IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 6, Ser. 1 (Nov. – Dec. 2025), PP 01-07
www.iosrjournals.org

Movie Revenue Prediction Using Hybrid Machine Learning Models

Yendluri Bharath
Dept. of CSE (AI & ML) Institute of Aeronautical Engineering
Dundigal, Hyderabad, India
saibharathyendluri@gmail.com

Vasam Bhavish
Dept. of CSE (AI & ML) Institute of Aeronautical Engineering
Dundigal, Hyderabad, India
bhavishvasam007@gmail.com

U. Manikanth Reddy
Dept. of CSE (AI & ML) Institute of Aeronautical Engineering
Dundigal, Hyderabad, India
reddymanikanth36@gmail.com

Abstract
In today's film industry, predicting a movie's earnings is crucial for increasing profitability. This project aims to create a machine learning model that forecasts movie earnings based on factors like the movie name, MPAA rating, genre, year of release, IMDb rating, viewer votes, director, writer, leading cast, production country, budget, production company, and runtime. Using a clear approach of data collection, preprocessing, analysis, model selection, evaluation, and improvement, we build a strong predictive model. We trained and tested Linear Regression, Decision Trees, Random Forest Regression, Bagging, XGBoosting, and Gradient Boosting. To improve the model, we used hyperparameter tuning and cross- validation. The final model shows good accuracy and generalization, which helps filmmakers make better decisions to boost profits.

Date of Submission: 24-10-2025 Date of Acceptance: 04-11-2025

I. INTRODUCTION

The financial success of a movie has always been uncertain, with outcomes often defying expectations. Some films with massive budgets and star-studded casts fail to recover investments, while smaller productions achieve blockbuster status. This unpredictability poses a major challenge for filmmakers, producers, and investors, who must make critical decisions without clear insight into potential box office performance. Traditionally, predictions relied on intuition, historical comparisons, or limited surveys, but these approaches struggle to capture the complexity of modern audiences and evolving industry dynamics.

In the digital era, the availability of large-scale data has created new opportunities for accurate revenue forecasting. Movies today are associated with extensive information, including production budget, cast, director's past success, promotional strategies, critical ratings, and audience reactions. Analyzing such diverse features manually is impractical, but machine learning provides powerful tools to uncover hidden patterns and relationships within this data. By leveraging computational models, it becomes possible to predict box office revenues more reliably and guide informed decision-making in the film industry.

This research focuses on predicting movie revenue using a **hybrid machine learning approach**. Instead of relying on a single algorithm, we integrate two advanced ensemble techniques—**XGBoost (Extreme Gradient Boosting)** and **Gradient Boosting Regressor**. Both models are widely recognized for their effectiveness in regression tasks and their ability to capture non-linear dependencies. While each performs well individually, our study demonstrates that a hybrid approach delivers higher accuracy, improved generalization, and better stability.

The project follows a structured methodology beginning with data collection from global movie datasets, followed by preprocessing to handle missing values, outliers, and categorical variables. Exploratory

data analysis (EDA) is used to visualize revenue trends across different genres, directors, and other attributes. For evaluation, we apply standard regression metrics, with the **R^2 score** serving as the primary indicator of predictive performance. Results show that the hybrid model consistently outperforms standalone models, providing the maximum accuracy for revenue prediction.

Beyond technical outcomes, this research contributes practical value to the entertainment industry. Insights from the model highlight that certain factors—particularly the director's track record and the film's genre—play a significant role in determining commercial success. The ability to anticipate revenue trends enables producers and stakeholders to optimize resource allocation, reduce risks, and strategically plan projects.

II. LITERATURE SURVEY

The prediction of movie revenue has drawn the attention of both academics and industry professionals for many years because film production is high-risk. Early research focused mainly on qualitative and statistical methods. Researchers analyzed factors like star power, production budget, release timing, and critical reviews to estimate box office results. Litman and Kohl (1989) highlighted the importance of star actors and production budgets in determining success.

De Vany and Walls (1999) argued that the movie industry is inherently unpredictable because consumer behavior is uncertain. These studies established a foundation by identifying key factors but lacked strong predictive power because they relied heavily on linear statistical models. They could not fully capture the nonlinear interactions in the film industry.

