PREDICTING FILM PROFITABILITY USING PRE-RELEASE DATA: A COMPARISON OF REGRESSION MODELS

ABSTRACT

The film industry involves substantial financial risk, with 15–20% of movies failing at the box office despite average production budgets of $100–150 million. This project aims to predict a film's return on investment (ROI) using only pre-release attributes, such as budget, runtime, and cast size. ROI is used instead of raw revenue to better reflect profitability and normalize for production scale and inflation. To address the skewness of ROI, a log transformation is applied. Three statistical learning models, multiple linear regression, random forest, and XGBoost, are trained to predict log(ROI) and evaluated using RMSE, MAE, and R². Of these models, the tuned XGBoost achieved the best performance with an RMSE of 0.7391 and an R² of 0.4227. However, all models showed a tendancy to under predict films with very high ROI. Budget and popularity were found to be the most important predictors of ROI. These findings highlight both the potential and the limitations of data-driven modeling for pre-release film investment decisions.

604760

INTRODUCTION

The global film industry generates billions of dollars annually, with over $40 billion in box office revenue reported in 2019 (Statista). However, this revenue comes with substantial financial risk. An estimated 15–20% of films fail at the box office (Stephen Follows), and with average production budgets ranging from $100 to $150 million (Nashville Film Institute), a single movie flop can lead to significant financial losses. As a result, forecasting a film's financial performance before release has become an important area of applied statistical modeling.

This study investigates whether return on investment (ROI), profit divided by budget, can be accurately predicted using only information available before a film's release. The primary objective is to evaluate how well different regression-based models predict ROI and to determine which pre-release attributes are most strongly associated with film profitability.

The contribution of this work lies in its focus on ROI, a normalized measure that accounts for production costs and better reflects a film's financial outcome than raw revenue. Additionally, this study compares the performance of traditional and modern regression methods, highlighting both their predictive power and limitations in this dataset.

The remainder of the poster summarizes the data preparation process, the modeling techniques used, key results including model evaluation metrics, and important graphs. Conclusions and suggestions for future work are also provided.

METHODS

STUDY DESIGN AND DATA DESCRIPTION

This study uses an observational dataset, found on Kaggle, of feature films, compiled from publicly available movie data sources. The dataset includes information about individual films released primarily between the early 2000s and late 2010s.

- Units of analysis: Individual films
- Sample size: Approximately 4,800
- Pre-release variables used: 6 pre-release features (e.g., budget, runtime, cast size) and calculated ROI

Preprocessing Steps:

- Films with missing or zero values for budget or revenue were removed.
- A new variable, Return on Investment (ROI), was defined as:

$$ROI=(Budget-Revenue)/Budget$$

Due to the highly skewed distribution of ROI, a log transformation was applied:

$$log\_ROI=\log(1+ROI)$$

- Additional features were engineered:
- num_cast: number of listed cast members
- num_production_companies: number of production companies
- has_tagline: binary indicator for presence of a tagline
- All processing and modeling were conducted in R using the tidyverse, randomForest, xgboost, and caret packages.

STATISTICAL MODELS AND ESTIMATION

Three models were used to predict log(ROI):

(a) **Multiple Linear Regression (MLR)**

A traditional linear model assuming additive and linear effects:

$$\log(ROI)=\beta 0+\beta 1budget+\beta 2popularity+\ldots+\beta kxk+\varepsilon$$

Estimation was performed using Ordinary Least Squares (OLS) via lm().

Assumes linearity, independence, and homoscedasticity.

(b) **Random Forest**

An ensemble method based on decision trees, which reduces variance through bootstrapping and random feature selection. Fit using randomForest() with 500 trees.

(c) **XGBoost (Tuned)**

A gradient-boosted tree model trained with squared error loss, implemented via xgboost() with the following hyperparameters:

- nrounds = 300, eta = 0.05, max_depth = 4
- subsample = 0.8, colsample_bytree = 0.8

This configuration was chosen to minimize overfitting while maintaining strong predictive performance.

MODEL EVALUATION

The dataset was split into training (80%) and test (20%) subsets using createDataPartition() from the caret package.

