Unlocking Box Office Success: Predicting Movie Popularity

UCB Data Analytics Bootcamp - Project 4

Group 6:
Gursimran Kaur (Simran)
Jeff Kim
Rose Mary Rios

Project Overview

ETL Highlights

EDA Highlights

Machine Learning Data Preparations

ML - Baseline Models → Performant Model

Machine Optimizations

Project Overview

Objective: This project seeks to **predict movie popularity scores** by leveraging advanced machine learning models by analyzing a rich dataset of historical movie data.

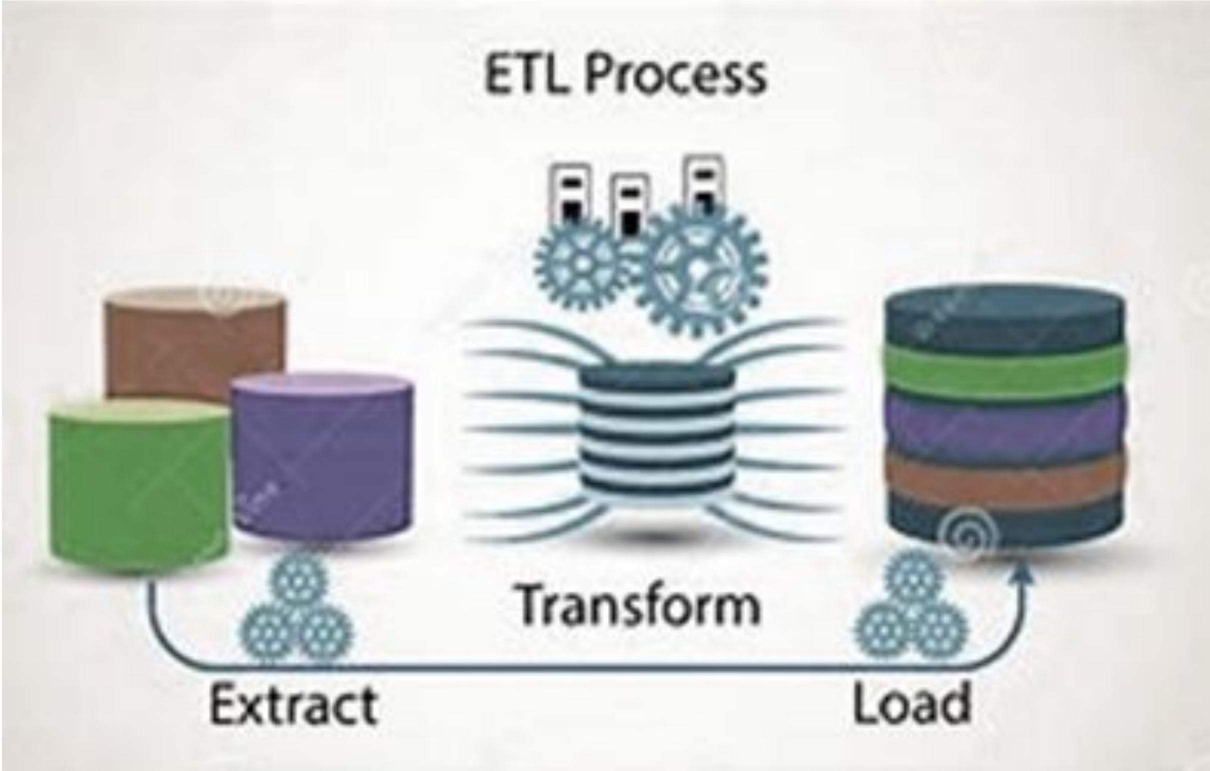
Data Source: Kaggle's TMDB 5000 Movie , Credits and Oscar Best Picture Movies Datasets

Assumptions:
- Winning or being nominated for Oscars boosts popularity
- Big-name directors and famous actors make movies more popular
- Higher budgets lead to more popular movies
- Popular genres draw more audiences

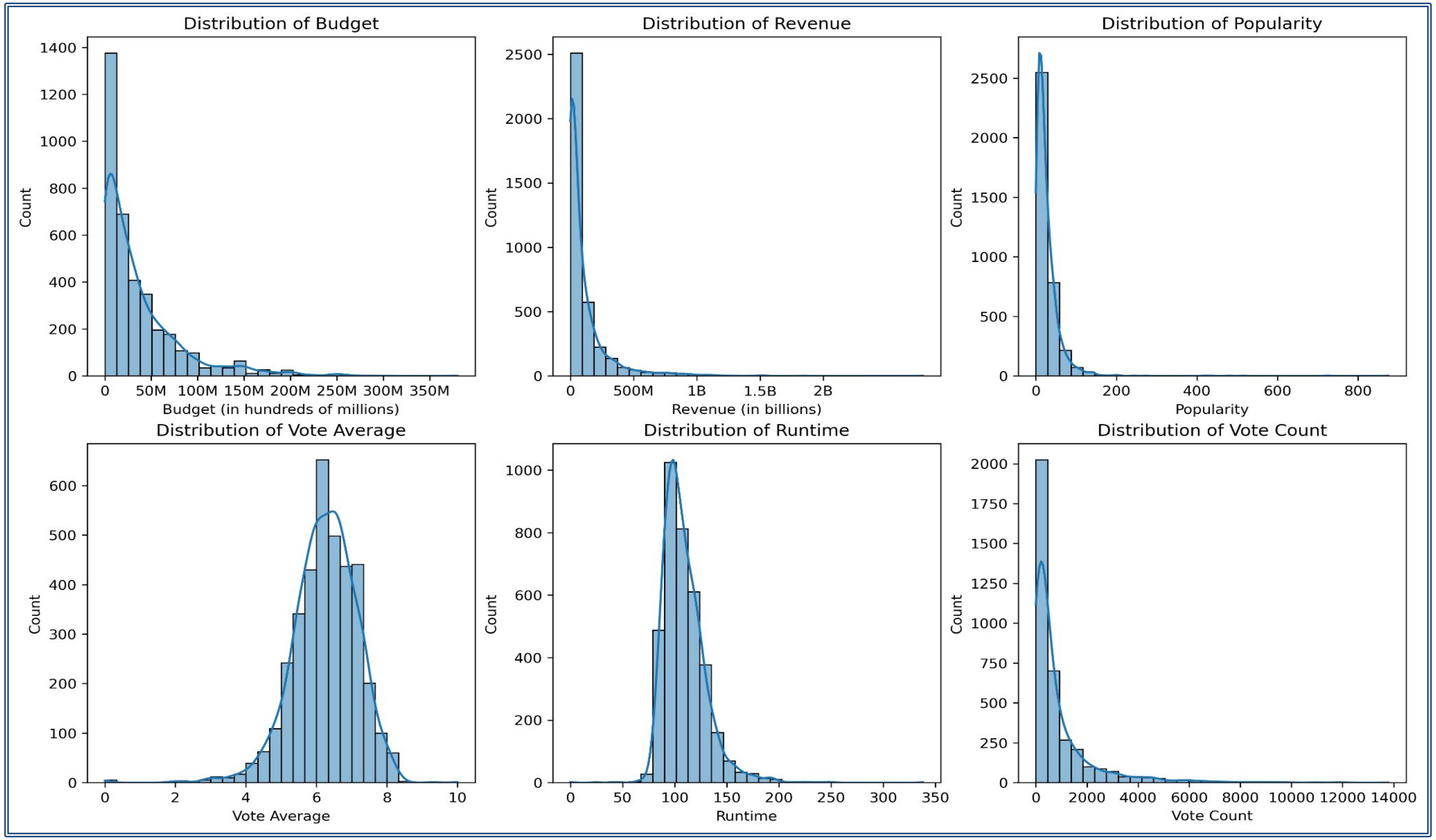
ETL Highlights

Key Activities:

- ✓ **Data Loading:** Successfully imported and loaded the necessary datasets into a Pandas dataframe.
- ✓ **Feature Selection & Parsing:** Identified and extracted relevant features and parsed complex data into structured, Pandas dataframes.
- ✓ **Data Cleaning:** Thoroughly addressed data quality issues, including handling missing values, outliers, inconsistencies, and duplicates.
- ✓ **Data Merging:** Combined multiple datasets into a unified dataset for analysis, ensuring data integrity and alignment.
- ✓ **Feature Engineering:** Created new features or transformed existing ones to improve model performance and capture relevant relationships within the data.
- ✓ **Data Formatting:** Standardized data formats and ensured consistency across features, making the data suitable for machine learning algorithms.
- ✓ **Exporting Cleaned Data:** Successfully exported the cleaned and prepared data into a format suitable for further analysis or modeling.