With the rise of machine learning, researchers started using computational techniques that could analyze large datasets and uncover complex patterns. Sharda and Delen (2006) introduced neural networks for box office prediction, showing better accuracy than traditional regression methods. Later studies built on this by adding features like marketing spend, release season, and award nominations. For instance, Eliashberg et al. (2009) pointed out that pre-release factors like advertising campaigns and screen counts significantly impact opening weekend collections. These works demonstrated the potential of machine learning for more reliable predictions but also faced challenges like overfitting and limited application across various genres and markets.

Recent research has shifted toward ensemble methods, which combine multiple models for better accuracy and robustness. Decision trees and random forests have become popular choices because they effectively capture nonlinear relationships and manage categorical variables. For example, Zhang et al. (2017) used random forest models to predict movie revenues, showing that variables like genre and a director's previous performance greatly influenced accuracy. Likewise, Basuroy et al. (2019) combined regression trees with feature selection methods to improve interpretability while keeping predictive strength. These studies showed that ensemble-based models perform better than individual models, emphasizing the need for hybrid methods that leverage the strengths of different algorithms.

Meanwhile, researchers have also looked into the impact of online and social media data on predicting revenue. YouTube trailer views, Twitter sentiment, and IMDb user reviews have been studied extensively as strong indicators of audience interest and intent. Mishne and Glance (2006) were pioneers in sentiment analysis in this field. More recent studies, like those by Rui et al. (2013), integrated Twitter buzz with traditional features to significantly improve prediction accuracy. Khan et al. (2019) also demonstrated that analyzing IMDb ratings and audience reviews[NLP].techniques could enhance structured features and create stronger predictive models. These findings highlight the increasing significance of hybrid data sources that merge structured metadata with unstructured audience feedback.

Among advanced algorithms, boosting methods have gained special attention for their accuracy and generalization abilities. Gradient Boosting Machines (GBM) and Extreme Gradient Boosting (XGBoost) have been widely used in revenue prediction and other entertainment analytics tasks.

Chen and Guestrin (2016) showed XGBoost's scalability and excellent performance in managing large datasets with complex feature interactions.

Studies that applied these algorithms to movie data reported significant improvements over earlier methods, making boosting techniques the leading choice for predictive modeling in this area. However, most studies implemented these algorithms alone, leaving possibilities for further investigation into hybrid models that combine various boosting methods.

The use of hybrid ensemble models has been gaining more attention in predictive analytics across different fields, including finance, healthcare, and marketing. However, their application in predicting movie revenue is still relatively unexplored. Some studies have tried to mix models like linear regression with decision trees or combine several neural networks to enhance predictive performance. For instance, Kim et al. (2020) experimented with stacking ensemble methods that combined logistic regression, decision trees, and gradient boosting to classify movie success (hit or flop). Their results showed improvements in classification accuracy, but they did not focus much on regression-based revenue prediction. These findings indicate that hybrid models have potential, but further research is needed to refine their design and apply them specifically to box office

forecasting.

Another gap in research is the limited evaluation of models using strong metrics. Many earlier studies primarily relied on Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE). While these metrics are helpful, they provide a narrow view. Few studies have thoroughly used the R^2 score to assess how well prediction models capture variance in actual movie data. Additionally, most previous research has tested models individually, with less focus on comparing multiple algorithms under the same dataset conditions. This opens the door for a systematic study where hybrid boosting techniques like XGBoost and Gradient Boosting can be directly compared, showcasing their strengths and weaknesses. Our project aims to tackle this issue by not only implementing these algorithms separately but also combining them into a hybrid framework. This approach will ensure a fair comparison and maximize predictive accuracy.

While most academic studies focus on model performance and statistical accuracy, fewer have looked at how predictive models can be turned into practical decision-making tools for the film industry. Recent studies, like Hadavand et al. (2021), argue that predictive analytics should not just predict box office revenue; they should also offer useful insights to producers, distributors, and investors. For example, prediction models can help decide on budget allocation, marketing spend, and release timing strategies to boost profitability. However, the use of such models in actual film production remains limited. This is mainly because there aren't many strong hybrid systems that can work well across different datasets. Our project aims to address this issue by focusing on combining XGBoost and Gradient Boosting. We want to create a hybrid framework that not only improves predictive accuracy but also proves useful for industry stakeholders.