Models were evaluated on the test set using:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- R² (coefficient of determination)

Visual diagnostics included:

- Predicted vs. actual plots (on both log and ROI scales)
- Residual plots to assess bias, underfitting, or overfitting

FIGURE 1

Model	RMSE	MAE	R²
Multiple Linear Regression	0.9094	0.6729	0.1260
Random Forest	0.7409	0.5564	0.4198
XGBoost	0.7391	0.5639	0.4227

FIGURE 2

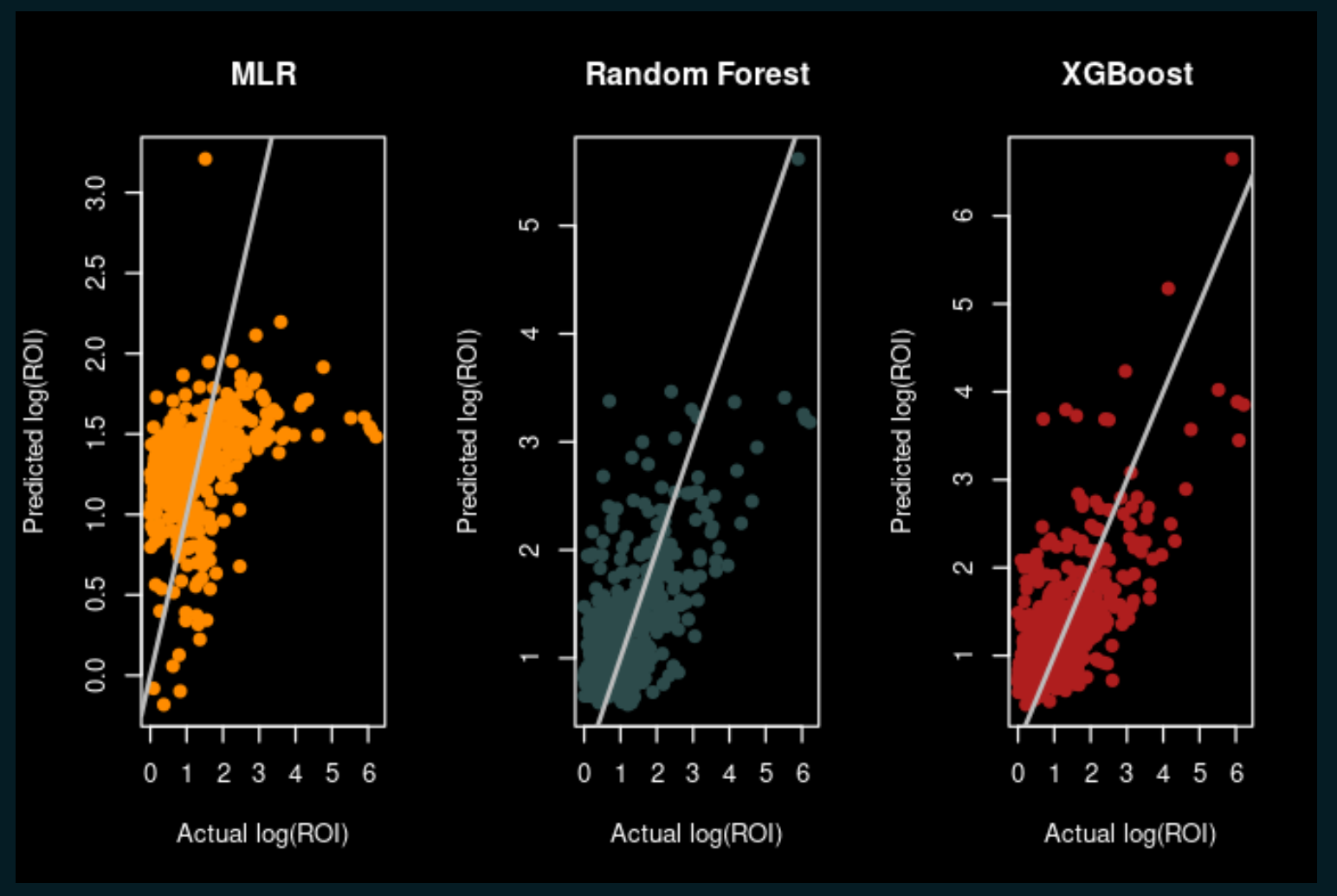


FIGURE 3

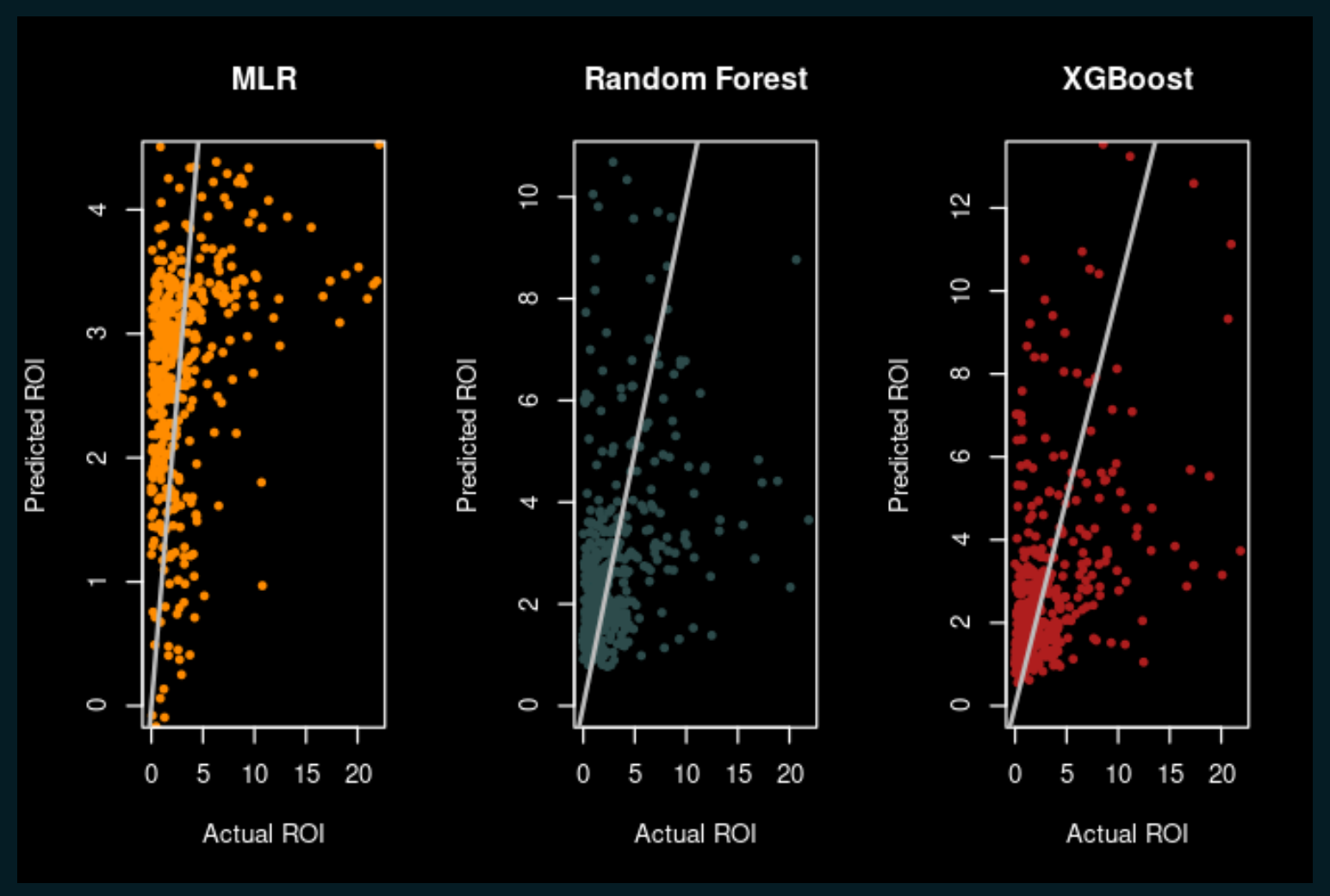


FIGURE 4

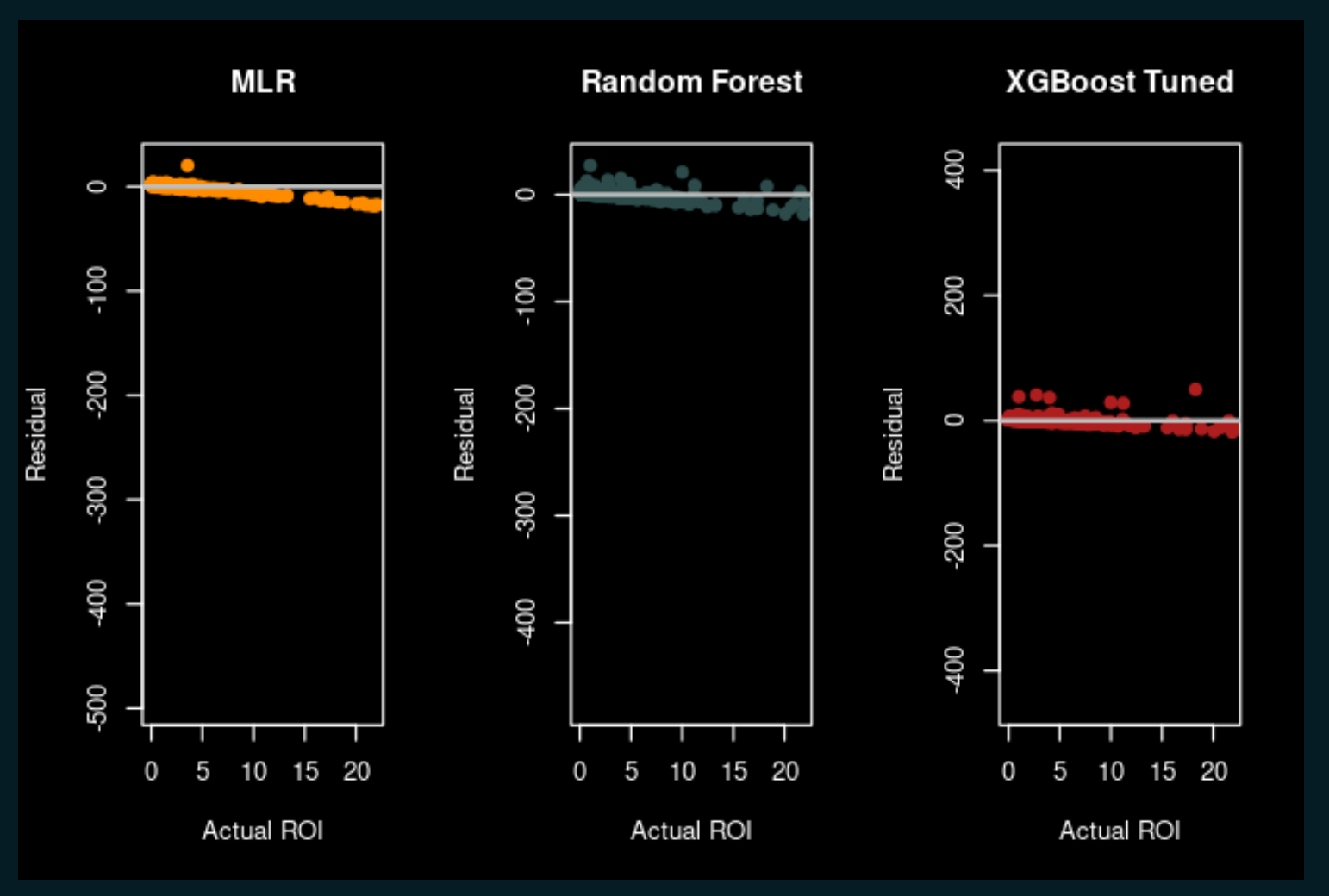
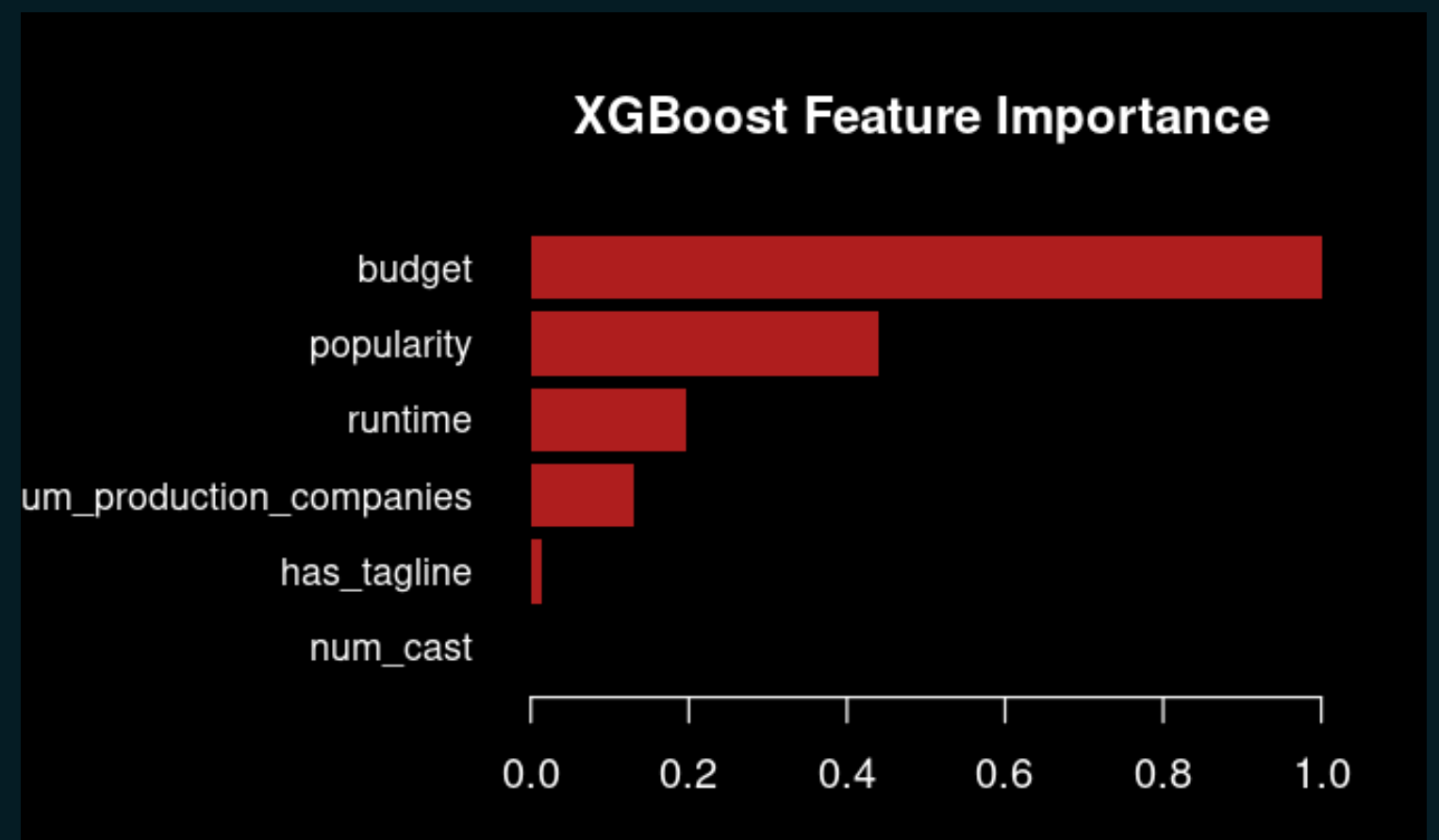


