Predicting Box Office Success: A Data-Driven Approach for Movie Performance Forecasting

1st Mohammad Tayyab
Department of Computer Science
University of Engineering & Technology, Lahore
Lahore, Pakistan
tayyabashraf629@gmail.com

2nd Mohammad Ahmad
Department of Computer Science
University of Engineering & Technology, Lahore
Lahore, Pakistan
muhammadahmadmughal0987@gmail.com

3rd Mohammad Saad Akmal
Department of Computer Science
University of Engineering & Technology, Lahore
Lahore, Pakistan
saadakmal460@gmail.com

4th Robass Atif
Department of Computer Science
University of Engineering & Technology, Lahore
Lahore, Pakistan
mohammadrobass@gmail.com

***Abstract*—Movie success is influenced by a multiple factors. Only high budget alone does not guarantee a film's success at the box office. Elements such as genre combination, genre popularity, and box office competition can significantly affect a movie's performance. Understanding these factors can provide valuable insights into predicting a movie's success. In this study, we explore how machine learning can be utilized to predict box office outcomes by considering these various elements.To enhance the prediction of movie success, we incorporated several machine learning models, including Linear Regression, Random Forest, Decision Tree, XGBoost, and Gradient Boosting. Additionally, we employed a Stacked Regressor model combining Gradient Boosting and Random Forest, optimized through Grid Search. Our approach achieved an impressive accuracy of 92%, demonstrating the effectiveness of these models in forecasting movie success.**

I. Introduction

The movie industry is a highly competitive and risky business, with production costs for major films often reaching hundreds of millions of dollars. According to a report by Variety, the average production budget for a Hollywood film in 2022 was around $65 million, while marketing expenses can double this figure [1]. Despite these significant investments, many films fail to recoup their costs at the box office. In fact, it is estimated that nearly 60% of films released in the U.S. do not make a profit [2]. These financial risks highlight the challenges faced by movie studios and production companies in predicting a film's success before its release. Traditional methods of forecasting success, such as relying on star power or budget, are no longer sufficient due to the increasing complexity of audience preferences, competition, and market dynamics. As a result, movie studios are increasingly turning to data-driven approaches, including machine learning and artificial intelligence, to predict box office success more accurately and mitigate financial risks. These technologies have the potential to analyze vast amounts of data, uncover hidden patterns, and provide valuable insights that help filmmakers make more informed decisions.

The role of Artificial Intelligence (AI) and Machine Learning (ML)in predicting movie success has become increasingly significant in recent years. These advanced technologies are used to analyze vast amounts of data and detect patterns that contribute to a movie's performance at the box office [3]. AI and ML models have the potential to identify hidden relationships between various factors such as genre, cast, budget, marketing strategies, and audience reception, which traditional methods might overlook. For example, AI can process historical box office data, viewer preferences, and social media trends to predict a film's likelihood of success before its release [4].

In this paper, we propose a machine learning model to predict the success of a movie at the box office. In this study, we used a dataset that contains various movie attributes such as Genre, Genre Combination, Box Office Competition, Release Date, and Popularity during the release period to predict the movie's financial performance. Various data cleaning and scaling techniques were applied to prepare the dataset. Afterward, different machine learning models were applied to predict the box office success. Our work contributes to predicting a movie's success based on its pre-release attributes and providing valuable insights to movie studios to optimize their investments and marketing strategies.

The contributions of this article includes the following:

1) Identifying crucial factors behind movie sucess.
2) Pre-Processing the dataset for sucess prediction.
3) Apply machine learning algorithms on the prepared dataset to predict movie sucess.
4) Evaluating the model performance.

The rest of the article is organized as follows. Section "Related Work" discusses existing research and methods related to movie success prediction. The preparation of the dataset for movie success prediction is detailed in Section "Preparation

of the Dataset." Section "Predictive Methods" describes the machine learning models and algorithms used for prediction. The results of the predictive models are presented in Section "Results." Finally, conclusions are drawn, and potential future work is discussed in Section "Conclusion and Future Work."

