An Analysis of Factors Influencing the Box Office of Movies and Optimization of the XGBoost Model for Prediction

Yiheng Fang[1], **Xinyuan Zhang**[2, *]

[1]School of Statistics, Jiangxi Normal University, Nanchang, Jiangxi, 330022, China
[2] School of Statistics, Changchun University of Technology, Changchun, Jilin,130012, China
*Corresponding author: 20223858@stu.ccut.edu.cn

Abstract:

Box office prediction is crucial for the film industry, but traditional models struggle to capture the non-linear relationships and interactions among factors such as budget and release timing. Recent advancements have introduced machine learning approaches, including Random Forest (RF) and Extreme Gradient Boosting (XGBoost). XGBoost the good performance due to its exceptional feature selection capability and nonlinear modeling. Based on the data of 1,313 films from 1995-2016, this study systematically compares three models – linear regression, RF, and XGBoost. By adding new features like director's film count and optimizing hyperparameters, the results show XGBoost achieves the best performance (Coefficient of Determination (R^2) =0.69, Root Mean Squared Error (RMSE)=0.78, Mean Absolute Percentage Error MAPE=3.12%, Mean Absolute Error (MAE)=0.55), significantly outperforming RF and linear regression. Feature importance analysis reveals budget and the total number of audience ratings (NAR) as the most important variables, while summer releases also significantly impact revenue. Therefore, high-budget productions, enhanced audience engagement, and strategic release time can maximize box office revenue. The study shows that XGBoost works well for predicting box office success and also provides useful data to help the film industry make better decisions.

Keywords: Box office prediction, Machine learning, XGBoost, Film industry analytics.

1. Introduction

The analysis of factors influencing movie box office performance has garnered significant attention in research. Various studies have identified a range of variables that contribute to a film's financial success, including budget, ratings, runtime, and release timing.

Mendez and Mendoza conducted a study that emphasizes the positive correlation between movie budgets and box office income. Their research utilized data from the International Movie Database (IMDb) covering popular films from 2019 to 2022. They found that while budget is a significant factor, it is not the only predictor of a movie‘s revenue success, suggesting that other variables also play crucial roles in determining box office performance [1]. Chen analyzed 240 blockbuster films released by major distributors from 2014 to 2023. His findings indicate a significant positive correlation between production budget and global box office performance. Additionally, Wasserman et al. found that user ratings were found to positively influence box office outcomes, reinforcing the idea that audience perception is critical [2].

Lu and Feng developed a comprehensive indicator system to analyze the factors affecting box office performance, utilizing multiple linear regression to construct a box office prediction model. Their research emphasizes the importance of a structured approach to understanding box office dynamics in the Chinese film industry [3]. Chen and Huang et al. both utilized advanced statistical techniques, including multiple linear regression and machine learning models such as Random Forests (RF) and eXtreme Gradient Boosting (XGBoost). Chen‘s research incorporated a range of features, including user ratings and production budgets, to assess their impact on box office performance, while Huang et al. focused on visualizing data and selecting relevant variables for their predictive models [4, 5]. Huang et al. found that the RF model provided a more accurate fit compared to traditional regression methods [5]. Additionally, Chang utilized both RF and XGBoost models, demonstrating that XGBoost outperformed RFs in terms of prediction accuracy, thus showcasing the effectiveness of advanced machine learning techniques in analyzing complex datasets related to film performance [6]. Ahmad et al. developed a mathematical model and Chi-Squared (X^2) to predict movie success based on various attributes, including budget, cast, and release timing. Their research underscores the complexity of predicting box office success, as it cannot be attributed to a single factor [7]. Lastly, Subramaniyaswamy et al. examined the effectiveness of multiple linear regression and support vector machine classification in predicting box office success. Their work highlights the need for advanced modeling techniques to accurately predict box office outcomes in the dynamic film industry [8].

In summary, the complexity of box office prediction lies in the nonlinear interactions among factors like budget, audience ratings, and release timing, which traditional linear regression models fail to capture effectively. Existing solutions have evolved from basic statistical methods to machine learning models, where RF partially handles nonlinear features. XGBoost, with its gradient boosting algorithm and feature importance evaluation, has been proven by Chen and Huang et al.to exhibit superior predictive performance [4,5]. This research tries to make a better and more practical model by adding new factors (like director experience) and adjusting XGBoost settings, while also creating a method for adding new things like social media data in the future.

