Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/75/20240555

A refined approach to early movie box office prediction leveraging ensemble learning and feature encoding

Chuang Xie

School of Software, Hefei University of Technology, Hefei, China

tsechong@mail.hfut.edu.cn

Abstract. Predicting the revenue of a movie prior to its release presents a significant challenge. The ability to predict pre-release revenue enables movie production companies to devise effective marketing strategies and mitigate the risks associated with potential box office failures. The primary hurdles in this endeavor stem from managing the myriad factors influencing box office outcomes and accurately forecasting a movie's revenue before it becomes available to the public. To overcome these challenges, we introduce a sophisticated Early Movie Box Office Prediction Model that incorporates Ensemble Learning and Feature Encoding techniques. This model amalgamates multiple foundational models, utilizing regression and decision trees to forecast box office revenues. Our composite model demonstrates superior performance over its constituent models, achieving an impressive accuracy rate of 91.4%.

Keywords: Early Movie Box Office Prediction, Ensemble Learning, XGBoost, GDBT.

1. Introduction

In recent years, the film industry has experienced significant growth, establishing itself as a vital component of contemporary entertainment. The escalating volume of movies produced and released each year has underscored the importance of precise box office predictions for filmmakers, production companies, and distributors. Such predictive insights are essential, enabling stakeholders to make informed decisions regarding marketing strategies, thereby enhancing profitability.

Despite its importance, the accurate prediction of a movie's box office performance prior to its release remains a formidable challenge. Various factors, including pre-sale figures, audience surveys, and social media data, offer predictive insights but are often marred by uncertainties. For instance, audience word-of-mouth and preferences, being highly subjective, are challenging to quantify with accuracy. Additionally, market competition factors such as release timing, genre, and advertising strategies can significantly influence a movie's financial success.

Moreover, existing prediction models each come with their unique set of characteristics and limitations, complicating the task of addressing the multifaceted nature of box office predictions. Linear regression models, for example, presuppose a linear correlation between movie attributes and their box office returns but fall short in capturing the complex interrelations among diverse features. Similarly, Multilayer Perceptron (MLP) models, while adept at learning non-linear relationships, demand extensive data for training and hyperparameter tuning, and are prone to overfitting in the face of limited data, thus failing to discern the fundamental patterns among features. Given the myriad information and

DOI: 10.54254/2755-2721/75/20240555

features derived from movies, models may vary in their effectiveness, excelling with certain types of data while underperforming with others.

This paper proposes an innovative approach that harnesses the predictive power of multiple pre-release movie features, combining machine learning with ensemble learning techniques to create a comprehensive box office prediction model. By integrating the predictions from various models, this approach capitalizes on the strengths of each while mitigating their individual weaknesses. Consequently, this model more adeptly navigates the complex dynamics influencing box office outcomes, delivering predictions with enhanced accuracy.

2. Related Works

The endeavor to forecast a movie's box office performance prior to its release has long posed a substantial challenge. Initial studies primarily leveraged statistical models for predictions. For instance, linear regression emerged as a favored approach due to its ability to interpret the influence of individual factors on a movie's revenue. Mestyán *et al.* employed a linear regression model to predict the opening weekend box office for a collection of 312 movies, integrating multiple movie features[1]. This approach also included an analysis of the movies' popularity through Wikipedia page views and editing activities related to the movies. Dey's work utilized a linear regression model incorporating both post-release variables (such as opening week revenue and number of screenings) and pre-production variables to estimate a movie's final revenue[2]. Similarly, Chakraborty *et al.* applied linear regression to establish a correlation between movie success and factors like IMDb ratings and Tomatometer scores[3]. Verma *et al.* expanded the predictive toolkit by incorporating logistic regression and introducing music ratings as a predictive feature, acknowledging the distinct characteristics of Bollywood cinema[4].

However, the complex, often non-linear relationships between myriad factors and movie box office performance necessitated the adoption of more sophisticated machine learning models. Agarwal *et al.* evaluated the efficacy of diverse models, including machine learning algorithms, time series analysis, and neural networks[5]. Apala *et al.* utilized k-means clustering to categorize movies based on data from Twitter, YouTube, and IMDb, followed by the development of a decision tree classifier prediction model[6]. Galvão and Henriques applied a combination of multi-layer perceptron neural networks, decision trees, and multiple regression analyses for revenue prediction[7]. Quader *et al.*'s studies highlighted the superior performance of multi-layer perceptron models over support vector machines (SVM) in forecasting movie revenues using data from IMDb and Rotten Tomatoes[8], [9].

