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, our models will uncover the key factors driving a movie's success, offering critical insights for producers and marketers.

Data Source: Kaggle’s TMDB 5000 Movie and Oscar Best Picture Movies Datasets

Challenge: Identifying which features had the most impact on predicting popularity.

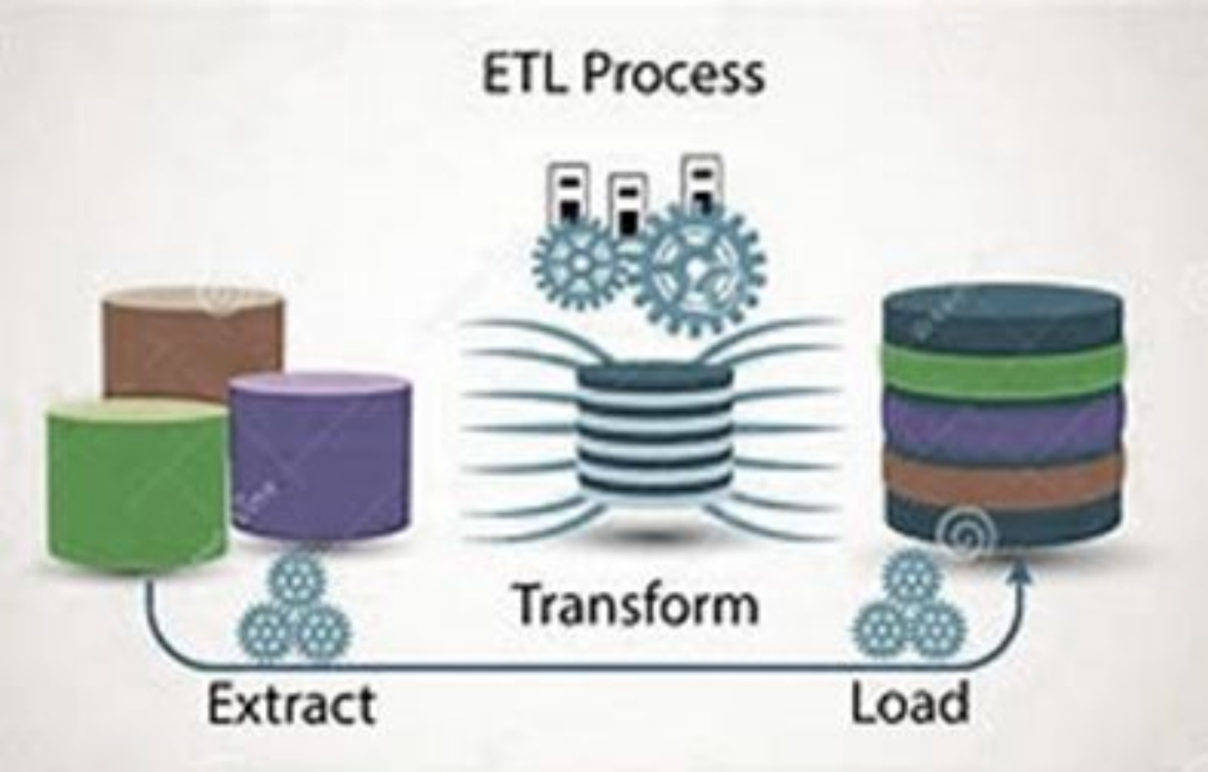
Assumptions:

- 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

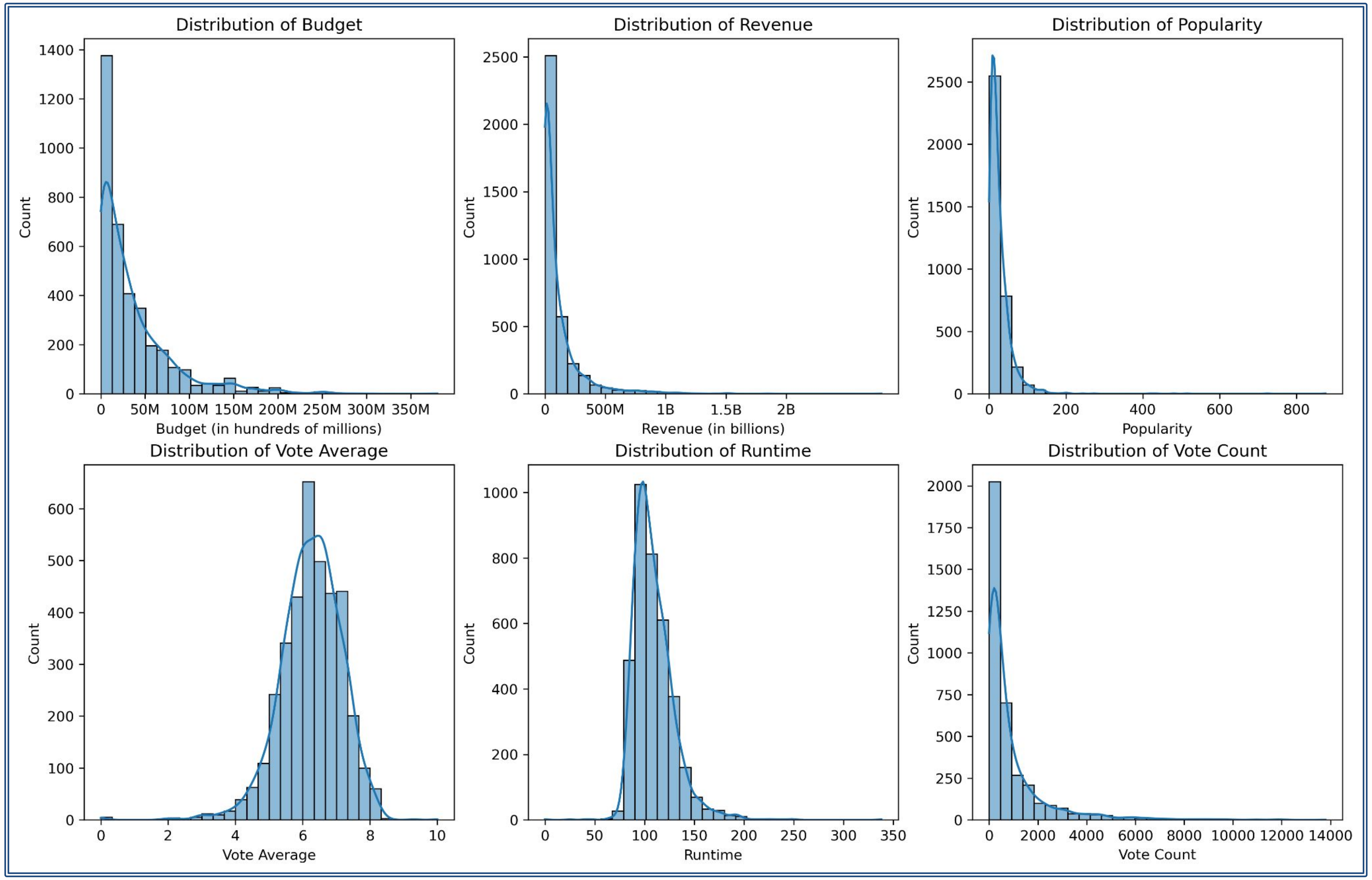
Key Activities:

- ✓ **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.



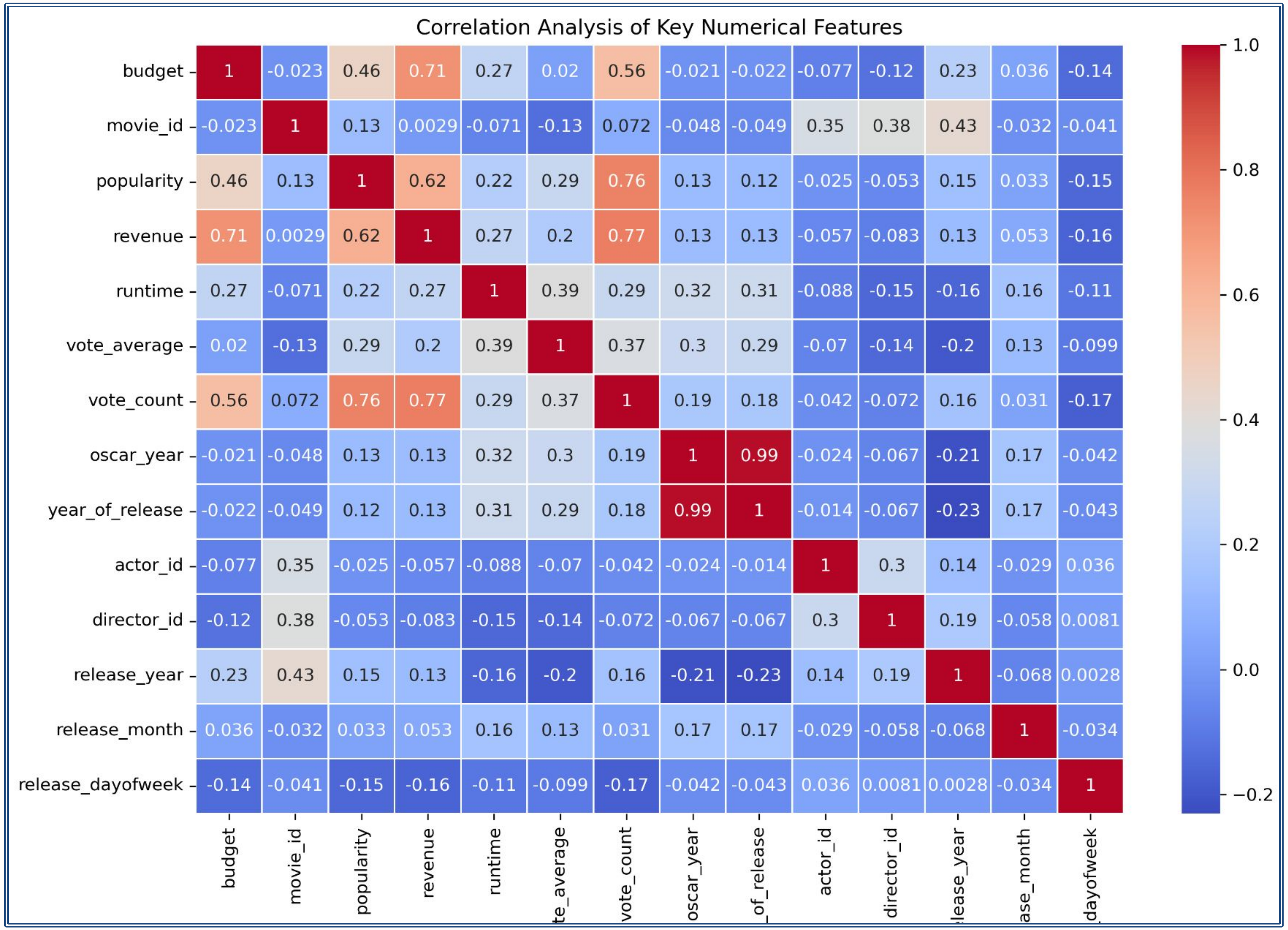
EDA Highlights

- Budget, revenue, popularity, and vote count have a right skewed distribution
- Shows extreme outliers
- Vote average and runtime have a normal distribution



