Copyleaks

Scan details

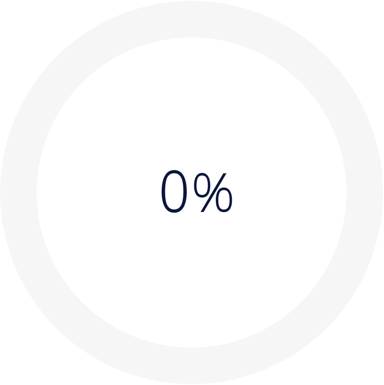
Scan Time	Total Pages	Total Words
August 30th, 2025 at 10:22 UTC	11	2664

Analysis Report

Plagiarism Detection and AI Detection Report

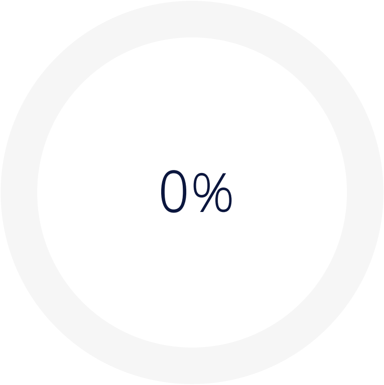
DA_RP (1).pdf

Plagiarism Detection



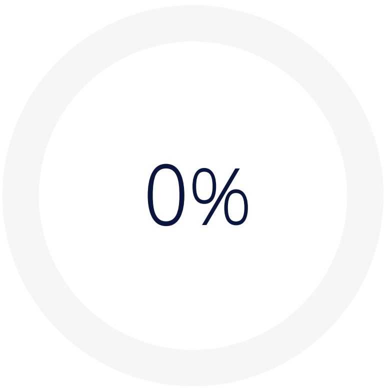
Plagiarism Types	Text Coverage	Words
● Identical	0%	0
● Minor Changes	0%	0
● Paraphrased	0%	0
Excluded		
○ Omitted Words		340

AI Detection



	Text Coverage	Words
○ AI Text	0%	0
○ Human Text	100%	2,324
Excluded		
○ Omitted Words		340

Plagiarism



Results (0)

*Results may not appear because the feature has been disabled.

Private Cloud Hub 0	**Shared Data Hub** 0	**Filtered / Excluded** 0
Internet Sources 0	**AI Source Match** 0	**Current Batch** 0

Plagiarism Types	Text Coverage	Words
● Identical	0%	0
● Minor Changes	0%	0
● Paraphrased	0%	0
Excluded		
◌ Omitted Words		340

About Plagiarism Detection

Our AI-powered plagiarism scans offer three layers of text similarity detection: Identical, Minor Changes, and Paraphrased. Based on your scan settings we also provide insight on how much of the text you are not scanning for plagiarism (Omitted words).

● Identical

One to one exact word matches. Learn more

● Minor Changes

Words that hold nearly the same meaning but have a change to their form (e.g. "large" becomes "largely"). Learn more

● Paraphrased

Different words that hold the same meaning that replace the original content (e.g. 'large' becomes 'big') Learn more

◌ Omitted Words

The portion of text that is not being scanned for plagiarism based on the scan settings. (e.g. the 'Ignore quotations' setting is enabled and the document is 20% quotations making the omitted words percentage 20%) Learn more

Copyleaks Shared Data Hub

Our Shared Data Hub is a collection of millions of user-submitted documents that you can utilize as a scan resource and choose whether or not you would like to submit the file you are scanning into the Shared Data Hub. Learn more

Filtered and Excluded Results

The report will generate a complete list of results. There is always the option to exclude specific results that are not relevant. Note, by unchecking certain results, the similarity percentage may change. Learn more

Current Batch Results

These are the results displayed from the collection, or batch, of files uploaded for a scan at the same time. Learn more

Box Office Revenue Prediction: An Ensemble Learning Approach with Feature Engineering and Comprehensive Evaluation

Jash Ladhani[1,2*], Manav Bhuta[2,3†], Dhairya Shah[1,2†], Aaryan Lunis[1,2†]

[1*]Computer Engineering, NMIMS, Vile Parle, Mumbai, 400056, Maharashtra, India.
[2]Computer Engineering, NMIMS, Vile Parle, Mumbai, 400056, Maharashtra, India.
[3]Computer Engineering, NMIMS, Vile Parle, Mumbai, 400056, Maharashtra, India.
[4]Computer Engineering, NMIMS, Vile Parle, Mumbai, 400056, Maharashtra, India.