II. Related Work

Several studies have explored the use of machine learning techniques to predict movie box-office success, utilizing a variety of approaches and feature sets. Quader et al. [5] evaluated seven machine learning classification techniques but achieved an accuracy of less than 70%, suggesting significant room for improvement. Their dataset lacked advanced features that could enhance prediction performance. Similarly, Sharda and Delen [6] applied neural networks to predict box-office success but relied on outdated data, limiting their model's applicability to current trends and audience behaviors.

Quader et al. [7] proposed another machine learning approach for predicting movie success, though they did not incorporate newer or more advanced features, reducing the robustness of their model. Their reliance on traditional data sources, such as box-office numbers and ratings, meant that newer influences, such as social media buzz or streaming service data, were not captured. Abidi et al. [8] used machine learning to predict movie popularity but faced challenges related to data sparsity and model accuracy, as their features were not extensive enough to reflect modern trends shaping movie success.

Razeen et al. [9] employed regression techniques for movie success prediction but did not leverage more sophisticated machine learning methods or integrate newer features, limiting the effectiveness of their model. Madongo and Zhongjun [10] proposed a deep multimodal feature-based model, but their model struggled with data sparsity and lacked up-to-date features, raising concerns about its applicability in the modern movie industry. Similarly, Memon and Hussain [11] focused on pre-release features for predicting movie success, yet their model did not account for critical post-release events or real-time audience feedback, making it less reliable for predicting success after a movie's release.

Liao et al. [19] developed a stacking fusion model for early box office prediction in China's film market, achieving moderate success but facing challenges with generalization due to market-specific features. Verma and Verma [13] compared the performance of supervised machine learning algorithms for Bollywood movies, but their approach lacked integration of multimodal features, such as audience sentiment or streaming platform trends, reducing its robustness.

Lee et al. [14] explored the performance of ensemble methods in predicting box office revenue but primarily focused on traditional features like cast and crew, neglecting newer sources of data, such as social media influence. Ahmad et al. [12] worked on predicting pre-production success using YouTube trailer reviews but did not explore deeper multimodal features or post-release data.

Awan et al. [20] proposed a recommendation engine for predicting movie ratings using big data approaches but did not specifically address box office prediction or integrate advanced genre-related features. Similarly, Bogaert et al. [21] examined the impact of social media on box office sales but focused on a cross-platform comparison rather than integrating these influences into a prediction model.

Lastly, Sahu et al. [17] predicted movie popularity and audience targeting using content-based recommender systems, but their approach lacked advanced feature engineering and model optimization, which are crucial for accurate box office predictions.

TABLE I
Comparison of Features

Paper	Genre Combination	Genre Combination Popularity	Box Office Competition
Bogaert et al. (2021) [21]	No	No	No
Awan et al. (2021) [20]	No	No	No
Liao et al. (2022) [19]	No	No	No
Verma and Verma (2020) [13]	No	No	No
Proposed	Yes	Yes	Yes

The comparison of studies reveals significant differences in the features considered for movie success prediction. While earlier works did not examine factors such as genre combination, genre combination popularity, and box office competition, this study identifies these features as critical for improving predictive accuracy. By integrating advanced and up-to-date features, our model addresses the gaps highlighted in prior research and demonstrates the potential for enhanced performance in predicting movie success.

III. Methodology

The methodology incorporates several important steps to ensure correct results. The first of these is the description of the dataset, where we explain the source, size, and main features of the data. The next is data pre-processing, which involves cleaning the data, fixing missing information, and preparing the data for analysis. This is followed by model selection, where we choose the best machine learning or deep learning model for the task by trying different options. Finally, we check which model is better to use with the evaluation of the model in terms of its performance using a R2 score, Mean Square Error and Mean Absolute error.

A. Data Collection

The first step in the methodology is the collection of relevant data. For this study, historical data on movies is gathered from publicly available sources such as IMDb from year 1990. The data spans several years to account for varying market conditions and industry trends.

B. Dataset Overview

The dataset includes features such as movie genre, budget, release date, runtime, and box office revenue. Additionally, popularity, vote average, box office competition, vote count are also included to provide a more comprehensive understanding of the factors influencing a movie's success.