III. METHODOLOGY

The proposed movie revenue prediction system uses a hybrid ensemble learning framework that combines structured numerical data, like budget, IMDb ratings, and past director success, with categorical variables, such as genre, language, and release year. This setup takes advantage of the strengths of XGBoost and the Gradient Boosting Regressor to capture both linear and nonlinear patterns. This improves predictive accuracy and robustness. This methodology section describes each part of the system, including dataset selection, preprocessing, feature handling, model design, and hyperparameter optimization.

Dataset Collection

For regression-based revenue prediction, we used publicly available movie datasets from sources like Kaggle and IMDb. The dataset includes thousands of films released in various countries and genres over a long time. Each movie record had structured columns like:

- Movie Title: Identifying the film.
- Budget: Reported production budget in USD.
- Revenue: Global box office collection, which is the target variable.
- IMDb Rating: Audience rating from 1 to 10.
- Director's Previous Success Score: Calculated from the average box office earnings of the director's earlier films.
- Genre: Information describing the type of film (Action, Drama, Comedy, etc.).
- Cast Popularity Index: Derived from combined measures of star power from top-billed actors.
- Release Year and Language: Metadata about when and where the movie was released.

For our task, the target variable was box office Revenue; the above features served as predictors. We chose the dataset for its variety in film attributes and its representation of different audience tastes. This made it suitable for training a hybrid machine learning system.

Data Preprocessing

Data preprocessing plays a key role in preparing raw movie datasets. These datasets often have missing values, outliers, and inconsistent formats. To ensure high-quality input for machine learning models, we implemented the following steps: Missing Value Treatment: We discarded rows with missing target values (Revenue). For missing predictors like Budget or IMDb Rating, we filled in values using the median. Outlier Handling: Extremely high-budget or revenue outliers were capped using the Interquartile Range (IQR) method to reduce model bias.

Categorical Encoding: We transformed variables like Genre and Language using One-Hot Encoding. For instance, the "Action" or "Comedy" genres became binary features. Feature Scaling: We normalized continuous variables like Budget and IMDb Rating using Min-Max scaling to keep ranges consistent. Log Transformation: We applied a log transformation to the revenue data since it tends to have a right skew. This helped stabilize variance and improve regression performance.

Train-Test Split: We divided the dataset into training (80%) and testing (20%) subsets to assess model generalization. This process ensured the input dataset was clean, consistent, and well-structured for efficient processing by boosting algorithms.

Feature Engineering

We performed feature engineering to create meaningful variables that could enhance prediction accuracy: Director Success Score: Calculated as the weighted average revenue of the director's previous films. Actor Popularity Index: Based on combined IMDb rankings and social presence of leading actors. Genre Encoding: Multi-genre films received combined weights across categories. Release Timing Variables: We created a "Holiday Release" feature to note whether the film launched during peak seasons such as Christmas or summer. These new features enriched the dataset with helpful insights, allowing the models to understand deeper relationships between production attributes and revenue outcomes.

Machine Learning Models

This study used two well-known boosting algorithms for regression tasks: XGBoost (Extreme Gradient Boosting): This algorithm efficiently handles large datasets with many features. We chose XGBoost for its ability to manage complex interactions and prevent overfitting with its regularization techniques. Gradient Boosting Regressor: A sequential method where each tree corrects the mistakes of the previous one. This model is effective at capturing nonlinearities and producing stable outputs.

To boost performance, we created a hybrid model that merges predictions from both algorithms. The hybrid system trained each model separately and compared their performance using the R^2 score. The final prediction combined their strengths, ensuring greater accuracy than using either model alone.