*Corresponding author(s). E-mail(s): jashladhani07@gmail.com;
Contributing authors: bhutamanav@gmail.com;
dhairyamihir2005@gmail.com; lunisaaryan@gmail.com;
†These authors contributed equally to this work.

Abstract

This paper presents a group learning method to predict movie box office revenue by using extensive feature engineering and strong model evaluation techniques. Employing a wide ranging dataset of 4,880 movies, we extracted and engineered key aspects such as vote counts, budget metrics, and runtime, which strongly correlate with revenue. Multiple models, including Linear Regression, Random Forest, Support Vector Regression, and Gradient Boosting, were trained and evaluated via cross-validation, with Gradient Boosting achieving the highest prediction accuracy, explaining 76.4% of the variance in box office revenue. Statistical significance testing validated the superiority of the Gradient Boosting model over alternatives. This research shows that ensemble methods, along with well-designed features, can accurately predict box office performance. This offers important insights for producers, distributors, and marketers in the film industry.The methods and results build on earlier studies in machine learning-based movie revenue prediction [1–4].

Keywords: Movie box office prediction, Machine learning models, Ensemble methods, Gradient Boosting algorithm, Feature importance analysis

1 Introduction

Globally, the film industry is flourishing. It's critical for marketers, distributors, and producers to forecast box office performance. Budget decisions, release schedules, advertising campaigns, and distribution strategies are all aided by early estimates. However, a number of factors make it challenging to forecast box office revenue.[5, 6]. Actors, genre, budget, audience preferences, and market conditions are some of these variables. Better modelling approaches that can handle intricate relationships and interactions among numerous variables have been produced by recent developments in machine learning. [7, 8].

Traditional linear regression models frequently fall short of capturing this complexity, despite their seeming simplicity. Because they provide higher predictive accuracy in a variety of domains, techniques like Random Forests and Gradient Boosting have grown in popularity.[9], they often struggle with this complexity. Consequently, ensemble learning approaches such as Random Forests and Gradient Boosting have gained popularity, demonstrating improved predictive accuracy across numerous domains [2, 10]. These techniques have been successfully applied in various box office prediction studies, leveraging diverse datasets and methodologies [11, 12].

To tackle these challenges, this study introduces a framework that combines feature engineering, data analysis, and thorough model evaluation with cross-validation. We analyze a dataset of 4,880 films, test various machine learning algorithms, and pinpoint key predictive indicators. Our results reveal that Gradient Boosting consistently performs better than other models, supporting previous findings on the effectiveness of ensemble methods and well-designed features in forecasting revenue.[1, 13, 14].

The contributions of this paper are threefold: (1) developing a detailed feature set capturing timing, budget, and audience engagement; (2) conducting an extensive comparison and tuning of machine learning models using robust cross-validation; and (3) statistically validating the performance of the best-performing model. Together, these efforts establish a solid framework for practical and accurate box office forecasting.

2 Exploratory Data Analysis Results

Exploratory data analysis was performed on 4,880 movies to inform feature selection and model development. The dataset shows a wide range of budgets and revenues, with an average budget of $33.8 million and median revenue of $33.5 million, reflecting a skewed revenue distribution dominated by a few high earners.

Drama, Comedy, and Action are the most common genres, ensuring diverse representation. Temporal trends reveal more releases in the 2000s and 2010s, reflecting industry growth. Correlation analysis highlighted strong relationships between revenue and features such as vote count (0.773) and budget (0.734), confirming their predictive importance.

These insights validated the dataset's quality and helped focus on key features for modeling, supporting the development of effective box office prediction models [1, 7, 8].