2. Method

2.1 Data Source

The dataset used in this study is Movies and Directors: Dataset for Film Analytics, provided by user Elza on the public data platform Kaggle (2024) [9]. It contains box office information for 4,769 movies, covering key dimensions such as budget, director, release date, and ratings from 1995 to 2016.

2.2 Variable Selection and Description

The variable selection process is critical to building an effective forecasting model. The selected variables should be relevant and informative to the target variable (movie revenue in this study). The study selected several key features as independent variables for modeling movie box office revenue. These include: Budget (The production budget of the movie), Popularity (A platform-specific algorithm-generated popularity index), the average audience rating (AAR), the total number of audience ratings (NAR), Year (The release year of the movie), and NDF.

To ensure data quality before analysis, began with initial cleaning of the dataset to remove records with zero budget or income. To quantify directors‘ professional experience impact on box office performance, it introduced the Number of Directed Films (NDF) as a novel feature variable capturing both directorial proficiency and market appeal. For categorical variable treatment, we implemented one-hot encoding to convert temporal features (release month and day of week) into orthogonal binary dummy variables, with month-prefixed columns representing different months and day-prefixed columns indicating weekdays, thereby eliminating potential ordinal bias from numerical label encoding.

During the data preprocessing stage, these features were standardized. After applying the natural logarithm transformation to box office revenue, the aforementioned features were used to train the model. These variables collectively incorporate the film‘s production cost, popularity level, evaluation metrics, release timing, as well as the director‘s experience and activity level, with the aim of predicting a movie‘s box office performance.

2.3 Indicator Selection and Description

To comprehensively evaluate the predictive performance of the model, we selected four commonly used regression evaluation indicators: Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). R^2 offers an intuitive assessment of how well the model explains the data. RMSE emphasizes larger errors by squaring the residuals, making it sensitive to outliers. MAPE expresses prediction accuracy as a percentage, enabling scale-free error comparison. MAE provides a direct measure of the average size of prediction errors. The combination of these four metrics helps us more accurately select and optimize models.

2.4 Methodology

This study uses three regression models for the prediction of box office performance as follows. First, the Linear Regression Model captures a linear association between the features and the target variable. Second, RF Regressor is an ensemble learning approach that enhances model generalizability and accuracy by generating multiple decision trees. Each tree is built using randomly selected features and instances during training. Predictions can be determined via averaging or through the majority voting process. Last, XGBoost Regression. It is one of the gradient boosting-based learning techniques. Supports model improvement by iterative optimization. Has the strength of dealing with intricate data relationships and feature selection.

In the current work, to achieve better performance of the model, XGBoost was used with hyperparameter tuning by Grid Search to determine the best set of hyperparameters.

3. Results and Discussion

3.1 Results

The relationship between budget and box office is shown in Fig. 1. Overall, budget and revenue show a positive correlation trend. However, there is a phenomenon that some high-budget movies underperform at the box office, showing some volatility.

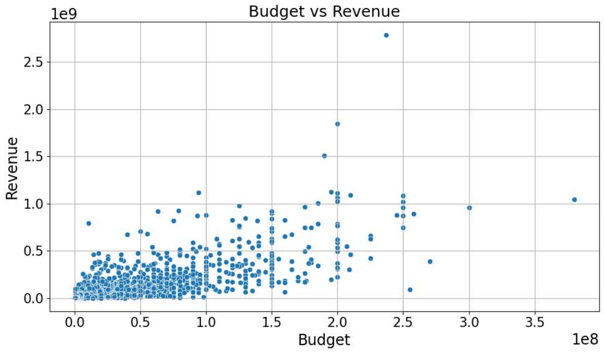

Fig. 1 Budget vs Revenue (Picture credit: Original)

The relationship between AAR and box office is shown in Fig. 2. Movies with higher average ratings (e.g., above 7.0) typically have higher revenues, though some high-rated outliers with low revenue exist, indicating that favorable ratings do not guarantee high revenue. Films with medium average ratings (e.g., 5.0-6.0) show a wider revenue distribution, suggesting greater variability in revenue within this rating range.