Notably, ensemble models based on decision trees, such as Gradient Boosting Decision Trees (GBDT) and Random Forest (RF), have demonstrated exceptional accuracy in box office predictions. These models utilize ensemble techniques like bagging and boosting to create multiple decision trees, thereby enhancing prediction accuracy. Liu and Xie applied bagging to develop a larger dataset for training decision tree models[10]. Wu *et al.* used traditional decision tree algorithms alongside RF and GBDT for model construction, with GBDT showing the best performance[11]. Lee *et al.* confirmed that decision tree-based ensemble methods outperformed both linear regression and other non-ensemble techniques[12].

In light of these developments, ensemble learning models are increasingly preferred over singular machine learning models. Yet, many studies still rely on post-release data, such as opening weekend revenue and IMDb indices, for predictions. This paper introduces a novel model that aims for precise pre-release predictions. By selecting a range of pre-release data features—including genres, spoken languages, cast, crew, production, and keywords—and employing One-Hot encoding for variable types, we prioritize the most indicative variables for our predictions[13]. Subsequently, we devise a model rooted in an ensemble framework that employs a voting system to amalgamate multiple regression models. This approach enables the prediction of a movie's box office performance by harnessing pre-release data and aggregating outcomes from various foundational models through voting or averaging.

DOI: 10.54254/2755-2721/75/20240555

3. Methodology

3.1. Feature Selecting and Encoding

Table 1. THE PRIMARY FEATURES OF MOVIES

Feature	Element Type	Description
Budget	Integer	The budget of a movie
Popularity	Float	An index of movie popularity
Runtime	Integer	The length of the movie(minutes)
Genre	String	The genre of a movie
Release Date	Date	The date a movie was first released
Spoken Language	String	The languages spoken in the movie
Cast	String	The cast of a movie
Title	String	The title of a movie
Crew	String	The crew of a movie
Production	String	The information about the production company
Original Language	String	The original language of a movie
Keywords	String	The keywords of a movie

The efficacy of box office revenue prediction can be significantly enhanced through judicious feature selection. The data utilized in this study, as detailed in Table 1, encompass a variety of features that are not inherently numerical, such as movie genres, cast and crew lists, and other attributes represented as strings. To accommodate these multifaceted feature variables, we employ One-Hot encoding[13]. This method involves selecting the most frequent string variables within the dataset, converting these into new binary features—encoded as either 1 or 0—and integrating them into a sparse matrix to represent the feature space effectively. Utilizing a dataset from Kaggle's box office prediction competition as a case study, we conduct feature selection and extraction from diverse data types, including genres, spoken languages, cast, crew, production companies, and keywords[14].

3.1.1. Genres and Keywords. Genres play a pivotal role in identifying a movie's narrative style and target demographic, making them essential for categorizing films. Conversely, keywords, while also integral to the content, offer a granular insight into the specific elements and characteristics of the storyline. In this study, we identify and extract the top 20 genres and the top 30 keywords based on their occurrence frequency within the dataset. These are then sorted and catalogued in Table 2 for use in subsequent analyses. It is important to underscore that these attributes do not have a direct one-to-one correlation with films; a single movie can be associated with multiple genres or possess several keywords, reflecting the complexity and richness of its thematic and narrative elements.

Table 2. TOP 20 GENRES AND TOP 30 KEYWORDS IN THE DATASET

Ranking	Genre	Keyword
1	('Comedy': 2605)	('woman director', 457)
2	('Drama': 3676)	('independent film', 384)
3	('Family': 675)	('duringcreditsstinger', 350)
4	('Romance': 1435)	('based on novel', 312)
5	('Thriller': 1869)	('murder', 305)
6	('Action': 1735)	('violence', 245)
7	('Animation': 382)	('love', 190)
8	('Adventure': 1116)	('revenge', 188)

DOI: 10.54254/2755-2721/75/20240555

Table 2. (continued).