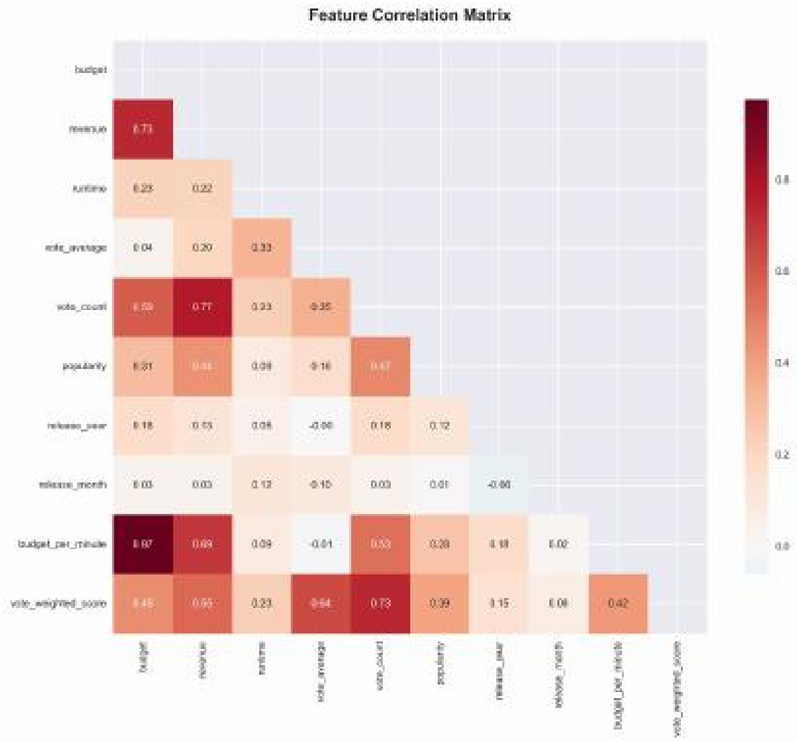


Fig. 1: Feature distribution and importance analysis conducted during exploratory data analysis. This visualization highlights key predictors such as vote count, budget metrics, and runtime that exhibit strong correlations with box office revenue.

3 Model Performance Comparison

Multiple machine learning models were trained and evaluated to predict movie box office revenue. These models included Linear Regression, Ridge Regression, Lasso Regression, Random Forest, Gradient Boosting, and Support Vector Regression (SVR). We split the dataset into training (3,904 movies) and test (976 movies) sets. Careful cross-validation helped fine-tune the hyperparameters.

Among the models tested, Gradient Boosting was the best performer. It achieved a test R2 of 0.7639, which means it explains over 76% of the variance in box office revenue. It also had the lowest test RMSE at around $93.7 million and a MAE of about $45.1 million. Random Forest, both tuned and untuned, showed competitive performance with test R2 scores around 0.74 and RMSE close to $98 million.

Linear models, including Linear, Ridge, and Lasso Regression, performed similarly. They reached a test R2 near 0.726 and had higher RMSE values of around $100 million.

The feature importance analysis showed that vote count and budget-related features were key predictors across the ensemble models. These results demonstrate how well ensemble learning methods can capture complex interactions in predicting box office revenue. A summary of model performance metrics is provided in Table 1.

Table 1: Model Performance Metrics

Model	Test R^2	Test RMSE	Test MAE	CV RMSE Mean
Gradient Boosting	0.7639	$93,678,947	$45,133,838	$85,595,256
Random Forest (Tuned)	0.7412	$98,070,891	$45,070,750	$80,906,132
Gradient Boosting (Tuned)	0.7365	$98,955,651	$44,881,441	$83,264,262
Random Forest	0.7363	$99,003,586	$45,810,673	$80,944,907
Ridge Regression	0.7262	$100,874,773	$51,358,293	$87,389,058
Lasso Regression	0.7262	$100,884,057	$51,375,764	$87,391,113
Linear Regression	0.7262	$100,884,058	$51,375,764	$87,391,113

