ISSN: 3107-6513

Movie Success Prediction Using Machine Learning and Financial Performance Analysis

Shivam Yadav[1], Sandhya Kaprawan[2]

[1]M.S.(Data Analytics), [2]Assistant Professor

[1]shivamyadav26037@gmail.com, [2]sandhya.kaprawan@udit.mu.ac.in

Abstract

The film industry is one of the most uncertain and competitive sectors where substantial investments do not always guarantee commercial success. Movie performance depends on multiple interconnected factors such as production budget, marketing expenditure, audience engagement, actor popularity, director reputation, ticket pricing, and release scale. Traditional methods of evaluating movie success are often based on intuition, experience, and historical observations, which may not accurately capture the complex relationships among these variables. This research presents a Machine Learning-based Movie Success Prediction System designed to classify movies as either HIT or FLOP while simultaneously evaluating their financial performance. Due to the limitations of publicly available datasets, a structured synthetic dataset was developed using logical relationships between critical movie-related features. The proposed system employs a Random Forest Classifier to learn patterns from historical and simulated data and generate prediction outcomes. In addition to classification, the system performs financial analysis by estimating revenue, total cost, and profitability based on audience occupancy, seating capacity, ticket price, and marketing investment. A Streamlit-based dashboard was developed to provide an interactive and user friendly platform where users can experiment with different movie scenarios and observe both predictive and financial outcomes in real time. Experimental evaluation demonstrated an overall accuracy of 86.40%, with a precision of 90%, recall of 74%, and F1-score of 81%. The results indicate that integrating machine learning prediction with financial evaluation provides a more comprehensive decision-support framework than traditional prediction-only systems. The proposed approach can assist producers, investors, and distributors in making more informed decisions during movie planning and release stages.

Keywords—Machine Learning, Movie Success Prediction, Random Forest, Financial Analysis, Predictive Analytics, Streamlit, Film Industry, Data Analytics

I. INTRODUCTION

The global entertainment industry has experienced significant growth over the past few decades, with the film sector becoming one of the largest contributors to media and entertainment revenues. Every year, production companies invest millions of dollars in movie development, marketing campaigns, distribution strategies, and promotional activities. Despite these investments, predicting whether a movie will become successful remains a highly challenging task. Movie success is influenced by a wide range of factors that interact in complex ways. Production budget determines the scale and quality of production, while marketing expenditure influences audience awareness and engagement. Similarly, actor popularity, director reputation, genre selection, ticket pricing, and release scale contribute significantly to the overall performance of a movie. However, these variables do not operate independently.

Their combined influence creates intricate relationships that are difficult to analyze using traditional methods. Historically, movie success prediction has relied heavily on expert judgment, industry experience, and historical box-office trends. Producers and investors often make decisions based on intuition and previous success patterns. While such approaches provide some guidance, they frequently fail to account for nonlinear relationships and hidden interactions among multiple influencing variables. Consequently, many high-budget movies fail commercially, whereas low budget productions occasionally achieve remarkable success. The emergence of Machine Learning (ML) has transformed decision-making across numerous

ISSN: 3107-6513

industries by enabling systems to learn patterns directly from data [2], [3], [5].

ML algorithms can analyze large volumes of information, identify hidden relationships, and generate predictive outcomes with greater consistency than traditional approaches. In the context of movie success prediction, machine learning offers the potential to evaluate multiple variables simultaneously and provide objective assessments of future performance [6], [8].

Several studies have explored the use of predictive analytics for box-office forecasting and movie performance evaluation. However, many existing systems focus exclusively on classification outcomes such as successful or unsuccessful movies without considering financial viability. In practical business environments, profitability often plays a more critical role than simple success classification. A movie may receive a positive prediction but still fail to generate sufficient profit, whereas another movie may achieve profitability despite modest predictive indicators [5], [12].

