Predicting Movie Revenue Using Machine Learning

Ava Bender
Data Science Capstone
Florida Atlantic University
Boca Raton, Florida
Abender2022@fau.edu

I. INTRODUCTION

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. As highlighted in the article *Why Movies Succeed and Fail*, factors such as timing, audience preferences, and marketing representation play significant roles in determining box office outcomes. Even well-made films can underperform due to poor timing or mismarketing [1]. This report focuses on developing a machine learning based approach to predict movie revenues using various data factors like budget, popularity, genres, and release date.

The motivation for this project comes from the challenging conditions of the film industry, where large investments are often at risk. As addressed by *MovieMaker Magazine*, box office success depends on unpredictable factors such as star power, timing, audience trends, and genre appeal, making revenue forecasting an essential challenge [2]. Accurate predictions can help reduce losses and guide informed decision making. Existing methods, including regression, and clustering models, have attempted to address this issue. Research has shown that techniques like Linear Regression and Support Vector Machines can capture relationships between features and revenue, though their accuracy is limited by data complexity [3][4].

This project proposes using ensemble models like Gradient Boosting and XGBoost, which have shown superior performance in handling non-linear relationships and large datasets [5][6]. This procedure involves preprocessing skewed data, feature selection, and model tuning to achieve high predictive accuracy. This approach aims to address data challenges and outperform traditional models in predicting movie revenue.

II. MAIN BODY

The primary goal of this project is to accurately predict movie revenues using machine learning techniques. Traditional methods like simple regression models often fail to capture the complex relationships between variables such as budget, popularity, and audience engagement. The motivation for this project focuses in addressing these limitations

by using advanced ensemble methods like XGBoost and Gradient Boosting. These methods excel in handling nonlinear relationships, skewed data distributions, and high dimensional datasets. Important challenges addressed in the design include skewed data, feature selection, and enhancing model performance through advanced algorithms and preprocessing techniques.

Exploratory Data Analysis (EDA) is important in understanding the dataset and preparing it for modeling. EDA revealed skewed distributions in budget and revenue, which were normalized using log transformations. Correlation analysis identified strong relationships between revenue and features like budget, vote count, and IMDB votes, which guided feature selection. Visualizations like heatmaps, scatter plots, and box plots helped identify trends, relationships, and areas requiring preprocessing, such as missing values and outliers.

Building off the insights from EDA, the framework for this project consisted of three main stages: data preprocessing, feature selection, and model training/evaluation. Data preprocessing involved several important steps to ensure high quality input data for the models. Rows with missing or zero values in critical columns like budget or revenue were removed. Skewed distributions in budget and revenue were normalized using log transformations, making the data more suitable for machine learning models. Categorical values, such as genres were encoded into numeric formats using one-hot encoding. Additionally, the release dates column was in a string format which was not directly usable, so it was processed into more meaningful components, creating separate columns for year, month, weekday in order to capture temporal patterns. Finally, outlier removal further refined the dataset by eliminating extreme values that could distort predictions.

Feature selection focused on using insights from EDA to identify the most important predictors of movie revenue. Feature selection focused on identifying predictors like vote count, budget, and popularity using Random Forest and XGBoost models. Irrelevant features were removed to reduce noise and improve accuracy. Model training tested algorithms including Linear Regression, Random Forest, Gradient Boosting, and XGBoost. The models were evaluated using R-Square score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE Hyperparameter tuning using Grid Search further enhanced performance.

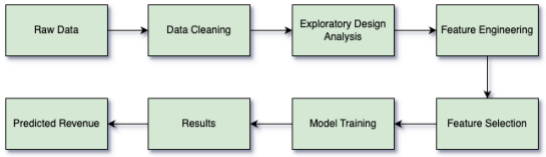

Fig. 1. System framework flowchart

The system architecture as shown in the flowchart, outlines the process from raw data preprocessing to revenue prediction. Starting with data cleaning and feature engineering informed by EDA. The refined dataset was used to train the models, which predicts movie revenues based on the most relevant features. Ensemble methods ensure better accuracy, addressing challenges posed by simpler models.

This project ultimately aims to demonstrate that ensemble methods outperform simpler models, emphasize the important role of EDA and preprocessing, and provide practical insights into the factors driving movie revenue. These objectives ensure the findings are reliable and useful for stakeholders in the film industry.

III. EXPERIMENTS

The primary goal of the experimental studies was to evaluate the effectiveness of machine learning models in predicting movie revenue based on a cleaned and feature engineered dataset. This involved comparing the performance of several models, including baseline methods, to determine which approach provided the most accurate and reliable predictions. The experiments also aimed to validate the hypothesis that ensemble methods like XGBoost outperform simpler models in handling complex relationships in movie data.

The experiments were conducted using Jupyter Notebook. Python libraires such as Scitkit-learn, 3XGBoost, and LightGBM were used to design and implement the models. Data preprocessing and visualization were performed using Pandas, NumPy, Matplotlib, and Seaborn. Linear Regression and Random Forest were selected as baseline methods to provide a benchmark for performance comparisons. Gradient Boosting and XGBoost served as advanced models due to their ability to minimize errors continually and capture nonlinear relationships. Hyperparameter tuning for the XGBoost model was performed using Grid Search, optimizing parameters like learning rate, maximum depth, and the number of estimators to enhance accuracy.