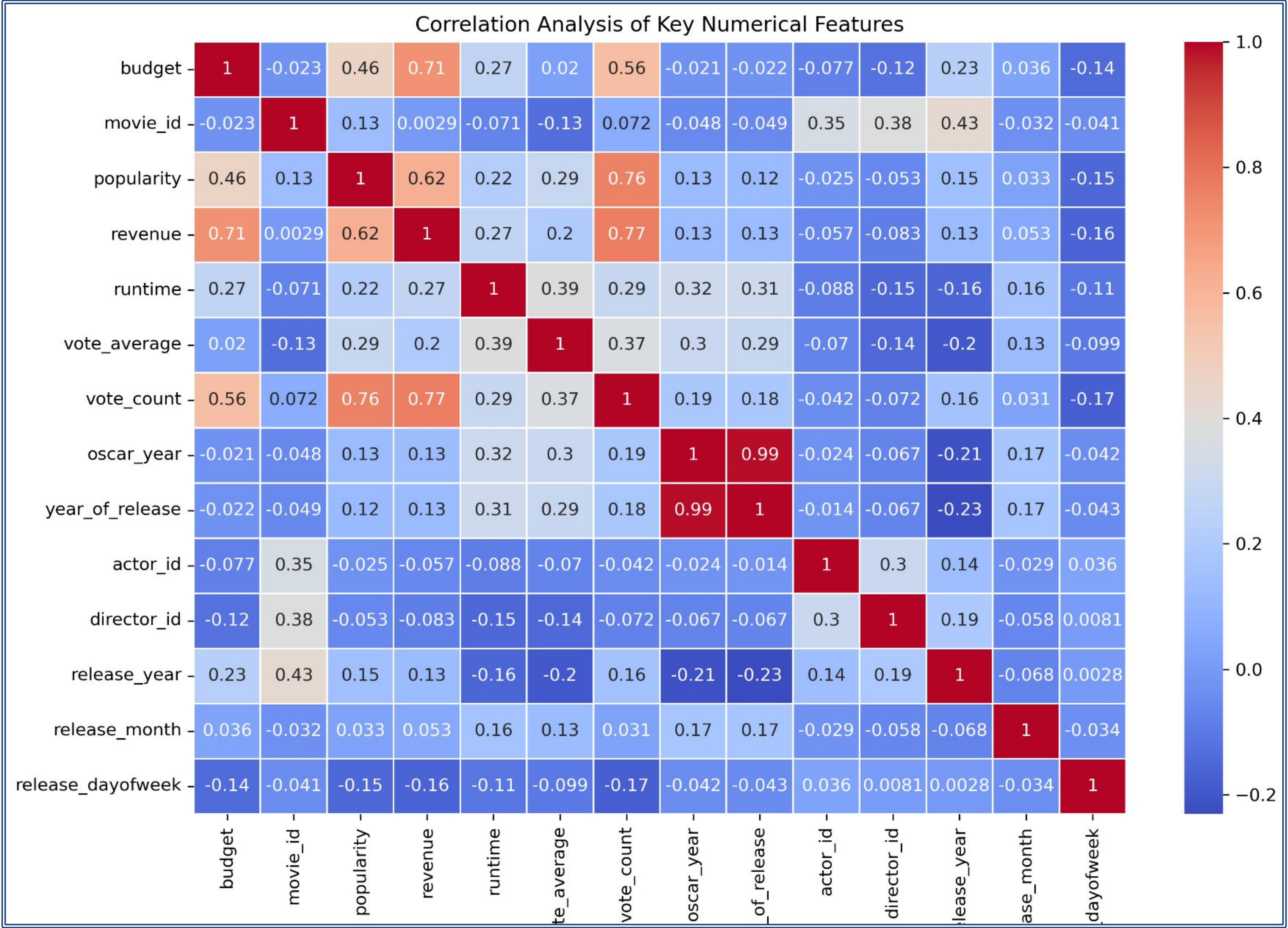
EDA Highlights



EDA Highlights

Key Insights:

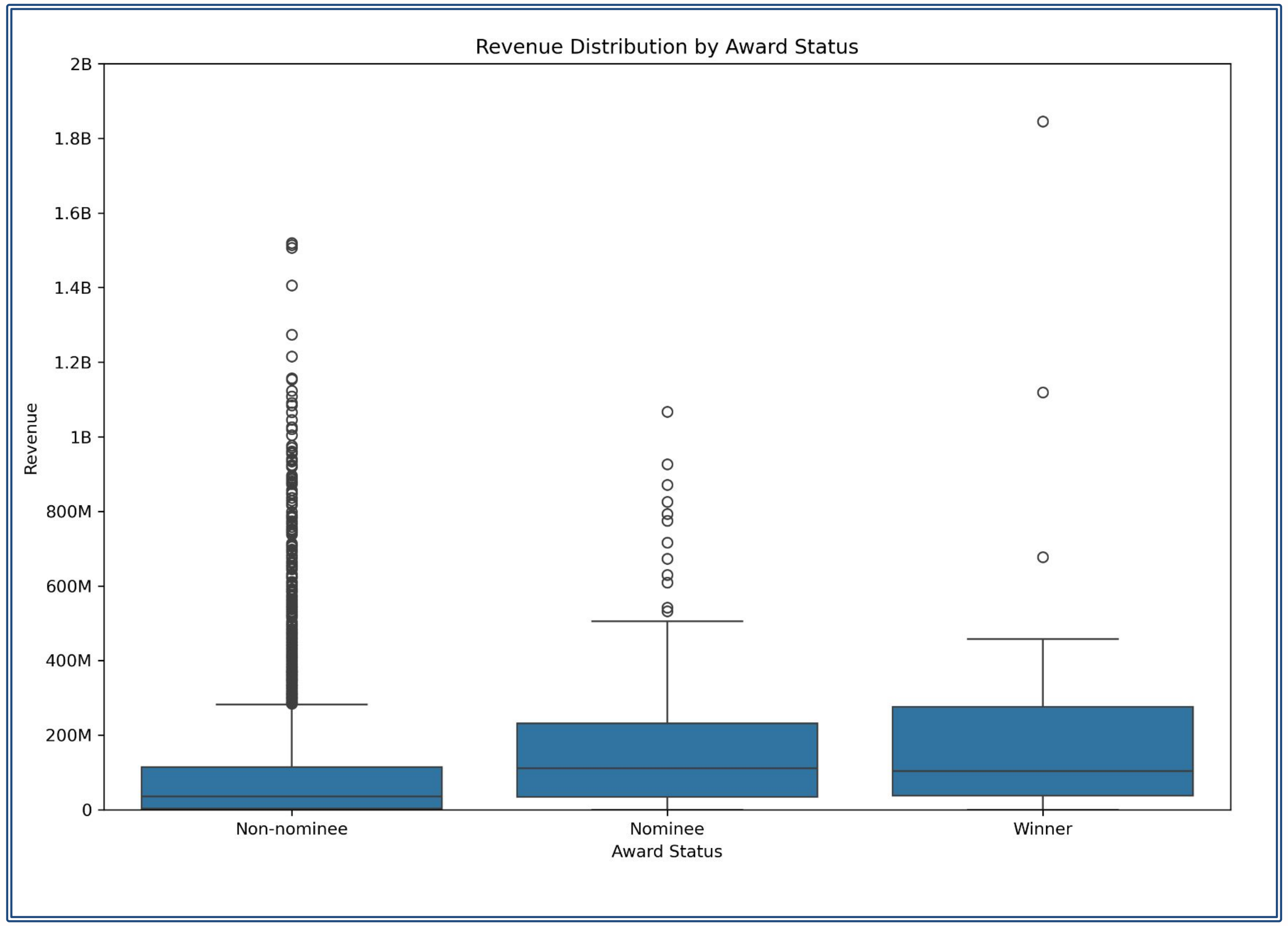
- **Strong Positive Correlations:** Budget, popularity, vote count, and revenue are positively correlated.
- **Moderate Positive Correlations:** Runtime and revenue have a moderate positive correlation.
- **Negative Correlations:** Oscar wins don't necessarily correlate strongly with budget or release year.

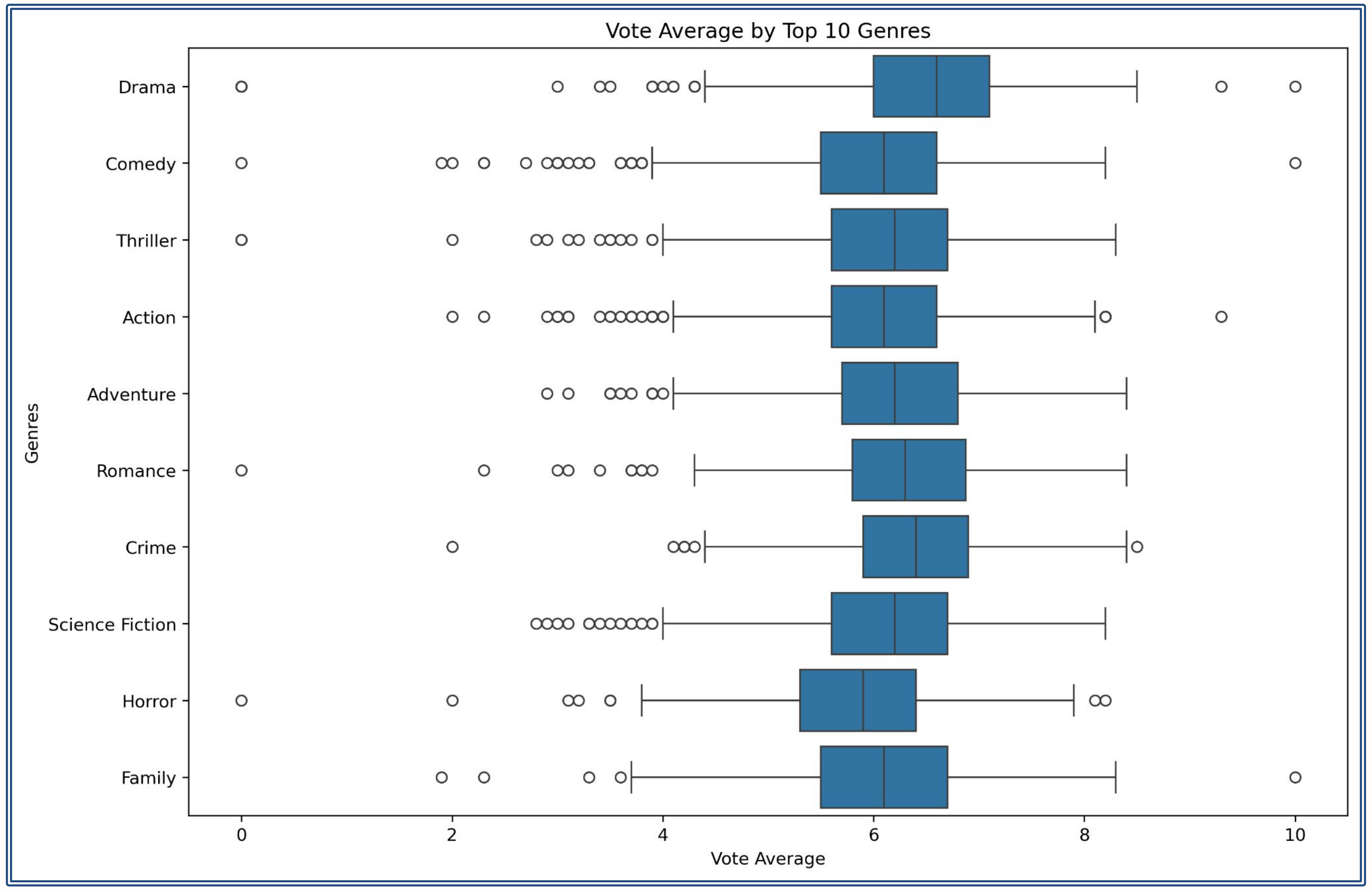


EDA Highlights

- Award Impacts
- Non-Award Factors:

- Genre Differences
- Genre Outliers

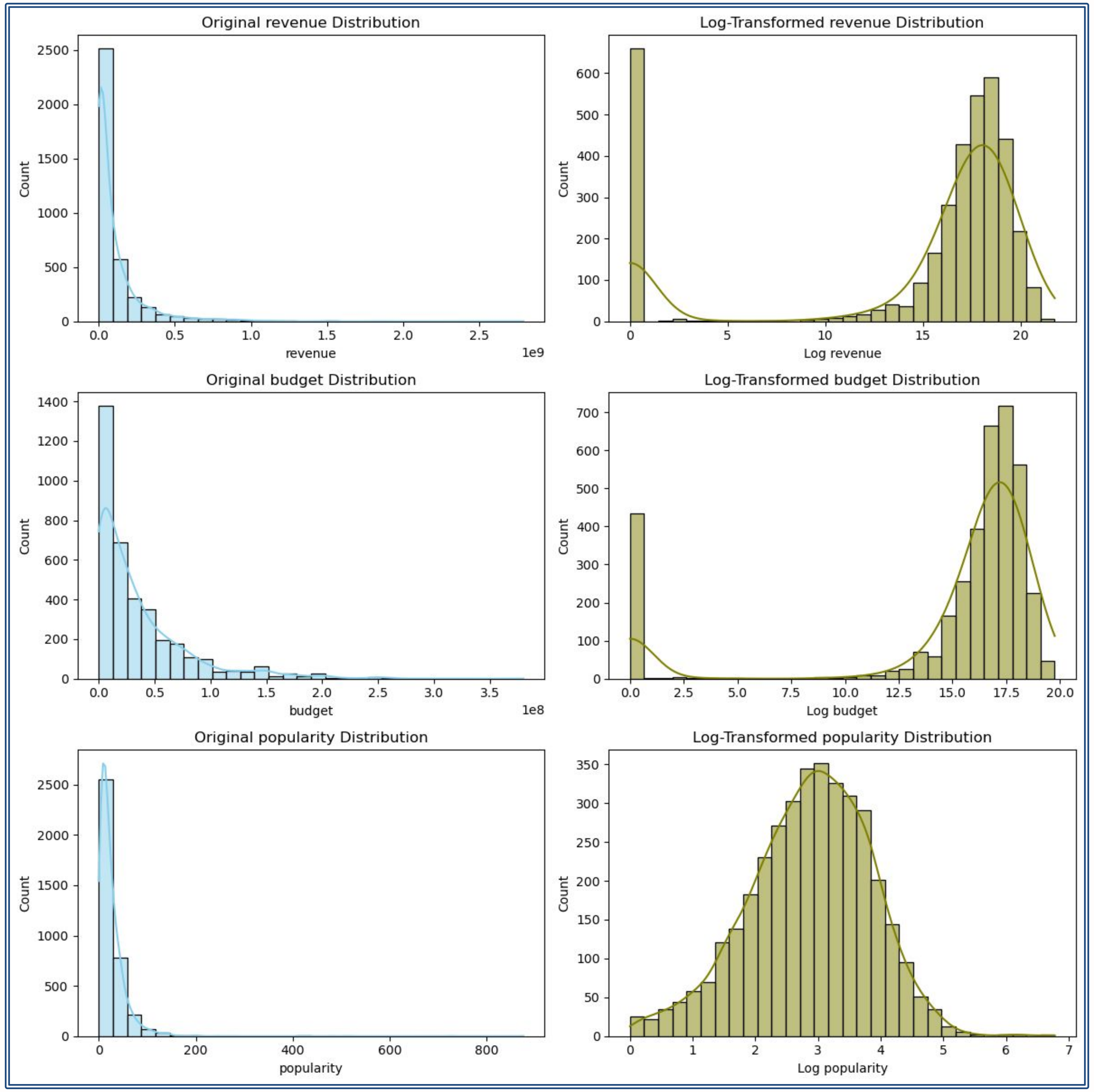




ML Data Preparation

Data Preparation for Modeling

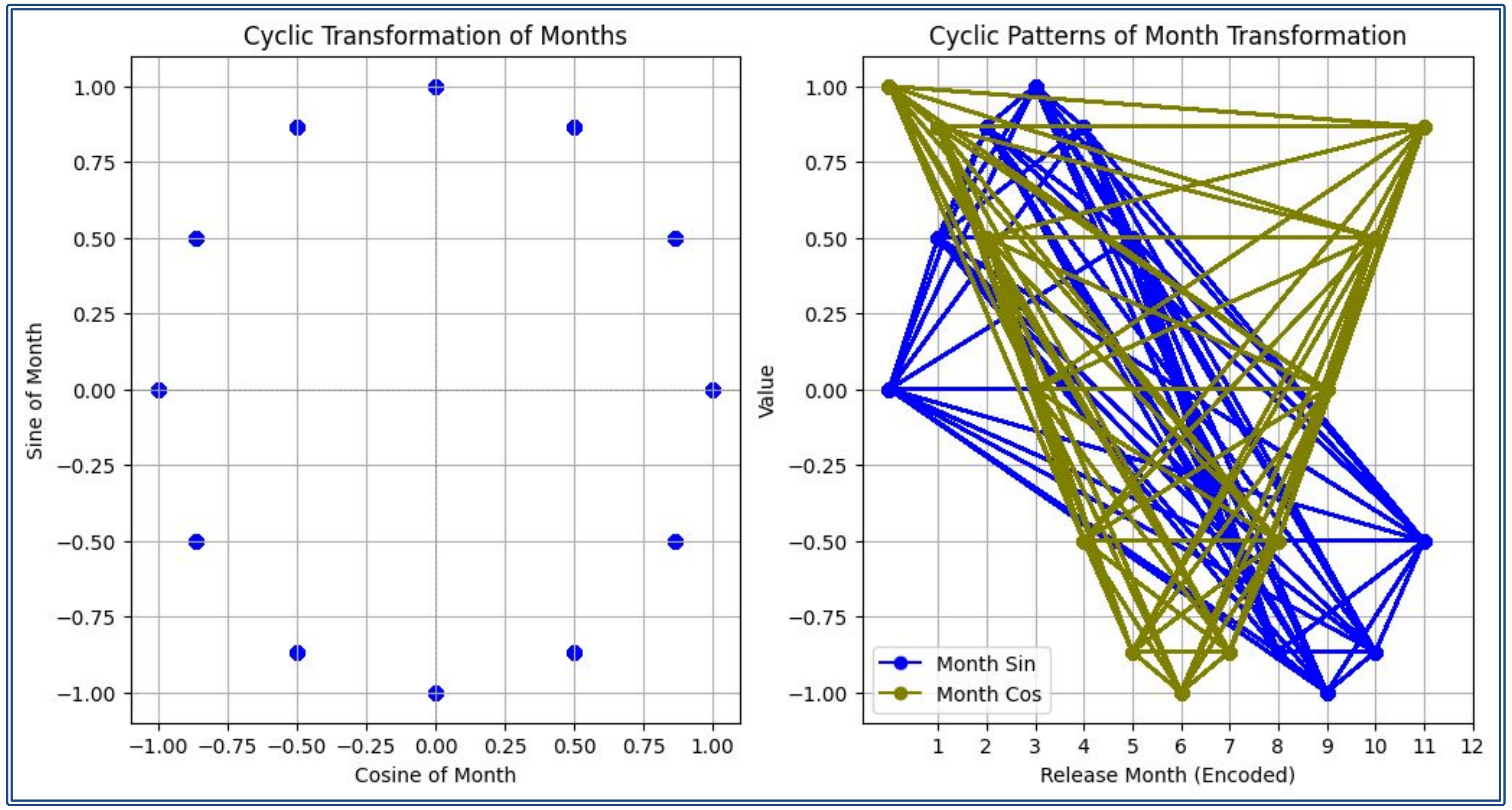
- **Feature Engineering:**
  - **Log Transformations** to handle skewed data and outliers
  - **Interactive Feature:** to capture combined relationship (budget x runtime)
  - **Label Encoding** on categorical features
  - **Cyclic Transformation** of month and day to represent their circular nature