9	('Horror': 735)	('sex', 186)
10	('Documentary': 221)	('aftercreditsstinger', 183)
11	('Music': 267)	('biography', 176)
12	('Crime': 1084)	('sport', 175)
13	('Science Fiction': 744)	('friendship', 168)
14	('Mystery': 550)	('dystopia', 166)
15	('Foreign': 84)	('police', 160)
16	('Fantasy': 628)	('suspense', 159)
17	('War': 243)	('sequel', 158)
18	('Western': 117)	('nudity', 153)
19	('History': 295)	('musical', 147)
20	('TV Movie': 1)	('teenager', 145)
21		('female nudity', 130)
22		('drug', 130)
23		('los angeles', 127)
24		('new york', 123)
25		('prison', 115)
26		('3d', 113)
27		('high school', 111)
28		('family', 111)
29		('alien', 100)
30		('world war ii', 98)

3.1.2. Spoken Languages. The languages spoken in a movie are intricately linked to its cultural context, significantly influencing its reception and market performance post-release. Recognizing this, we identify the top 15 spoken languages from our dataset as key features. The ranking of these languages is delineated in the following Table 3.

Table 3. TOP 15 SPOKEN LANGUAGES IN THE DATASET

Ranking	spoken language
1	('en':6351)
2	('fr':199)
3	('hi':118)
4	('ru':109)
5	('es':95)
6	('ja':90)
7	('it':56)
8	('ko':49)
9	('de':49)
10	('zh':46)
11	('cn':41)
12	('ta':31)
13	('sv':20)
14	('da':17)
15	('pt':13)

DOI: 10.54254/2755-2721/75/20240555

3.1.3. Cast and Crew. The cast and crew are pivotal in the filmmaking process, directly influencing a movie's success. Accordingly, we quantify the presence of actors and various crew members for each movie in our dataset, integrating these figures as features listed in Table 4.

Table 4. THE FEATURES OF THE CREW AND CAST

Feature Number	Feature
1	Gender 0 of Cast
2	Gender 1 of Cast
3	Gender 2 of Cast
4	Art
5	Camera
6	Costume & Make-Up
7	Crew
8	Directing
9	Editing
10	Lighting
11	Production
12	Sound
13	Visual Effects
14	Writing
15	Gender 0 of Crew
16	Gender 1 of Crew
17	Gender 2 of Crew

3.1.4. Production. The capabilities and experience of film production companies can greatly vary, often reflected in the quality of their cinematic outputs. Films produced by companies with a rich history and originating from developed regions are generally perceived as being of higher quality. For example, the globally successful Avengers series was produced by the Walt Disney Corporation, a media conglomerate[15]. Furthermore, the geographical location of a film production company is a critical aspect, given these entities often cater first to their domestic markets. This study, therefore, accounts for both the names of production companies and their geographical locations as detailed in Table 5.

Table 5. TOP 30 PRODUCTION COMPANIES AND TOP 20 PRODUCTION COUNTRIES IN THE DATASET

Ranking	Production Company	Production Country
1	('Warner Bros.', 491)	('United States of America', 5617)
2	('Universal Pictures', 463)	('United Kingdom', 917)
3	('Paramount Pictures', 393)	('France', 570)
4	('Twentieth Century Fox Film Corporation', 341)	('Germany', 411)
5	('Columbia Pictures', 236)	('Canada', 323)
6	('Metro-Goldwyn-Mayer (MGM)', 207)	('India', 220)
7	('New Line Cinema', 198)	('Italy', 160)
8	('Touchstone Pictures', 158)	('Japan', 157)
9	('Walt Disney Pictures', 147)	('Australia', 148)
10	('Columbia Pictures Corporation', 140)	('Spain', 139)
11	('Canal+', 130)	('Russia', 132)
12	('TriStar Pictures', 121)	('China', 99)

DOI: 10.54254/2755-2721/75/20240555

Table 5. (continued).