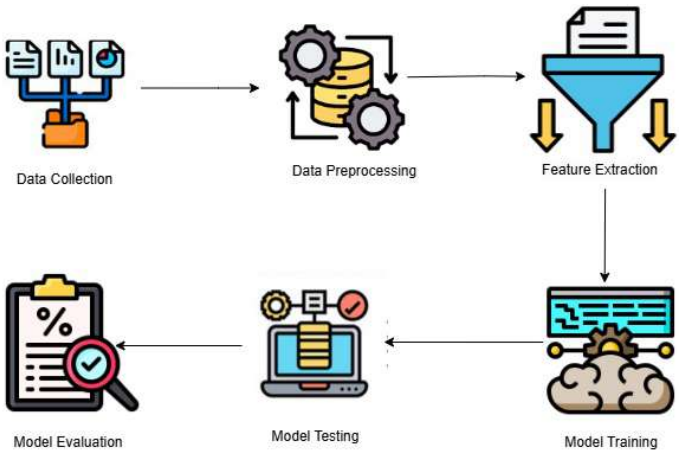

Fig. 1. Proposed Methodology Flow Chart

TABLE II
FEATURE INPUT TABLE FOR MOVIE SUCCESS PREDICTION

Feature	Type	Range/Values	Description
ID	Numerical	Unique ID	Unique identifier for each movie.
Title	Categorical	Movie Name	Title of the movie.
Vote Average	Numerical	0–10	Average rating by users.
Vote Count	Numerical	Integer	Number of votes received.
Status	Categorical	Released, etc.	Movie's release status.
Release Date	Date	YYYY-MM-DD	Official release date.
Revenue	Numerical	0–billions	Box office revenue.
Runtime	Numerical	Minutes	Length of the movie.
Adult	Categorical	Yes, No	Rated for adults.
Budget	Numerical	Millions	Production budget.
Homepage	Categorical	URL/None	Official movie website.
IMDB ID	Categorical	Unique ID	IMDB identifier.
Original Language	Categorical	Language code	Original production language.
Overview	Text	Summary	Plot description.
Popularity	Numerical	0–100	Popularity score.
Genres	Categorical	Action, etc.	Categorization genres.
Production Companies	Categorical	Names	Companies involved.
Production Countries	Categorical	Names	Countries of production.
Spoken Languages	Categorical	Names	Languages spoken.
Keywords	Categorical	Keywords	Related themes.
Release Month	Numerical	1–12	Month of release.
Season	Categorical	Summer, etc.	Release season.
Genre Combination	Categorical	Multiple genres	Combination of genres.
Genre Combination Popularity	Numerical	0–100	Genre combo popularity.
Release Year	Numerical	Year	Year of release.
Box Office Competition	Numerical	0–100	Competition level.

In figure 2, genres combination "Fantasy, Romance, Family" is the most popular, followed by "Action, Science Fiction, Adventure" and "Action, Comedy, Fantasy, Adventure." These genre groupings demonstrate varying audience preferences, with "Fantasy" and "Adventure" genres appearing frequently among the top combinations.

The line chart in figure 3 shows the trend of average movie popularity over the years. The graph remains relatively stable

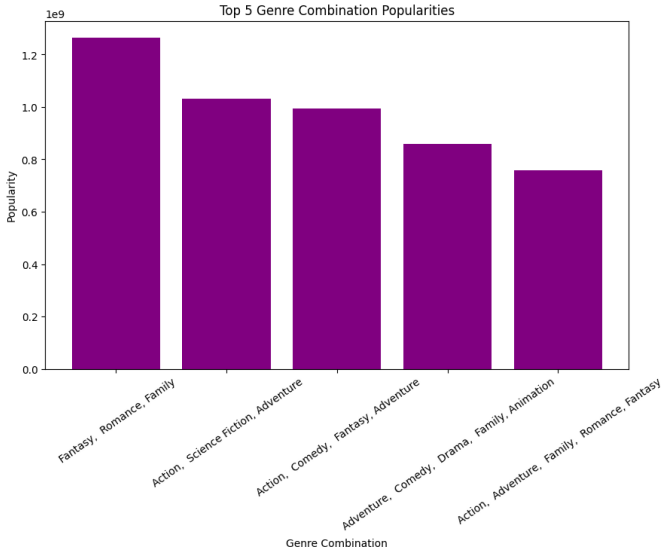

Fig. 2. Top 5 genres combination by popularity.