EDA Highlights

- **Budget and popularity:** Movies with higher budgets and popularity tend to have higher revenues.
- **Vote count and revenue:** A strong correlation exists between vote count and revenue, indicating the importance of audience engagement.
- **Runtime and revenue:** While the correlation is moderate, longer movies might have higher revenue potential.
- **Oscar wins:** Winning an Oscar doesn't necessarily correlate strongly with budget or release year.



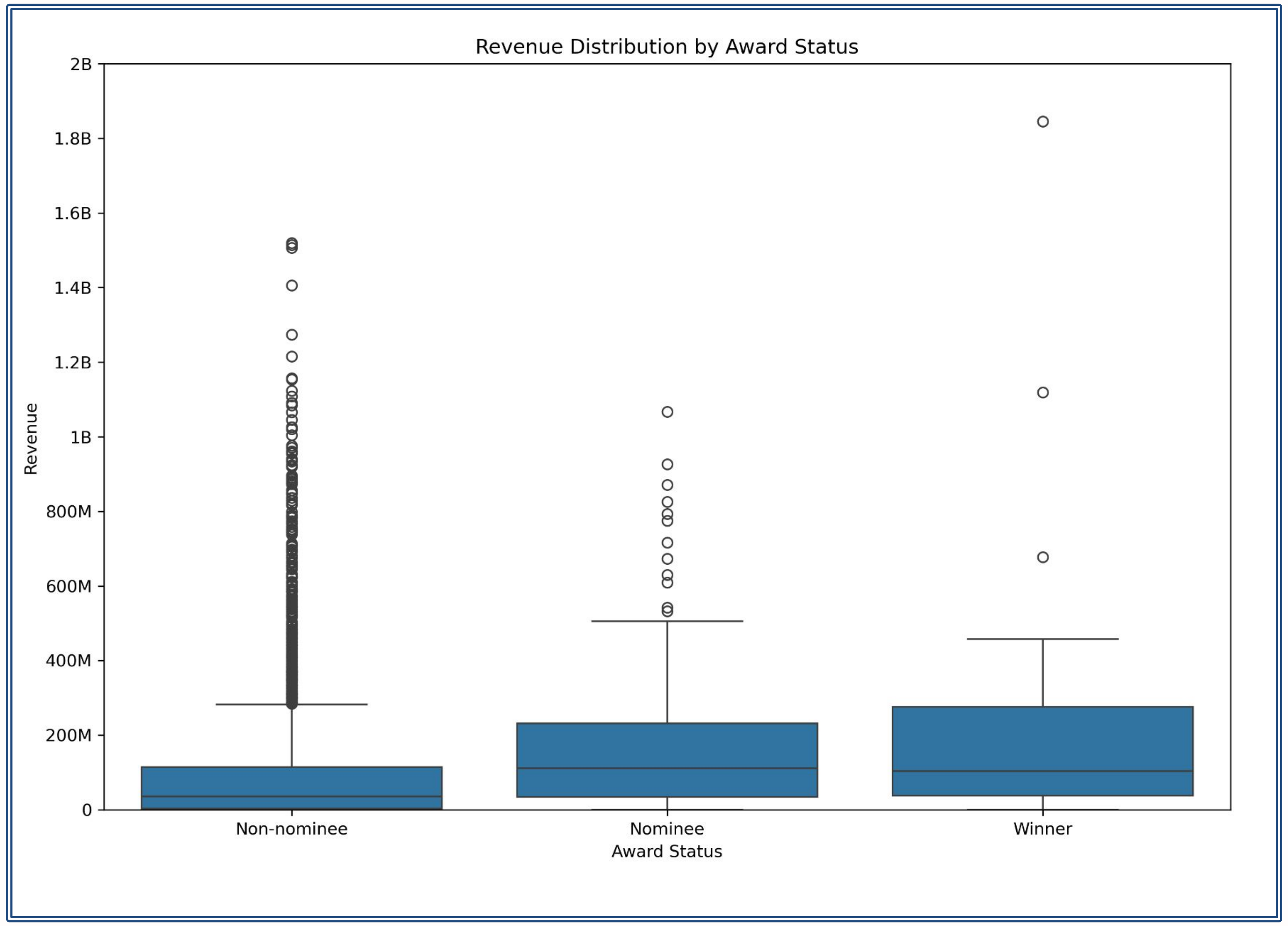
Correlation Analysis of Key Numerical Features

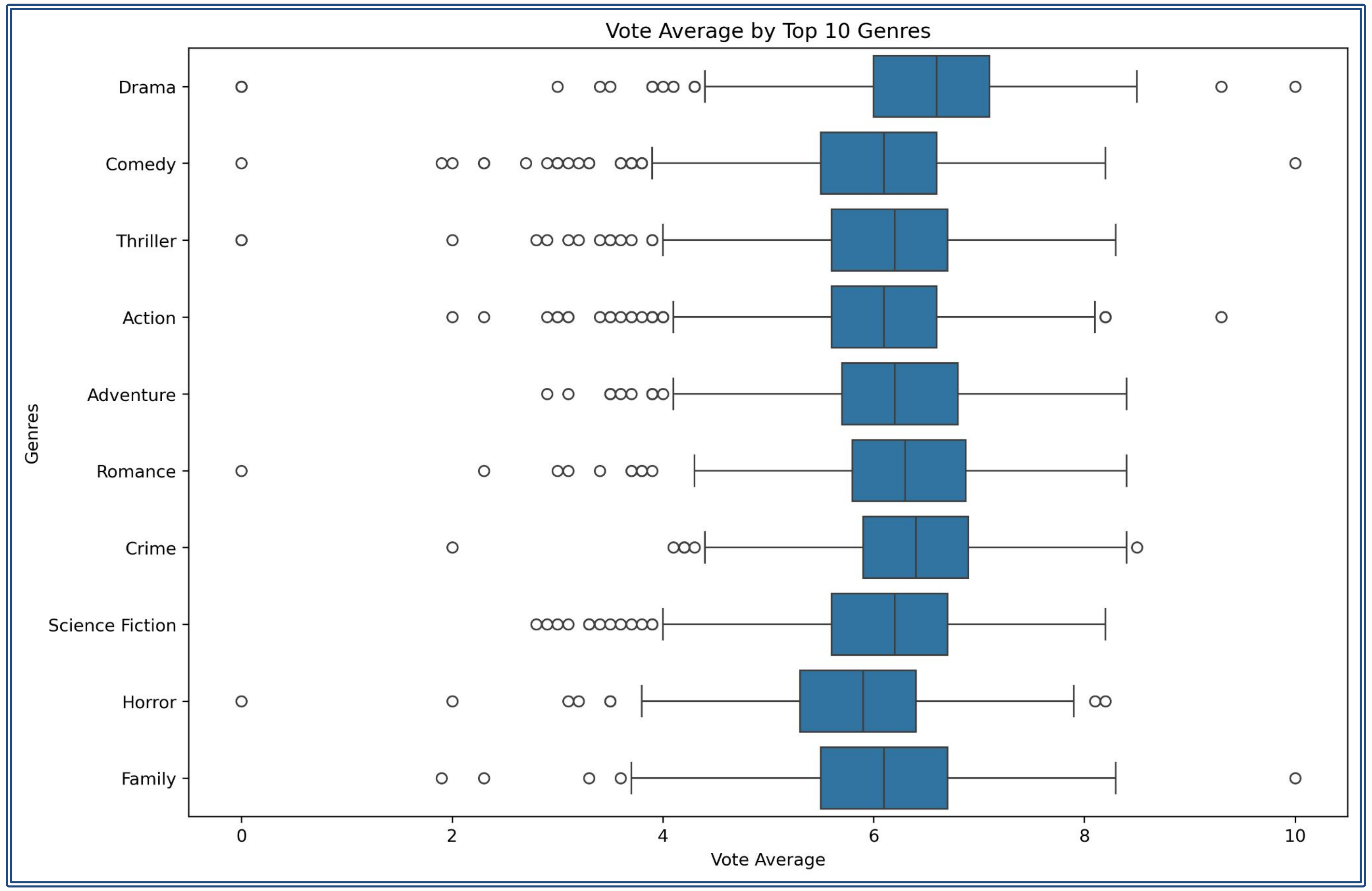
	budget	movie_id	popularity	revenue	runtime	te_average	vote_count	oscar_year	_of_release	actor_id	director_id	lease_year	ase_month	dayofweek
budget	1	-0.023	0.46	0.71	0.27	0.02	0.56	-0.021	-0.022	-0.077	-0.12	0.23	0.036	-0.14
movie_id	-0.023	1	0.13	0.0029	-0.071	-0.13	0.072	-0.048	-0.049	0.35	0.38	0.43	-0.032	-0.041
popularity	0.46	0.13	1	0.62	0.22	0.29	0.76	0.13	0.12	-0.025	-0.053	0.15	0.033	-0.15
revenue	0.71	0.0029	0.62	1	0.27	0.2	0.77	0.13	0.13	-0.057	-0.083	0.13	0.053	-0.16
runtime	0.27	-0.071	0.22	0.27	1	0.39	0.29	0.32	0.31	-0.088	-0.15	-0.16	0.16	-0.11
vote_average	0.02	-0.13	0.29	0.2	0.39	1	0.37	0.3	0.29	-0.07	-0.14	-0.2	0.13	-0.099
vote_count	0.56	0.072	0.76	0.77	0.29	0.37	1	0.19	0.18	-0.042	-0.072	0.16	0.031	-0.17
oscar_year	-0.021	-0.048	0.13	0.13	0.32	0.3	0.19	1	0.99	-0.024	-0.067	-0.21	0.17	-0.042
year_of_release	-0.022	-0.049	0.12	0.13	0.31	0.29	0.18	0.99	1	-0.014	-0.067	-0.23	0.17	-0.043
actor_id	-0.077	0.35	-0.025	-0.057	-0.088	-0.07	-0.042	-0.024	-0.014	1	0.3	0.14	-0.029	0.036
director_id	-0.12	0.38	-0.053	-0.083	-0.15	-0.14	-0.072	-0.067	-0.067	0.3	1	0.19	-0.058	0.0081
release_year	0.23	0.43	0.15	0.13	-0.16	-0.2	0.16	-0.21	-0.23	0.14	0.19	1	-0.068	0.0028
release_month	0.036	-0.032	0.033	0.053	0.16	0.13	0.031	0.17	0.17	-0.029	-0.058	-0.068	1	-0.034
release_dayofweek	-0.14	-0.041	-0.15	-0.16	-0.11	-0.099	-0.17	-0.042	-0.043	0.036	0.0081	0.0028	-0.034	1