Another major challenge in this domain is the availability of high-quality datasets. Public movie datasets often contain incomplete information, inconsistent structures, missing values, and limited feature diversity. During the preliminary stages of this research, attempts were made to develop prediction models using existing movie datasets. However, the resulting performance was unsatisfactory due to inconsistencies and insufficient feature representation [2], [10]. To overcome these limitations, this research adopts a structured data-generation approach. A synthetic dataset was developed using domain knowledge and logical relationships among critical movie-related variables. Instead of random value assignment, a scoring mechanism was implemented to simulate realistic interactions between factors such as budget allocation, audience occupancy, actor popularity, marketing effectiveness, and screen distribution. The proposed framework extends beyond traditional prediction systems by integrating financial performance analysis with machine learning classification. The system first predicts whether a movie is likely to become a HIT or FLOP and then independently estimates revenue, total cost, and profit/loss outcomes. This dual-analysis approach provides users with a broader understanding of potential movie performance and supports more informed decision-making. Furthermore, a web-based application was developed using Streamlit to improve accessibility and usability. The interface allows users to modify input parameters dynamically and observe changes in both predictive and financial outcomes. This feature enables experimentation with different movie scenarios and provides valuable insights into how individual factors affect overall performance. The primary objectives of this research are:

• To develop a machine learning model capable of predicting movie success based on multiple influencing factors.

• To design a structured dataset representing realistic movie industry conditions.

• To evaluate financial performance through revenue, cost, and profitability estimation.

• To provide an interactive decision-support system for stakeholders in the film industry. The remainder of this paper is organized as follows.

Section II presents the literature review and discusses existing approaches to movie success prediction. Section III explains the proposed methodology and dataset development process. Section IV describes the system architecture and implementation details. Section V presents experimental results and performance evaluation. Section VI discusses financial analysis and interpretation. Section VII highlights limitations and future work. Finally, Section VIII concludes the research.

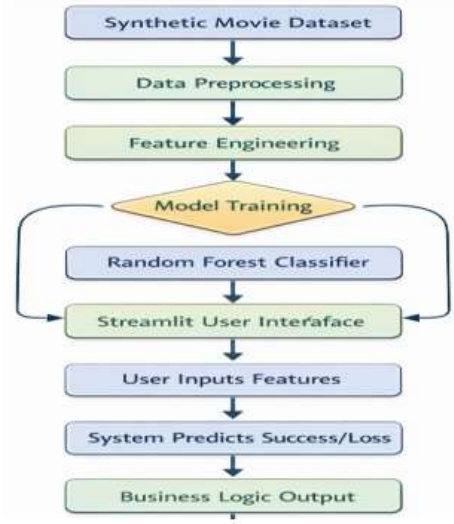

Figure 1: System Architecture Diagram

ISSN: 3107-6513

II. LITERATURE REVIEW

Movie success prediction has attracted considerable attention from researchers due to the highly uncertain and dynamic nature of the entertainment industry. Unlike many traditional prediction problems, movie performance depends on a complex combination of financial, creative, operational, and audience-related factors. The success of a movie cannot be attributed to a single variable; rather, it is the result of interactions among production quality, marketing effectiveness, distribution strategy, audience engagement, and market competition [5], [12].

Early studies on movie performance prediction primarily relied on descriptive statistics and historical trend analysis. Researchers examined factors such as genre popularity, actor reputation, director success history, and previous box-office performance to estimate the likelihood of success [7], [10]. While these approaches provided useful insights, they were limited in their ability to capture nonlinear relationships among multiple variables. As a result, prediction accuracy remained relatively low, especially when applied to diverse movie categories and changing market conditions [8], [9]. With the advancement of computational techniques and the availability of large datasets, researchers gradually shifted toward data-driven approaches. Machine Learning algorithms emerged as powerful tools for identifying hidden patterns and relationships within complex datasets [3], [6]. Unlike traditional statistical methods, machine learning models can simultaneously analyze multiple influencing factors and generate predictions based on learned patterns rather than predefined assumptions [2], [8]. Several machine learning techniques have been explored in movie success prediction research. Logistic Regression has been widely used because of its simplicity and interpretability [7].

However, its performance is often limited when relationships among variables become highly nonlinear. Decision Tree models provide better interpretability and can handle mixed data types, but they are susceptible to overfitting when trained on complex datasets [10]. Support Vector Machines (SVM) have also been applied to movie prediction problems due to their ability to perform classification in high-dimensional feature spaces. Although SVM models often achieve good predictive performance, they require careful parameter tuning and may become computationally expensive when dealing with large datasets [7], [9].