FIGURE 5



RESULTS

Figure 1 compares the performance of the three regression models—Multiple Linear Regression (MLR), Random Forest (RF), and Tuned XGBoost—using RMSE, MAE, and R² on the log-transformed ROI. Tuned XGBoost achieved the best balance of low error and high explained variance, with the lowest RMSE (0.7391) and highest R² (0.4227). Random Forest performed similarly, while MLR had the weakest predictive accuracy overall.

Figure 2 shows predicted vs. actual log(ROI) for all three models. In these plots, closer grouping along the reference line means more accurate predictions. XGBoost and RF showed tighter clustering along the line, especially in the mid-range, while MLR struggled to capture non-linear patterns and often underfit the data.

Figure 3 shows the same comparison but on the original ROI scale, allowing for real-world interpretation. Once back-transformed, the models consistently underpredicted extreme hits with very high ROI. Nonetheless, XGBoost again showed the strongest alignment with actual values, suggesting better real-world applicability despite the skewed outcome distribution.

Figure 4 shows residual plots (errors vs. actual ROI). Ideally, residuals should scatter randomly around zero, with no clear patterns. XGBoost and RF showed better balance with tighter residual lines. MLR residuals showed more structure, particularly a tendency to overpredict low-ROI films and underpredict profitable ones, indicating bias and limited flexibility in the model.

Figure 5 displays the feature/variable importance from the XGBoost model. The most influential predictor by far was budget, followed by popularity and runtime. These three variables contributed over 90% of the model's predictive gain, demonstrating a strong relationship with financial outcomes. Other features, such as tagline presence, number of production companies, and cast size, contributed little, suggesting they add minimal predictive value.

CONCLUSION

This study demonstrates that movie ROI can be predicted with moderate accuracy using pre-release features. Among the models tested, XGBoost outperformed both Random Forest and Multiple Linear Regression. For production companies, these findings offer actionable insights. Budget and popularity emerged as the most important predictors of ROI. While budget is a constraint in many cases, it can be adjusted to avoid over-investing in projects with weak pre-release indicators. Similarly, increasing a film's pre-release popularity through marketing, social media campaigns, or trailers, may significantly boost financial returns. By identifying which features matter most and modeling their impact, studios can use tools like XGBoost to screen projects, forecast profitability, and optimize investment strategies before production even begins. This predictive approach can help reduce financial risk, especially in an industry where a single flop can cost tens of millions. Ultimately, data-driven forecasting can support more confident, evidence-based decisions in a high-stakes industry where small changes in early planning may yield substantial differences in outcome.