EDA Highlights

- Nominees/Winners show consistent, higher medians revenue compared to non-nominees
- Non-nominated movies have extreme outliers with very high revenue

- Drama and Romance have a higher vote average = stronger audience approval
- Comedy and Horror show wider variation and lower medians in vote average
- Significant outliers in Family and Horror genres



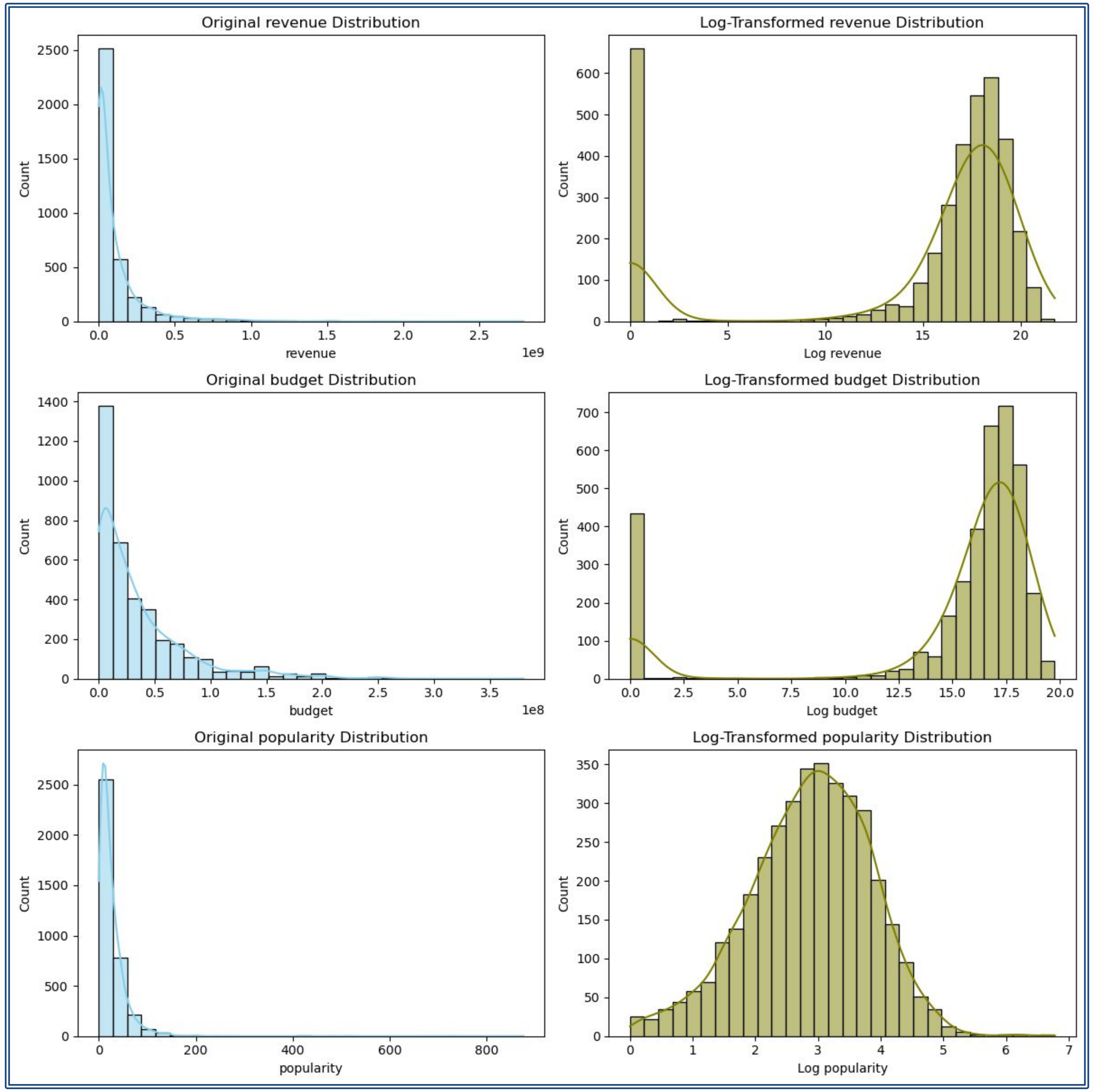


ML Data Preparation

Baseline Features: ‘budget’, ‘runtime’, ‘vote_average’, ‘release_month’, ‘release_dayofweek’, and ‘revenue’
Target: ‘popularity’

Data Preparation for Modeling

- **Feature Engineering:**
  - **Log Transformations** to handle skewed data and outliers
  - **Interaction Terms** multiplying budget and runtime to capture combined relationship
  - **Label Encoding** on categorical features by transforming release_month and release_dayofweek into numerical values
  - **Cyclic Transformation** of month and day to represent their circular nature