In recent years, ensemble learning methods such as Random Forest have gained significant popularity. Random Forest combines multiple decision trees and aggregates their outputs through a majority voting mechanism. This approach improves prediction stability, reduces overfitting, and enhances generalization performance. Additionally, Random Forest models can effectively handle both categorical and numerical features, making them particularly suitable for movie performance prediction tasks [1].

Another important aspect discussed in existing literature is feature selection. Researchers have consistently reported that prediction accuracy depends not only on the choice of algorithm but also on the relevance of selected features. Traditional features such as genre, budget, and cast information provide a basic understanding of movie performance. However, modern studies emphasize the importance of incorporating additional variables such as marketing expenditure, audience occupancy, release scale, social influence, and performance indicators of actors and directors. Marketing investment has been identified as one of the most influential factors affecting movie success. Effective promotional campaigns increase audience awareness and contribute directly to ticket sales. Similarly, audience occupancy serves as an important indicator of public response and market acceptance. Higher occupancy rates generally correlate with stronger financial performance and improved box-office outcomes [1], [11].

A major challenge highlighted by researchers is the lack of comprehensive and consistent datasets. Publicly available movie datasets frequently contain missing values, incomplete records, and inconsistent feature definitions. Many datasets focus only on specific genres, geographic regions, or streaming platforms, limiting their usefulness for generalized prediction systems. During preliminary experimentation for this research, attempts were made to develop prediction models using limited animation-based datasets. However, the resulting accuracy remained below 50%, demonstrating the limitations of restricted and inconsistent data sources. To overcome such challenges, several studies recommend generating structured datasets or performing extensive preprocessing and feature engineering. Data quality has repeatedly been identified as a critical factor influencing model

ISSN: 3107-6513

performance. Even sophisticated algorithms may fail when trained on incomplete or poorly structured datasets [10], [11].

Another limitation observed in existing research is the exclusive focus on classification outcomes such as HIT or FLOP prediction [5], [12]. While these predictions provide useful information, they do not necessarily reflect financial success. A movie predicted as successful may still fail to generate sufficient profit if production and marketing costs exceed revenue. Conversely, a movie predicted as unsuccessful may achieve acceptable financial returns under certain market conditions[5], [12].

For this reason, recent research has emphasized the importance of integrating financial evaluation with predictive modeling. Revenue estimation, profitability analysis, and return-on investment calculations provide additional business-oriented insights that improve decision making. Combining machine learning prediction with financial assessment creates a more practical framework for stakeholders such as producers, distributors, and investors. Based on the findings of previous studies, three key observations can be identified. First, prediction accuracy depends heavily on the quality and structure of the dataset. Second, ensemble learning techniques such as Random Forest provide stable and reliable classification performance. Third, financial analysis is essential for interpreting movie success in practical business environments.

These observations form the foundation of the proposed Movie Success Prediction System. The current research addresses the limitations of previous approaches by utilizing a structured dataset, a Random Forest classification model, and an integrated financial evaluation framework. This combination enables both predictive and business-oriented analysis, providing a more comprehensive decision-support system for the film industry.

III. Proposed Methodology

The proposed Movie Success Prediction System is designed to provide a comprehensive framework for predicting movie performance and evaluating financial outcomes before release. Unlike traditional approaches that focus only on classification, the proposed system integrates machine learning-based prediction with financial analysis to provide a more practical decision support mechanism for stakeholders in the film industry.

The overall workflow of the proposed framework consists of dataset generation, data preprocessing, model development, prediction generation, financial evaluation, and user interaction through a Streamlit-based dashboard. Each component is designed to operate independently while contributing to the overall functionality of the system.

