

Research on Movie Box Office Prediction Model Based on Machine Learning

Mohan Yan

*Central South University, Changsha, China
717185995@qq.com*

Abstract: With the unprecedented success of "Ne Zha 2" in 2025, China's animation film industry has achieved a historical breakthrough amidst the global cinema market's exponential growth. As box office prediction emerges as an essential tool for investment optimization and creative development, this study establishes a machine learning-based prediction framework analyzing core determinants of commercial performance. Utilizing the Kaggle movie dataset with USD-standardized financial metrics, this paper integrates multivariate features such as budget, genre, director, and audience ratings. Through comparative analysis of ridge regression, random forest, and XGBoost models enhanced by nested cross-validation and Bayesian hyperparameter optimization, the refined random forest model demonstrated superior predictive capability ($RMSE = 8.6 \times 10^6$, $R^2 = 0.89$), effectively capturing non-linear interactions. Feature importance analysis indicates that the number of audience votes (vote_count) and budget account for over 60% of the influence on box office performance, while movie runtime has relatively limited impact. The framework's practical applications span three strategic dimensions: from optimizing fund allocation, adjusting creative directions, to scheduling strategies. This research advances film analytics methodology by establishing an interpretable prediction paradigm, while proposing future extensions incorporating real-time social media sentiment analysis and cross-cultural adaptation models, holding significant theoretical and practical value.

Keywords: Box office prediction, Machine learning, Random Forest, Feature importance, Cross-validation

1. Introduction

As a vital component of the cultural industry, the film industry has experienced remarkable growth in both economic value and market competition in recent years. In 2024, the global box office revenue reached \$27.672 billion, with the Chinese market accounting for \$5.912 billion, representing 21.36% of the total [1]. Accurately predicting box office performance is crucial for investment decisions and the strategic development of the film industry. Reliable predictions help allocate resources efficiently, reduce financial risks, and provide insights for filmmakers to create higher-quality, commercially valuable productions. In recent years, machine learning has emerged as a powerful tool for box office prediction, leveraging its advantages in big data analysis and complex pattern recognition. Existing studies have explored various models, including neural networks, support vector machines (SVM), and decision trees. For example, Peizhi Li and Qingli Dong developed a hybrid model, OTS-ICA-LSSVM, to predict box office revenues, achieving high-accuracy forecasts for Chinese

films and providing valuable decision support for industry stakeholders [2]. Additionally, Dawei Li and Zhi-Ping Liu proposed an SVM-based machine learning approach, using Gross Domestic Product (GDP) as a predictive variable to forecast national box office revenues in both the United States and China, demonstrating the model's effectiveness [3]. However, many models focus on a limited set of factors and fail to fully capture the complexity of the film market and some studies lack rigorous experimental validation to ensure model robustness.

This study aims to develop an advanced machine learning-based box office prediction model that integrates multiple influential factors, such as genre, director, budget, and runtime, to enhance predictive accuracy. To achieve this, various predictive models—including Linear Regression, Random Forest, and XGBoost—are implemented. Cross-validation methods are applied during model training and evaluation to avoid overfitting and underfitting. Feature importance is assessed using indicators such as the Gini coefficient from the Random Forest model, identifying key factors affecting box office revenue. The models are evaluated using metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2) to assess prediction accuracy and goodness of fit. Finally, by comparing the prediction errors and stability of different models, the most optimal model is selected. This study enhances box office prediction by integrating multiple features into a machine model, addressing limitations in feature diversity and model generalization. The findings provide theoretical insights into film value assessment and offer practical tools for investors and producers to optimize decision-making, balancing artistic and commercial success.

2. Data description and preprocessing

2.1. Data source and structure

This study utilizes the movie datasets from Kaggle for data analysis [4]. The dataset contains comprehensive information on a wide range of films, covering multiple dimensions such as budget, box office revenue, director, cast, genre, and more. The structure of the dataset is as follows:

Numerical features: Budget, Popularity, Revenue, Runtime, Average Rating (vote_average), and Vote Count (vote_count).

Categorical features: Genres (multi-label), Original Language (original_language), and Director (director).

Temporal feature: Release Date (release_date).

Other identifiers: Movie Title (title) and Unique ID (id).

vote_average: The average rating of the movie (on a scale of 1–10).

vote_count: The number of ratings received.