13	('Relativity Media', 115)	('Hong Kong', 96)
14	('United Artists', 105)	('Belgium', 64)
15	('Miramax Films', 104)	('Ireland', 62)
16	('Village Roadshow Pictures', 89)	('South Korea', 58)
17	('Regency Enterprises', 81)	('Sweden', 50)
18	('DreamWorks SKG', 78)	('Mexico', 44)
19	('Fox Searchlight Pictures', 69)	('Netherlands', 43)
20	('Amblin Entertainment', 68)	('Denmark', 40)
21	('Lionsgate', 68)	
22	('StudioCanal', 65)	
23	('Working Title Films', 63)	
24	('Dune Entertainment', 62)	
25	('Summit Entertainment', 61)	
26	('Dimension Films', 60)	
27	('BBC Films', 56)	
28	('Orion Pictures', 56)	
29	('Hollywood Pictures', 55)	
30	('Fox 2000 Pictures', 52)	

3.1.5. Budget and Revenue. Moreover, the budget and revenue figures from Kaggle's dataset necessitate preprocessing. Notably, there exists a vast disparity in these financial metrics across films—for instance, the difference between a blockbuster and a low-budget comedy can be staggering. Reflecting on this, we categorize movies into five revenue-based classes, from "class 1" to "class 5", as informed by the research conducted by Quader et al. and Delen et al.[9], [16]. As illustrated in Table 6, a movie's classification is indicative of its financial success; a lower class suggests poor box office results, whereas class 5 denotes blockbuster status with substantial popularity. Furthermore, to mitigate potential adverse effects on model training—which could result from the significant disparities in budget and revenue—we apply feature scaling to these variables within the training set. Specifically, we employ logarithmic transformation for both budget and revenue data. This approach aims to reduce noise and anomalies within the dataset, thereby facilitating a more uniform representation of these features, enhancing model stability and prediction accuracy.

Table 6. THE CLASSIFICATION OF MOVIE IN THE DATASET

Movie Class	**Revenue Range (USD)**
1(Flop)	≤ 500,000
2	From 500,000 to 1,000,000
3	From 1,000,000 to 40,000,000
4	From 40,000,000 to 150,000,000
5(Blockbuster)	>150,000,000

3.2. Base Model Selection

In this study, we select five machine learning algorithms as base learners for our ensemble model, aiming to achieve accurate and stable prediction outcomes. This section briefly introduces each selected algorithm and evaluates their predictive performance using uniform data in subsequent analyses.

DOI: 10.54254/2755-2721/75/20240555

3.2.1. Random Forest. Random Forest employs the bagging method of ensemble learning, operating through the training of multiple decision trees and averaging their predictions to yield final results[17]. Each decision tree within the Random Forest model is trained on a randomly chosen subset of the total features, effectively minimizing the impact of data noise. This strategy endows the model with substantial generalization capabilities and robustness, particularly against outliers and anomalies within the dataset.

3.2.2. Gradient Boosting Decision Tree (GBDT). Unlike Random Forest, Gradient Boosting Decision Trees (GBDT) utilize the boosting approach of ensemble learning[18]. This model sequentially trains decision trees, where each tree incrementally corrects errors made by the previous ones, thus optimizing the loss function, minimizing residuals, and successively enhancing overall model performance.

3.2.3. XGBoost. XGBoost stands as an advanced iteration of GBDT, notable for its utilization of the second-order derivatives of the loss function for more precise data fitting and iterative performance improvements[19]. XGBoost further incorporates L1 and L2 regularization techniques, providing a mechanism to effectively manage model complexity through the adjustment of regularization parameters, thus preventing overfitting.

3.2.4. Linear Regression. Linear regression models estimate the relationship between variables by fitting a multidimensional linear equation to the data[20]. It is especially known for its rapid training process and efficiency with smaller datasets. The model is valued for its simplicity and ease of interpretation, making it particularly well-suited for analyzing linear relationships in low-dimensional datasets. Nonetheless, its assumption of linear interdependence between variables may limit its ability to accurately model datasets with prominent non-linear characteristics.

3.2.5. Lasso Regression. Lasso regression extends linear regression by incorporating L1 regularization, which has the effect of reducing certain coefficients to zero[21]. This property enables Lasso to highlight the most significant features, thus simplifying the model and enhancing its performance in high-dimensional spaces. This characteristic makes Lasso regression a powerful tool for feature selection and model simplification in complex datasets.