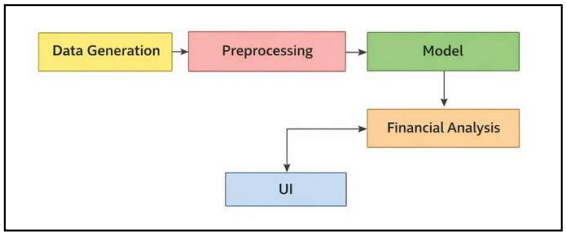

Figure 2: Proposed System Architecture

A. Dataset Development

One of the major challenges encountered during this research was the lack of a comprehensive and reliable movie dataset containing all required features. Initial experimentation using publicly available datasets resulted in poor prediction performance due to missing values, inconsistent feature structures, and limited feature diversity. In particular, early experiments conducted using animation-only datasets achieved prediction accuracy below 50%, indicating insufficient generalization capability. To overcome these limitations, a synthetic dataset generation approach was adopted. Instead of relying entirely on external data sources, a structured dataset was developed using logical relationships among critical movie-related variables. The dataset includes features such as:

- Genre
- Production Budget
- Marketing Investment
- Number of Screens
- Seating Capacity
- Ticket Price
- Audience Occupancy
- Actor Popularity
- Director Success Rate
- Screen Stability

Each feature was assigned realistic value ranges and interconnected using domain knowledge to simulate real-world movie industry conditions.

ISSN: **3107-6513**

The generated dataset was designed to reflect practical relationships among production, promotion, distribution, and audience engagement factors. A weighted scoring mechanism was implemented to determine whether a movie should be classified as HIT or FLOP.

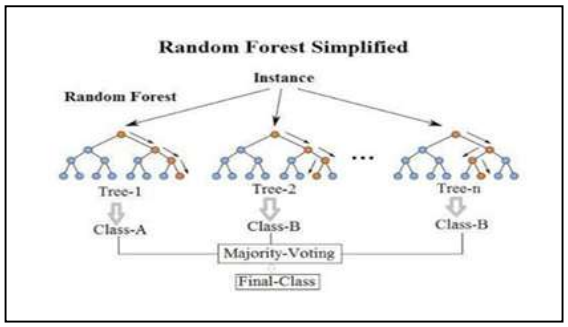

Figure 3. Random Forest Classification Framework

The model was trained using the generated dataset and evaluated using multiple performance metrics, including Accuracy, Precision, Recall, and F1-Score. Class balancing techniques were also incorporated to improve prediction performance across both HIT and FLOP categories.

B. *Prediction Framework*

The prediction module is responsible for generating movie success predictions based on user provided input parameters. The prediction workflow consists of the following steps:

1. User inputs movie-related parameters.
2. Input values are validated.
3. Data preprocessing is applied.
4. Encoded values are passed to the trained model.
5. The Random Forest classifier generates a prediction.
6. Confidence probability is calculated.
7. Final results are displayed.

The prediction output is presented in the form of:

- Predicted Class (HIT/FLOP)
- Prediction Confidence Score *Equations*

C. *System Workflow*

The workflow begins with dataset preparation and preprocessing, followed by model training and prediction generation. The resulting prediction is integrated with financial calculations and displayed through the user interface.

This structured approach ensures that the system remains scalable, interpretable, and practical for real-world movie performance evaluation.

IV. PERFORMANCE METRICS

The model achieved an overall accuracy of 86.40%, indicating its ability to correctly classify the majority of movie instances. A precision value of 90.00% demonstrates that movies predicted as successful were highly likely to be genuinely successful. The recall score of 74.00% indicates that the model successfully identified a significant proportion of successful movies. Furthermore, the F1-score of 81.00% reflects a balanced trade-off between precision and recall. These results indicate that the Random Forest classifier provides reliable predictive performance for movie success prediction.

Metric	Value
Accurracy	86.40%
Precision	90.00%
Recall	74.00%
F1-Score	81.00%

Table 1. Performace Evaluation Metrics

A. *Confusion Matrix Analysis*

The confusion matrix shows that the model correctly classified 567 FLOP movies and 297 HIT movies. Only 33 FLOP movies were incorrectly classified as HIT, while 103 HIT movies were incorrectly classified as FLOP. The relatively low number of false positive classifications demonstrates the model's effectiveness in avoiding incorrect success predictions. This characteristic is particularly valuable for producers and investors because it reduces the likelihood of overestimating movie performance. Overall, the confusion matrix confirms the robustness and consistency of the proposed prediction framework.