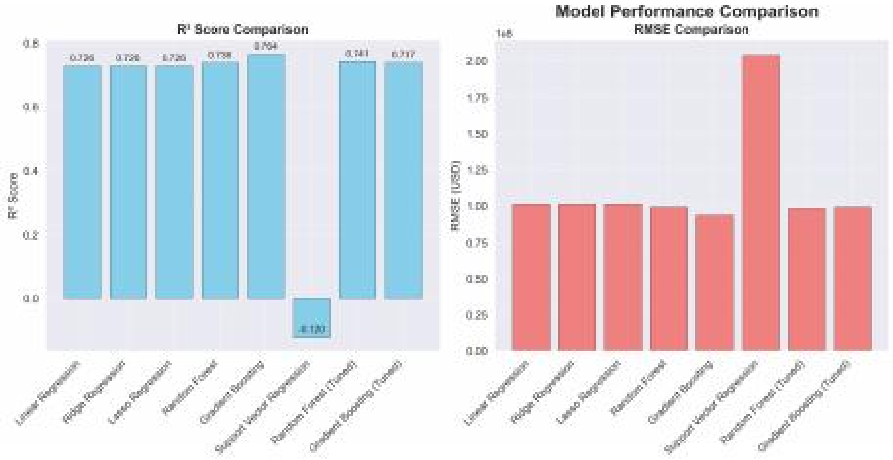


Fig. 2: Comparison of model performance based on R^2 score and RMSE across different algorithms. Gradient Boosting achieves the highest R^2 and lowest RMSE, highlighting its superior accuracy for predicting box office revenue.

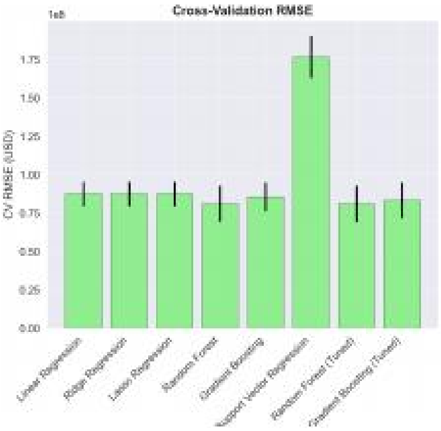


Fig. 3: Mean Absolute Error (MAE) comparison among all models.

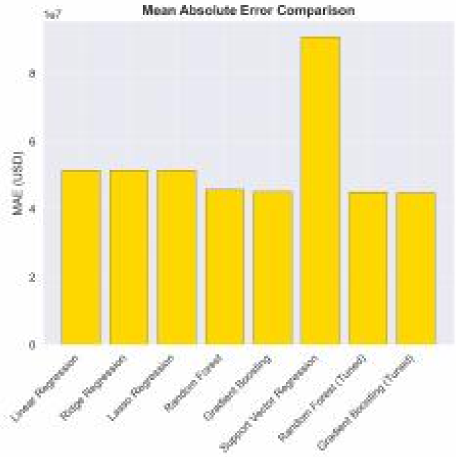


Fig. 4: Cross-validation RMSE for different models.Gradient Boosting and Random Forest variants exhibit robust error control.

These findings confirm the superiority of Gradient Boosting for box office revenue prediction, aligning with similar studies emphasizing ensemble methods' accuracy and robustness [1–3].

4 Figures

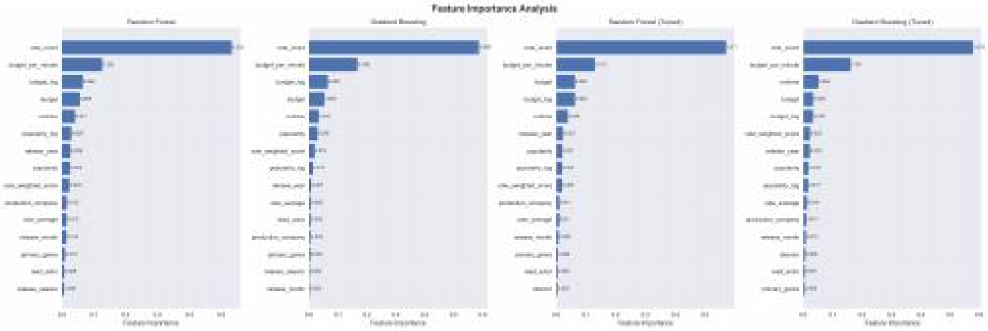


Fig. 5: Feature importance analysis for box office revenue prediction models. The plots show that vote count and budget-related features are consistently the most influential predictors across Random Forest and Gradient Boosting models.