3.3. Model Construction—Voting Model

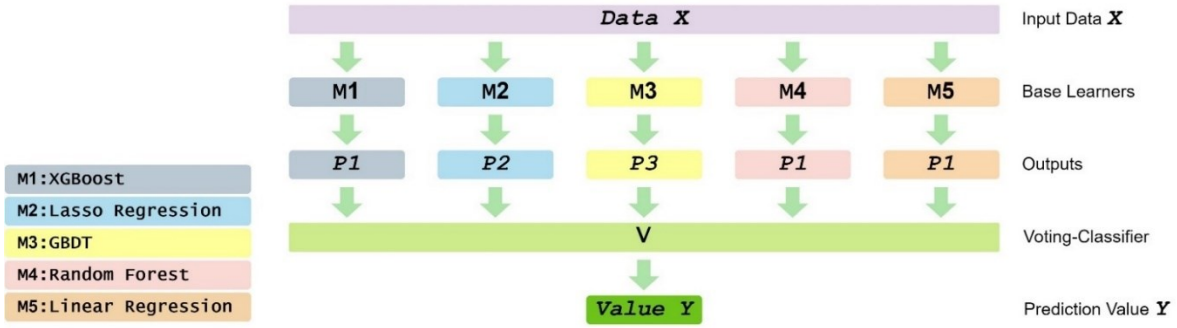

Figure 1. Box Office Prediction Model based on Voting Framework

The Voting model, depicted in Figure 1, embodies a fusion model within a voting framework, integrating five foundational algorithms: XGBoost, Lasso Regression, GBDT, Random Forest, and Linear Regression. Initially, each model **M** independently assesses the input movie features, **Data X**, to produce a distinct prediction result **P**. Subsequently, the voting classifier **V** aggregates these predictions from the five base models, determining the final prediction value **Y** based on the most frequently occurring class among the predictions. The operation of the voting classifier can be succinctly represented as follows:

DOI: 10.54254/2755-2721/75/20240555

$C(x)$ denotes the classification operation on input values, $Mode(x)$ signifies the mode of the classification outcomes, and P_i represents the predicted value from each base model.

$$Y = Mode(C(P_1), C(P_2), C(P_3), C(P_4), C(P_5)) \quad (1)$$

In instances where predictions are evenly split between two classes—e.g., two base models forecast the box office for a movie to be in class B, and another two predict class E—the voting classifier **V** adopts a different approach. As delineated in formula (2), **V** calculates the average of all base model predictions. The movie is then classified according to this averaged value, ensuring a balanced resolution to prediction ties.

$$Y = C(\frac{1}{n}\sum_{i=1}^{n} P_i) \quad (2)$$

Here, **n** represents the total number of base models involved in the prediction process. This methodological approach facilitates a comprehensive and balanced utilization of diverse predictive insights, aiming to enhance the accuracy and reliability of the box office forecast.

3.4. Models Training

3.4.1. 10-Fold Cross-Validation. Given the dataset's scope, the models described in Section 3.3 are trained employing 10-fold cross-validation[22]. This approach guarantees that every data point in the training set undergoes validation, thus ensuring the dataset is fully leveraged. Moreover, it facilitates a precise evaluation of the models' performance by averaging the validation results across all 10 folds.

In this process, illustrated in Figure 2, the dataset is segmented into 10 equally sized partitions. Nine of these partitions are amalgamated to form the training set, while the remaining partition serves as the validation set. This procedure is iterated 10 times, with each iteration featuring a unique combination of training and validation sets. Consequently, all models undergo training and validation across these varied subsets, yielding a comprehensive set of experimental outcomes. Such a methodical approach to training and validation not only minimizes errors attributable to dataset randomness but also bolsters the credibility of the evaluation outcomes.

