - Hyperparameter C was tuned on the validation set.
    - Easy to interpret through feature coefficients.
2. **Random Forest Classifier**
    - Ensemble of decision trees built on bootstrap samples with random feature subsets.
    - Captures non-linear relationships and feature interactions.
    - Tuned n_estimators, max_depth, min_samples_split, and min_samples_leaf.
    - Provides feature importance scores that show which variables drive predictions.

For both models I used accuracy, precision, recall, F1 score, and ROC AUC as evaluation metrics, plus confusion matrices and ROC curves on the test set.

Results

Model Setup and Metrics

After tuning on the validation set, I retrained each model on the combined train+validation data and evaluated on the **test set**. Key results (Figure 12):

- **Logistic Regression**
  - Accuracy: 0.733
  - Precision: 0.792
  - Recall: 0.748
  - F1 Score: 0.769
  - ROC AUC: 0.792
- **Random Forest**
  - Accuracy: 0.746
  - Precision: 0.772
  - Recall: 0.813
  - F1 Score: 0.792
  - ROC AUC: 0.812

The Random Forest slightly outperforms Logistic Regression on accuracy, recall, F1 score, and AUC. The ROC curves (Figure 9) show both models are clearly better than random, with the Random Forest curve sitting consistently above the Logistic Regression curve.

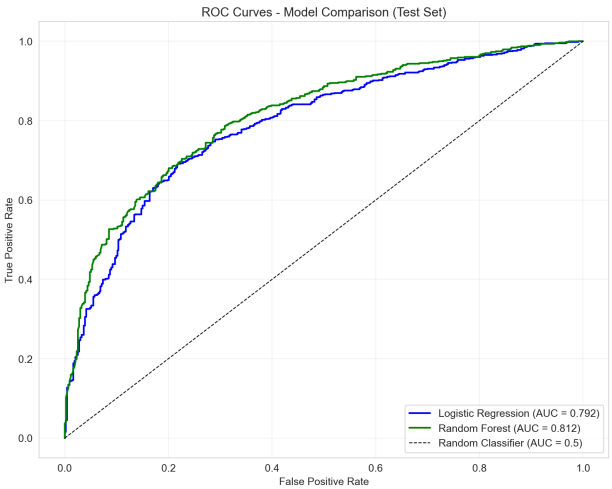

Figure 9

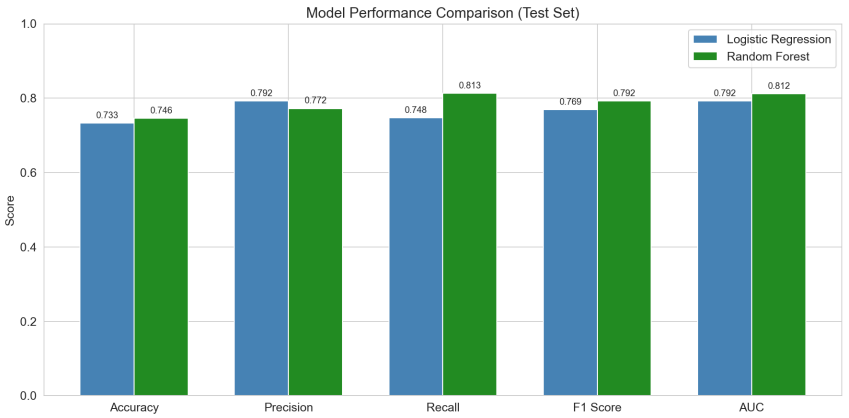

Figure 12

Confusion Matrices and Trade-offs

Confusion matrices (Figure 8) show:

- Logistic Regression:
    - 477 true positives, 309 true negatives
    - 161 false negatives, 125 false positives
- Random Forest:
    - 519 true positives, 281 true negatives
    - 119 false negatives, 153 false positives

Random Forest **catches more profitable movies** (higher recall) but also predicts profitability for more movies that are not actually profitable. In a real studio setting, this might be acceptable if the cost of missing a potential hit is higher than over-estimating a few weaker projects.

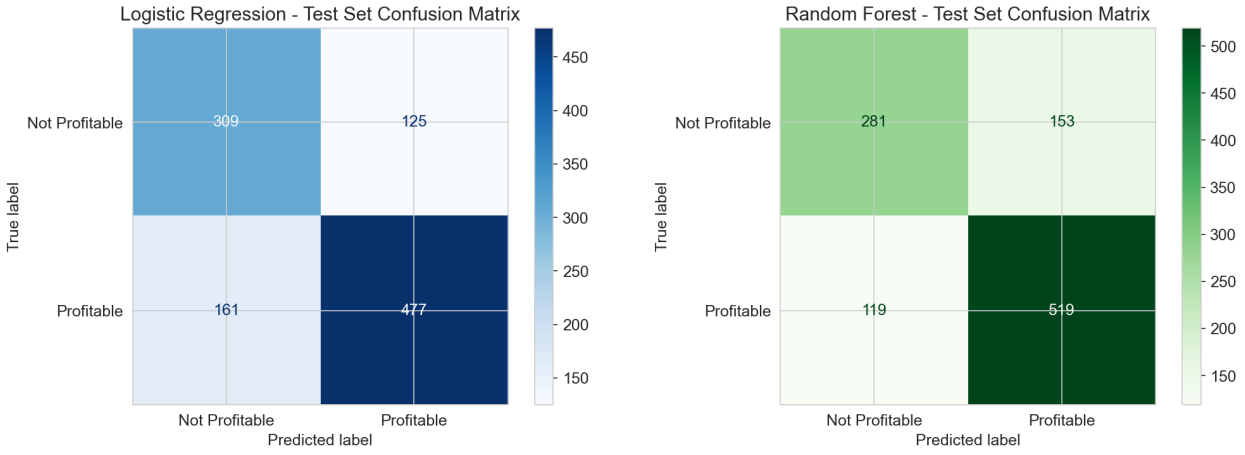

Figure 8

Feature Importance and Coefficients

Random Forest feature importance (Figure 10) shows:

- Top numeric features: vote_count_log, popularity, budget_log, release_year, vote_average, and runtime.
- Genre features have smaller but non-zero importance; Action, Comedy, Drama, and Horror are the most influential genres.

Logistic Regression coefficients (Figure 11) show a similar story:

- Positive: vote_count_log, Animation, Horror, Romance, and Comedy increase the probability of profitability.
- Negative: budget_log, Mystery, later release_year, and some genres like Drama and Science Fiction decrease the probability.

Overall, both models agree that audience reach (votes and popularity), decent ratings, and certain genres are key drivers of movie profitability.

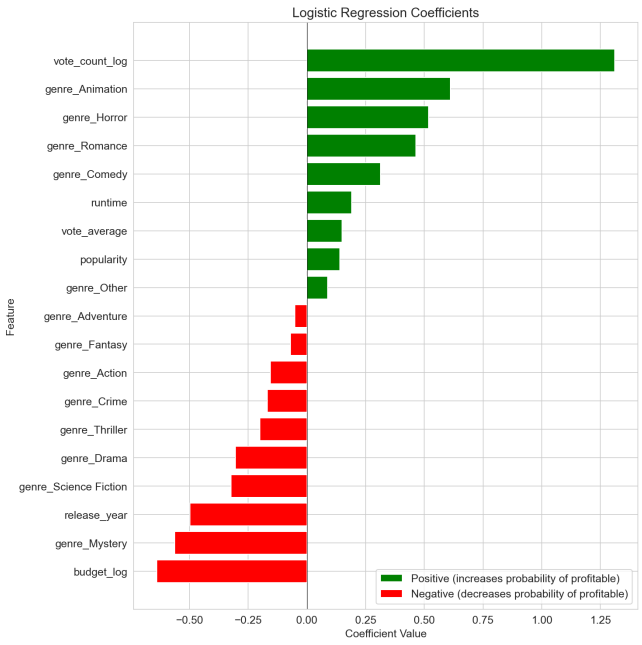

Figure 10

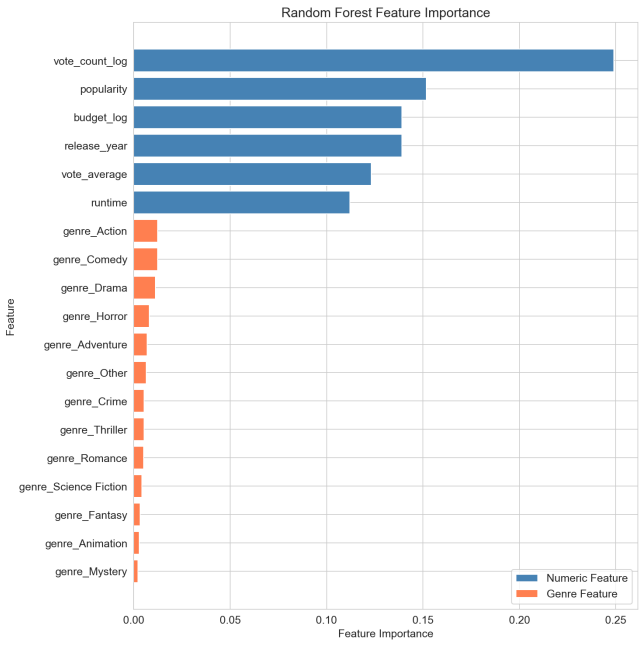

Figure 11

Conclusions

In this project I built a full machine learning pipeline to predict movie profitability using basic metadata. Both Logistic Regression and Random Forest perform reasonably well, with test AUC values around 0.8. The Random Forest model performs best overall, especially in recall and F1 score, and is better at identifying profitable movies.

From the analysis I learned that:

- Engagement features like vote counts and popularity are more predictive of profitability than budget alone.
- Some genres (Horror, Animation, Science Fiction) have higher profitability rates than others.
- A simple linear model captures a lot of the signal, but a non-linear ensemble can squeeze out extra performance.

At the same time, the project has limitations. The profitability label ignores marketing costs and long-term revenue, and the dataset may be biased toward films with complete metadata. Future work could include more detailed features (cast, director, studio) and additional models such as gradient boosting.

References

Banik, Rounak. *The Movies Dataset*. Kaggle, 2017.

Acknowledgement

In this project I used online resources and AI tools (such as ChatGPT) to help brainstorm ideas, debug code, and improve wording in my report. All final decisions about data cleaning, model design, and interpretation of the results were made by me, and the code in the GitHub repository was written and run by me.

Source Code

GitHub link: https://github.com/AustinProfenius/3156-Final-Project