Original revenue Distribution

Log-Transformed revenue Distribution

Original budget Distribution

Log-Transformed budget Distribution

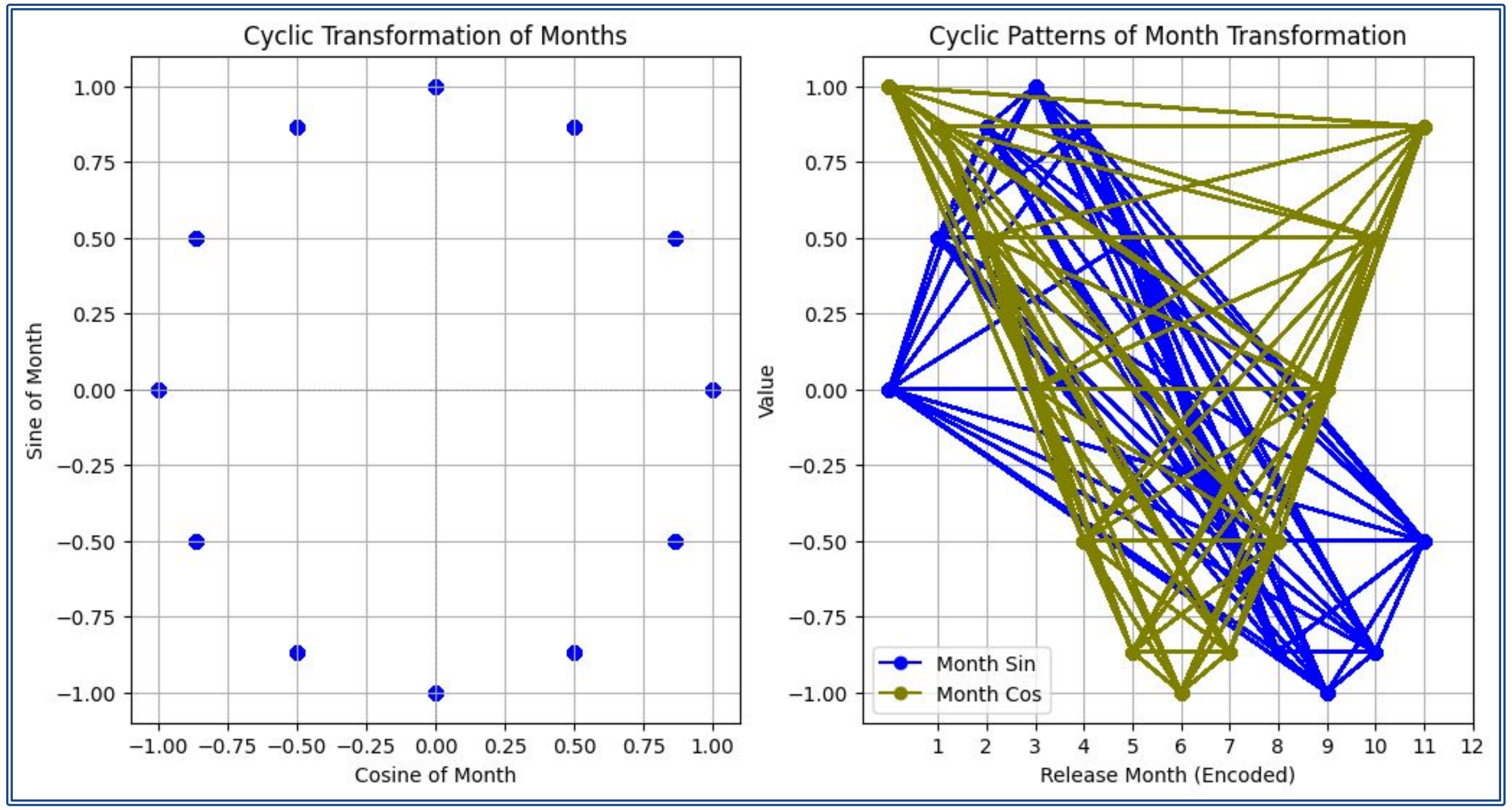
Original popularity Distribution

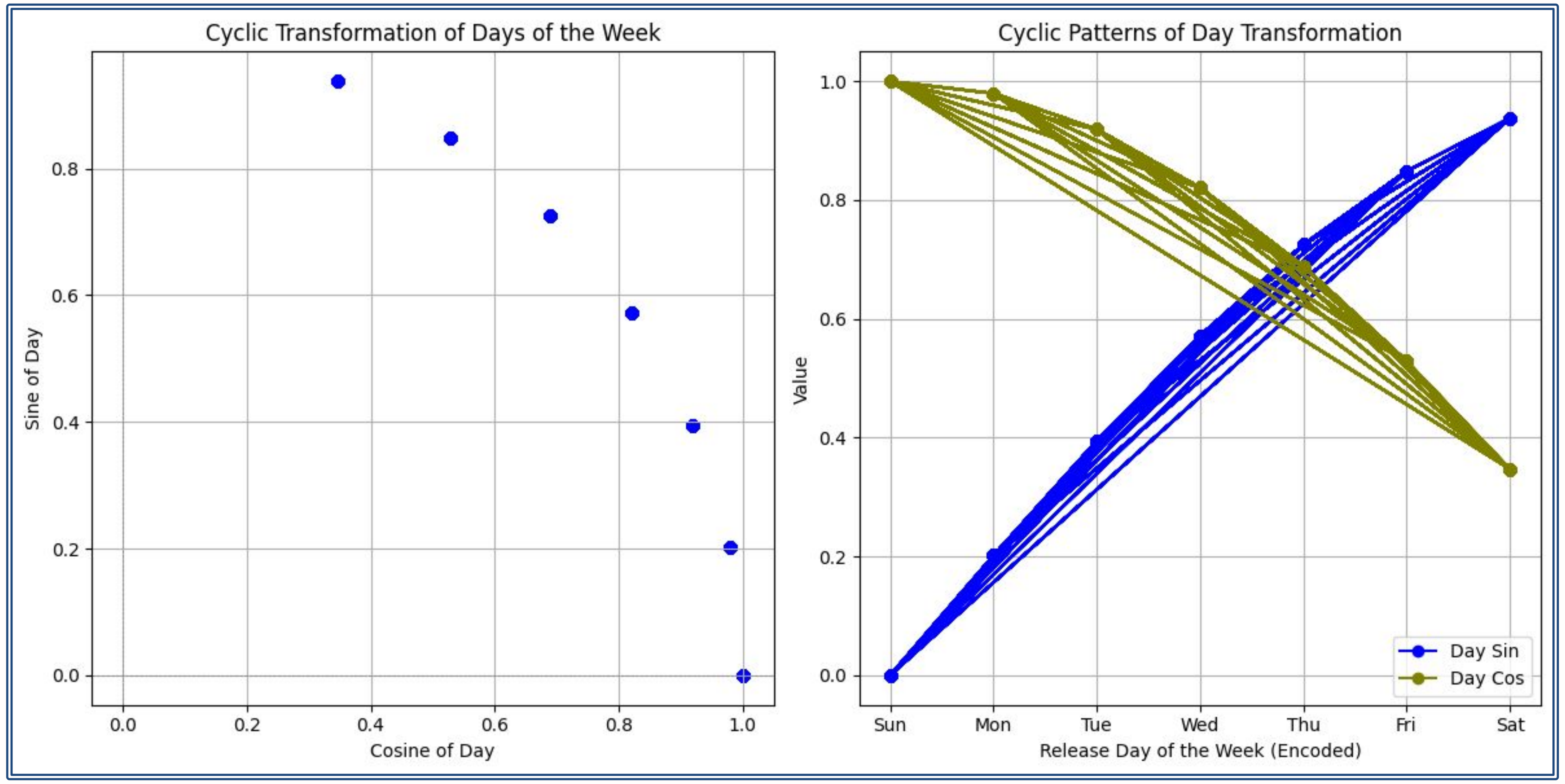
Log-Transformed popularity Distribution

ML Data Preparation

Cyclic Encoding was used to :

- **Teach the model patterns:** We anticipated it to help the model recognize patterns data that repeat over and over, like the days of the week.
- **Make better predictions:** The intent was for better understanding of patterns. The model was enabled to make better guesses about what might happen next. For example, knowing movie sales are usually higher on weekends, it can predict higher sales for a Saturday.