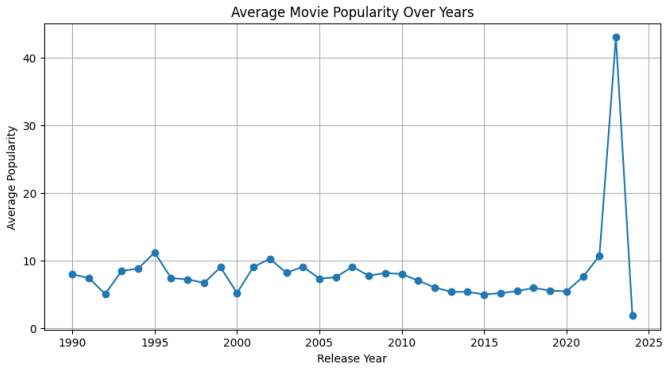

Fig. 3. Movies Popularity over the years.

from 1990 to 2020, with minor fluctuations. However, a sharp spike in popularity is observed around 2023, indicating a significant increase in average movie popularity during that year. Moreover the box office competition also affects the popularity.

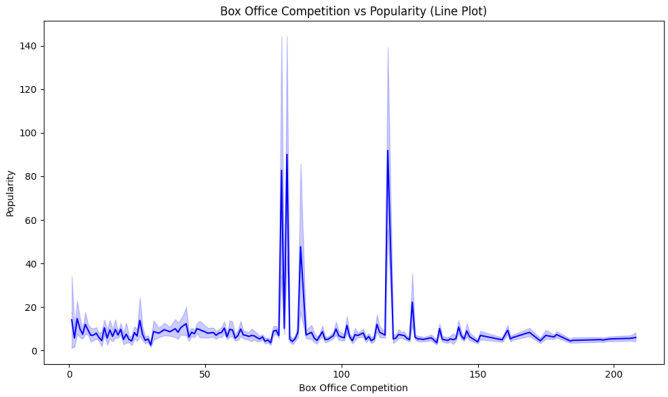

Fig. 4. Movies Popularity over the years.

The figure 4 shows fluctuations in popularity, with most values remaining relatively low and steady. However, there are a few prominent spikes, particularly around the middle of the range, where popularity increases dramatically before quickly dropping back to baseline levels. These spikes likely correspond to specific events or factors driving sudden interest.

C. Data Preprocessing

Data Cleaning: Firstly, missing values in the dataset were handeled to maintain the quality.For numerical features such as popularity, vote counts were replaced with their medians.

Label Encoding: Since machine learning algorithms only compute numerical values he categorical columns were encoded. Features such as movie release season , isAdult level wereencoded ordinaly to preserve their natural ordering for example the seasons Spring, Summer, Fall, Winter were encoded as 1, 2, 3 and 4.

Feature Engineering: Feature engineering was employed to derive new features to improve the predictions. For example box office competition was calculated from the movies released date. Similarly the the genere combination popularity was calculated from the generes and their seperate popularity. These features improved the depth of the dataset and thus contributed in enhancing model performance.

Standardization: To ensure that numerical features are on comparable scale the dataset was standardized.The scaled dataset ensured that features like revenue, budget,popularity and vote average contribute equally to training process preventing any single feature with a larger range from dominating the learning process.

Correlation Analysis: The correlation matrix in figure 5 shows the relationships between various variables, with values ranging from -1 to 1. Strong positive correlations, such as between revenue and vote count or revenue and budget, indicate that higher values in one variable are associated with higher values in the other. Weak or negligible correlations, such as those involving seasons mapped, suggest little to no relationship. Negative correlations, though minimal here, suggest inverse relationships between variables. Overall, the matrix highlights that variables like budget, vote count, and revenue are closely related, while others, such as box office competition and runtime, have weaker or more variable relationships with the rest.

Finally, the dataset was split into training and testing subsets, with 80% of the data allocated for training and 20% for testing. This ensured that model performance could be validated on unseen data, assessing its generalization capabilities.