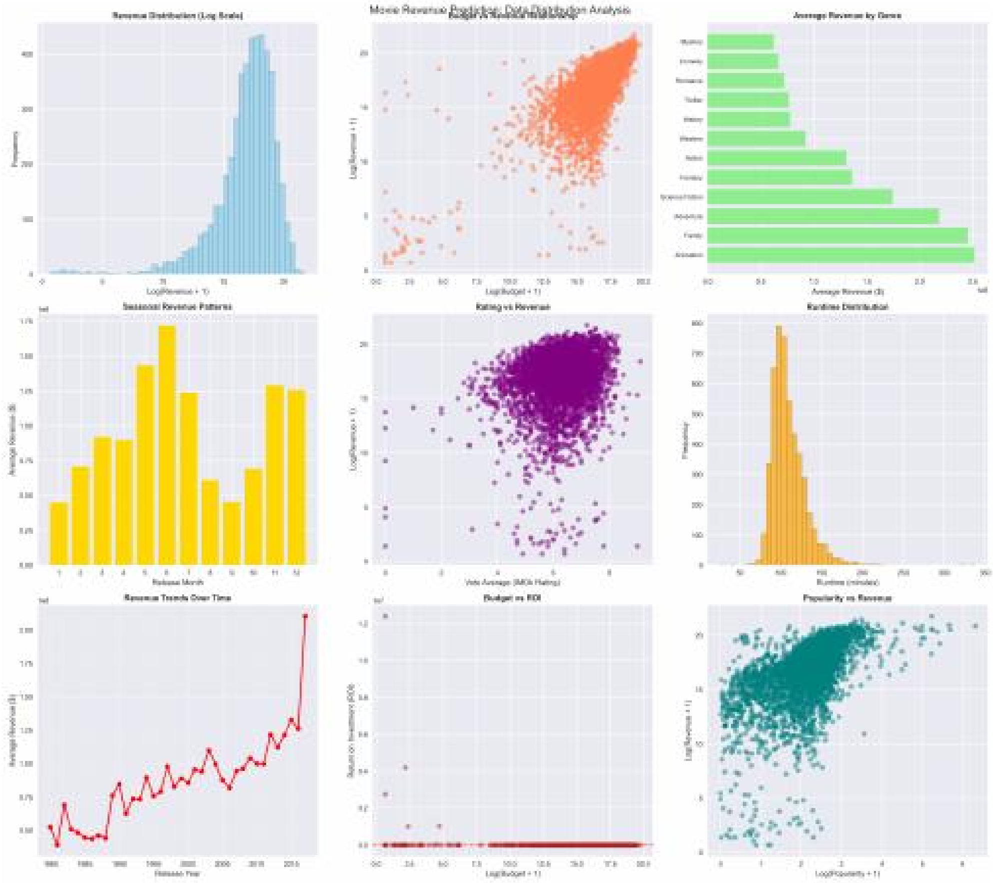


Fig. 6: Exploratory data analysis visualizations, including revenue distributions, feature relationships, genre comparisons, and temporal patterns. These plots reveal key patterns and correlations that informed model feature selection and engineering.

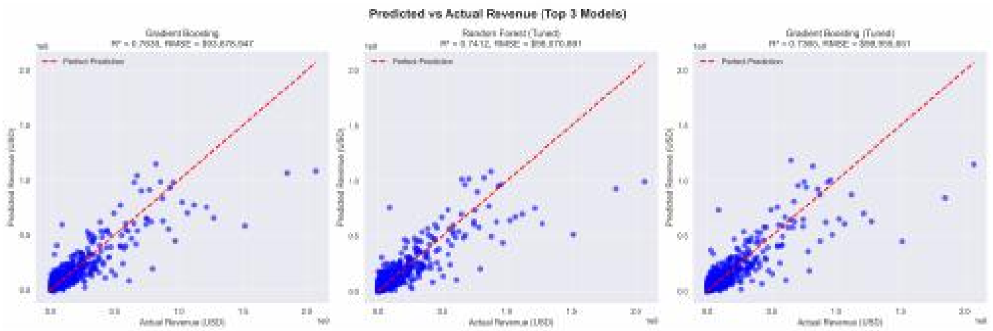


Fig. 7: Comparison of predicted versus actual box office revenue for the top three models. The plots demonstrate that Gradient Boosting provides the highest prediction accuracy, closely aligning predicted values with actual revenue and outperforming Random Forest and its tuned variants.