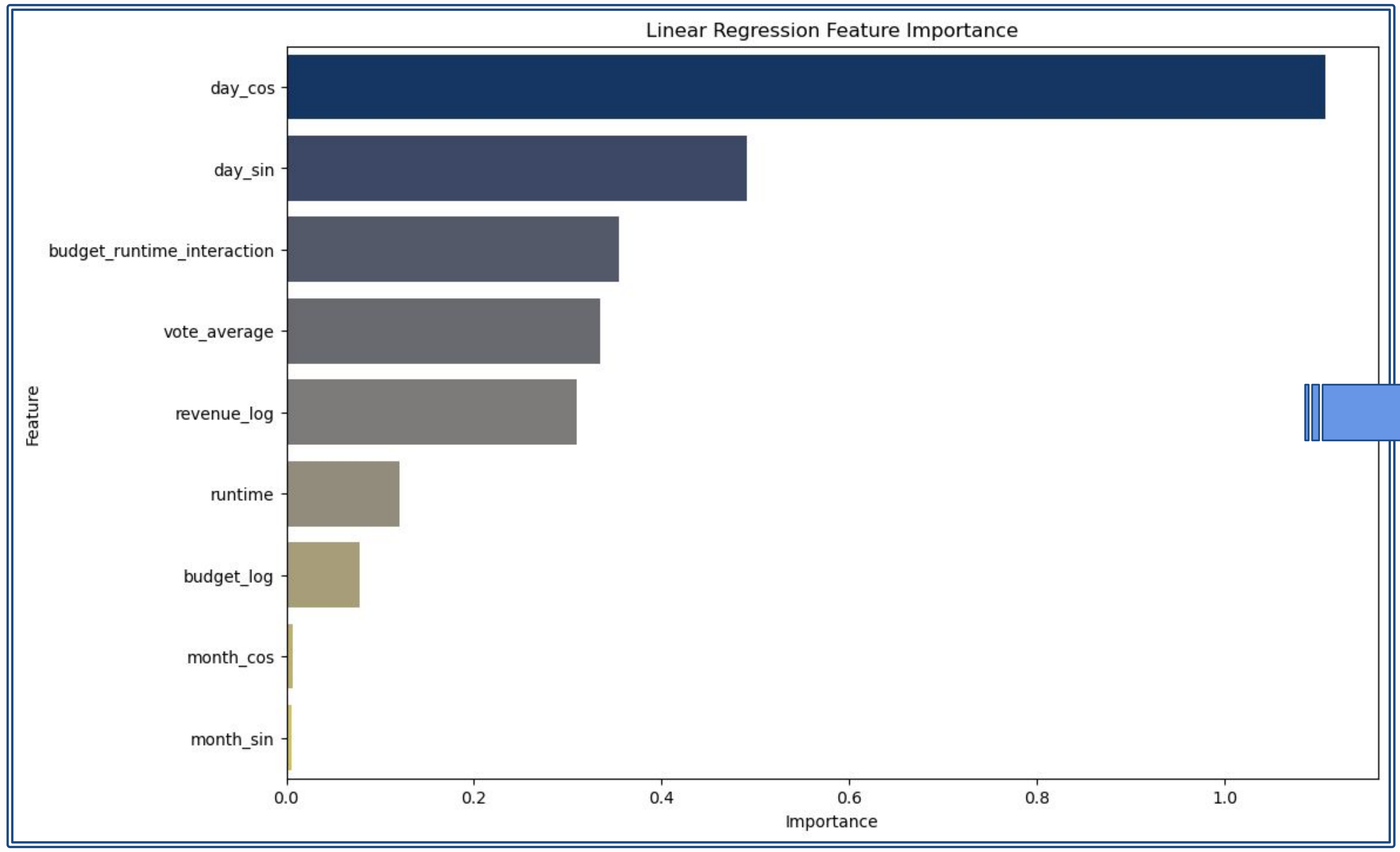


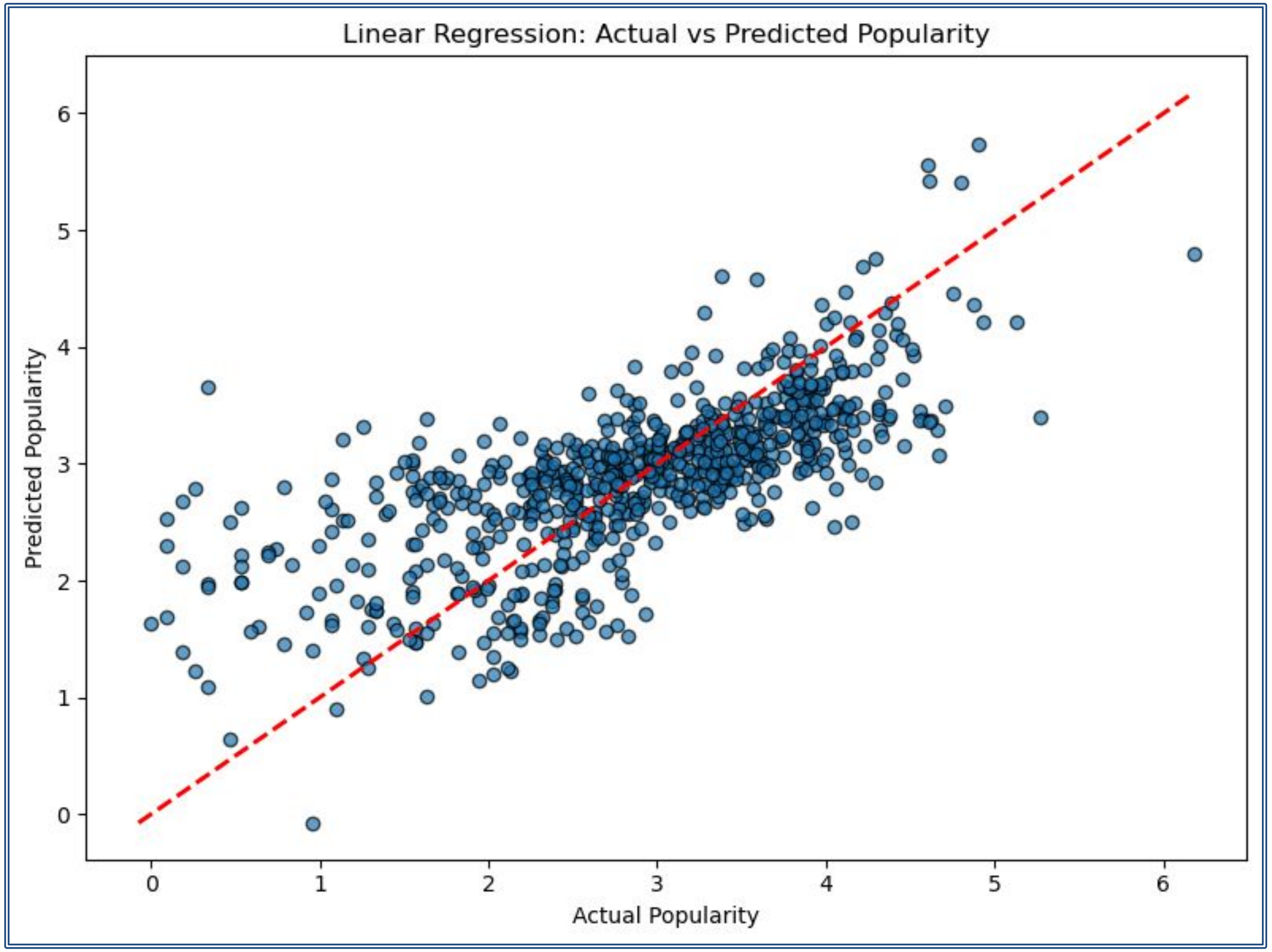
Model 1: Linear Regression

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

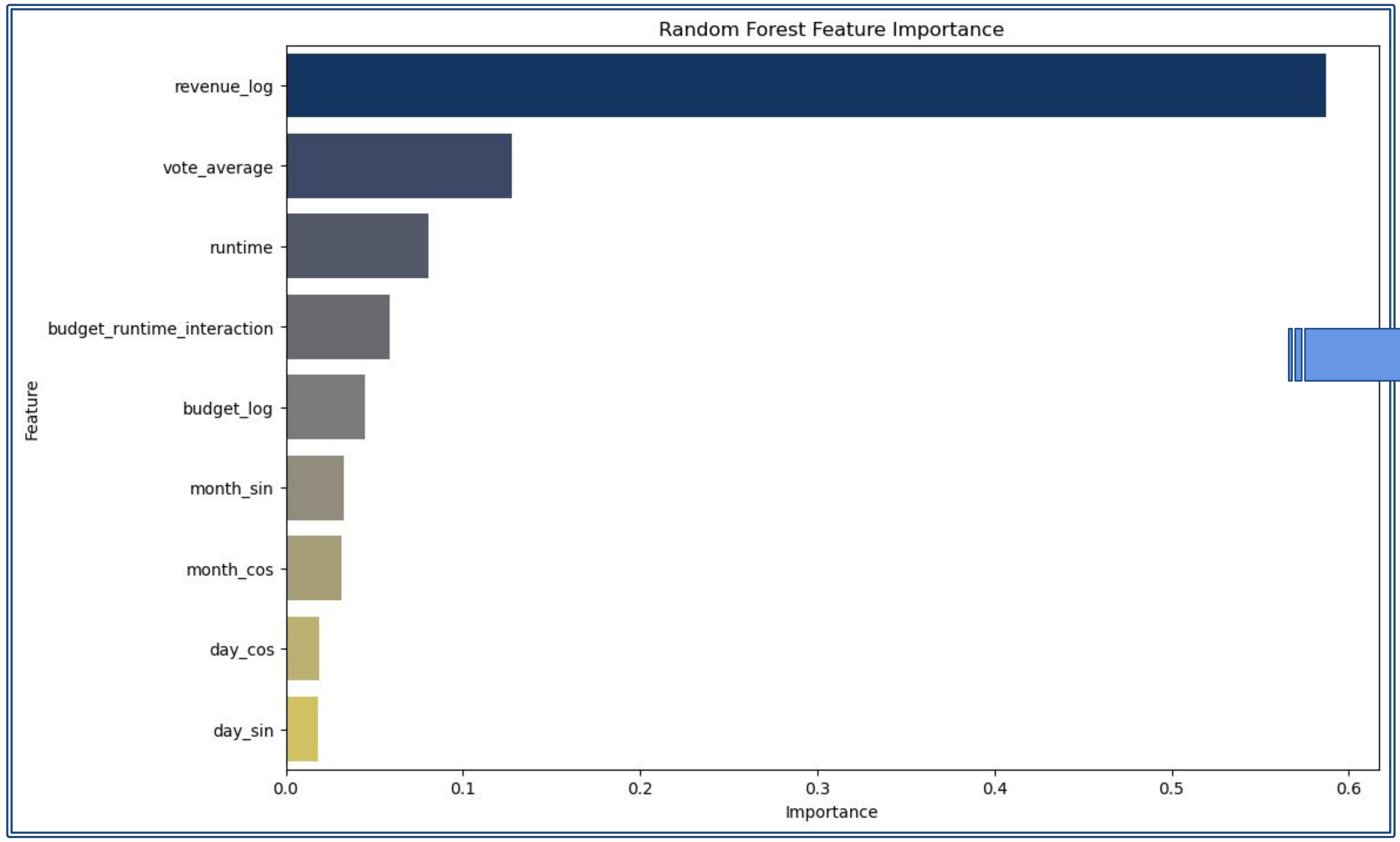
R-squared (R^2): 0.5066330266530592

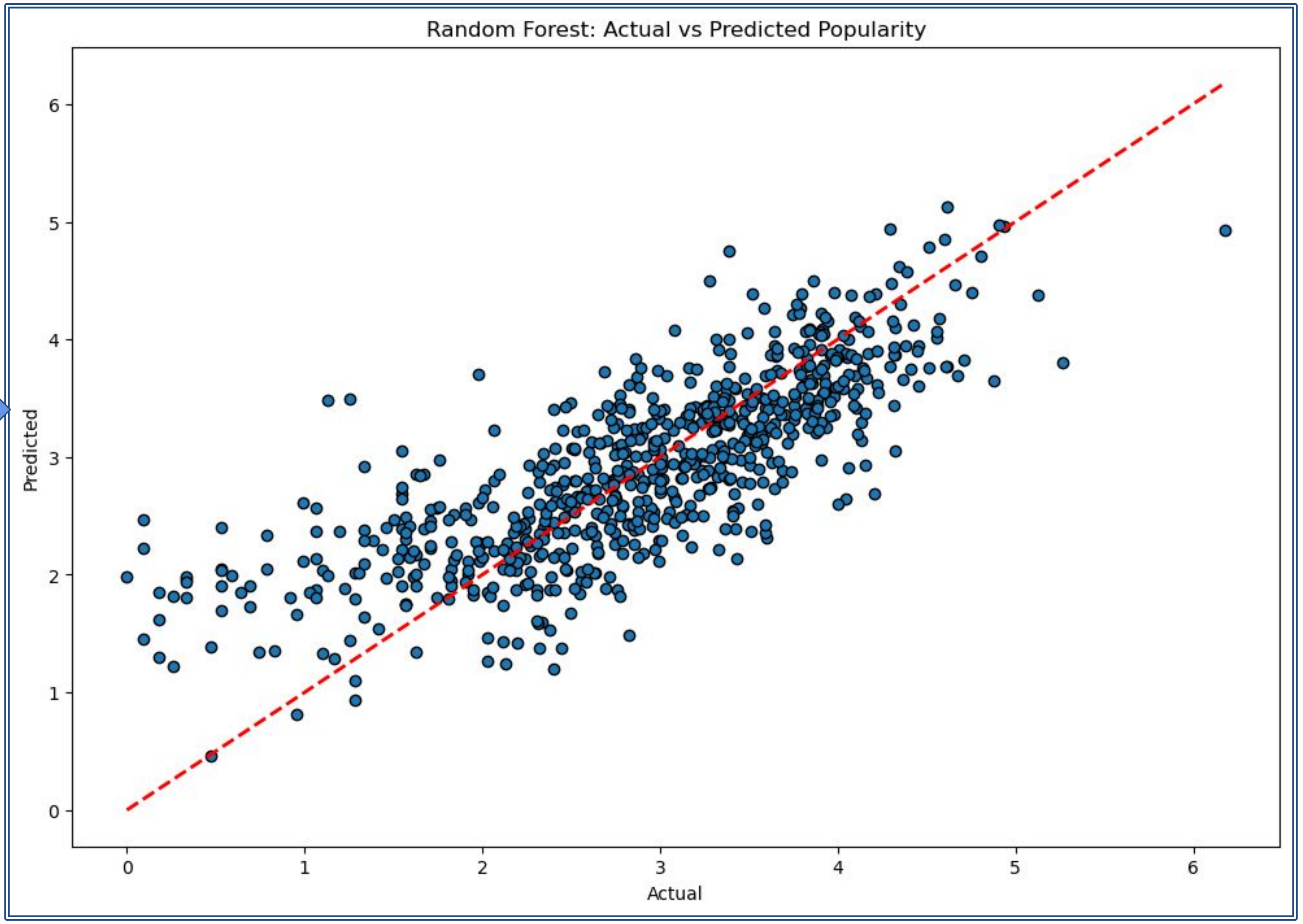




Model 2: Random Forest

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





RANDOM FOREST

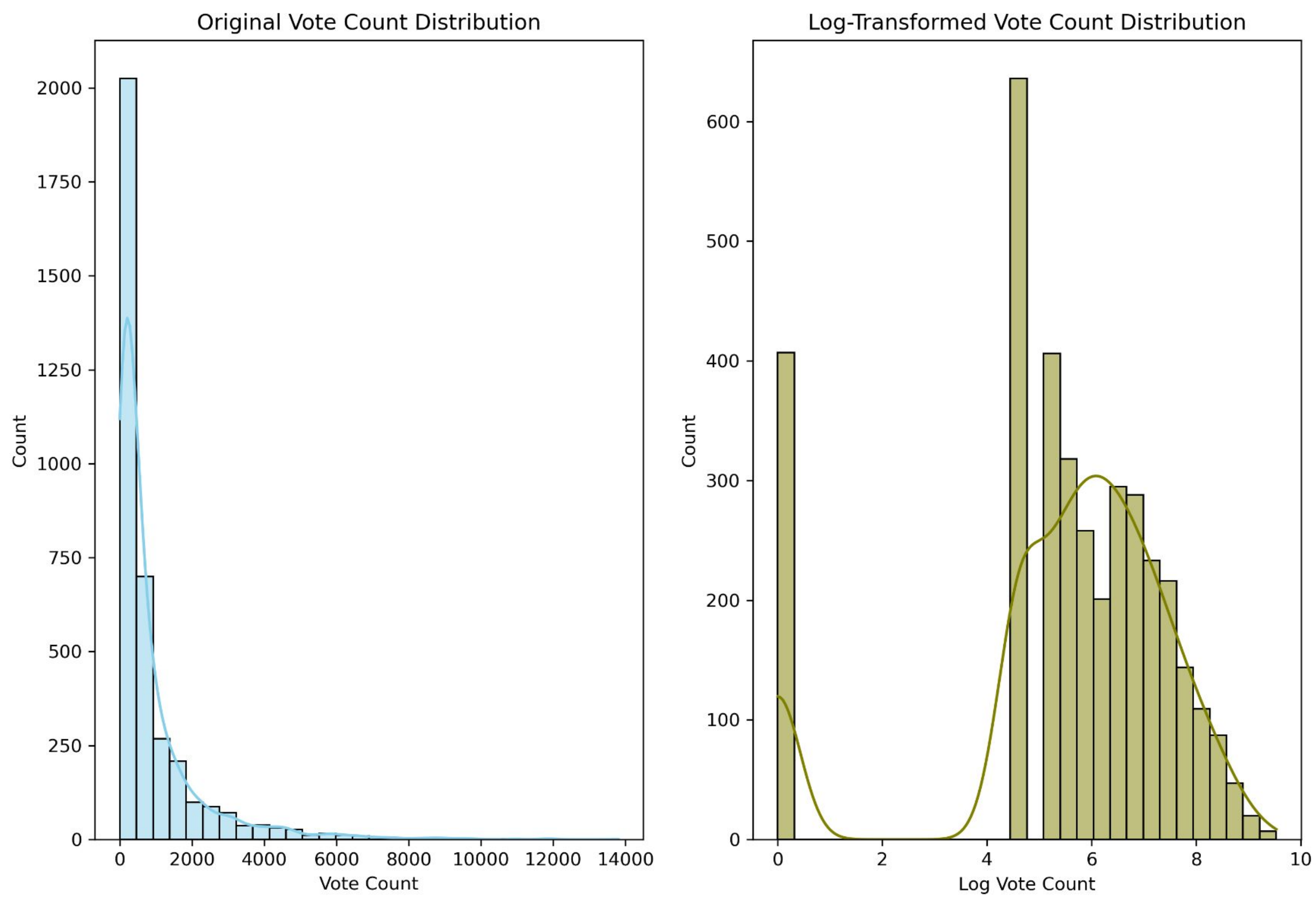
- Complex Relationships handling
- Feature Importance recognition
- Robustness handling of outlier’s noisy data