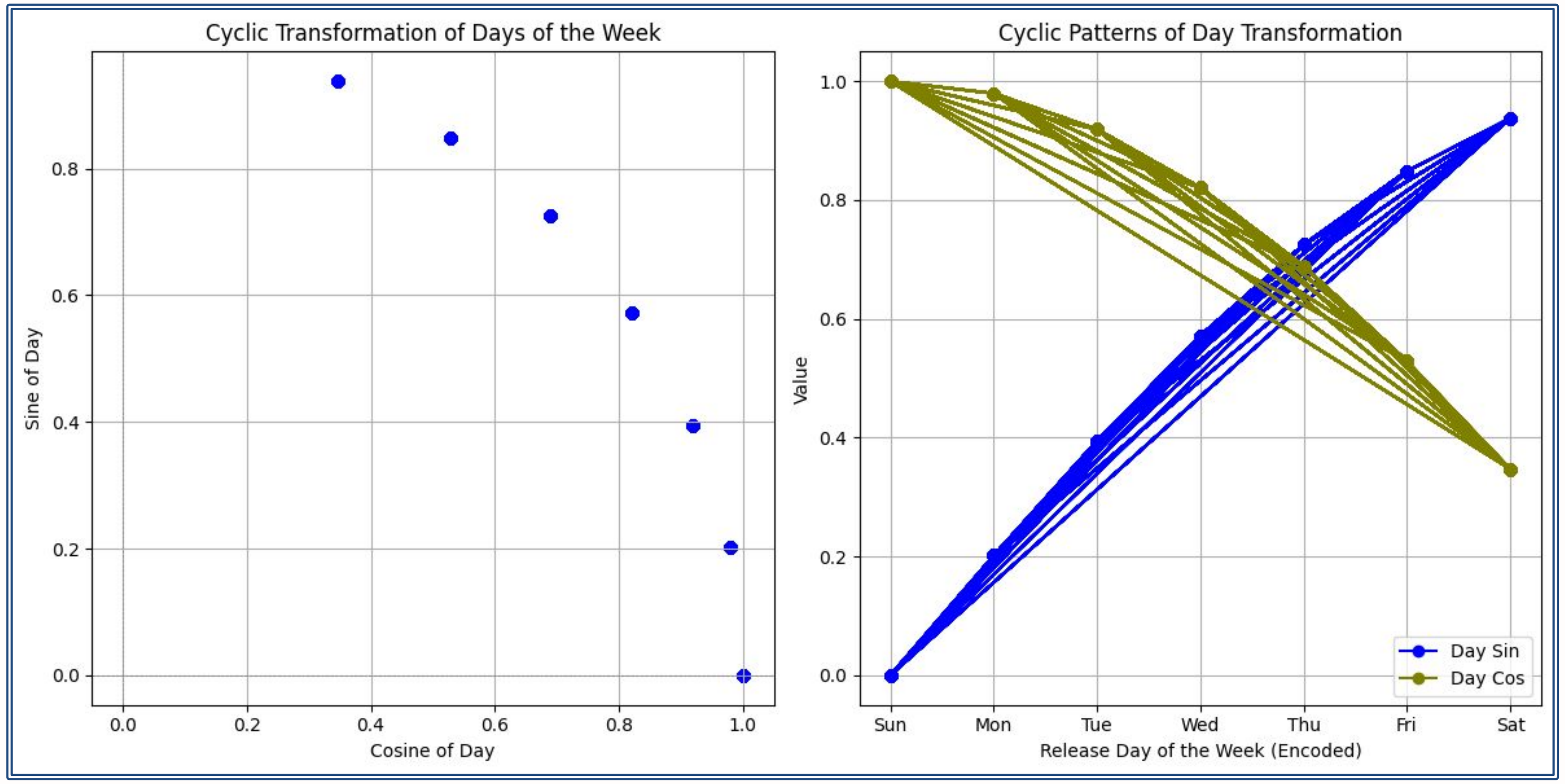


ML Data Preparation

Cyclic Encoding was used to:

- Help the model identify repeating patterns
- Improve prediction accuracy



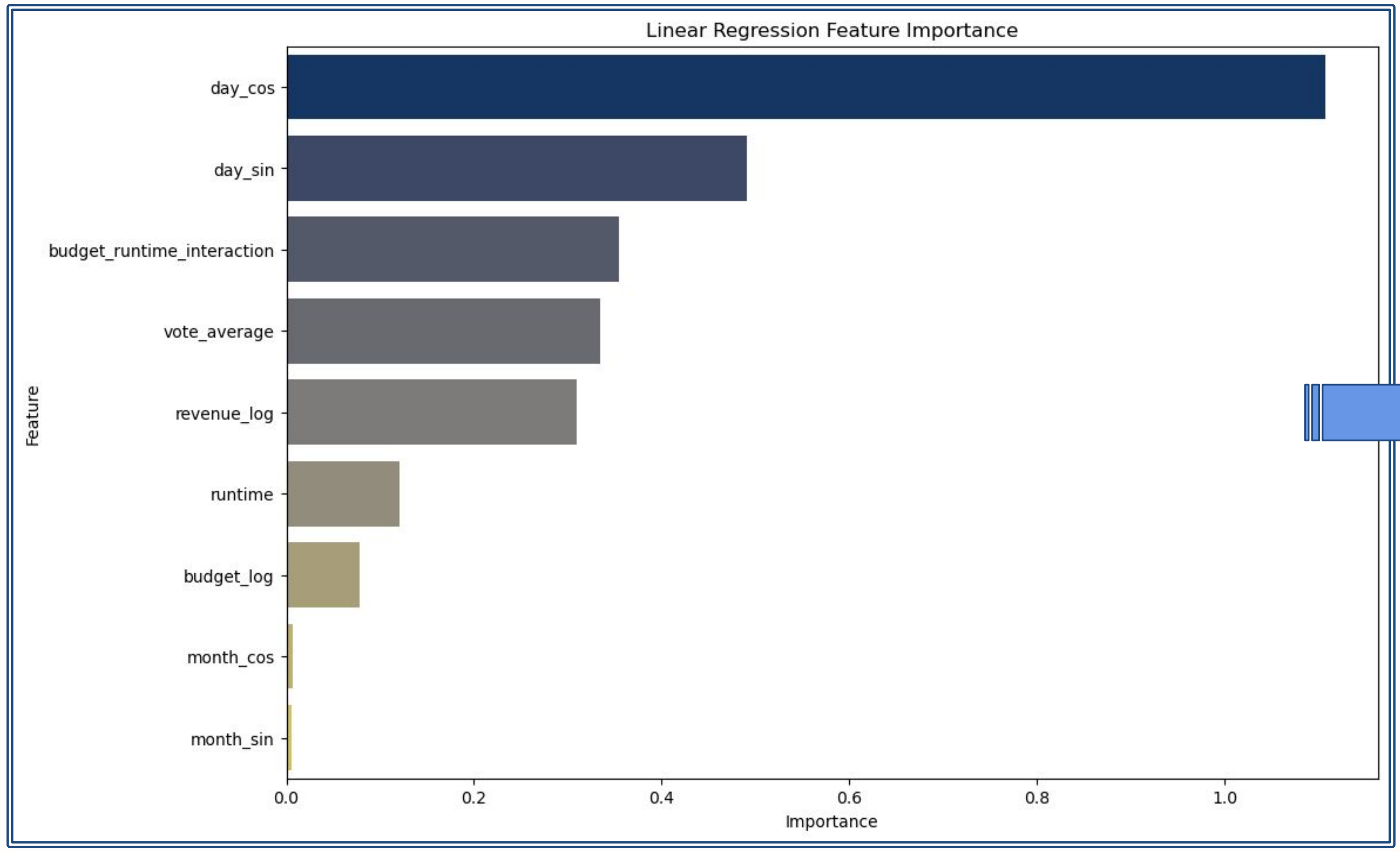


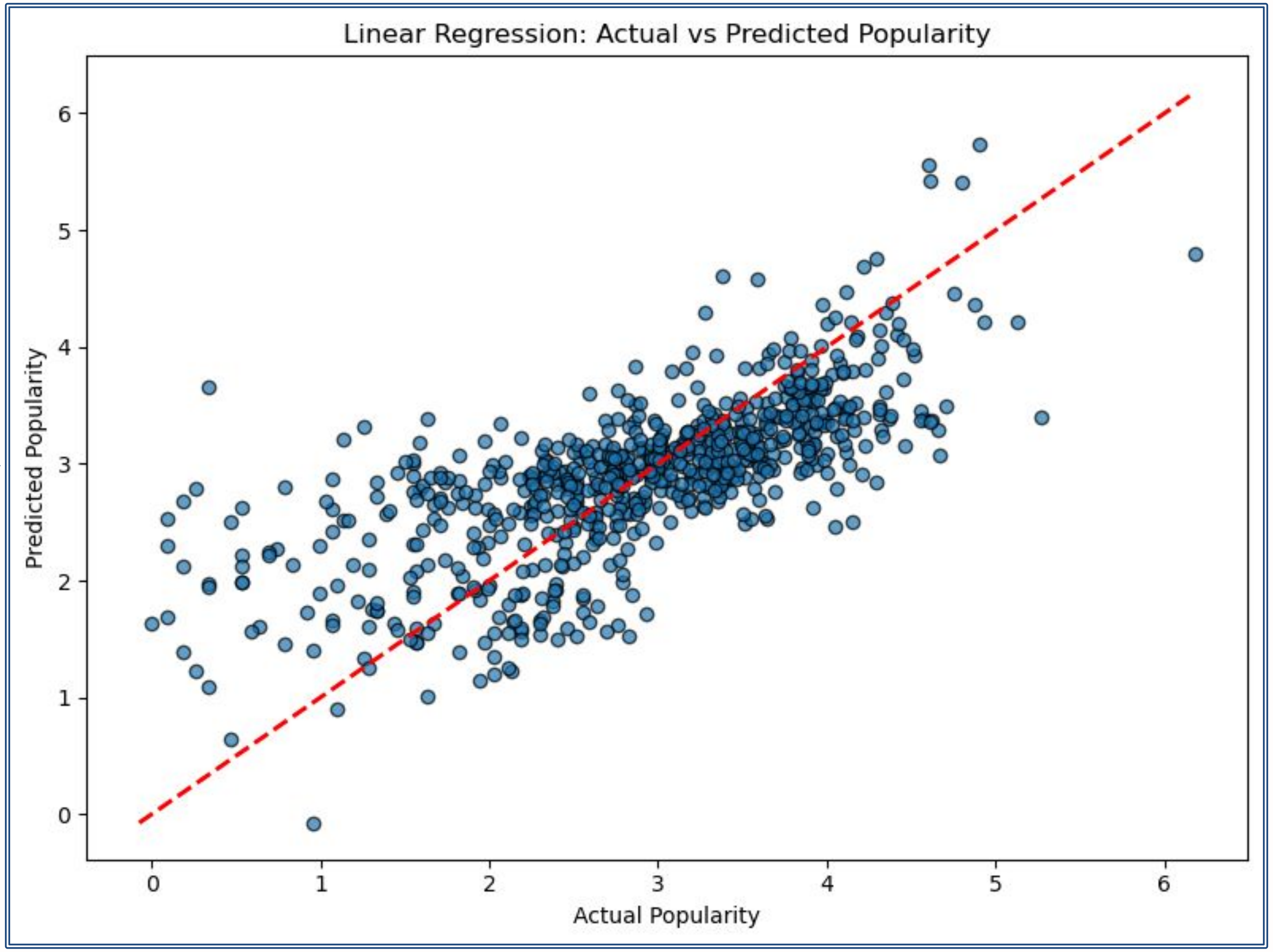
Model 1: Linear Regression

Updated Features: ['budget_log', 'runtime', 'vote_average', 'month_sin', 'month_cos', 'day_sin', 'day_cos', 'revenue_log', 'budget_runtime_interaction']

Root Mean Squared Error (RMSE): 0.6884275084348853

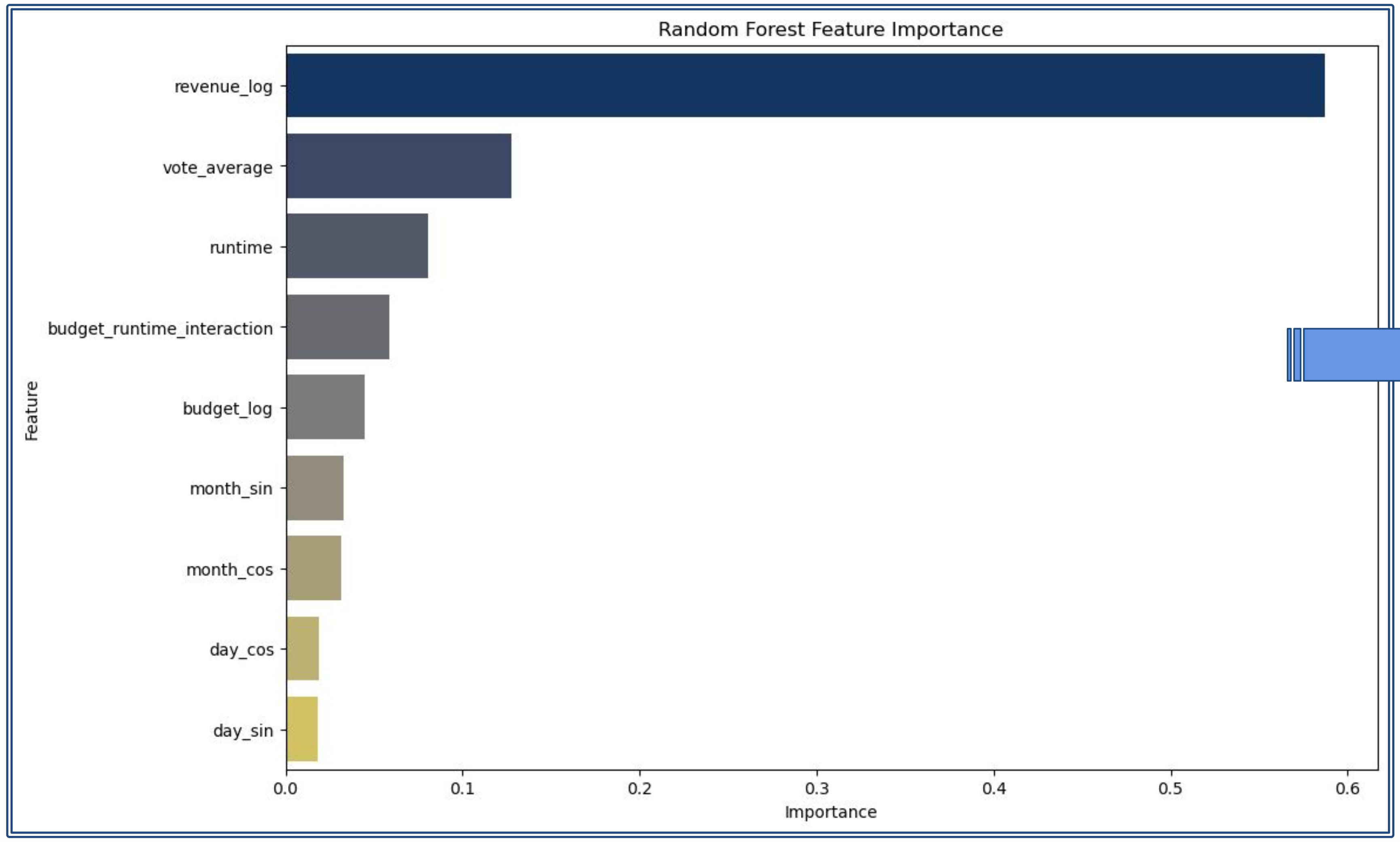
R-squared (R^2): 0.5066330266530592

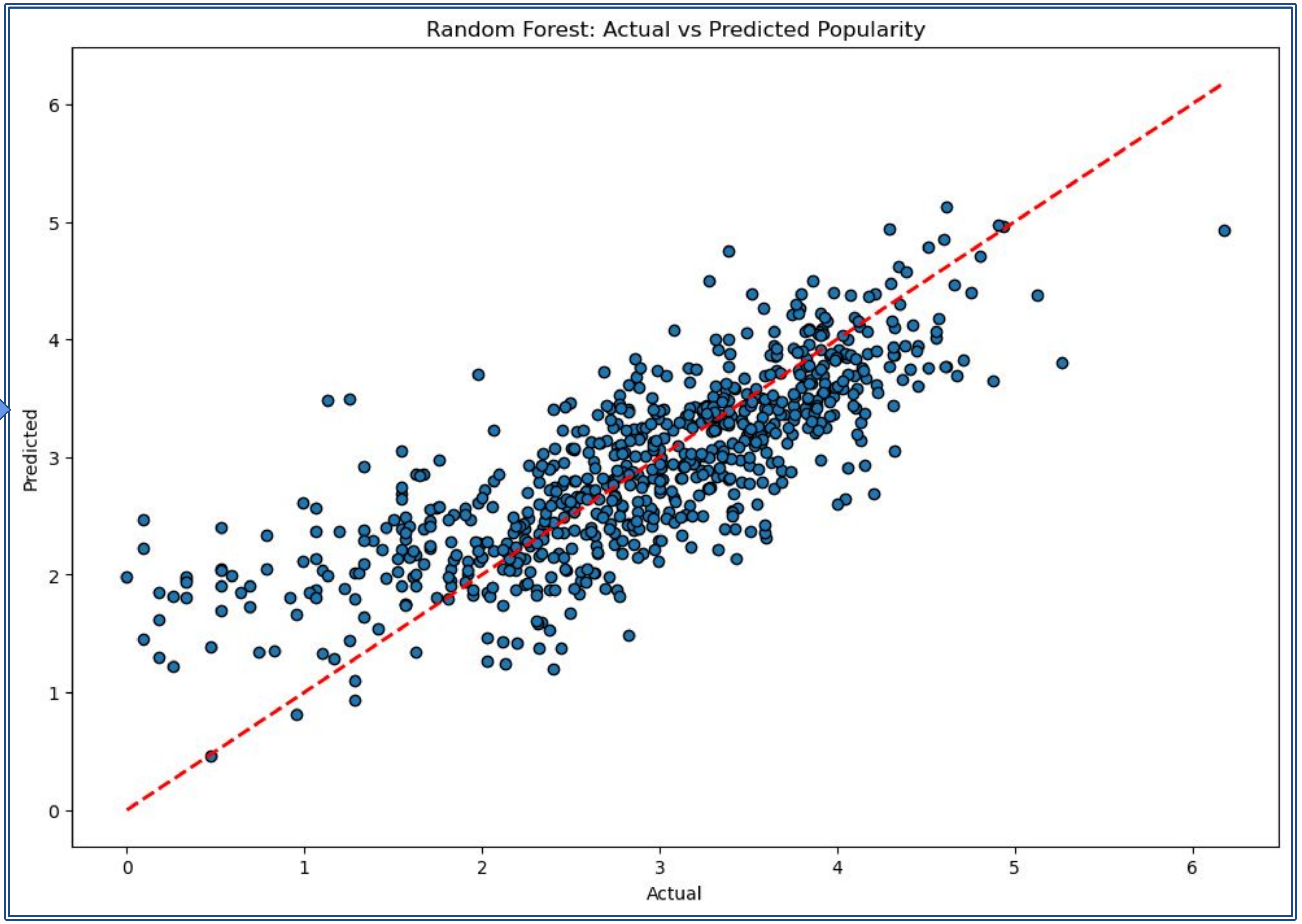




Model 2: Random Forest

Root Mean Squared Error (RMSE): 0.6055005301511441 R-Squared (R^2): 0.6183346814644578