D. Model Fitting

Linear Regression (LR) refers to a regression algorithm, which falls under the category of statistical supervised machine learning. Its task is to model the linear relationship between dependent and independent variables [22]. The model aims to minimize the residual sum of squares between observed and predicted outputs. This LR model is represented mathematically in Equation 1, where y is the predicted value, b_0 is the bias term, and b_1 is the coefficient used against the input variable x.

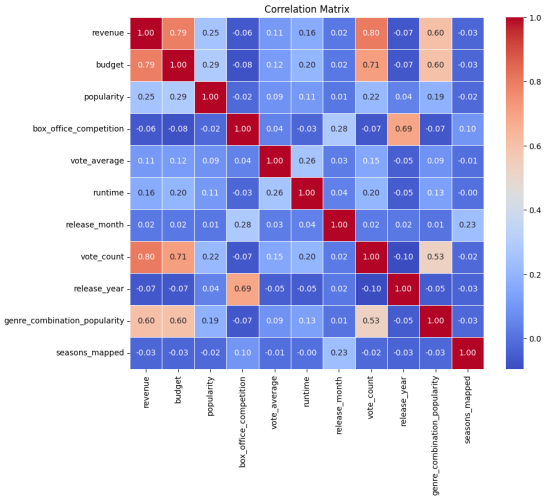

Fig. 5. Correlation Heatmap.

$$y = b_0 + b_1 x \tag{1}$$

Decision Tree Regressor (DTR) is a non-linear regression algorithm that falls under the category of supervised machine learning. It predicts the dependent variable by learning simple decision rules inferred from the data features [23]. The DTR model divides the input space recursively into regions and predicts a constant value for each region. This can be expressed as Equation 2, where R_i represents a region, and c_i is the constant value predicted for that region.

$$y = \sum_{i=1}^{n} c_i I(x \in R_i) \tag{2}$$

Random Forest Regressor (RFR) is an ensemble-based regression algorithm that combines the predictions of multiple decision trees to improve accuracy and reduce overfitting. It predicts the dependent variable by averaging the outputs of all decision trees in the ensemble [24]. This RFR model is mathematically represented in Equation 3, where T is the number of trees and y_t is the prediction from tree t.

$$y = \frac{1}{T} \sum_{t=1}^{T} y_t \tag{3}$$

XGBoost Regressor (XGBR) is a gradient-boosting regression algorithm that builds an ensemble of weak learners (decision trees) by optimizing a regularized loss function. The predictions are obtained by summing the contributions of each tree [25]. This XGBR model is represented mathematically in Equation 4, where f_t is the prediction of the t^{th} tree.

$$y = \sum_{k=1}^{K} f_k(x), \quad f_k \in \mathcal{F} \tag{4}$$

Gradient Boosting Regressor (GBR) is an ensemble regression algorithm that builds models sequentially by minimizing

the residuals of previous models using gradient descent. The final prediction is the sum of the contributions of all weak learners [26]. This GBR model is represented mathematically in Equation 5, where f_t is the prediction from the t^{th} model.

$$y = \sum_{m=1}^{M} h_m(x) \tag{5}$$

A stack regressor is an ensemble learning technique that combines multiple regression models to improve prediction accuracy. The idea behind stacking is to train several base models (often different types of models) on the same dataset, and then use another model (called the meta-model) to combine the predictions of these base models.In a stack regressor, the base models can be varied (e.g., Random Forest, Gradient Boosting, Linear Regression, etc.), and the meta-model is typically a simpler model, such as linear regression, that learns how to best combine the predictions of the base models.