- Data Versatility with numerical & categorical
- Predictive Power for unseen data

= Ideal for predicting future movie trends!

Optimization One: Feature Engineering

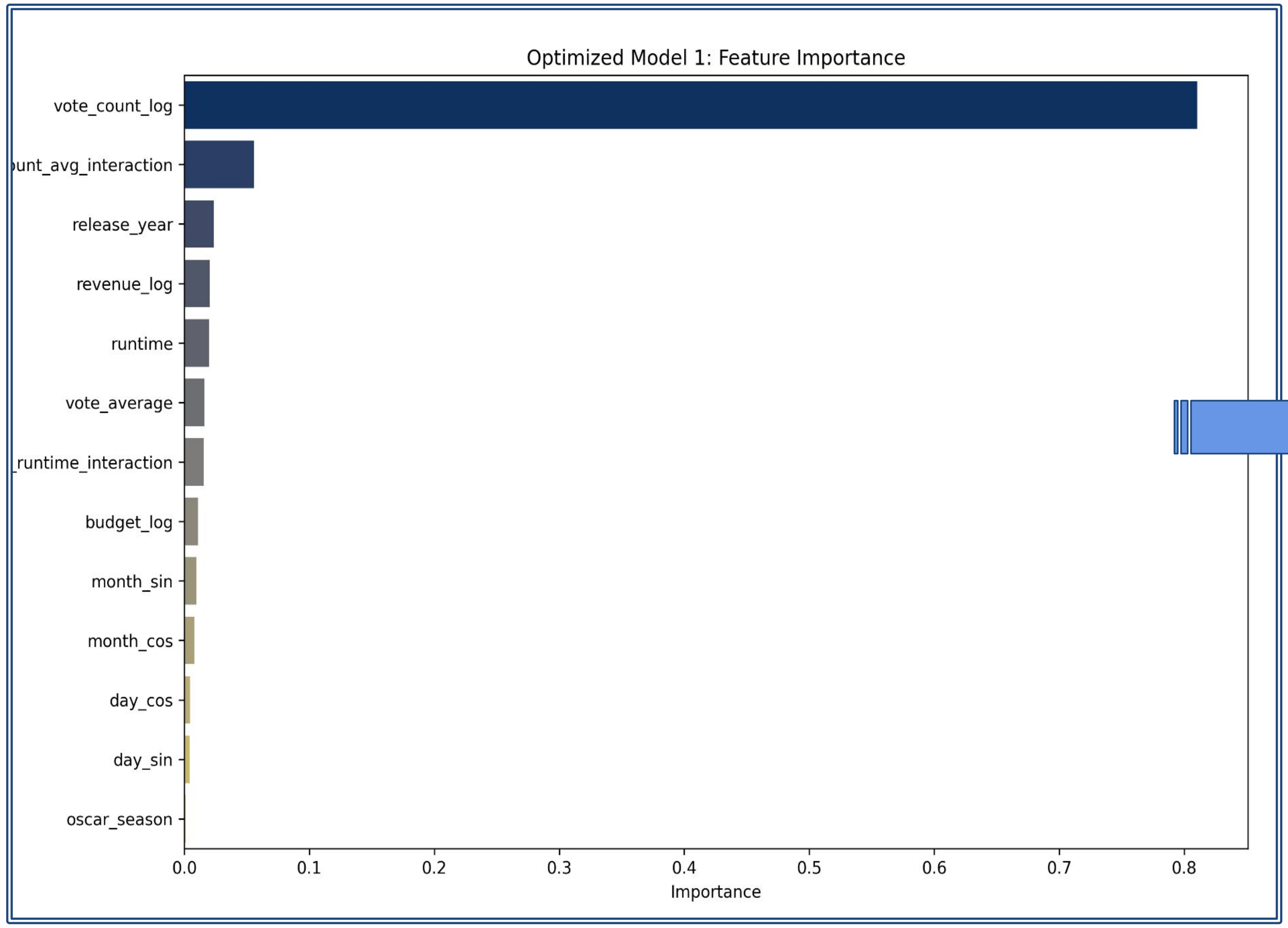
- **Log Transformation:** on ‘vote_count’ feature to normalize its distribution
- **Interaction Features** combining ‘vote_count_log’ and ‘vote_average’ to capture the relationship
- **Seasonal Feature** that flags movies released during the Oscar season, in the fall and winter months