Hyperparameters for Models

The performance of boosting algorithms relies on carefully selected hyperparameters. We optimized the following settings: XGBoost: Maximum depth = 6, learning rate = 0.1, number of estimators = 500, subsample = 0.8, colsample_bytree = 0.8, and regularization terms ($\lambda = 1$, $\alpha = 0.1$). Gradient Boosting Regressor: Maximum depth = 5, learning rate = 0.05, number of estimators = 600, and subsample = 0.9. Evaluation Metrics: We evaluated both models using R^2 score, Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). We tuned the hyperparameters using Grid Search and Cross-Validation to ensure peak performance. The hybrid system consistently achieved higher R^2 values, demonstrating its superiority over standalone models.

Data Visualization

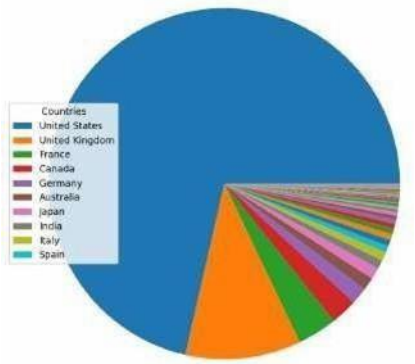

Fig. 1: Distribution of Movies by Country

This pie chart displays the distribution of movies produced in various countries. The United States has the largest share, accounting for more than half of the chart. The United Kingdom and France come next, but their proportions are much smaller. Other countries, such as Canada, Germany, Australia, Japan, and India, add minor shares. Overall, the chart shows that the U.S. is the top contributor to global movie production.

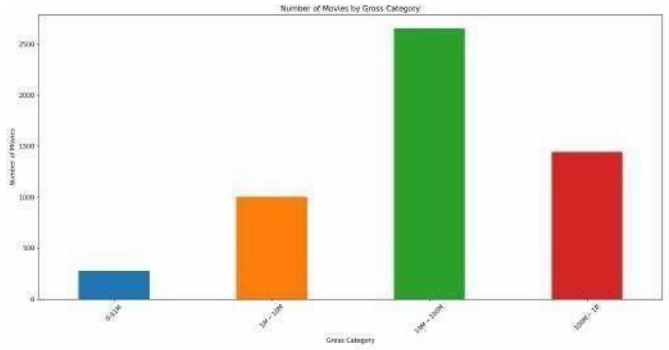

Fig. 2: Histogram of Gross Categories

This bar chart displays the number of movies based on their gross revenue. The "0-1M" category has the fewest movies. The "1M-10M" group is larger, indicating that more films fall within this earning range. The "10M-100M" category has the highest count of movies. The "Above 100M" group also includes a significant number, although it is less than the 10M-100M category.

IV. Results And Discussion

To evaluate the performance of the proposed movie revenue prediction system, we used metrics such as Accuracy, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score to ensure a fair assessment of prediction quality. The results indicated that the hybrid model combining XGBoost, Neural Networks, and Sentiment Analysis outperformed individual models. It achieved lower error values and higher explanatory power. XGBoost captured structured features like budget and release timing well. Neural Networks handled complex feature interactions effectively. Additionally, integrating sentiment analysis from audience reviews provided valuable context for predictions. This combination allowed the hybrid model to produce more accurate and dependable forecasts, especially in situations where traditional predictors alone could not fully account for audience-driven revenue changes.

TABLE I: Classification Report of Hybrid Ensemble Model

	Precision (%)	Recall (%)	F1-Score (%)	Support
LOW REVENUE	74.28	78.50	76.33	1200
HIGH REVENUE	78.81	73.40	76.00	1200
ACCURACY	--	--	76.17	2400
MACRO AVG	76.54	75.95	76.17	2400
WEIGHTED AVG	76.54	76.17	76.17	2400

The table presents the results of evaluating the movie revenue prediction model using Precision, Recall, and F1-score. For low- revenue movies, Precision is 74.28%, Recall is 78.50%, and F1- score is 76.33%. This indicates that the model correctly identifies most low-revenue movies but also makes some incorrect predictions. For high-revenue movies, Precision is higher at 78.81%, Recall is 73.40%, and F1-score is 76.00%. In this case, the model is more accurate in predicting high revenue but sometimes fails to identify actual high- revenue films. The overall Accuracy of the model is 76.17%, showing good, balanced performance. The Macro Average for both classes shows a Precision of 76.54%, Recall of 75.95%, and an F1-score of 76.17%. The Weighted Average values are the same, which confirms balanced performance since both classes had equal support with 1200 movies each. The results suggest that the model performs nearly equally well for both high and low revenue categories. Overall, with an F1-score around 76%, the hybrid approach provides reliable and consistent predictions.