E. Model Evaluation

Once each model was trained on the data, their performances were evaluated based on avariety of metrics, including Mean Square Error (MSE), Mean Absolute Error (MAE), R2 score, and Root Mean Square Error (RMSE), to assess the quality of predictions on the test set.The following table summarizes the performance metrices for each model:

TABLE III
COMPARISON OF MODELS EVALUATION METRICS

Model	**R2**	**MSE**	**RMSE**	**MAE**
Linear Regressor	0.716	2.57×10^{15}	5.07×10^{7}	1.45×10^{7}
Decision Tree Regressor	0.602	3.61×10^{15}	6.00×10^{7}	1.11×10^{7}
XGBoost Regressor	0.572	3.88×10^{15}	6.22×10^{7}	1.57×10^{7}
Random Forest Regressor	0.801	1.80×10^{15}	4.24×10^{7}	7.95×10^{6}
Gradient Boost Regressor	0.790	8.28×10^{7}	9.10×10^{3}	5.87×10^{6}
Stacking Regressor (RF & GB)	0.922	8.28×10^{7}	9.10×10^{3}	5.87×10^{6}

The given tables shows that the Stacking Regressor with Grid Search hyper parameter tuning delivered the highest accuracy among all the models.With an accuracy of 92% , it outperformed other models in term of prediction.

This figure 6 compares predicted values against actual values to evaluate the performance of a predictive model. The red dashed line represents the ideal case where predicted values perfectly match the actual values (y = x). Most data points cluster closely around this line, indicating that the model performs well. However, some outliers deviate from the line, suggesting instances where the predictions differ significantly from the actual outcomes. Overall, the model demonstrates a strong predictive capability with some variability.

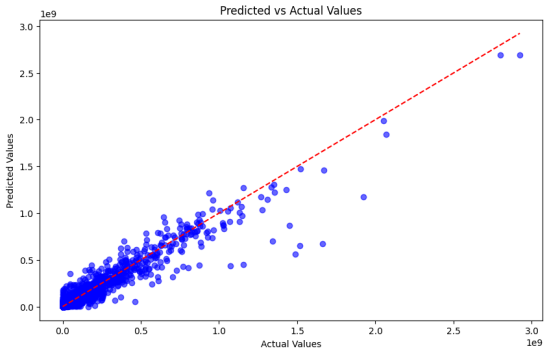

Fig. 6. Predicted vs Actual Values.

IV. CONCLUSION

In this study, we developed a predictive model for forecasting the success of movies at the box office using various features such as genre combination, genre combination popularity, box office competition etc. The results demonstrate the potential of data-driven approaches in the entertainment industry, offering valuable insights for stakeholders to optimize investment strategies and decision-making processes. Through the application of machine learning techniques, the model was able to identify key predictors of success and achieve a reasonable level of accuracy. However, it is important to note that while the model provides useful predictions, external factors such as market conditions, competition, and audience preferences can also influence a movie's performance. Future research can focus on integrating more complex features, including social media sentiment analysis and post-release performance data, to further improve the model's predictive power. Overall, this study contributes to the growing body of knowledge on data analytics in the entertainment sector, showing the potential for further exploration and development in this field.

REFERENCES

[1] Variety. "Average Production Budget of a Hollywood Film." *Variety*, 2022. https://variety.com/2022/08/production-budget-hollywood-film.

[2] Finke, P. "60% of Movies Lose Money at the Box Office." *Deadline*, 2020. https://deadline.com/2020/02/60-percent-of-movies-lose-money-at-the-box-office-1202856490/.

[3] AI in Screen Trade. "The role of AI in predicting box office success and audience preferences." *AI in Screen Trade*, July 17, 2024. https://aiinscreentrade.com/2024/07/17/the-role-of-ai-in-predicting-box-office-success-and-audience-preferences/.

[4] Science. "Artificial intelligence predicts which movies will succeed and fail, simply by the plot." *Science*, 2024. https://www.science.org/content/article/artificial-intelligence-predicts-which-movies-will-succeed-and-fail-simply-plot.

[5] Quader, Nahid, Md Osman Gani, and Dipankar Chaki. "Performance evaluation of seven machine learning classification techniques for movie box office success prediction." *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, pp. 1–6, 2017. IEEE.

[6] Sharda, Ramesh, and Dursun Delen. "Predicting box-office success of motion pictures with neural networks." *Expert Systems with Applications*, vol. 30, no. 2, pp. 243–254, 2006. Elsevier.

[7] Quader, Nahid, Md Osman Gani, Dipankar Chaki, and Md Haider Ali. "A machine learning approach to predict movie box-office success." *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pp. 1–7, 2017. IEEE.