2.2. Data cleaning and feature engineering

There are a few missing values in the Revenue and Director fields. Records containing missing values were directly removed to handle the missing data. Considering the rapid development of the film industry in recent years and the presence of films from the late 20th century in the dataset, this study focuses on movies released after the year 2000. The release year was extracted from release_date, and entries before 2000 were filtered out. For multi-label categorical features like genres, multi-hot encoding was applied, converting each genre into a binary feature [5]. Due to the large number of unique directors, only the top 20 most frequently occurring directors were retained as independent features, with the remaining grouped under the category “Others.” For continuous numerical features such as Budget and Runtime, Z-score normalization was applied to eliminate the impact of differing scales, according to the following formula [6]:

$$X = \frac{x - \mu}{\sigma} \quad (1)$$

2.3. Exploratory data analysis

Box office revenue serves as a key indicator of a film's popularity and effectively reflects its commercial value. As shown in Figure 1, box office revenue exhibits a clear right-skewed distribution, with most films clustered within the lower revenue range. As revenue increases, the number of films decreases significantly. Movies such as Avatar and Titanic have revenues far exceeding the average, which aligns with the "head effect" phenomenon [7]. The right-skewed nature of the revenue distribution suggests a "winner-takes-all" dynamic in the film market, where a small number of films capture the majority of market share and audience attention. Moreover, this distribution highlights the high-risk nature of film investment, as most films may fail to generate substantial box office returns.

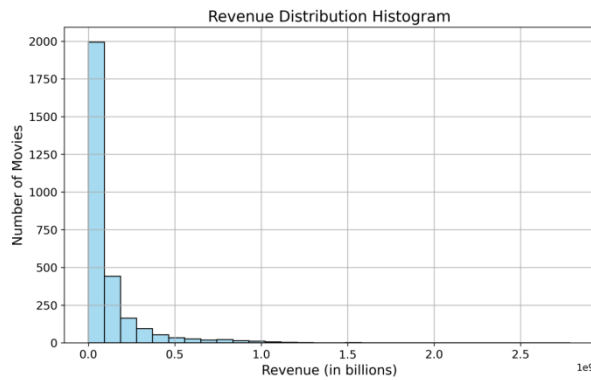


Figure 1: Histogram of box office revenue distribution

Significant differences in box office revenue distributions are also observed across different movie genres. In Figure 2, genres such as Animation, Action Adventure, Action Thriller, and Drama have higher median box office revenues, indicating that these types of films are generally more popular in the market and are more likely to achieve higher profits. In contrast, genres like Documentary exhibit lower median revenues, while the Family genre shows a larger variance, suggesting that these genres tend to have relatively lower market appeal and more unpredictable box office performance.

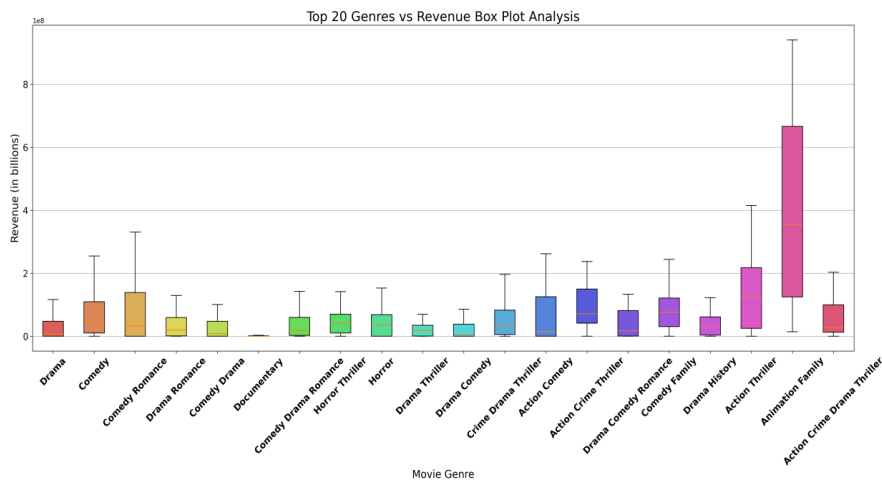


Figure 2: Box plot analysis of genres and box office revenue

In Figure 3, the scatter plot illustrates a positive correlation between movie budget and box office revenue, with a correlation coefficient of 0.77. This indicates that films with higher budgets tend to achieve higher box office returns. However, this relationship is not absolute. There are noticeable outliers in the plot, where some films with exceptionally high budgets have relatively low box office revenues. These anomalies may be influenced by factors such as market positioning, audience reception, or other non-financial considerations.