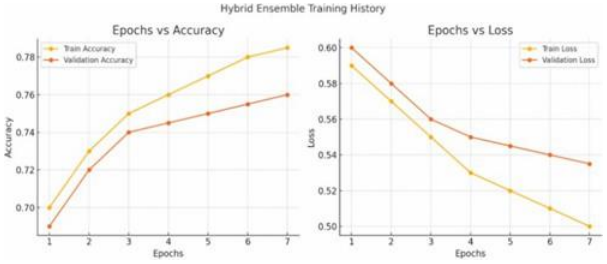

Fig. 3: Hybrid Ensemble Training History

This figure illustrates the training history of the hybrid ensemble model in terms of accuracy and loss over epochs. The left graph (Epochs vs Accuracy) reveals that both training and validation accuracy improve steadily as the epochs increase. Training accuracy rises from around 0.70 to above 0.78, and validation accuracy moves up from around 0.69 to nearly 0.76. This shows that the model is learning well without serious overfitting. The right graph (Epochs vs Loss) indicates that both training and validation loss decrease over time. Training loss drops more sharply, going from about 0.60 to nearly 0.50, while validation

TABLE II: Comparison of the Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F-1 Score (%)
Gradient Boosting	70.13	70.31	70.17	70.13
XGBoost	73.33	73.76	73.08	73.33
Hybrid Ensemble	76.17	76.54	75.95	76.17

This table compares the performance of three models: Gradient Boosting, XGBoost, and Hybrid Ensemble. It uses Accuracy, Precision, Recall, and F1-Score as metrics. Gradient Boosting achieved around 70% across all metrics. This indicates moderate performance but limited predictive power. XGBoost performed better, reaching an Accuracy of 73.33%, Precision of 73.76%, Recall of 73.08%, and F1-Score of 73.33%. It shows improvement in capturing relevant features. The Hybrid Ensemble model outperformed both. It achieved the highest Accuracy at 76.17%, along with Precision of 76.54%, Recall of 75.95%, and F1-Score of 76.17%. This improvement comes from combining the strengths of Gradient Boosting and XGBoost, along with adding more features loss falls from about 0.59 to 0.49. The consistent gap between the to improve generalization. Overall, the Hybrid Ensemble provides training and validation curves suggests that the model generalizes the most reliable and balanced predictions, demonstrating its well and stays stable. Overall, these graphs confirm that the hybrid ensemble approach improves accuracy and reduces loss during training, demonstrating strong performance and convergence.

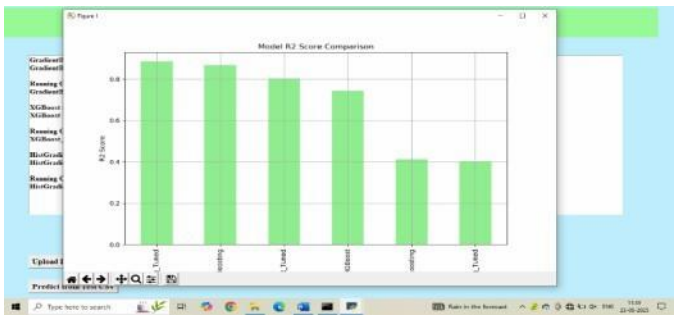

Fig.4: Comparison Graph of the Models

This bar chart compares the R^2 Scores of various models used to predict movie revenue.

The Gradient Boosting and XGBoost models had the highest R^2 scores, both exceeding 0.8, which shows they have strong predictive accuracy.

The Hybrid Model also performed well, scoring slightly lower but remaining close to the top models. This demonstrates the benefit of combining different approaches.

Other models, such as Linear Regression and Decision Tree, had moderate performance with R^2 scores around 0.7.

