Predicting Movie Revenue using ML

Ava Bender

FLORIDA ATLANTIC UNIVERSITY

INTRODUCTION

Problem:

- Predicting movie revenues before release is a complex but important problem in the entertainment industry. The financial success of a movie heavily influences production, marketing, and distribution strategies.

Motivation:

- The motivation for this project comes from the challenging conditions of the film industry, where large investments are often at risk making it important for accurate revenue forecasting to minimize financial risks.

PROPOSED SOLUTION

Machine Learning Models:

- Advanced ensemble methods, such as XGBoost and Gradient Boosting, were used to predict movie revenues. These models are well suited for handling nonlinear relationships and complex datasets.
- Simpler Models like Linear Regression and Random Forest were used as baselines for comparisons.

Data Preprocessing:

- Log transformations were applied to normalize skewed features like budget and revenue.
- Temporal features (release year, month, weekday) were extracted from release dates.
- Genres were one-hot encoded to handle multi category data.

Figure 1 illustrates the step-by-step process followed to build and evaluate the movie revenue prediction system.

DATA FOR EVALUATION

Dataset Characteristics:

- Started with 1 million + rows, reduced to around 14,500 after cleaning.
- Expanded columns from 28 to 54 using feature engineering.

Figure 2 illustrates the relative importance of different features in predicting movie revenue.

Top Features:

- Budget, Vote Count, Popularity, Runtime.

Moderate Features:

- Release Year, IMDB Votes, Vote Average, Release Day and Month.

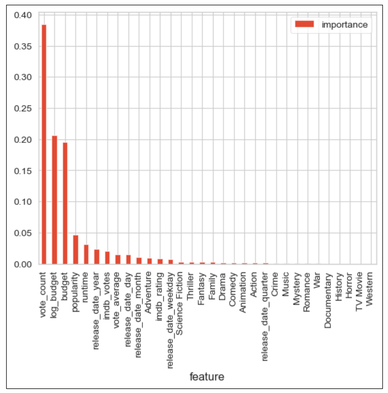

Figure 2. Feature Importance

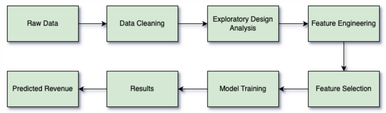

Figure 1. System Framework Flowchart

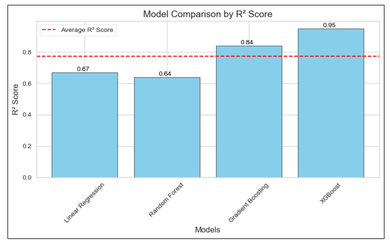

Figure 3. Model Comparison R-Square Score

Model	*R-Square Score*	*MAE*	*RMSE*
Linear Regression	0.67	$34.19M	$81.10M
Random Forest	0.64	$26.73M	$73.87M
Gradient Boosting	0.84	$23.84M	$55.55M
XGBoost	0.95	$12.98M	$32.11M

Table 1. Model Performance

Movie Title	*Predicted Revenue*	*Actual Revenue*	*Prediction Error*	*Relative Prediction Error*
How to Train Your Dragon	$ 491,014,656.00	$494,879,471.00	$ 3,864,815.00	0.78 %
Violent Night	$77,883,713.55	$75,734,910.00	$2,148,803.55	2.84%
Inception	$852,675,637.72	$825,532,764.00	$27,142,873.72	3.29%
Avatar	$2,627,122,975.37	$2,923,706,026.00	$296,583,050.63	10.14%

Table 2. Predicted Vs. Actual Revenue Selected Movies

RESULTS

Performance Metrics:

- As seen in Table 1, XGBoost achieved the best results with R-Square = 0.95, Mean Absolute Error (MAE), = 12.98M, and Root Mean Squared Error (RMSE) = $32.11M.
- Gradient Boosting also performed well with R-Square = 0.84 and MAE = $23.84M.
- The comparison of the R-Square scores amongst all models can be better visualized by looking at Figure 3.

Table 2 compares predicted and actual revenues for randomly selected movies, showing the accuracy of the XGBoost model. The relative prediction errors for movies like Inception (3.29%), Violent Night (2.84%), and How to Train Your Dragon (0.78%) highlights the models predictive capabilities. Whereas Avatar (10.14%) emphasize areas for improvement. The overall results validate the models effectives.

CONCLUSION

Key Findings: XGBoost achieved the best accuracy out of the models, with an R-Square score of 0.95, demonstrating the effectiveness of advanced ensemble methods.

Insights: Emphasized the importance of data preparation and using advanced models for complex datasets.

Summary: This project successfully developed a machine learning approach to predict movie revenues with relatively high accuracy. By addressing challenges like skewed data distributions, complex relationships between features, and the need for precise feature selection, the process ensures reliable and effective predictions. The use of advanced models like XGBoost and Gradient Boosting, combined with thorough data preprocessing and feature engineering, improves overall performance.