Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/73/20240419

A stacking ensemble model based on Elastic Net and Random Forest for box office prediction

Weichen Fu

Institute of International Education, Guangdong University of Finance, 527 Guangzhou, 510000, China

631402070301@mails.cqjtu.edu.cn

Abstract. The rapid growth of the film industry has made accurate box office predictions crucial. A reliable prediction model can assist producers and insiders in optimizing resources by adjusting strategies based on forecast outcomes. This paper aims to compare the predictive capabilities of various models and proposes a stacking ensemble learning-based box office prediction model. We start by selecting relevant IMDB data factors such as movie duration, director details, cast lineup, and movie region for data cleaning and screening, thereby enhancing the input for box office predictions. Subsequently, we construct multiple forecasting models for comparison purposes and select the most superior ones as base models. We then create a secondary comparison stacking model. Furthermore, we utilize 10-fold cross-validation to fine-tune the model parameters, resulting in more precise evaluation results. Experimental findings demonstrate that XGBoost outperforms other single models, while different stacking ensemble models exhibit notable performance variations. The proposed stacking ensemble model, based on Elastic Net and Random Forest models, yields the best predictive outcome.

Keywords: box office prediction, ensemble learning, Elastic Net, Random Forest

1. Introduction

The past two decades have witnessed rapid growth in the global film industry, contributing significantly to the global economy. Despite the expansion of the film market, box office predictions before a movie's release have remained a topic of interest both within and outside the industry. With slowing audience growth rates and complex high-risk factors involved in production processes, accurate box office predictions are crucial. Previous studies have explored the relationship between social media activity and future economic performance, and similar approaches can be applied to box office forecasts by extracting data from pre-release social media-related activities. However, current box office prediction methods have limitations in two main aspects. Firstly, most methods rely on datasets that are only available a few weeks before a movie is released, which limits their effectiveness. For example, using internet popularity or search volume as evaluation criteria only becomes feasible when film production is nearing completion. At this stage, marketing strategies for the film have already been completed, and time passing may increase costs and extend innovation cycles for film projects. Secondly, producers' excessive reliance on box office data, coupled with misjudging audience preferences and needs, increases the difficulty of accurately predicting movie performances.

DOI: 10.54254/2755-2721/73/20240419

Therefore, it becomes crucial to predict the economic performance of a movie at an early stage. A reliable box office prediction model can not only provide early predictions during the initial stages of production, helping industry insiders identify potential problems earlier, but also allow producers more time and resources to adjust and optimize strategies to minimize losses or maximize profits. This facilitates rational resource planning and optimization.

The common method for predicting movie box office revenue is to establish a regression model, typically using linear or nonlinear regression. By analyzing and selecting key variables or indicators that affect box office revenue, predictions can be made. Another prediction method is based on machine learning and neural network models. These methods employ machine learning algorithms such as support vector machines, decision trees, random forests, and neural networks to improve the accuracy and stability of predictions through model learning and training. Additionally, multimodal analysis methods, such as using image recognition technology to extract features from movie posters or integrating multiple data sources for comprehensive analysis, can enhance prediction accuracy. However, these methods have limitations. While they employ advanced machine learning and deep learning models, box office revenues are influenced by various factors such as film quality, cast lineup, promotional efforts, release timing, and word-of-mouth spread. The interrelationships among these factors are complex, leaving room for improvement in terms of prediction accuracy. Moreover, the generalization ability of the models varies, with some performing well for certain types of films but not generalizing well to other types or markets. Furthermore, many predictive models primarily rely on statistical and machine-learning approaches, often neglecting prior industry knowledge and audience subjective evaluations. They also rarely consider dynamic changes after a film's release, such as word-of-mouth spread and emotional feedback from audiences, which could impact prediction accuracy.

This article proposes a stacking-based ensemble learning model that significantly improves prediction accuracy by screening based on the prediction effect of a single model. This model enables early box office predictions for movie production, which holds significant practical value for film producers, investors, and distributors. By making early box office predictions during film production, they can make more informed decisions to avoid potential investment mistakes and save millions of dollars. This model utilizes various movie-related factors, such as movie duration, director information, and cast lineup, to predict box office performance in advance.

2. Related Works

2.1. Regression Based Models

In the field of movie box office prediction, several regression-based models have been proposed. W. Lu et al. introduced an improved fruit fly optimization algorithm (IFOA-GRNN) to establish a generalized regression neural network model. By analyzing various factors affecting the box office, multiple classification models were established. The IFOA-GRNN method showed increased accuracy and precision compared to traditional machine learning methods, with an average improvement of about 20% [1]. D. Choudhery et al. utilized tweets on the Twitter platform to discover useful patterns and predict box office revenues of movies. They employed a polynomial regression model to analyze tweet sentiment and reduced mean square error by selecting an appropriate order of polynomials while avoiding overfitting the data [2]. T. Liu et al. explored the balance between minimizing model parameters and regularization terms using the linear regression model and support vector regression model [3]. Y. Liao et al. and others employed linear and nonlinear regression models based on the wisdom of social media users to predict box office revenue. They utilized linear regression and support vector regression to predict box office revenue before a movie's release and evaluated the effectiveness through cross-validation experiments. The results showed that there is a correlation between large-scale social media content and box office revenue, and user purchase intentions can more accurately predict box office income [4]. S. Lee et al. applied an ensemble method involving decision trees, k-NNs, and linear regression to predict Korean movie box office figures based on Korean film data. They compared this with non-ensemble methods and confirmed the superiority of the former. The results showed that

DOI: 10.54254/2755-2721/73/20240419

using an ensemble method with decision trees provided more accurate predictions for box-office revenue [5].