[8] Abidi, Syed Muhammad Raza, Yonglin Xu, Jianyue Ni, Xiangmeng Wang, and Wu Zhang. "Popularity prediction of movies: from statistical modeling to machine learning techniques." *Multimedia Tools and Applications*, vol. 79, pp. 35583–35617, 2020. Springer.

[9] Razeen, Faaez, Sharmila Sankar, W. Aisha Banu, and Sandhya Magesh. "Predicting movie success using regression techniques." *Intelligent Computing and Applications: Proceedings of ICICA 2019*, pp. 657–670, 2021. Springer.

[10] Madongo, Canaan Tinotenda, and Tang Zhongjun. "A movie box office revenue prediction model based on deep multimodal features." *Multimedia Tools and Applications*, vol. 82, no. 21, pp. 31981–32009, 2023. Springer.

[11] Memon, Zulfiqar Ali, and Syed Muneeb Hussain. "Predicting movie success based on pre-released features." *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 20975–20996, 2024. Springer.

[12] Ahmad, Ibrahim Said, Azuraliza Abu Bakar, Mohd Ridzwan Yaakub, and Shamsuddeen Hassan Muhammad. "A survey on machine learning techniques in movie revenue prediction." *SN Computer Science*, vol. 1, no. 4, p. 235, 2020. Springer.

[13] Verma, Hemraj, and Garima Verma. "Prediction model for Bollywood movie success: A comparative analysis of performance of supervised machine learning algorithms." *The Review of Socionetwork Strategies*, vol. 14, no. 1, pp. 1–17, 2020. Springer.

[14] Lee, Sangjae, Bikash KC, and Joon Yeon Choeh. "Comparing performance of ensemble methods in predicting movie box office revenue." *Heliyon*, vol. 6, no. 6, 2020. Elsevier.

[15] Ahmed, Usman, Humaira Waqas, and Muhammad Tanvir Afzal. "Pre-production box-office success quotient forecasting." *Soft Computing*, vol. 24, no. 9, pp. 6635–6653, 2020. Springer.

[16] Wang, Zhaoyuan, Junbo Zhang, Shenggong Ji, Chuishi Meng, Tianrui Li, and Yu Zheng. "Predicting and ranking box office revenue of movies based on big data." *Information Fusion*, vol. 60, pp. 25–40, 2020. Elsevier.

[17] Sahu, Sandipan, Raghvendra Kumar, Mohd Shafi Pathan, Jana Shafi, Yogesh Kumar, and Muhammad Fazal Ijaz. "Movie popularity and target audience prediction using the content-based recommender system." *IEEE Access*, vol. 10, pp. 42044–42060, 2022. IEEE.

[18] Ahmad, Ibrahim Said, Azuraliza Abu Bakar, and Mohd Ridzwan Yaakub. "Movie revenue prediction based on purchase intention mining using YouTube trailer reviews." *Information Processing & Management*, vol. 57, no. 5, p. 102278, 2020. Elsevier.

[19] Liao, Yi, Yuxuan Peng, Songlin Shi, Victor Shi, and Xiaohong Yu. "Early box office prediction in China's film market based on a stacking fusion model." *Annals of Operations Research*, 2022. Springer.

[20] Awan, Mazhar Javed, Rafia Asad Khan, Haitham Nobanee, Awais Yasin, Syed Muhammad Anwar, Usman Naseem, and Vishwa Pratap Singh. "A recommendation engine for predicting movie ratings using a big data approach." *Electronics*, vol. 10, no. 10, p. 1215, 2021. MDPI.

[21] Bogaert, Matthias, Michel Ballings, Dirk Van den Poel, and Asil Oztekin. "Box office sales and social media: A cross-platform comparison of predictive ability and mechanisms." *Decision Support Systems*, vol. 147, p. 113517, 2021. Elsevier.

[22] Scikit-learn. (2023). *Linear Regression*. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[23] Scikit-learn. (2023). *Decision Tree Regressor*. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

[24] Scikit-learn. (2023). *Random Forest Regressor*. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[25] XGBoost. (2023). *XGBoost Documentation*. Available: https://xgboost.readthedocs.io/en/stable/

[26] Scikit-learn. (2023). *Gradient Boosting Regressor*. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html