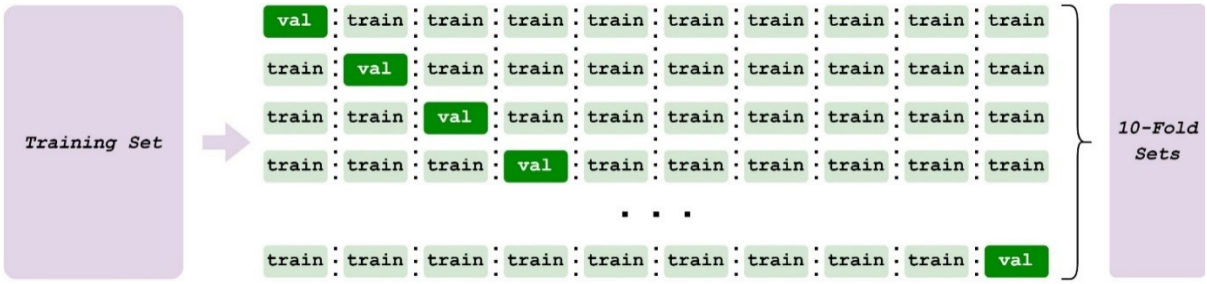

Figure 2. Split Training Set into 10 Folds

3.4.2. Hyperparameter Tuning with Grid Search. The subsequent stage in the training of the base models involves critical hyperparameter tuning and the selection of training methodologies. To enhance each base model's performance and generalizability, we implement a grid search strategy for hyperparameter tuning for the three decision tree-based models: XGBoost, Random Forest, and GBDT[23]. This technique systematically examines every possible combination of hyperparameters to identify the most effective settings for these models. As detailed in Table 7, the parameters adjusted include learning rate, max_depth, and n_estimators. It's important to note that the Random Forest model's hyperparameters do not encompass a learning rate. Furthermore, for the Lasso Regression model, the regularization coefficient is pragmatically set at 0.001, with the maximum number of iterations capped at 500.

DOI: 10.54254/2755-2721/75/20240555

Table 7. THE HYPERPARAMETERS OF BASE MODELS

Model	learning_rate	max_depth	n_estimators
XGBoost	0.1	4	150
Random Forest	-	9	300
GBDT	0.066	3	100

4. Experiments and Results

4.1. Dataset and Performance Metrics

4.1.1. Dataset. In this study, we utilize a dataset sourced from a Kaggle competition focused on box office predictions[14]. The dataset, compiled by the contest organizers, encompasses data for movies released between 1988 and 2018, retrieved from The Movie Database (TMDB). It includes detailed pre-release information and final revenue figures for 3000 movies, which are outlined in Table 1.

For the purpose of our experiments, the dataset is partitioned into a training set and a testing set. The training set comprises 2500 movies selected at random, inclusive of all pre-release features and their respective revenues. The remaining 500 movies are designated as the test set. We employ 10-fold cross-validation on the training set to refine and preliminarily assess the performance of each model. The final phase involves evaluating the voting model and the most effective base models using the test set to determine their predictive accuracy.

4.1.2. Performance Metrics. To accurately gauge the efficacy of the base models and the voting model developed in this research, we utilize the Average Percentage Hit Rate ($APHR$) as our primary metric[9], [24]. This measure calculates the proportion of samples correctly predicted by the model out of the total sample pool. Besides, two supplementary indices, $APHR_{bingo}$ and $APHR_{1-away}$, are introduced to provide a nuanced understanding of the prediction outcomes. $APHR_{bingo}$ refers to the precise classification of movies by the model, while $APHR_{1-away}$ expands on this by including movies whose predicted classifications deviate by no more than one rating class from their actual outcomes.

The calculation for the $APHR$ index is given by the following equation, where **n** denotes the total sample count, **L** the number of levels, and c_i the count of samples accurately classified into class i.

$$APHR = \frac{The\ number\ of\ samples\ accurately\ classified}{Total\ numberof\ samples} \tag{3}$$

$$APHR_{bingo} = \frac{1}{n}\sum_{i=1}^{L} c_i \tag{4}$$

$$APHR_{1-away} = \frac{1}{n}\sum_{i=1}^{L}(c_i + c_{i-1} + c_{i+1}) \tag{5}$$

4.2. Results Analysis

Table 8. ACCURACY OF MODELS IN 10-FOLD VALIDATION

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	**Final**
XGBoost	57.20%	57.60%	62.80%	62.80%	60.40%	58.00%	61.60%	58.40%	56.40%	58.00%	**59.32%**
Lasso Regression	62.40%	54.80%	61.20%	60.00%	57.60%	53.20%	57.20%	56.00%	52.80%	55.20%	**57.04%**
Random Forest	56.80%	60.00%	58.80%	60.80%	60.80%	57.60%	60.00%	60.80%	54.40%	54.40%	**58.44%**
GBDT	57.20%	59.60%	60.00%	62.00%	58.80%	58.00%	60.00%	62.40%	54.40%	56.00%	**58.84%**
Linear Regression	60.80%	53.60%	58.40%	60.00%	58.40%	54.40%	56.80%	54.00%	54.40%	54.40%	**56.52%**
Voting model	59.60%	60.40%	62.00%	62.80%	62.00%	57.60%	60.00%	62.40%	57.20%	56.00%	**60.00%**