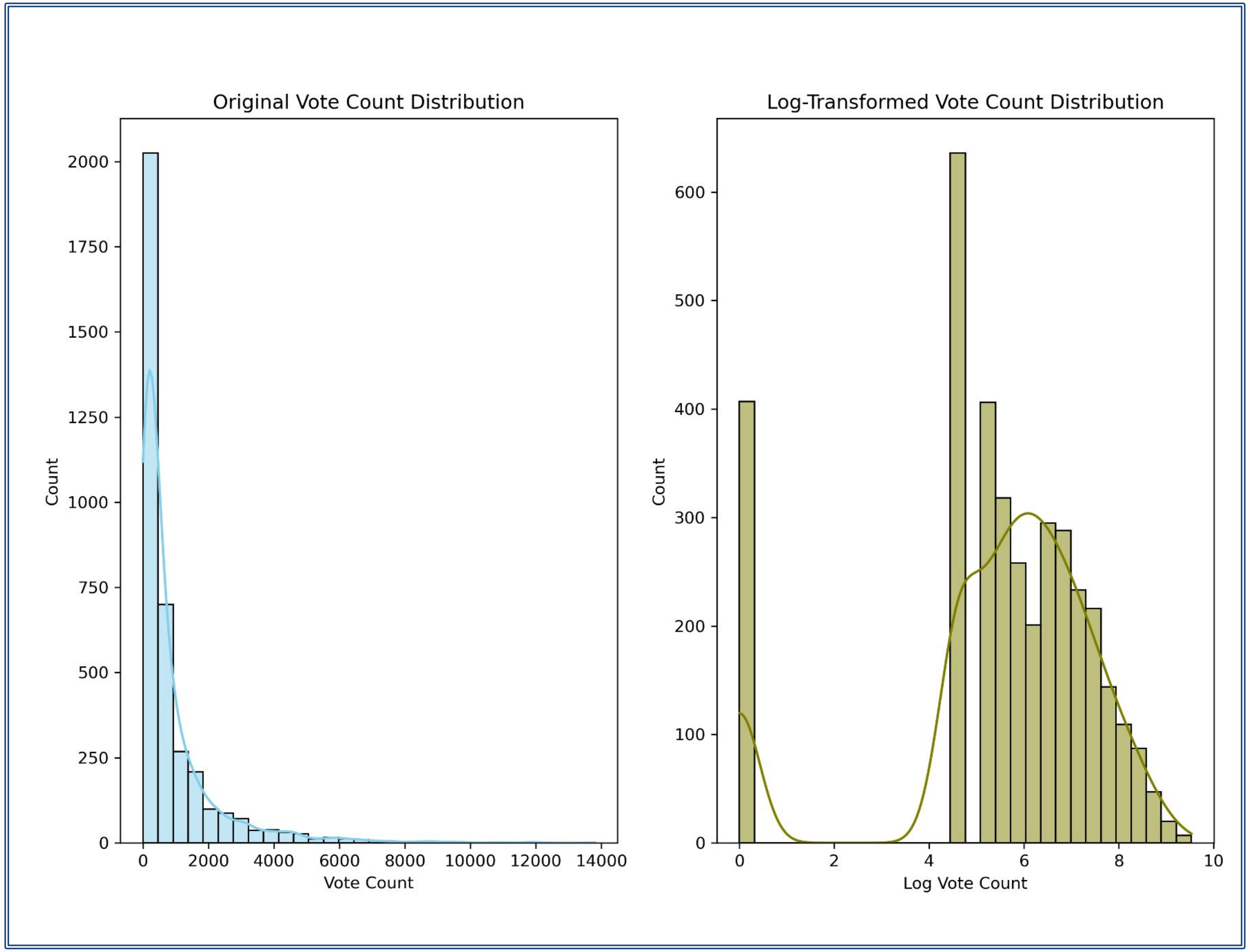
RANDOM FOREST

- Complex Relationships handling
- Feature Importance recognition
- Robustness handling of outlier's noisy data
- Data Versatility with numerical & categorical
- Predictive Power for unseen data

= Ideal for predicting future movie trends!

Optimization One: Feature Engineering

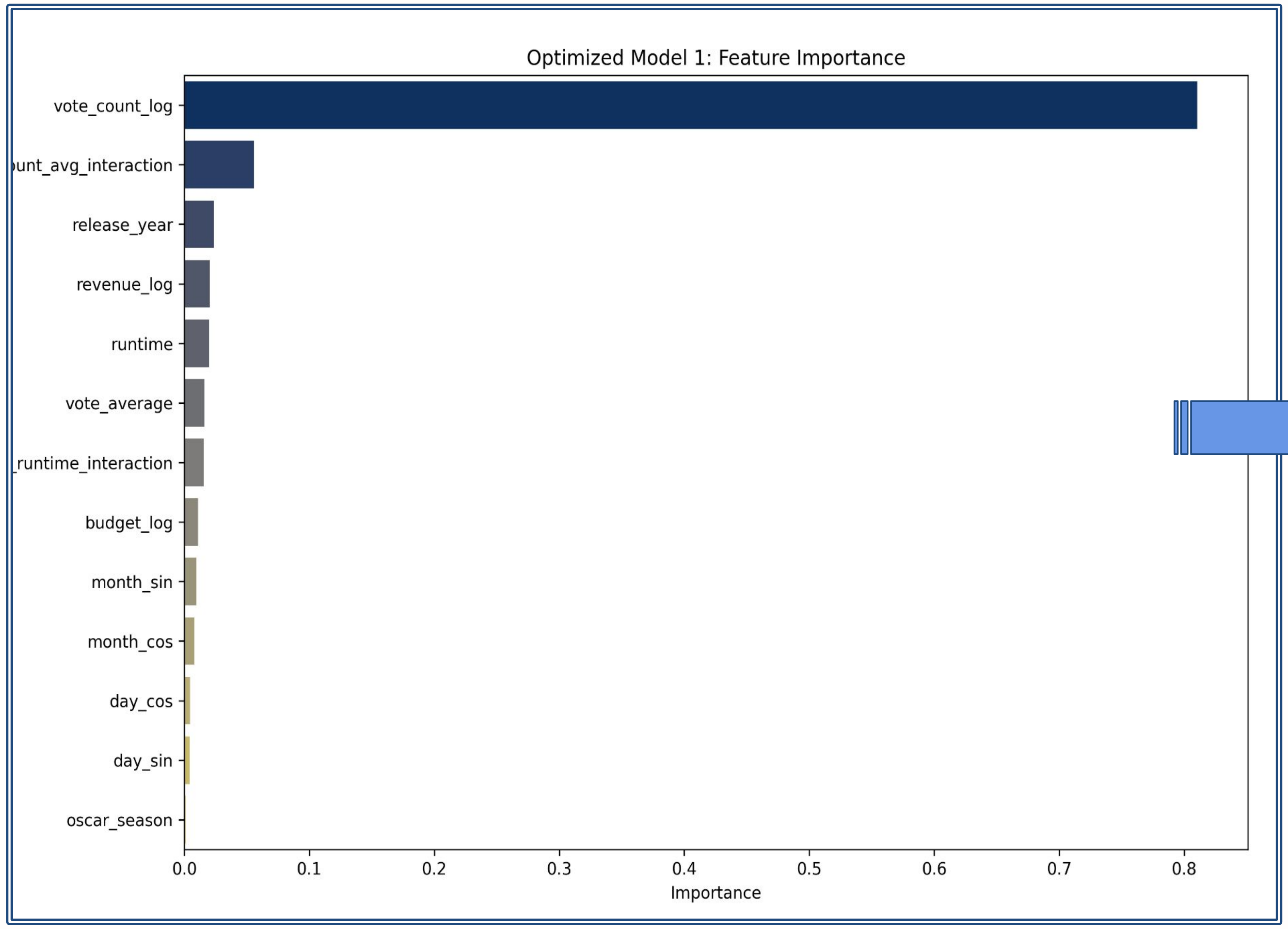
- **Log Transformation:** on ‘vote_count’ feature to normalize its distribution
- **Interaction Features** combining ‘vote_count_log’ and ‘vote_average’ to capture the relationship
- **Seasonal Feature** that flags movies released during the Oscar season, in the fall and winter months