ISSN: 3107-6513

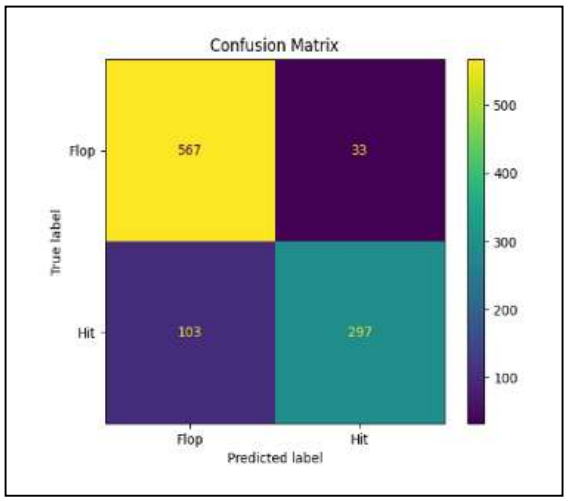

Figure 5: Confusion Matrix of the Proposed Random Forest Model.

V. SYSTEM IMPLEMENTATION

To improve usability and accessibility, the proposed model was integrated into a web-based application developed using Streamlit. The application provides a simple and interactive interface through which users can enter movie-related parameters and instantly obtain prediction and financial analysis results. The interface was designed to accommodate both technical and non-technical users, enabling efficient scenario evaluation without requiring machine learning expertise.

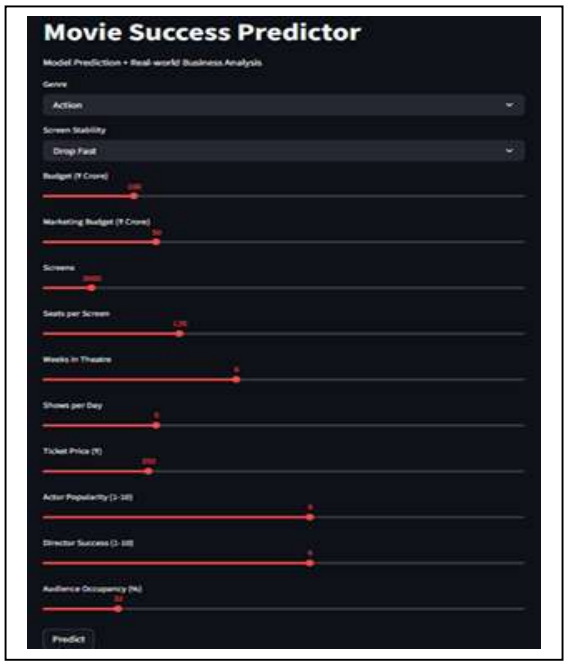

Figure 6. Streamlit-Based User Interface

The dashboard allows users to specify various movie-related attributes including production budget, marketing expenditure, audience occupancy, number of screens, ticket price, actor popularity, and director success rate. After receiving user inputs, the system performs preprocessing, generates a prediction using the trained Random Forest model, and calculates financial outcomes. The final output includes movie classification, prediction confidence, estimated revenue, total expenditure, and profitability assessment. The implementation demonstrates how predictive analytics can be transformed into a practical decision-support tool for real-world applications.

VI. DISCUSSION

The experimental results demonstrate that machine learning can effectively model the complex relationships that influence movie performance. Unlike traditional approaches that rely primarily on intuition and historical assumptions, the proposed framework utilizes data-driven learning to identify patterns among multiple variables simultaneously. One of the major strengths of the proposed system is the integration of predictive analysis with financial evaluation. Existing movie prediction studies frequently focus only on classification outcomes. However, from a business perspective, profitability often plays a more significant role than simple success classification. The achieved accuracy of 86.40% confirms that the selected features contribute meaningfully toward predicting movie outcomes. The high precision value further indicates that the model provides reliable success predictions, reducing the risk of incorrect investment decisions. The Streamlit implementation enhances practical usability by enabling users to experiment with different movie scenarios and observe corresponding outcomes in real time. These findings suggest that the proposed framework can serve as a valuable decision-support tool for producers, distributors, and investors within the film industry.