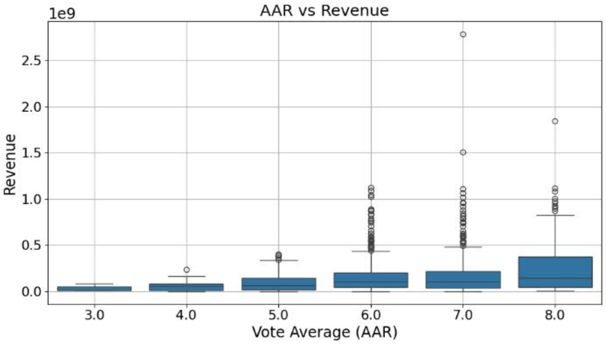

Fig. 2 Vote Average vs Revenue (Picture credit: Original)

The relationship between Year and box office is shown in the Fig. 3. The Fig. 3 is a line chart illustrating the trend of movie revenues from 1995 to 2016. Overall, movie revenues show an upward trend. Despite this general upward movement, there are substantial annual fluctuations in revenue. These variations may be attributed to factors such as big films release in specific years, economic conditions, and changes in audience preferences.

ISSN 2959-6157

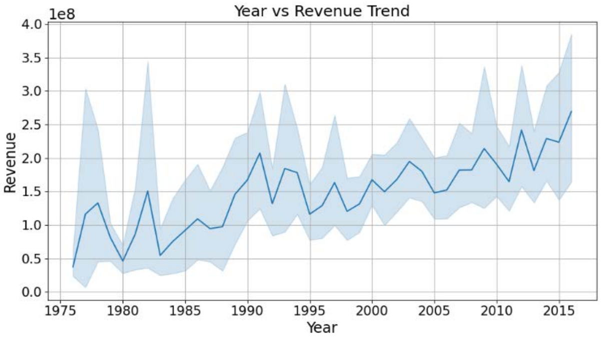

Fig. 3 Year vs Revenue (Picture credit: Original)

3.2 Methodology

3.2.1 Linear Regression

This text initially employed a linear regression model to predict movie box office revenue. Linear regression is a widely-used statistical method particularly suitable for predicting continuous target variables.

The comparison between the actual and predicted values of the linear regression model shows that the predicted values are relatively scattered compared to the actual values, particularly in the higher prediction range, where several outliers exist. While there is an overall increasing trend where predicted values rise with actual values, the linear regression model demonstrates limited prediction accuracy, especially for higher actual values, where predictions deviate substantially from reality.

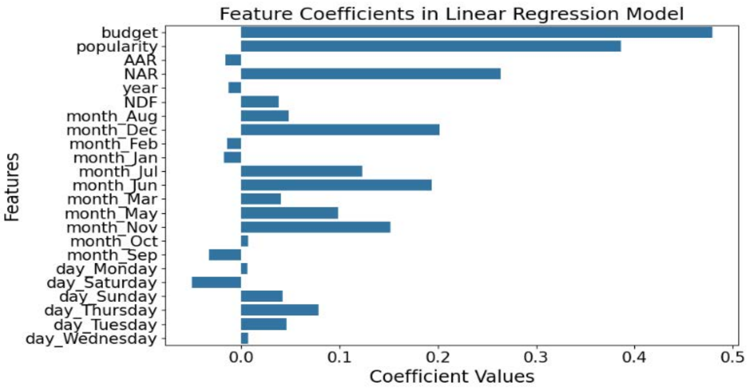

Fig. 4 Linear Regression Feature Coefficients (Picture credit: Original)

Fig. 4 shows the bar chart of feature coefficients from the linear regression model. Budget has the largest coefficient, indicating it's the most important feature for predicting box office revenue. Popularity ranks second, showing that a movie's popularity significantly affects its revenue.

3.2.2 RF Model

Next, this text selected the RF model for further investigation (Fig. 5). The predicted values from the RF model are close to the actual values, with most points distributed near the perfect fit line. The RF model demonstrates better performance in capturing nonlinear relationships in the data, with predictions being significantly closer to actual values, indicating better model fit.

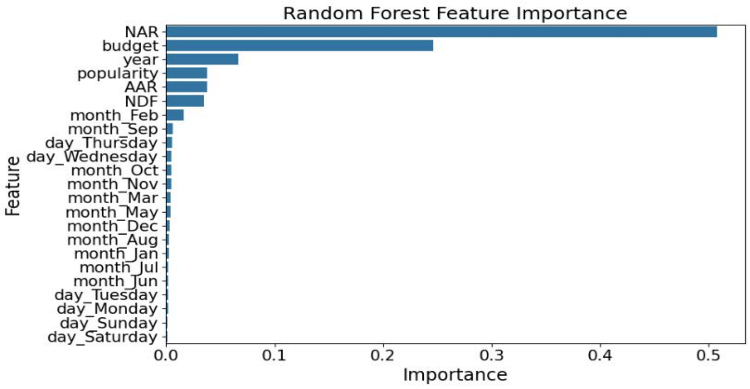

Fig. 5 RF Feature Importance (Picture credit: Original)

3.2.3 XGBoost

Finally, use the XGBoost regression model for box office analysis. The predicted and actual values of the XGBoost model are most tightly clustered, with most points closely distributed near the perfect fit line. Among the three models, XGBoost demonstrates the best fit, with predictions highly consistent with actual values, particularly maintaining strong predictive accuracy even for higher actual values.