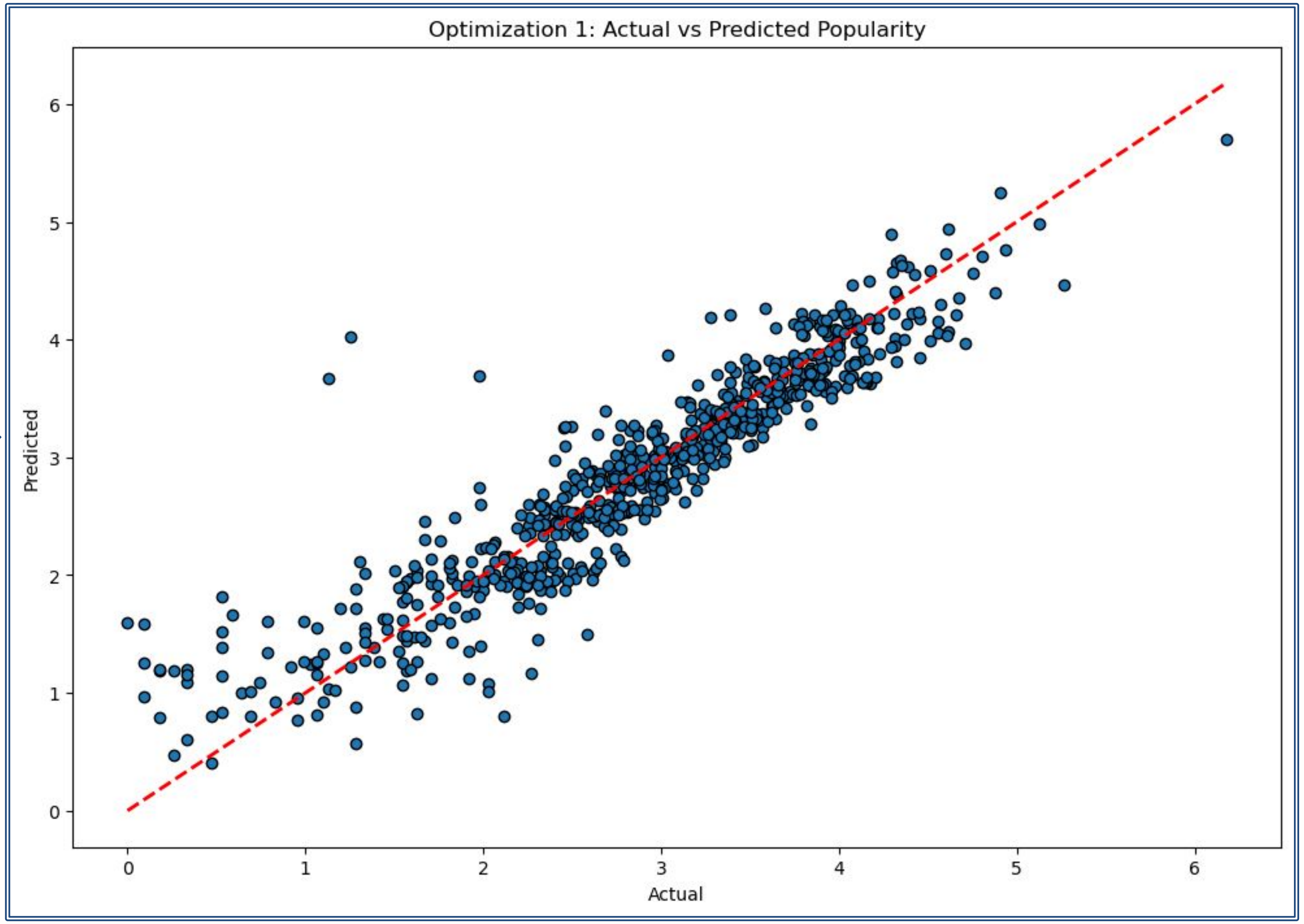


Optimization One Output

Root Mean Squared Error (RMSE): 0.3629928041279302

New R-Squared: 0.8628329893191401

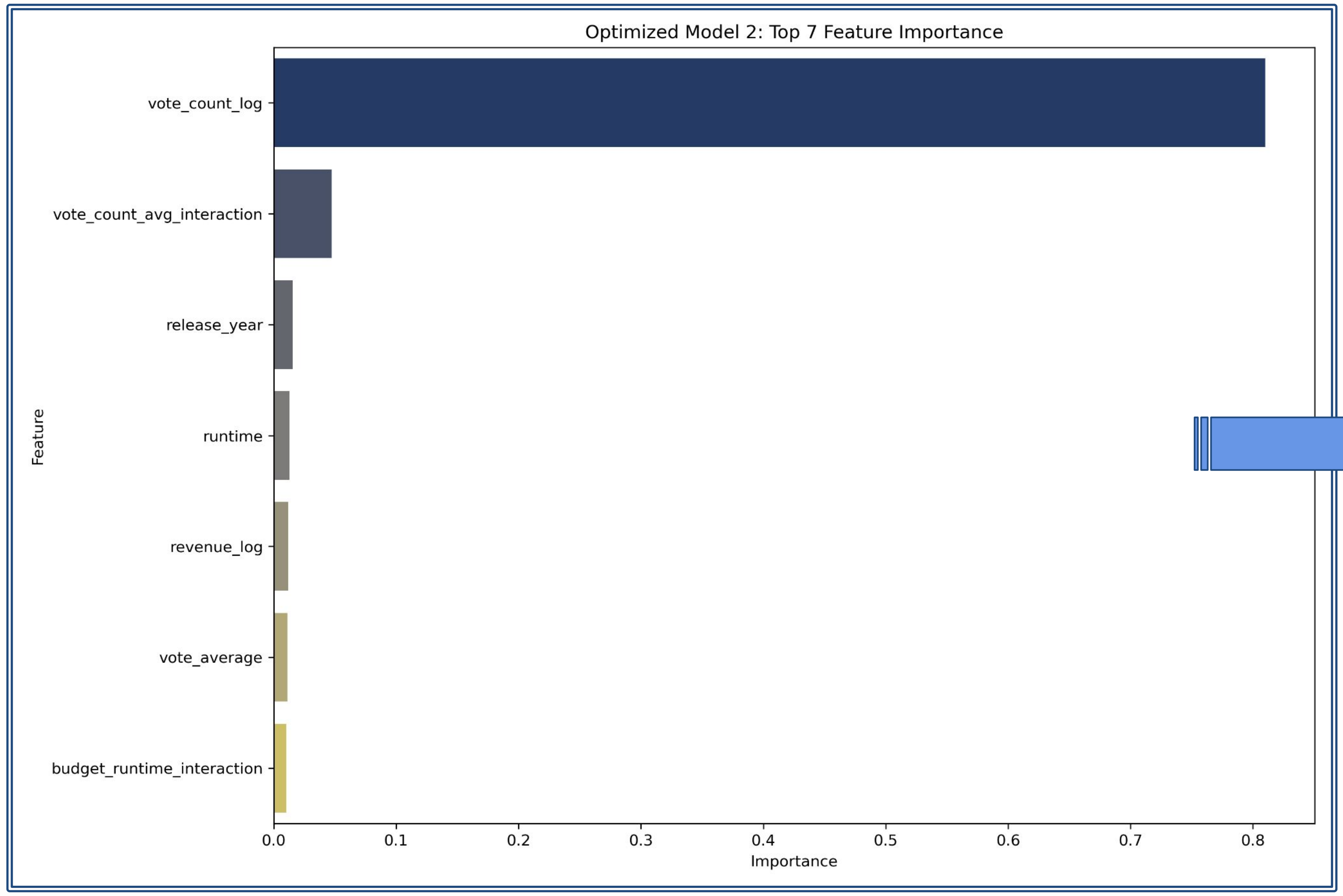


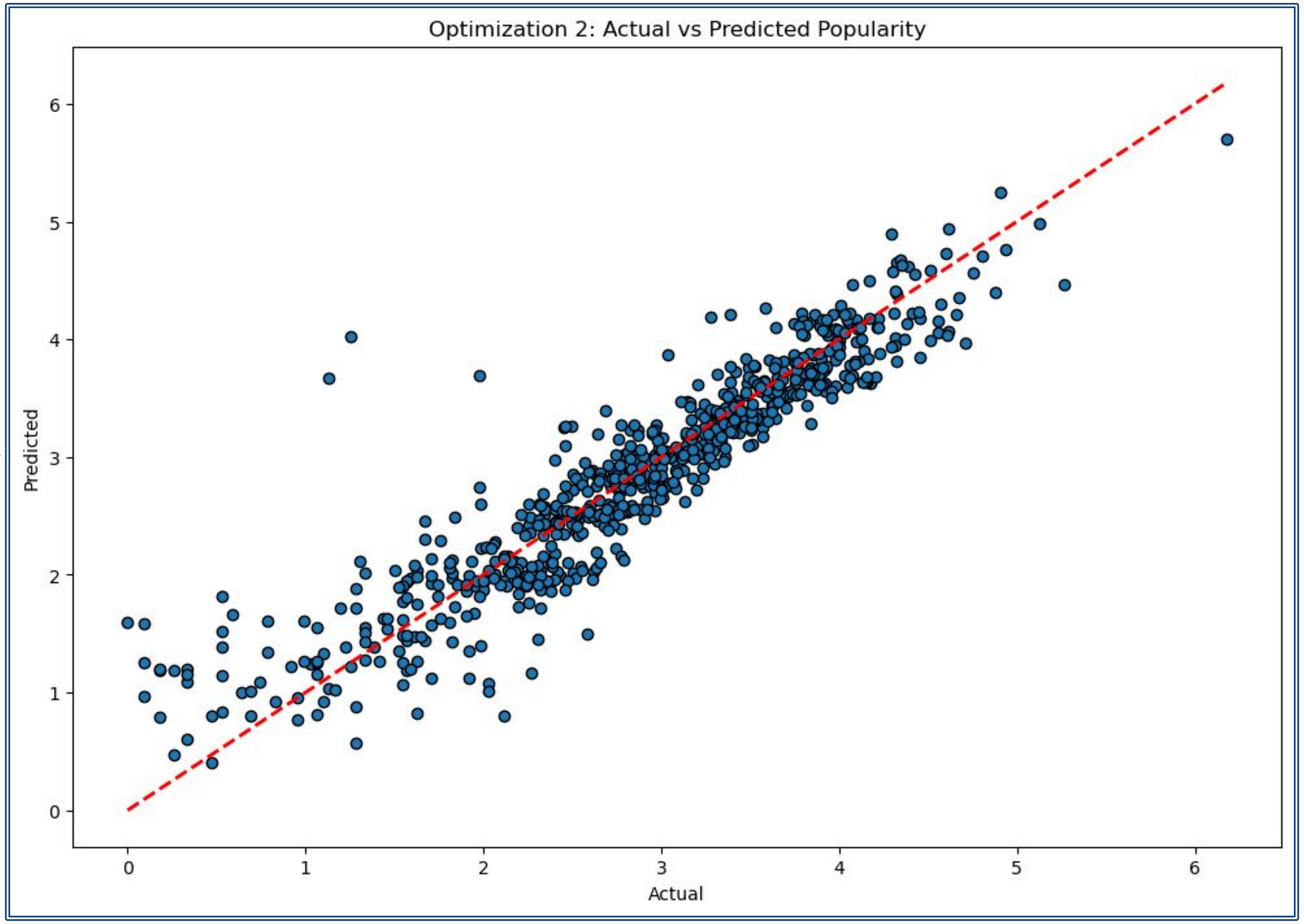


Optimization Two: One-Hot Encoding

One-hot encoding to ‘genres’ and ‘award’ columns

Root Mean Squared Error (RMSE): 0.35880964055083064 R-Squared (R^2): 0.8659762242238038