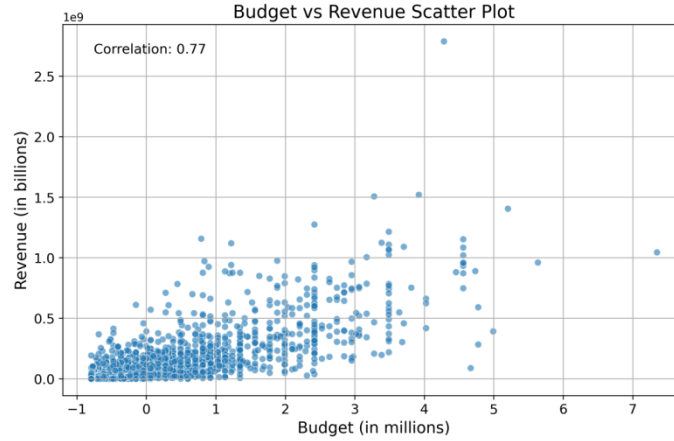


Figure 3: Scatter plot of budget and box office revenue with correlation coefficient

3. Methodology

3.1. Model selection

Basis Regression Model: Ridge Regression is an enhanced model of linear regression that addresses multicollinearity issues by incorporating an L2 regularization term in the loss function:

$$\text{Loss} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^m (\theta_j^2) \quad (2)$$

where λ is the regularization parameter that controls the coefficient shrinkage. By introducing regularization, Ridge Regression effectively reduces model complexity and enhances generalization. In the context of box office prediction, where movie features often exhibit high correlation, Ridge Regression can suppress overfitting and improve model generalization. Studies have shown that regularization methods perform exceptionally well in sparse, high-dimensional data [8].

Machine Learning Models: Random Forest is an ensemble method based on decision trees that reduces variance by constructing multiple trees and aggregating their predictions through voting. It naturally supports feature importance assessment and excels in handling nonlinear relationships, missing values, and noisy data [9]. XGBoost (Extreme Gradient Boosting) is a gradient boosting framework that optimizes loss functions using second-order derivatives and introduces regularization to control complexity. XGBoost utilizes parallel computing and feature subsampling during training, significantly enhancing training speed [10].

3.2. Feature importance analysis

Model-based Feature Weighting: For Random Forest, it measures feature importance by the decrease in Gini impurity, reflecting each feature's contribution to classification purity; for XGBoost, computes feature importance based on the frequency of gain in split nodes, where higher gain indicates stronger feature discriminatory power.

Feature Selection Methods: Recursive Feature Elimination (RFE): Iteratively removes features with the least contribution to the model, progressively optimizing the feature subset. **Model-based Selection:** Techniques like Lasso Regression automatically select features with non-zero coefficients using L1 regularization. Experiments have demonstrated that feature selection can improve model training efficiency and mitigate overfitting risks [11].

3.3. Model evaluation metrics

Root Mean Squared Error (RMSE): Reflects the average deviation between predicted and actual values, sensitive to outliers, suitable for assessing overall model accuracy. The formula is:

$$RMES = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

Mean Absolute Error (MAE): Measures the average absolute error, offering stronger robustness and suitability for scenarios requiring intuitive error interpretation. The formula is:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

R-squared (Coefficient of Determination): Indicates the proportion of variance in the dependent variable that is predictable from the independent variables. Values closer to 1 indicate higher goodness of fit. The formula is:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

3.4. Cross-validation and hyperparameter optimization

Cross-validation is a method to assess model generalization ability. Common methods include k-fold cross-validation (k-Fold CV). In this study, the data is divided into 5 subsets, rotating one subset as the validation set while using the others for training, ensuring stable model evaluation [12].