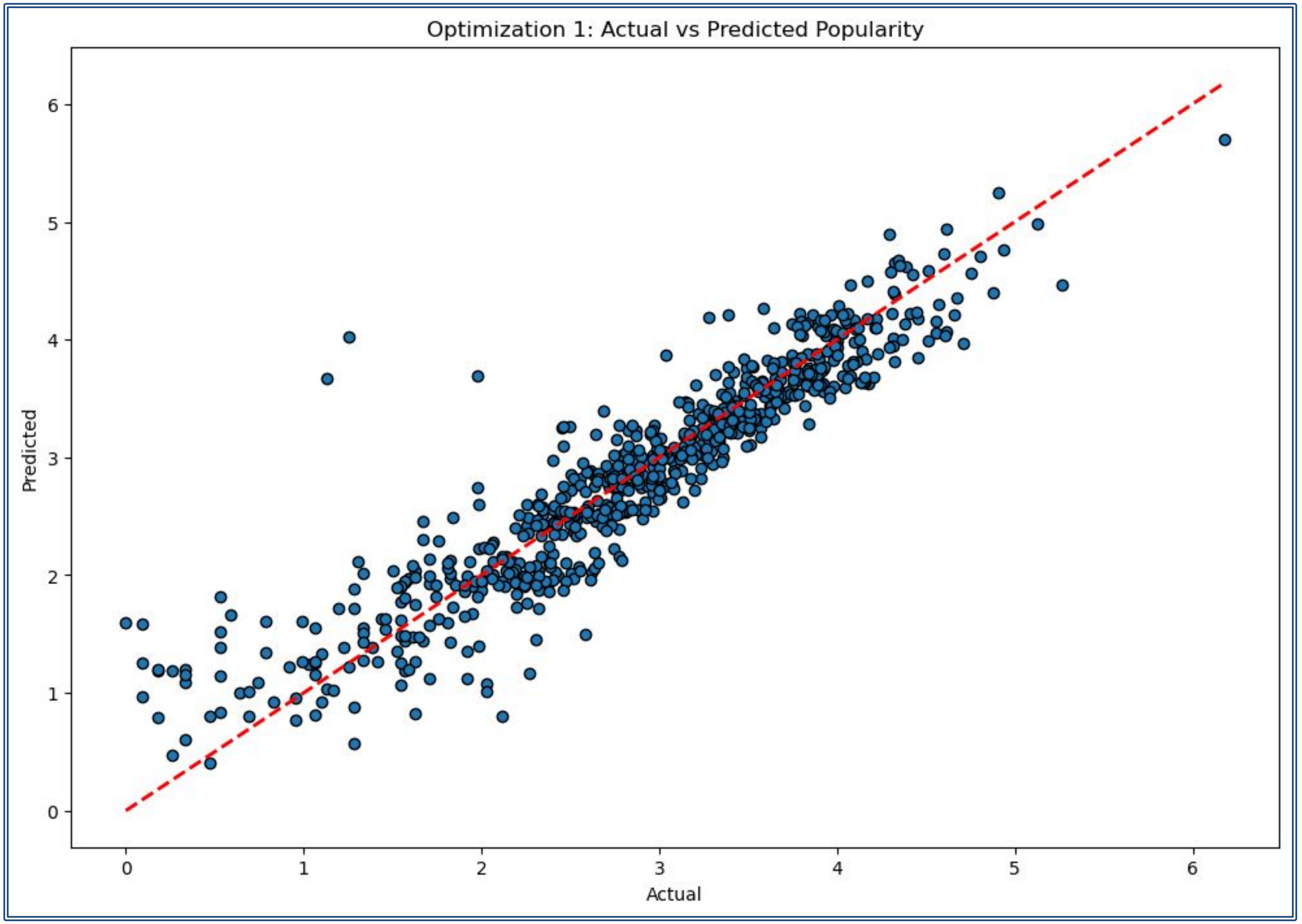
Updated Features: ['budget_log', 'revenue_log', 'runtime', 'vote_average', 'release_year', 'budget_runtime_interaction', 'month_sin', 'month_cos', 'day_sin', 'day_cos', 'vote_count_log', 'vote_count_avg_interaction', 'oscar_season']



Optimization One Output

Root Mean Squared Error (RMSE): 0.3629928041279302   New R-Squared: 0.8628329893191401





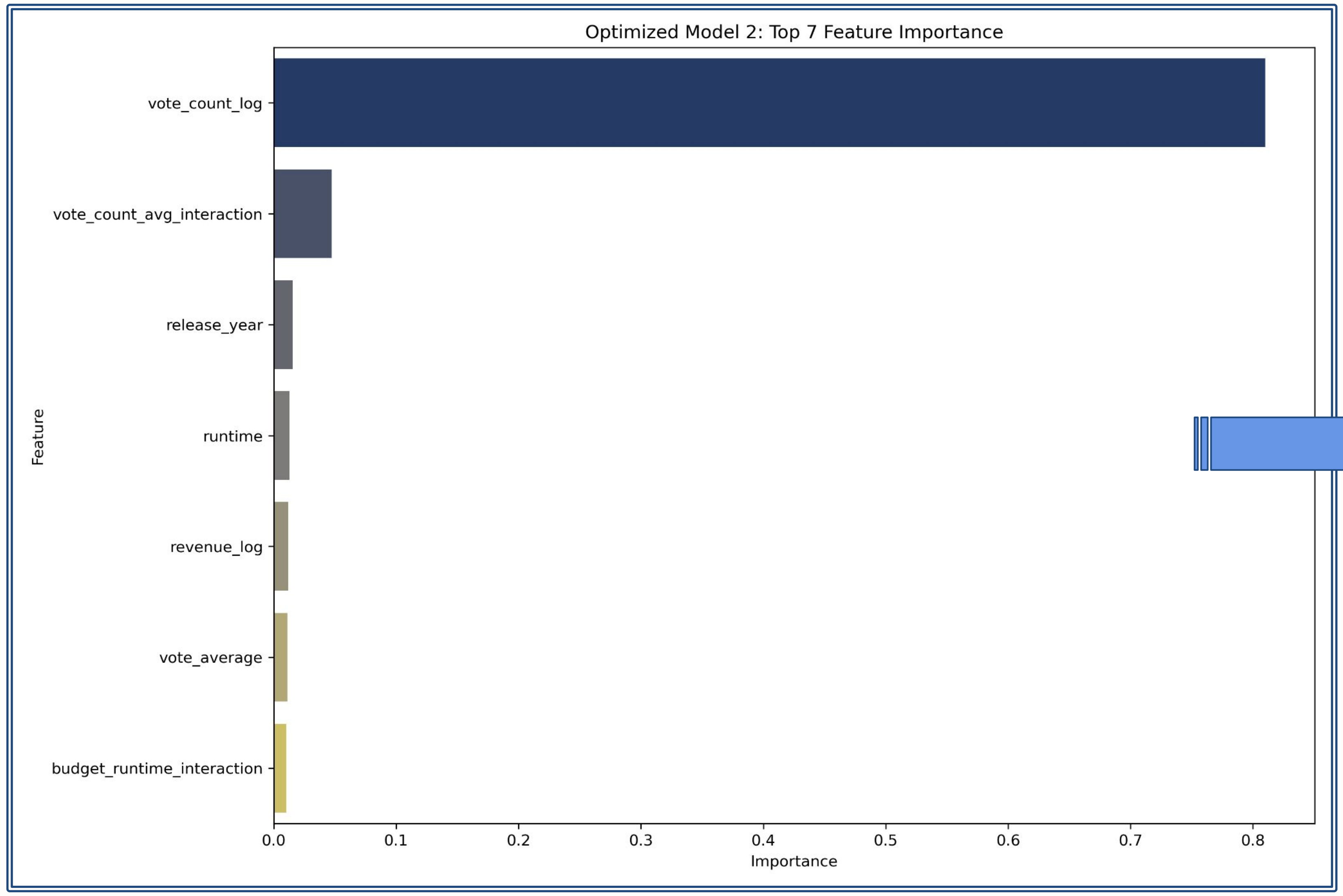
Optimization Two: One-Hot Encoding

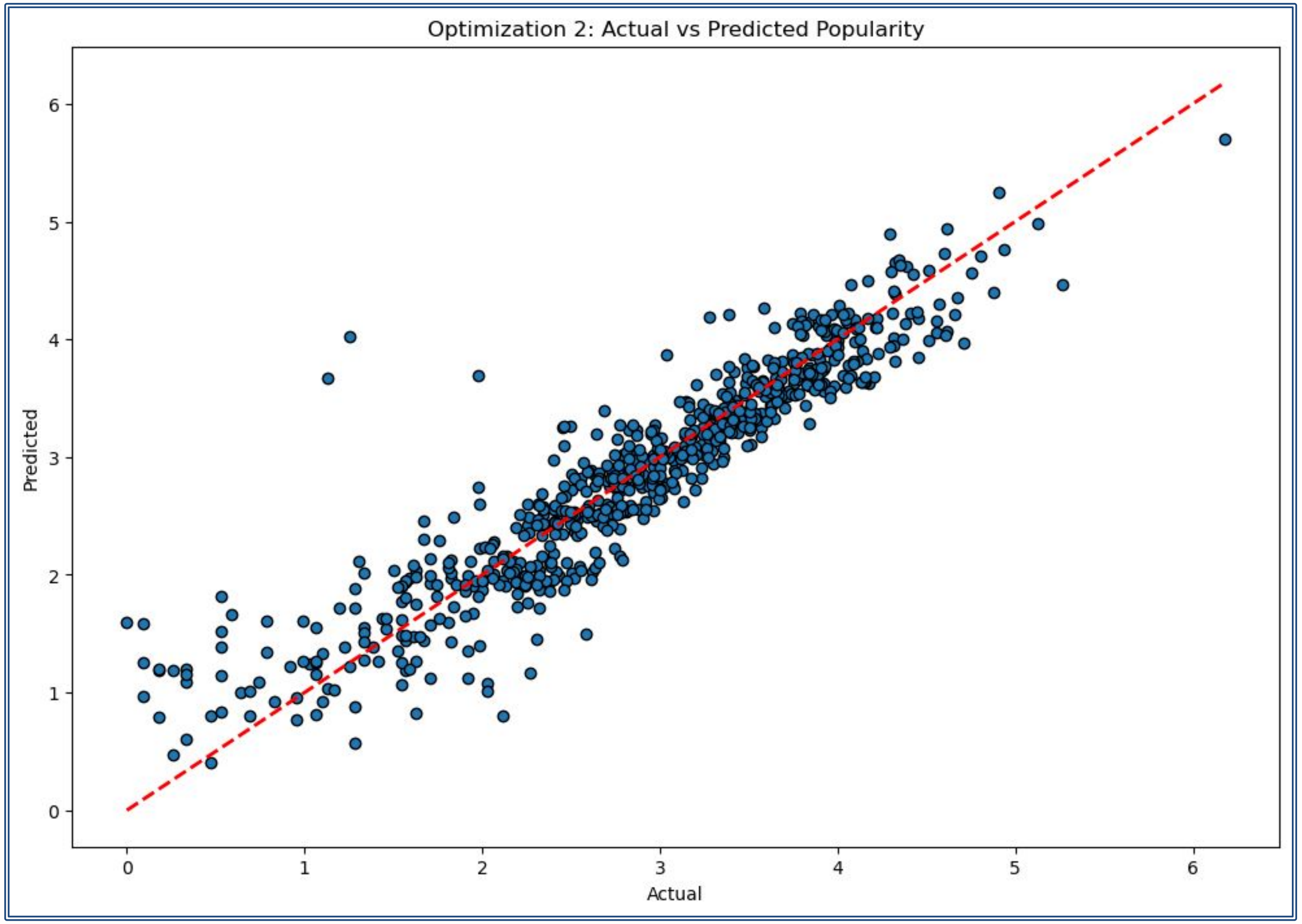
One-hot encoding to ‘genres’ and ‘award’ columns