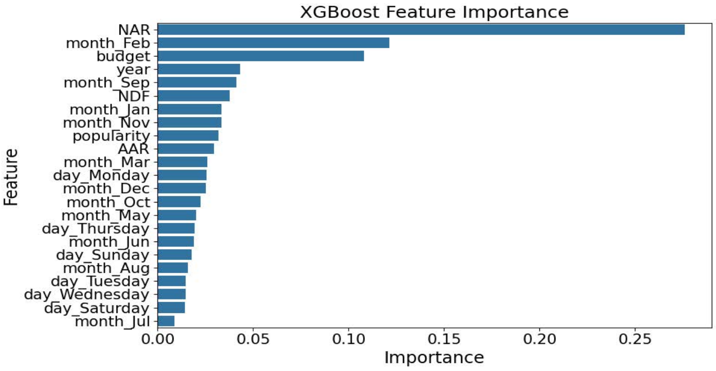

Fig. 6 XGBoost Feature Importance (Picture credit: Original)

Fig. 6 reveals that NAR and budget remain the most important features, consistent with the RF model's findings. Other features such as year, $month_{Jun}$, and popularity also demonstrate measurable importance, but their contributions are relatively lower.

3.3 Comparison and Analysis of Three Predictive Models

The essay previously applied Linear Regression, RF Regression, and tuned XGBoost Regression models to predict box office revenue, evaluating each model's performance using R^2, RMSE, MAPE, and MAE. Table 1 is the comparative performance.

Table 1. Comparison of Three Predictive Models (Table credit: Original)

Model	R^2	RMSE	MAPE	MAE	
Linear Regression	0.29	1.18	4.55%	0.80	
RF	0.65	0.83	3.35%	0.59	
XGBoost	0.63	0.85	3.47%	0.61	
Optimized XGBoost	0.69	0.78	3.12%	0.55	

Linear regression demonstrates the weakest performance among the tested models, with an R^2 of 0.29, implying that it accounts for merely 29% of the variance in box office revenue. This model's substantial prediction errors are evident in its high RMSE (1.18) and MAE (0.80), which likely stem from its failure to capture nonlinear relationships in the data.

In contrast, Random Forest (RF) substantially improves predictive accuracy, attaining an R^2 of 0.65 while reducing RMSE to 0.83. Its MAPE of 3.35% further indicates acceptable relative error magnitudes.

The baseline XGBoost model initially underperforms RF slightly, yielding an R^2 of 0.63 and RMSE of 0.85, though its competitive MAPE (3.47%) suggests untuned potential. After hyperparameter optimization, XGBoost emerges as the top-performing model, achieving the highest explanatory power (R^2=0.69) and minimal errors (RMSE=0.78, MAE=0.55). Its post-tuning MAPE of 3.12% underscores its reliability in practical applications.

3.4 Discussion

The exploratory data analysis (EDA) revealed critical points influencing box office revenue. A non-linear correlation exists, where higher budgets generally lead to higher revenues, but exceptions (e.g., high-budget failures) highlight the role of other factors like audience reception. Films with ratings above 7.0 tend to achieve higher revenues, though outliers (e.g., highly rated but low-revenue films) suggest that ratings alone are insufficient predictors. Revenue increased from 1995 to 2016, but annual fluctuations emphasize the impact of market dynamics and competition. These findings consistent with prior studies (e.g., Wasserman et al., 2015; Mendez & Mendoza, 2023), confirming that budget and audience engagement are primary factors, while release time affects revenues [1, 2].