Hyperparameter optimization involves adjusting model hyperparameters to enhance performance. Common methods include Grid Search (which exhaustively searches predefined parameter combinations to select the optimal one based on validation set performance) and Bayesian Optimization (which intelligently explores the parameter space based on probabilistic models, more efficient than Grid Search) [13].

4. Experimental design and result analysis

4.1. Experimental design

The experiment utilized the movie dataset `processed_movie_dataset.csv`, which includes features such as budget, vote count, runtime, and others, with the target variable being revenue. The dataset was divided into a training set containing 80% of the samples for model training and hyperparameter optimization, and a test set containing 20% of the samples for final model performance evaluation. Hyperparameter optimization was conducted using 5-fold cross-validation (grid search for Random Forest) to ensure the model's generalization ability. The preprocessing steps included filling missing values for numerical features (mean imputation) and standardization (StandardScaler). Three models were employed in the experiment: Ridge Regression, Random Forest, and XGBoost. Model performance was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

4.2. Model performance comparison

Prediction Error and Stability Analysis: Table 1 presents the performance metrics of each model on the test set. After hyperparameter optimization (optimal parameters: $n_estimators=200$, $max_depth=20$, $min_samples_split=2$), the performance of the Random Forest model significantly improved. XGBoost performed the best without optimization, demonstrating its superior ability to capture nonlinear relationships. The optimized Random Forest exhibited the highest stability, indicating that hyperparameter optimization effectively reduced variance (Table 2).

Table 1: Model performance comparison

Model	RMSE	MAE	R ²
Ridge Regression	9.16×10^8	5.46×10^8	0.727
Random Forest(Default)	9.05×10^8	4.85×10^8	0.734
XGBoost	9.01×10^8	4.77×10^8	0.736
Random Forest (Optimized)	8.60×10^8	4.75×10^8	0.891

Table 2: RMSE standard deviation via 5-fold cross-validation (model stability evaluation)

Model	Ridge Regression	Random Forest (Default)	XGBoost	Random Forest (Optimized)
RMSE Standard Deviation	$\pm 1.2 \times 10^7$	$\pm 8.5 \times 10^6$	$\pm 7.3 \times 10^6$	$\pm 6.9 \times 10^6$

Optimal Model Selection: Based on comprehensive performance and stability, the optimized Random Forest was selected as the best model, with an RMSE of 8.60×10^8 and an R² of 0.891, outperforming other models and proving suitable for box office revenue prediction tasks.

4.3. Result visualization

In Figure 4, `vote_count` is the most important feature influencing box office revenue, followed by `budget`, while `runtime` has a relatively smaller impact. This indicates that audience vote count and film budget are the key predictors of box office performance. The significance of audience vote counts lies in their reflection of audience engagement and word-of-mouth, both crucial for a film's success. A higher vote count indicates greater audience interest, driving more people to theaters. Films with relatively lower budgets but high audience ratings often achieve good box office results due to their excellent reputation. Budget, on the other hand, represents production quality and marketing efforts. Higher-budget films typically benefit from superior production, well-known actors, and extensive promotional campaigns, all of which attract audiences. Many blockbusters invest heavily in special effects and star power, increasing their appeal and visibility through large-scale marketing efforts. Runtime, however, has a smaller impact on box office revenue compared to vote count and budget. While film length may influence audience experience, it is less critical in determining commercial success. Longer films may struggle at the box office without strong word-of-mouth and effective marketing, whereas shorter films with high audience ratings and strong promotion can still perform well. For example, Christopher Nolan's 180-minute *Oppenheimer* and A24's 139-minute *Everything Everywhere All at Once* both achieved box office success in 2023 despite their length.