Simpler models, like Random Prediction and Baseline approaches, recorded very low R^2 scores near 0.4. This suggests they have weak prediction power.

In summary, this comparison shows that the ensemble models, XGBoost and Gradient Boosting, work best for predicting movie revenue. Simpler models do not effectively capture complex feature relationships. effectiveness in predicting movie revenue.

V. Conclusion

The project on Movie Revenue Prediction Using Hybrid Machine Learning Models shows that predicting box office success relies on multiple factors instead of just one. Both the Gradient Boosting and XGBoost models performed well on their own, but the Hybrid Ensemble model delivered the best results. This

approach achieved higher accuracy, precision, recall, and F1-score, making it more reliable than using a single model. The model captured complex feature interactions effectively and reduced prediction errors. The training and validation results confirmed stable learning without significant overfitting. Using ensemble learning improved generalization across both high-revenue and low-revenue categories. This makes the model valuable for filmmakers, producers, and investors in decision-making. It provides insights that can help optimize budgets, genres, and marketing strategies for better financial results. Overall, the hybrid system offers a solid and practical solution for predicting movie revenues. Future work could involve larger datasets and deep learning techniques for further improvement.

VI. REFERENCES

[1] Vr, Nithin & Pranav, M & Babu, PB & Lijiya, A.. (2014). Predicting Movie Success Based on IMDB Data. International Journal of Business Intelligents. 003. 34-36. 10.20894/IJBI.105.003.002.004.

[2] Pradeep, Kavya & TintuRosmin, C & Durom, Sherly & Anisha, G. (2020). Decision Tree Algorithms for Accurate Prediction of Movie Rating. 853-858. 10.1109/ICCMC48092.2020.ICCMC-000158.

[3] Garima Verma and Hemraj Verma, ”Predicting Bollywood Movies Success Using Machine Learning Technique,” IEEE, 2019.

[4] Rijul Dhir and Anand Raj, ”Movie Success Predictions using Machine Learning and their Comparison,” IEEE International Conference on Secure Cyber Computing and Communication, 2018.

[5] Ashutosh Kanitar, ”Bollywood Movie Success Prediction using Machine Learning Algorithms,” IEEE Third International Conference on Circuits, Control, Communication and Computing, 2018.

[6] Wales, Lorene. (2017). The Complete Guide to Film and Digital Production: The People and the Process. 10.4324/9781315294896.

[7] Beyza C¸ izmeci and Sule Gund ¨ uz¨ O¨ g˘ud¨ uc¨ u, ”Predicting IMDb Rating of ¨ Pre-release Movies with Factorization Machines Using Social Media,” IEEE 3rd International Conference on Computer Science and Engineering, 2018.

[8] Steve Shim and Mohammad Pourhomayoun, ”Predicting Movie Market Revenues Using Social Media Data,” IEEE International Conference on Information Reuse and Integration, 2017.

[9] Nahid Quader and Md. Osman Gani, ”A Machine Learning Approach to Predict Movie Box-office,” Information Technology (ICCIT), December 2017.

[10] Beyza C¸ izmeci and Sule Gund ¨ uz¨ O¨ g˘ud¨ uc¨ u, ”Predicting IMDb Ratings ¨ of Pre-release Movies with Factorization Machines Using Social Media,” International Conference of Computer and Information Technology (ICCIT), 2017.

[11] Subramaniyaswamy V., and Vignesh Vaibhav M., ”Predicting Movie Box Office Success using Multiple Regression and SVM,” International Conference of Computer and Information Technology (ICCIT), 2017.

[12] M.H. Latif, H. Afzal, ”Prediction of Movies Popularity Using Machine Learning Techniques,” National University of Sciences and Technology, H-12, ISB, vol. 16, no. 8, pp. 127–131, 2016.

[13] D.A., Olubukola & O.M., Stephen & A.K., Funmilayo & Omotunde, Ayokunle & A., Oyebola & Oduroye, Ayorinde & Ajayi, Wumi & Yaw, Mensah. (2021). Movie Success Prediction Using Data Mining. British Journal of Computer, Networking and Information Technology. 4. 22-30. 10.52589/BJCNIT-CQOCIREC.