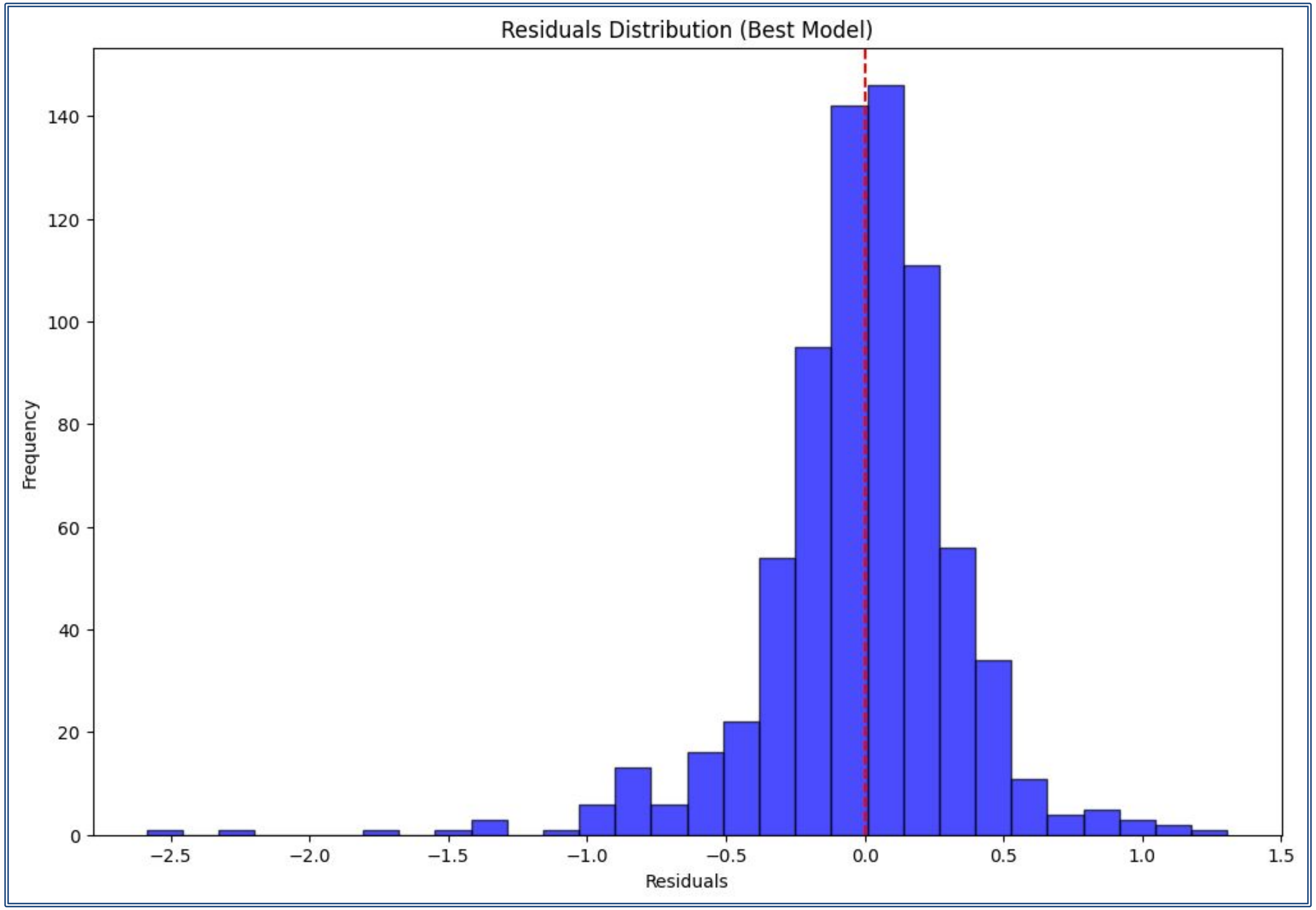




Optimized Residual Distribution

- Outliers: few outliers present → model struggled to accurately predict some data points

- Clustering: significant cluster are close to 0 → predictions are generally accurate



R-squared Results Based on Random Forest Optimization Techniques

1. Baseline Model

- Log Transformations
- Encoding and Cyclical Transformation
- Interaction Terms

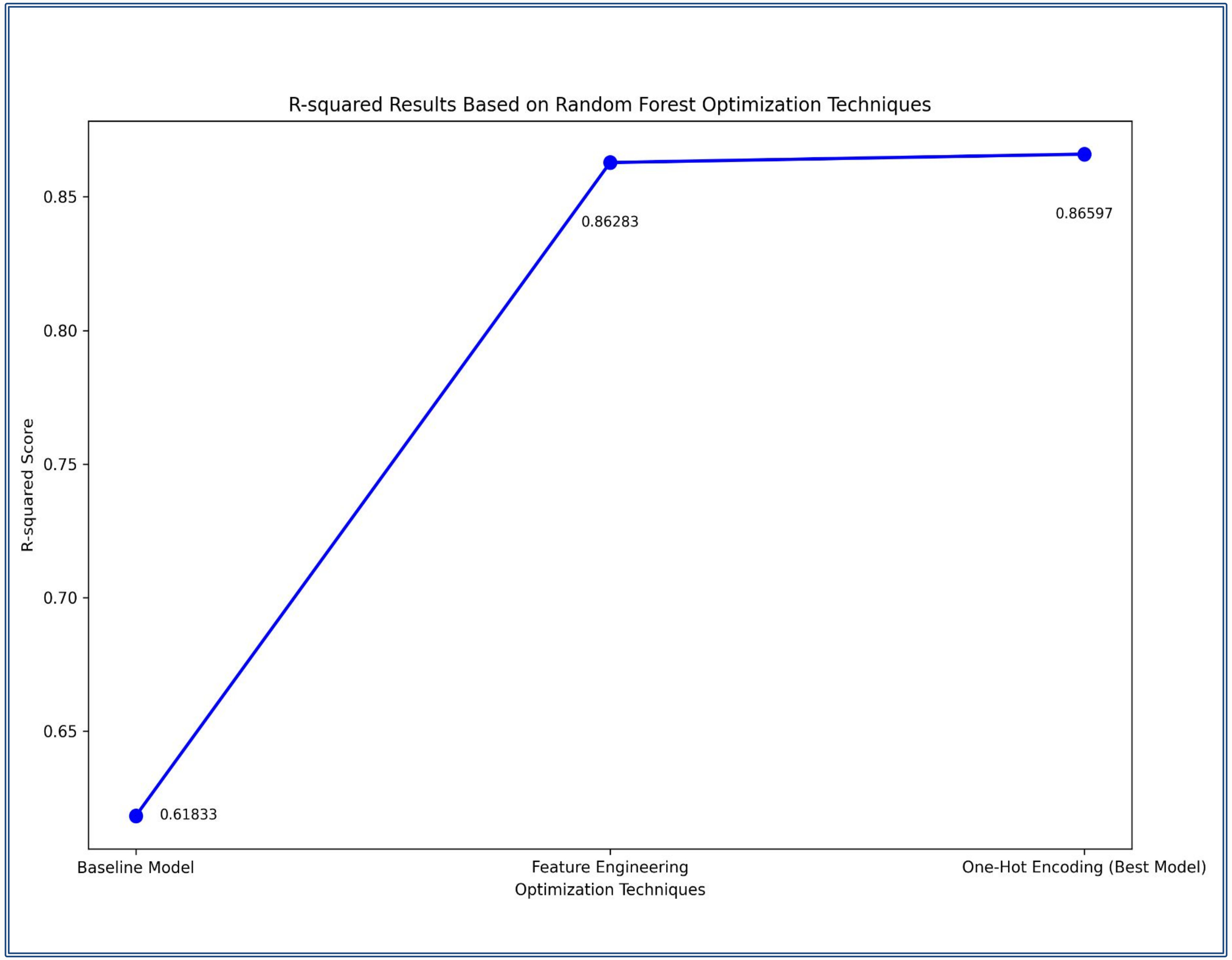
2. Feature Engineering

- Log Transformation
- Interaction Features
- Seasonal Features (sin/cos)

3. One-Hot Encoding

- One-Hot Encoding for Genres and Awards

The visual presentation showcases the progression and impact of each optimization technique on the model's accuracy.



Questions


Thank You!
Applause !!