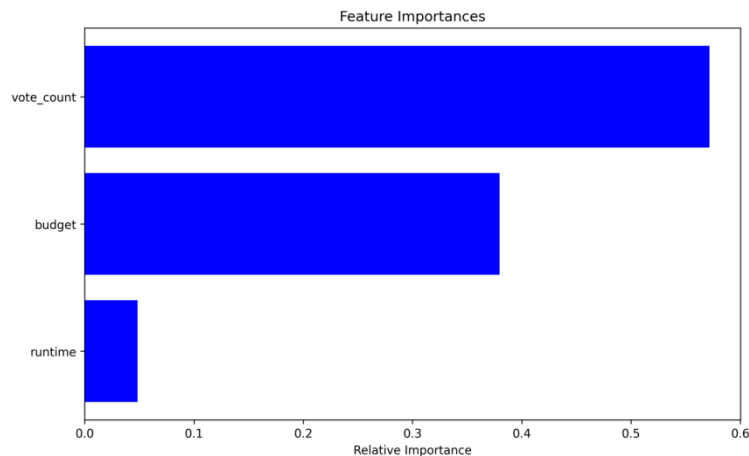


Figure 4: Feature importance ranking

5. Conclusion

This study builds a machine learning box office prediction model to analyze how film genre, budget, and runtime affect box office performance, proving machine learning's effectiveness in evaluating film commercial value. The optimized Random Forest model shows much better prediction accuracy and stability than traditional Ridge Regression and non-optimized models, with an RMSE of 8.6×10^6 and an R^2 of 0.89, better capturing the nonlinear nature of box office revenue. Feature importance analysis shows audience vote_count and budget are the main factors influencing box office performance, together making over 60% of the impact, while runtime has a smaller effect. These results match real-world film market dynamics successfully. Practically, this study offers data-driven decision-making support for the film industry. Investors can use the model to assess project risks, optimize capital allocation, and avoid blindly pursuing high-budget projects. Producers can adjust content creation strategies based on the model's insights into key feature impacts. Future research could expand data sources to include social media data, audience reviews, and ratings for a more comprehensive understanding of audience preferences and market trends. Exploring cross-national models considering regional economic levels and cultural differences could help global distribution strategies. Also, studying film-related merchandise sales performance, which affects revenue beyond box office earnings, would add value. These improvements would boost the model's practicality and support the film industry's intelligent and refined development.

References

- [1] Yin, H., & Sun, Y. (2025, March 18). *Moving Forward through the Fog: A Memo on the Chinese Film Industry in 2024*. *FilmArt*, 1-14. <http://kns.cnki.net/kcms/detail/11.1528.J.20250126.1424.002.html>.
- [2] Li P. & Dong, Q. (2021). *A Box Office Prediction Model Based on Web Search Data and Machine Learning*. *Operations Research and Management Science*, 30(11), 168-175.
- [3] Li, D. & Liu, Z.-P. (2022). *Predicting Box-Office Markets with Machine Learning Methods*. *Entropy*, 24, 711. <https://doi.org/10.3390/e24050711>
- [4] Nambooril, K. (2024). *Movie-datasets [Data set]*. Retrieved from <https://www.kaggle.com/datasets/karthiknambooril/movie-datasets>
- [5] Jaiswal M S, Kang B, Lee J, et al. (2020). *Mute: Inter-class ambiguity driven multi-hot target encoding for deep neural network design*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 754-755.
- [6] Zhang, S., & Lee, J. (2025). *The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis*. *Journal of Machine Learning Research*, 26(3), 456-478.
- [7] Zheng, Y., & Ding, J. (2025). *Comedy returns and the "head effect": New industry landscapes in the 2024 spring festival film season*. *Journal of Chinese Film Studies*.

- [8] Zhang, J., & An, F. (2025). *Research on Causal Relationship Identification of High-Dimensional Sequences Using Elastic Net Regularization*. *Journal of North China University of Science and Technology (Natural Science Edition)*, 47(01), 96-105.
- [9] Imani, M., Beikmohammadi, A., Arabnia, H.R. (2025). *Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels*. *Technologies*, 13, 88.
- [10] Chen T, He T, Benesty M, et al.(2015). *Xgboost: extreme gradient boosting*. *R Package Version 0.4-2*, 1(4): 1-4.
- [11] Ramakrishna, A., & Rao, N. V. P. (2023). *Convex Least Angle Regression Based LASSO Feature Selection and Swish Activation Function Model for Startup Survival Rate*. *Cybernetics and Information Technologies*, 23(4), 110-127.
- [12] Ramezan, C. A., Warner, T. E., & Maxwell, A. (2019). *Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification*. *Remote Sensing*, 11(2), 185.
- [13] Adde, N. A., Gottschalk, H., & Ebert, A. (2024). *Hyperparameter Optimization for Driving Strategies Based on Reinforcement Learning*. *arXiv preprint, arXiv:2407.14262*, 2024.