DOI: 10.54254/2755-2721/75/20240555

Our evaluation of the model performances utilized 10-fold cross-validation alongside the Average Percentage Hit Rate (APHR) index, as delineated in Section 4.1. The Bingo accuracy of the six models across 10 folds, along with their overall averages, is detailed in Table 8. Among the base models, the XGBoost model distinguished itself with the highest Bingo accuracy, reaching 59.32%. It was closely followed by the GBDT and Random Forest models, which exhibited Bingo accuracies of 58.84% and 58.44%, respectively. Notably, all decision tree-based models outperformed the linear regression-based models. The fusion model, employing a voting mechanism, surpassed all individual base models with an accuracy of 60.0%.

Table 9. ACCURACY OF SINGLE BASE MODELS AND VOTING MODEL

Performance	XGBoost	Lasso	Random Forest	GBDT	Linear	Voting Model
APHR (bingo)	62.40%	59.80%	61.20%	63.40%	56.80%	64.20%
APHR(1-away)	89.40%	86.40%	89.00%	89.60%	85.40%	91.40%

For the testing phase, the base models that demonstrated optimal performance during cross-validation were integrated to configure the ultimate Voting model. This model was subsequently assessed on the test set, with its performance measured against that of the individual base models and the methodologies proposed by Delen and Quader *et al*[9], [16]. As indicated in Table 9, the Voting Model achieved APHR accuracy indices of 64.20% and 91.40%, outstripping all single models and substantiating the efficacy of the voting-based approach. Additionally, Table 10 reveals that the Voting Model's performance excels in comparison to the approaches proposed by other researchers, even without the inclusion of post-release features from the opening weekend.

Table 10. COMPARISON AMONG RESEARCHS

Model	Data Type	BingoRate	1-AwayRate
Voting Model	Pre-release	64.20%	91.40%
MLP Model (Quader et al.)	All Feature	58.50%	89.67%
Hybrid model (Delen and Sharda)	Pre-release	56.07%	90.75%

5. Conclusion

Traditional approaches to box office prediction typically rely on singular models and utilize data released post-movie launch to estimate final revenues. Such methods offer limited utility for film production companies, which necessitate early predictions of a movie's financial performance to craft effective advertising strategies in the pre-release phase. In this study, we introduce a comprehensive fusion model that amalgamates the predictive powers of XGBoost, Random Forest, GBDT, Lasso Regression, and Linear Regression algorithms, utilizing pre-release data for forecasting movie revenues. Testing results affirm the fusion model's superior efficacy over individual models, achieving a notable Bingo accuracy of 64.2% and a 1-Away accuracy of 91.4% in revenue prediction.

Throughout the data preprocessing phase, we meticulously re-encoded six data categories, including genres, crew, and production, to refine the model's predictive accuracy. Nevertheless, this analysis predominantly centers on data associated with the production companies and the movies themselves, leaving potential insights from audience-generated pre-release social media data largely untapped. Future research could benefit from integrating multimodal learning approaches, such as incorporating movie poster visuals or images supplied by production companies, to further enhance the model's predictive accuracy. This direction not only promises to leverage a wider array of data sources but also opens avenues for a more nuanced understanding of audience preferences and market trends, potentially elevating the precision of box office predictions.

DOI: 10.54254/2755-2721/75/20240555

References

[1] M. Mestyán, T. Yasseri, and J. Kertész, "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data," PLoS One, vol. 8, no. 8, pp. 1–13, 2013, doi: 10.1371/journal.pone.0071226.

[2] S. Dey, "Predicting Gross Movie Revenue," Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.03565

[3] P. Chakraborty, M. Zahidur, and S. Rahman, "Movie Success Prediction using Historical and Current Data Mining," Int. J. Comput. Appl., vol. 178, no. 47, pp. 1–5, 2019, doi: 10.5120/ijca2019919415.

[4] G. Verma and H. Verma, "Predicting Bollywood Movies Success Using Machine Learning Technique," Proc. - 2019 Amity Int. Conf. Artif. Intell. AICAI 2019, no. June 2020, pp. 102–105, 2019, doi: 10.1109/AICAI.2019.8701239.