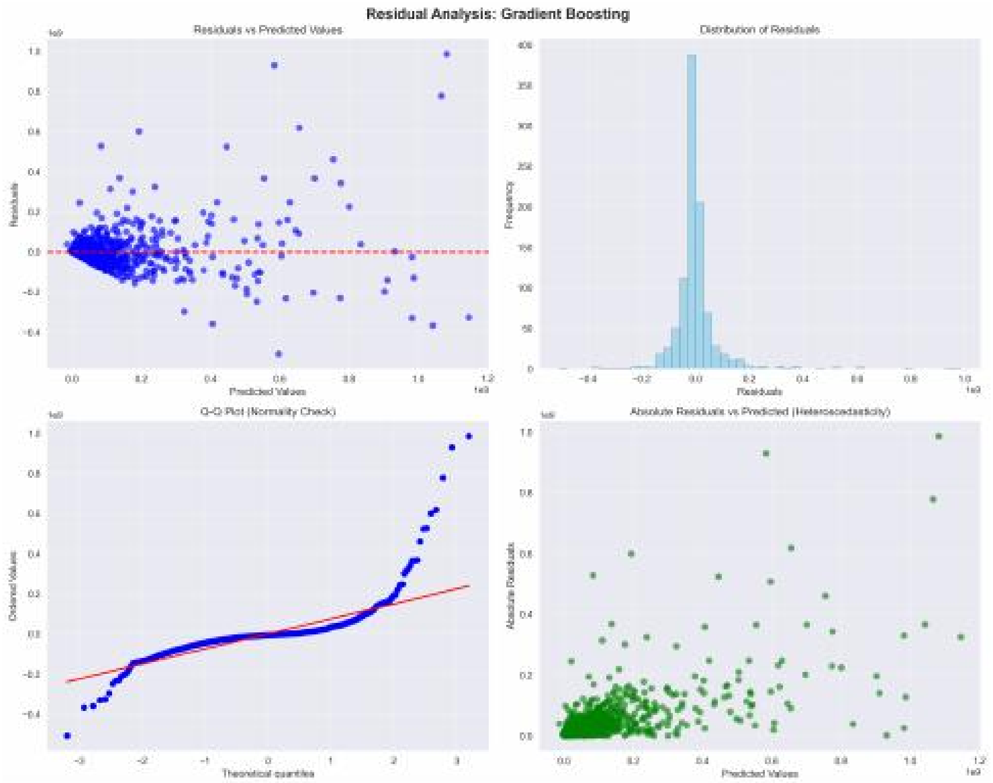


Fig. 8: Residual analysis for the Gradient Boosting model. The plots include residuals versus predicted values, residual distribution, normality check (Q-Q plot), and heteroscedasticity analysis, providing insights into model error patterns and distribution.

5 Methodology

This section describes the dataset used for analysis and the feature engineering techniques applied to enhance model performance.

5.1 Dataset Description

The study uses a carefully chosen dataset of 4,880 movies released from 1980 to 2017. This dataset came from merging metadata and credits information. The original raw data contained more than 45,000 entries for both movie metadata and credits. They filtered and combined these entries to concentrate on high-quality, relevant records. Key attributes include budget, revenue, vote counts, runtime, genres, and release year, among others. The dataset reveals a broad variety of financial and audience engagement metrics, showing the diversity of movie productions over nearly four decades.[1, 8].

5.2 Feature Engineering

To improve predictive capabilities, we conducted extensive feature engineering. We created features such as budget per minute and logarithmic transformations of budget to better capture nonlinear relationships with box office revenue. We included audience engagement indicators like vote count and weighted vote scores because they showed strong correlations with revenue during exploratory analysis. We also added temporal features, including release year and decade-based categorizations, to improve the model's understanding of industry trends. These engineered features reduced noise and improved signal quality, which is essential for the ensemble learning models used. [7, 15]. These steps together ensured a strong and informative feature set that served as the basis for effective training and evaluation of multiple machine learning models.

5.3 Model Training and Evaluation

The model training process involved splitting the dataset into training and test sets, with 3,904 movies used for training and 976 for testing. To ensure robust performance and avoid overfitting, hyperparameter tuning was performed using cross-validation, a technique that partitions the training data into multiple folds to iteratively train and validate the model. This approach helps estimate how well the model generalizes to unseen data [2].

Various machine learning algorithms including Linear Regression, Ridge and Lasso Regression, Random Forest, Gradient Boosting, and Support Vector Regression were trained. Cross-validation Root Mean Squared Error (RMSE) and test set metrics such as R^2, RMSE, and Mean Absolute Error (MAE) were used to evaluate model performance comprehensively.