Updated Features: optimized one features + encoded_columns

Root Mean Squared Error (RMSE): 0.6055005301511441

R-Squared (R^2): 0.8659762242238038


Optimized Model 2: Top 7 Feature Importance


Optimization 2: Actual vs Predicted Popularity

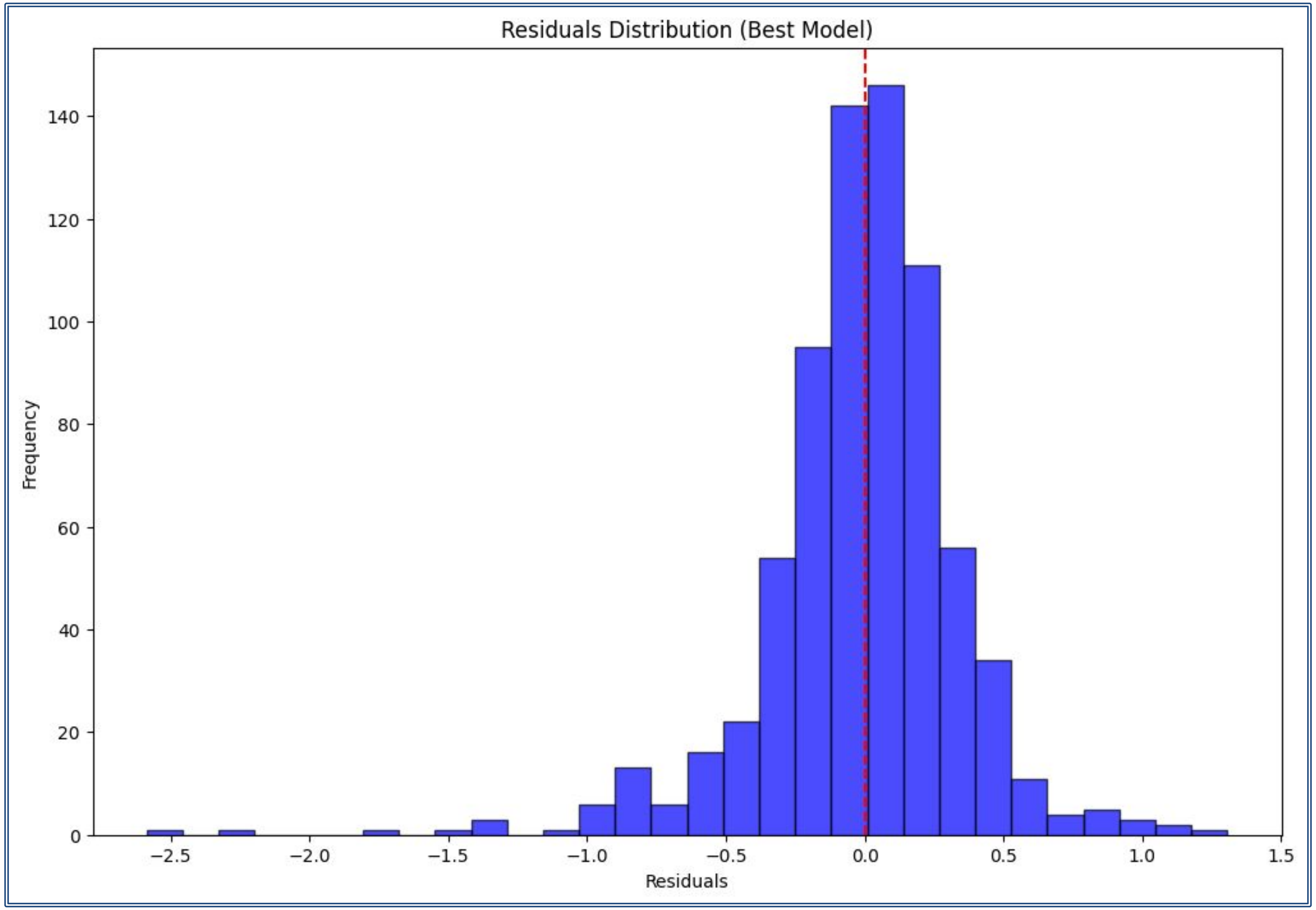
Optimized Residual Distribution

Key Observations:

- Skewness: The distribution is slightly right-skewed.
- Outliers: There are a few outliers present, suggesting that the model might have struggled to accurately predict some data points.
- Clustering around 0: A significant cluster of predictions close to 0, indicating that the model's predictions are generally accurate for many data points.

Overall Assessment:

While the model shows reasonable performance, the presence of skewness and outliers suggests that there might be room for improvement. Further investigation into the data, feature engineering, or model selection could help address these issues and enhance the model's accuracy.



R-squared Results Based on Random Forest Optimization Techniques

1. Baseline Model

- Techniques: Log Transformation, Encoding (One-hot or Label), Cyclical Transformation, Interaction Terms
- R-squared: 0.61833

2. Feature Engineering

- Techniques: Log Transformation, Interaction Features, Seasonal Features (sin/cos)
- R-squared: 0.86283

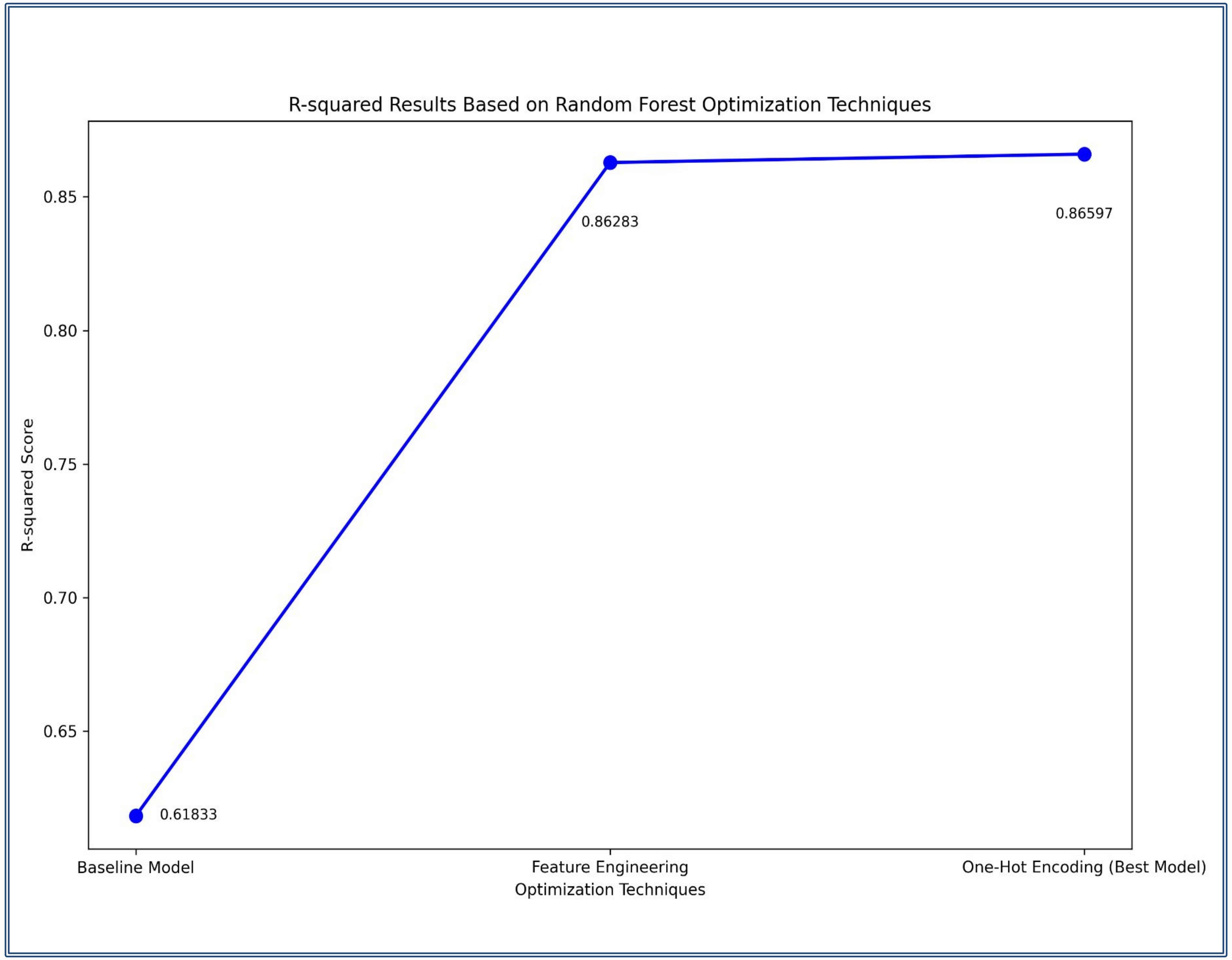
3. One-Hot Encoding

- Techniques: One-Hot Encoding for Genres and Awards
- R-squared: 0.86597

Visualization Insight:

The plot reveals a clear improvement in model performance as advanced feature engineering and encoding techniques are applied, achieving the highest R-squared score with the Best Model.

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



Questions

Thank You!

Applause !!