[5] M. Agarwal, S. Venugopal, R. Kashyap, and R. Bharathi, "A Comprehensive Study on Various Statistical Techniques for Prediction of Movie Success," pp. 17–30, 2021, doi: 10.5121/csit.2021.111802.

[6] K. R. Apala, M. Jose, S. Motnam, C. C. Chan, K. J. Liszka, and F. De Gregorio, "Prediction of movies box office performance using social media," Proc. 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2013, no. October 2015, pp. 1209–1214, 2013, doi: 10.1145/2492517.2500232.

[7] M. Galvão and R. Henriques, "Forecasting Movie Box Office Profitability," J. Inf. Syst. Eng. Manag., vol. 3, no. 3, 2018, doi: 10.20897/jisem/2658.

[8] N. Quader, M. O. Gani, D. Chaki, and M. H. Ali, "A machine learning approach to predict movie box-office success," 20th Int. Conf. Comput. Inf. Technol. ICCIT 2017, vol. 2018–Janua, pp. 1–7, 2018, doi: 10.1109/ICCITECHN.2017.8281839.

[9] N. Quader, M. O. Gani, and Di. Chaki, "Performance evaluation of seven machine learning classification techniques for movie box office success prediction," 3rd Int. Conf. Electr. Inf. Commun. Technol. EICT 2017, vol. 2018–Janua, no. December, pp. 1–6, 2018, doi: 10.1109/EICT.2017.8275242.

[10] Y. Liu and T. Xie, "Machine learning versus econometrics: prediction of box office," Appl. Econ. Lett., vol. 26, no. 2, pp. 124–130, 2019, doi: 10.1080/13504851.2018.1441499.

[11] S. Wu, Y. Zheng, Z. Lai, F. Wu, and C. Zhan, "Movie box office prediction based on ensemble learning," ISPCE-CN 2019 - IEEE Int. Symp. Prod. Compliance Eng. 2019, no. July, pp. 1–4, 2019, doi: 10.1109/ISPCE-CN48734.2019.8958631.

[12] S. Lee, B. KC, and J. Y. Choeh, "Comparing performance of ensemble methods in predicting movie box office revenue," Heliyon, vol. 6, no. 6, p. e04260, 2020, doi: 10.1016/j.heliyon.2020.e04260.

[13] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," J. Mach. Learn. Res., vol. 3, no. 6, pp. 1137–1155, 2003, doi: 10.1162/153244303322533223.

[14] "TMDB Box Office Prediction: Can you predict a movie's worldwide box office revenue?" Accessed: Oct. 12, 2023. [Online]. Available: https://www.kaggle.com/c/tmdb-box-office-prediction

[15] L. Jiang, "Research on Marvel Studios' Product Marketing Strategy in the New Media Environment," SHS Web Conf., vol. 181, no. 1, p. 04009, Jan. 2024, doi: 10.1051/shsconf/202418104009.

[16] D. Delen and R. Sharda, "Predicting the Financial Success of Hollywood Movies Using An Information Fusion Approach," Endüstri Mühendisligi Derg., vol. 21, no. 1, pp. 30–37, 2010.

[17] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[18] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, Mar. 2001, [Online]. Available: http://www.jstor.org/stable/2699986

DOI: 10.54254/2755-2721/75/20240555

[19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13–17–Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
[20] Y. Sun, X. Wang, C. Zhang, and M. Zuo, "Multiple Regression: Methodology and Applications," Highlights Sci. Eng. Technol., vol. 49, pp. 542–548, 2023, doi: 10.54097/hset.v49i.8611.
[21] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," J. R. Stat. Soc. Ser. B, vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.
[22] D. Berrar, "Cross-validation," Encycl. Bioinforma. Comput. Biol. ABC Bioinforma., vol. 1–3, no. April, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.
[23] X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," Knowledge-Based Syst., vol. 212, no. Dl, 2021, doi: 10.1016/j.knosys.2020.106622.
[24] Y. Liao, Y. Peng, S. Shi, V. Shi, and X. Yu, "Early box office prediction in China's film market based on a stacking fusion model," Ann. Oper. Res., vol. 308, no. 1–2, pp. 321–338, 2022, doi: 10.1007/s10479-020-03804-4.