Gradient Boosting emerged as the top-performing model based on these metrics, followed closely by tuned and untuned Random Forest models. The evaluation also included statistical significance testing to confirm whether observed performance differences were meaningful.

Model evaluation metrics were computed as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where y_i and $\hat{y}_i$ denote observed and predicted revenues, respectively, and ¯y is the mean observed revenue.

This rigorous evaluation framework ensured selection of robust and generalizable models suitable for accurate box office revenue forecasting. **Gradient Boosting**

The Gradient Boosting model was built iteratively, where each new decision tree fits the residuals of prior trees. Hyperparameters such as learning rate, tree depth, and number of estimators were carefully tuned using cross-validation to optimize performance. The pseudocode for the training process is outlined below:

```
Initialize model with constant prediction F0
for t = 1 to T:
    Compute residuals: r_i = y_i - F_{t-1}(x_i)
    Fit weak learner h_t to residuals r
    Compute multiplier gamma_t by line search
    Update model: F_t(x) = F_{t-1}(x) + gamma_t * h_t(x)
end
```

Random Forest Algorithm

Random Forest constructs multiple decision trees using bootstrap samples and random feature subsets. Predictions are averaged across trees to reduce variance. The following snippet demonstrates model initialization with hyperparameters:

```
from sklearn.ensemble import RandomForestRegressor

rf_model = RandomForestRegressor(
    n_estimators=200,
    max_depth=None,
    min_samples_split=10,
    min_samples_leaf=2,
    random_state=42
)
rf_model.fit(X_train, y_train)
```

Model Evaluation Code

Evaluation metrics including R^2, RMSE, and MAE were computed on test data using standard libraries:

```
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
import numpy as np
```

```
y_pred = model.predict(X_test)
r2 = r2_score(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
mae = mean_absolute_error(y_test, y_pred)
```

These code excerpts illustrate the practical implementation details of the machine learning pipeline, emphasizing reproducibility and transparency.

The full source code and scripts are included in the supplementary material for the interested reader.

References: Implementation and algorithmic methodology were based on established libraries and studies [3, 4, 13, 14].

6 Future Scope

This study forecasts movie box office revenue with accuracy using feature engineering and ensemble learning. More data sources, such as marketing budgets, social media sentiment analysis, and competitor release dates, should be incorporated into future research. This will assist in determining more general market elements that impact income. Analysis of multimedia content could increase prediction accuracy even more. More thorough examination of trailers, posters, and audience comments is made possible by machine vision and natural language processing tools. Investigating more complex deep learning models, such as transformers or hybrid models, may also be beneficial, particularly when paired with structured metadata. Furthermore, industry stakeholders may find use for real-time or early-warning prediction systems that dynamically update projections as new information—such as pre-release buzz and early ticket sales—becomes available. Lastly, examining explainability and interpretability of the model would contribute to the development of trust and practical insights for decision-makers in the ecosystem of film production and distribution. In an entertainment industry that is becoming more and more competitive, these expansions could greatly improve box office forecasting and allow for more intelligent, data-driven strategies.[7, 11, 12].

7 Conclusion

In conclusion the study shows how ensemble learning techniques, specifically Gradient Boosting, can be used to predict movie box office revenue. The model explains over 76% of the variance. This suggests that combining data from multiple sources—particularly vote counts and budgetary data—produces a powerful predictive signal. To improve the model's performance, we had to create better features. The conventional raw data did not reveal the patterns and nonlinear relationships that our method did. Previous research has shown how public perception affects financial results. This is supported by the ongoing significance of audience engagement metrics, such as vote count.Although the Random Forest model produced results that were competitive, Gradient Boosting's adaptability and iterative optimization made it possible to better handle the intricate dynamics of box office revenue. The poorer performance of simpler linear models indicates that stronger algorithms are needed to capture interactions in entertainment data. This study developed a robust ensemble learning framework for predicting movie box office revenue through careful feature engineering and rigorous model evaluation. By analyzing a dataset of 4,880 films, we demonstrated that Gradient Boosting outperforms other models, achieving the lowest prediction errors and the highest accuracy with an R2 score of 0.7639.The success of the model demonstrates how important it is to integrate time data, budget-related considerations, and audience engagement metrics in order to comprehend the factors influencing box office performance. To verify the effectiveness and dependability of the suggested solution, our approach combines statistical significance testing with a thorough exploratory analysis. Future research may be able to predict results more accurately if external factors like advertising campaigns, social media trends, and competing releases are taken into account. More thorough multimedia analysis, such as sentiment analysis and trailer content, could be added to the features to increase accuracy. [1, 2].