The results of this study indicate that the XGBoost model performs the best in predicting movie box office revenue. XGBoost achieved the highest R^2 (0.69) and the lowest errors (RMSE=0.78, MAE=0.55), better than RF (R^2 =0.65, RMSE=0.83, MAE=0.59) and Linear Regression (R^2 =0.29, RMSE=1.18, MAE=0.80). It can effectively capture non-linear relationships (e.g., diminishing returns on budget) and feature interactions (e.g., popularity × release timing), which simpler models failed to model. Linear Regression is easy to understand but inadequate for non-linear data. And RF has good balanced accuracy, but is less optimized than XGBoost. For the feature importance, budget and NAR are consistently top predictors across all models, confirming their dominance in revenue. Releasing time emerged as a secondary factor, suggesting seasonal trends (e.g., summer releases) impact revenue.

This study also has some limitations and shortcomings. The dataset used in this paper has a range of years between 1995 and 2016, with a long-time lag between the data and the present and a lack of relevant information on recent trends. The absence of consideration of qualitative factors that may affect the outcome of the projections (e.g., type of movie, cast, marketing investment).

Future research could further explore the impact of additional features, such as cast composition and promotional efforts on box office revenue. With the continuous evolution of the movie market and the acceleration of technological changes, the relationship between casting and box office revenue is undergoing a profound restructuring [10]. And attempt to use advanced models, like deep learning models, to further improve predictive accuracy. It integrates multidimensional data sources to form a more complete system of influencing factors [11].

4. Conclusion

This study systematically evaluated the performance of three regression models—Linear Regression, RF, and XGBoost—in predicting movie box office revenue using a comprehensive dataset spanning 1995 to 2016.

XGBoost outperforms competing models. Achieved the highest R^2 (0.69) and the lowest errors (RMSE=0.78, MAE=0.55), demonstrating good capability in capturing non-linear relationships and feature interactions (e.g., popularity × release timing).

Budget and NAR were consistently the top predictors across all models, emphasizing the critical roles of production investment and audience engagement in revenue. Time-related features (e.g., release year, summer months) and director experience (NDF) emerged as secondary but significant factors.

For filmmakers and studios, it is better to prioritize high-budget productions, audience engagement strategies (e.g., boosting NAR), and strategic release timing (e.g., summer months) could maximize revenue potential. Choosing experienced directors also helps to increase revenue. XGBoost's accuracy makes it a reliable tool for forecasting and risk assessment in the film industry.

Future research could further explore the impact of additional features, such as cast composition and promotional efforts, on box office revenue, and attempt to use advanced models, like deep learning models, to further improve predictive accuracy. Such approaches could incorporate multidimensional data sources, enabling a more comprehensive framework for analyzing box office success.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Fiorella Mendez,Luciana Mendoza. Does the budget of a movie have an impact on income? Fort Hays States University, SACAD, 2024, 103.

[2] Max Wasserman, Satyam Mukherjee, Konner Scott, et al. Correlations between user voting data, budget, and box office for films in the internet movie database. Journal of the Association for Information Science and Technology, 2015, (66): 858–868.

[3] Lu Wenjing, Feng Xiao. Confirmation analysis of the influence factors of movie box office in the era of big data. Journal of Communication University of China Natural Science Edition, 2017, (24):41-46.

[4] Joon Wei Chen. The influence factors of global movie box office and the correlation with the stock prices of movie and TV companies. Graduate Institute of Finance and Economics, School of Management, National Taiwan University, 2024.

[5] Huang Wenqing, Lai Jiajia, Ning Qiongmin, et al. Analysis and prediction of influencing factors of movie box office. Research and Discussion on Information Technology and Informatization, 2019.

[6] Chang Yaxin. Statistical analysis of influencing factors of domestic movie box office. Star River Film & TV, 2024, (6), 211-213.

[7] Javaria Ahmad, Prakash Duraisamy, Amr Yousef, et al. Movie success prediction using data mining. university of central missouri, University of Business and Technology, University of North Texas (Eds.),2023.

[8] Subramaniyaswamy V., Viginesh Vaibhav M., Vishnu Prasad R., et al. Predicting movie box office success using

multiple regression and SVM. Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2017.

[9] Kaggle user Elza, Movies and directors: dataset for film analytics (Kaggle,2024), published November 2024, cited May 2025 https://creativecommons.org/publicdomain/zero/1.0/.

[10] Beijing Normal University New Media Communication Research Center. Research report on social responsibility of Chinese film and TV stars. Beijing: Beijing Normal University Press, 2019,23-25.

[11] Chen Bangli, Xu Meiping. Research on movie gross box office prediction model based on LARS-SVR. Journal of Shaanxi Normal University (Natural Science Edition), 2018, 46(1): 10-15.