VII. CONCLUSION

This research presented a machine learning-based framework for predicting movie success and evaluating financial performance within the film industry. The study addressed several limitations observed in existing approaches, including dataset

ISSN: 3107-6513

inconsistency, limited feature representation, and the absence of financial analysis in prediction systems. A structured dataset containing multiple movie-related attributes was utilized to represent production, marketing, distribution, and audience engagement factors.

The Random Forest classifier was employed to model the complex relationships among these variables and generate movie success predictions. Experimental evaluation demonstrated promising performance, achieving an accuracy of 86.40%, precision of 90.00%, recall of 74.00%, and an F1-score of 81.00%. These results indicate that the proposed model can effectively identify patterns associated with movie success and failure. In addition to predictive classification, the integration of financial evaluation provided a broader perspective for decision-making.

By estimating revenue, cost, and profitability, the framework enabled users to assess both predictive and commercial outcomes simultaneously. The Streamlit-based implementation further enhanced the practical value of the research by providing an interactive environment for scenario analysis and experimentation. The developed system demonstrates how machine learning and business analytics can be combined to support strategic decision-making in the entertainment industry.

Overall, the proposed framework provides a reliable and practical approach for movie success prediction and serves as a foundation for future research in predictive analytics for the film sector.

VIII. FUTURE WORK

This research presented a machine learning-based framework for predicting movie success and evaluating financial performance within the film industry. The study addressed several limitations observed in existing approaches, including dataset inconsistency, limited feature representation, and the absence of financial analysis in prediction systems. A structured dataset containing multiple movie-related attributes was utilized to represent production, marketing, distribution, and audience engagement factors.

The Random Forest classifier was employed to model the complex relationships among these variables and generate movie success predictions. Experimental evaluation demonstrated promising performance, achieving an accuracy of 86.40%, precision of 90.00%, recall of 74.00%, and an F1-score of 81.00%. These results indicate that the proposed model can effectively identify patterns associated with movie success and failure. In addition to predictive classification, the integration of financial evaluation provided a broader perspective for decision-making.

By estimating revenue, cost, and profitability, the framework enabled users to assess both predictive and commercial outcomes simultaneously. The Streamlit-based implementation further enhanced the practical value of the research by providing an interactive environment for scenario analysis and experimentation. The developed system demonstrates how machine learning and business analytics can be combined to support strategic decision-making in the entertainment industry.

Overall, the proposed framework provides a reliable and practical approach for movie success prediction and serves as a foundation for future research in predictive analytics for the film sector.

REFERENCES

[1] Breiman, L., "Random Forests," Machine Learning, Vol. 45, No. 1, pp. 5–32, 2001.
[2] Han, J., Kamber, M., & Pei, J., Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufmann, 2011.
[3] Géron, A., Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, 2022.
[4] Goodfellow, I., Bengio, Y., & Courville, A., Deep Learning, MIT Press, 2016.
[5] Provost, F., & Fawcett, T., Data Science for Business, O'Reilly Media, 2013.
[6] Pedregosa, F., Varoquaux, G., Gramfort, A., et al., "Scikit-Learn: Machine Learning in Python," Journal of Machine Learning Research, Vol. 12, pp. 2825–2830, 2011.
[7] Bishop, C. M., Pattern Recognition and Machine Learning, Springer, 2006.
[8] James, G., Witten, D., Hastie, T., & Tibshirani, R., An Introduction to Statistical Learning, Springer, 2021.
[9] Hastie, T., Tibshirani, R., & Friedman, J., The Elements of Statistical Learning, Springer, 2017.
[10] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J., Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.

ISSN: **3107-6513**

[11] Kuhn, M., & Johnson, K., Applied Predictive Modeling, Springer, 2018.
[12] Sharda, R., Delen, D., & Turban, E., Analytics, Data Science, and Artificial Intelligence, Pearson, 2020.
[13] IMDb Dataset Repository, IMDb Developer Documentation.
[14] Box Office Mojo Database, Movie Revenue and Box Office Statistics.
[15] Streamlit Documentation, Streamlit Framework for Interactive Machine Learning Applications.
[16] NumPy Documentation, Numerical Computing with Python.
[17] Pandas Documentation, Data Analysis and Manipulation Tools.
[18] University of Mumbai, Movie Success Prediction Using Machine Learning, M.Sc. Data Analytics Project Report, 2026.