References

[1] Xu, J.: Long-range movie box office prediction based on machine learning. Highlights in Science, Engineering and Technology **92**, 308–315 (2024)

[2] Singh, R., *et al.*: Comparing performance of ensemble methods in predicting movie box office revenue. ScienceDirect Data Analytics **240**, 31–47 (2024)

[3] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)

[4] Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189–1232 (2001)

[5] Rao, A.: Movie revenue prediction using machine learning. International Journal of Research in Management, Economics and Technology Sciences **4**(4), 52–63 (2025)

[6] Agarwal, M., Venugopal, S., Kashyap, R., Bharathi, R.: Movie success prediction and performance comparison using various statistical approaches. International Journal of Artificial Intelligence and Applications **13**(1), 19–33 (2022)

[7] Lee, K., Park, J., Kim, I., Choi, Y.: Predicting movie success with machine learning techniques: Ways to improve accuracy. Expert Systems with Applications **165**, 113–124 (2021)

[8] Shahid, M.H., Islam, M.A.: Investigation of time series-based genre popularity features for box office success prediction. BMC Medical Informatics and Decision Making **23**, 103 (2023)

[9] Kumar, S.: Box office revenue prediction using linear regression in machine learning. International Journal of Creative Research Thoughts **9**(1), 1–8 (2025)

[10] Zheng, Y.: Predicting movie box office based on machine learning, deep learning, and statistical methods. Applied and Computational Engineering **31**, 45–52 (2024)

[11] Madongo, C.T., Tang, Z., Hassan, J.: Movie box-office revenue prediction model by mining deep features from trailers using recurrent neural networks. Journal of Advances in Information Technology **15**(6), 764–771 (2024)

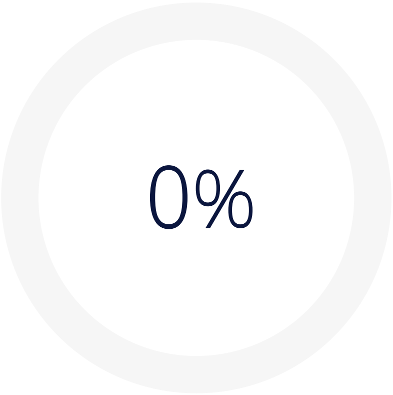
[12] Udandarao, V., Gupta, P.: Movie revenue prediction using machine learning models. arXiv preprint arXiv:2405.11651 (2024)

[13] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[14] Pedregosa, F., *et al.*: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[15] Liu, J.: A refined approach to early movie box office prediction leveraging ensemble learning and feature encoding. In: ACE Conference Proceedings, pp. 78–85 (2024). https://www.ewadirect.com/proceedings/ace/article/view/13734

AI Content



	Text Coverage	Words
○ AI Text	0%	0
○ Human Text	100%	2,324
Excluded		
◌ Omitted Words		340

About AI Detection

Our AI Detector is the only enterprise-level solution that can verify if the content was written by a human or generated by AI, including source code and text that has been plagiarized or modified. Learn more

○ AI Text

A body of text that has been generated or altered by AI technology. Learn more

○ Human Text

Any text that has been fully written by a human and has not been altered or generated by AI. Learn more

Copyleaks AI Detector Effectiveness

Credible data at scale, coupled with machine learning and widespread adoption, allows us to continually refine and improve our ability to understand complex text patterns, resulting in over 99% accuracy—far higher than any other AI detector—and improving daily. Learn more

Ideal Text Length

The higher the character count, the easier for our technology to determine irregular patterns, which results in a higher confidence rating for AI detection. Learn more

Reasons It Might Be AI When You Think It's Not

The AI Detector can detect a variety of AI-generated text, including tools that use AI technology to paraphrase content, auto-complete sentences, and more. Learn more

User AI Alert History

Historical data of how many times a user has been flagged for potentially having AI text within their content. Learn more

AI Logic

The number of times a phrase was found more frequently in AI vs human text is shown according to low, medium, and high frequency. Learn more