The dataset originally contained over 1 million rows and 28 columns but was reduced to around 14,500 rows after preprocessing. Through feature engineering, the number of columns increased from 28 to 54. This transformation involved extracting additional insights from existing data, such as processing release data as mentioned earlier. Genres which was originally a single column, were encoded into binary columns using one-hot encoding, allowing the model to recognize a movies association with multiple genres simultaneously. Key features included budget, popularity, vote count, genres, and temporal variable. The target variable, revenue, was normalized using a log transformation to address its skewed distribution. This preprocessing and feature engineering process ensured the dataset was clean and ready for model training and evaluation.

The results of the experiments are summarized in the table and figure below. The table compares key metrics, including R-Squared score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), for each model. XGBoost significantly outperformed all other models, achieving an R-Squared score of 0.95, MAE of $12.98M, and RMSE of $32.11M, showing its strong ability to capture complex relationships in the data. Gradient Boosting also performed well with an R-Squared score 0.84 and MAE of $23.84M, but it fell short of XGboosts accuracy.

TABLE I. MODEL PERFORMANCE

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

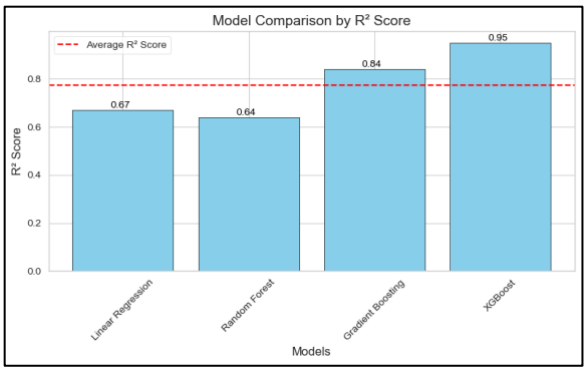

Figure 2. Model Comparison R-Square Score

Baseline models like Linear Regression and Random Forest struggled to achieve similar performance. Linear Regression for example had an R-Squared score of 0.67, MAE of $34.19M, and RMSE of $81.10M, highlighting its limitations and handling the nonlinear relationships present in the dataset. Random Forest performed slightly better than Linear Regression but was still significantly less accurate than XGBoost.

The accompanying bar chart provides a visual comparison of the R-Squared scores for all models emphasizing XGBoosts superior predictive capabilities. The chart also highlights the gap in performance between advanced ensemble methods and baseline models. These results validate the use of XGBoost as the best model for predicting movie revenues ensuring high accuracy and reliability.

To validate the model's performance, predictions were compared against actual revenues for random movies. The results demonstrate XGBoosts strong predictive capabilities. For instance, the model predicted a revenue of $491M for *How To Train Your Dragon*, with a relative error of 0.78%, and $852.68M for *Inception*, with a relative error of 3.29%. Similarly, the prediction for *Violent Night* closely matched its actual revenue with an error of 2.84% however for *Avatar* the model underestimated the revenue resulting in a relative error of 10.14%. These examples show the model's ability to generate accurate predictions for high revenue movies while highlighting areas for further refinement particularly for outliers like *Avatar*. The use of log transformed prediction significantly improved the model's performance by reducing skewness in the data.

TABLE II. PREDICTED VS. ACTUAL REVENUE

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%

IV. CONCLUSION

This project addressed the challenge of predicting movie revenue using machine learning. By preprocessing skewed data, engineering features, and applying advanced models like XGBoost, the project achieved high predictive accuracy with an R-Squared score of 0.95. The findings highlight the importance of effective preprocessing and ensemble methods, offering insights for future research on improving prediction models.

REFERENCES

[1] Chen, David. "Why Movies Succeed and Fail - the Life and Times of David Chen." *The Life and Times of David Chen*, 8 Mar. 2011, davechen.net/2011/03/why-movies-succeed-and-fail/.

[2] MM Writers. "Box Office Insights: Understanding Why Some Movies Make Money While Others Fail." *MovieMaker*, MovieMaker Magazine, 12 Aug. 2024, www.moviemaker.com/box-office-insights-understanding-why-some-movies-make-money-while-others-fail/.

[3] Pakom Walanaraya, et al. *Movie Revenue Prediction Using Regression and Clustering*. 5 July 2018, https://doi.org/10.1109/icei18.2018.8448610.

[4] Rachaell Nihalaani, et al. *Movie Success Prediction Using Naïve Bayes, Logistic Regression and Support Vector Machine*. 3 Sept. 2021, https://doi.org/10.1109/icrito51393.2021.9596138.

[5] Darapaneni, Narayana, et al. "Movie Success Prediction Using ML." *IEEE Xplore*, 1 Oct. 2020, ieeexplore.ieee.org/document/9298145.

[6] D Menaga, and Akshaya Lakshminarayanan. *A Method for Predicting Movie Box-Office Using Machine Learning*. 6 July 2023, https://doi.org/10.1109/icesc57686.2023.10192928.

[7] "Machine Learning Prediction - Javatpoint." *Www.javatpoint.com*, www.javatpoint.com/machine-learning-prediction.

[8] Teasdale, Aaron. "Film Industry Realities: Reviews and Box Office Impact." *Meer*, Meer.com, 3 Sept. 2024, www.meer.com/en/80240-film-industry-realities-reviews-and-box-office-impact.