2.2. Neural Networks Based Model

C. T. Madongo et al. proposed a neural network model based on multimodal feature classifiers, which predicts box office revenue by learning and extracting different features comprehensively. This method utilizes movie posters and related data to estimate the box office income of movies through neural network algorithms, thereby improving the accuracy of classification and prediction [6]. Y. Ru et al. employed the LSTM model and Deep-DBP model for film box office prediction, verifying their effectiveness and predictive accuracy through experiments. These models consider multiple factors, including dynamic and static data, and make predictions in the early stages after a movie's release to provide more accurate box office forecast results [7]. J.-Y. Chang collected static and dynamic movie-related data and generated and evaluated a model on the impact of box office performance using Naive Bayes classification and neural network models [8].

2.3. Research Using Specific Models and Special Statistical Methods

L. Liao et al. used the Elaboration Likelihood Model (ELM) as a theoretical basis and conducted empirical analysis using data from 304 movies. They differentiated between the factors influencing purchase intention and the different stages of purchase decision-making [9]. Y. Zhou et al. employed a multimodal DNN model to predict movie box office revenue, which includes a CNN for extracting features from movie posters. This method improves the predictive performance of movie box office revenue [10]. L. Jiang et al. used correlation analysis and machine learning models to study the relationship between film elements and box office/audience ratings. They achieved predictions of movie box office and audience ratings through machine learning methods [11]. K. Yan conducted research using the IMDB movie dataset, employing machine learning models such as decision trees and random forests for experimentation and analysis. The results confirmed the usability and transparency of the random forest algorithm. Additionally, the rating range on the IMDB website was predicted [12].

L. Haiyan et al. used factor analysis and Pearson correlation coefficient as the main research methods. They identified the potential factors affecting movie box office through factor analysis and quantified the linear correlation between movie box office and various indicators using Pearson correlation coefficient [13]. N. D. Roy et al. mentioned two types of fare prediction models based on machine learning algorithms and weighted average ensemble. The experimental results showed that the prediction system based on machine learning algorithms performs well in flight fare prediction, and the weighted average ensemble model can further improve the accuracy of prediction [14]. V. Gupta et al. utilized ensemble learning algorithms, specifically Random Forests and XGBoost, to predict movie ratings and box office revenues using data from IMDb, YouTube comments, and Wikipedia. They evaluated the accuracy of their models through confusion matrices and ROC curves, finding that gradient boosting performed best with a success rate of 84.1297% [15]. K. Lee et al. employed an ensemble learning approach called Cinema Ensemble Model (CEM) to forecast movie box office results. They found that decision trees had the best predictive performance within the ensemble methods and introduced new features based on cross-media narrative theory [16]. S. Sahu et al. used a k-fold hybrid deep ensemble model (k-HDEM) to analyze raw Indian film data obtained from IMDb in order to predict movie popularity levels. They introduced newly derived features to enhance model performance and constructed a new Indian film dataset upon which they proposed an effective feature extraction method for improving model performance [17]. S. Wu et al. built a movie box-office prediction model using machine-learning techniques such as ensemble learning algorithms and decision trees. They demonstrated that Gradient Boosting Decision Tree (GBDT) significantly outperformed other models and was superior in predicting cinema ticket sales compared to traditional machine-learning algorithms [18].

D. Carr analyzed and predicted movie box office revenues by employing Support Vector Machines (SVM) and Random Forest (RF). The study found that analyzing script data took longer, but current

DOI: 10.54254/2755-2721/73/20240419

prediction methods did not consider the order of words in scripts, which could lead to inaccuracies [19]. D. P. O and J. Paik used Random Forests and Gradient Boosting models for prediction. They collected multi-platform data to train and improve their models, gradually improving overall accuracy by giving higher weightage to previously incorrectly predicted data [20].

3. Methodology

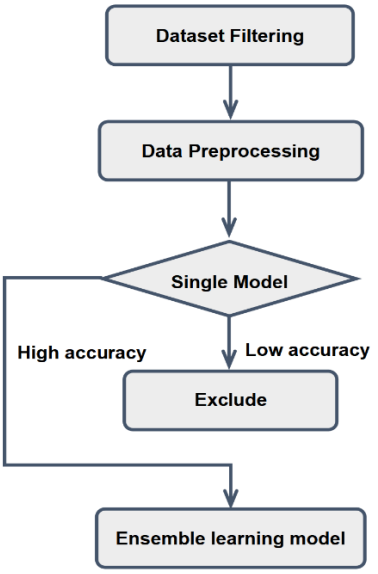

Figure 1. Research Method Process.

The research method process is shown in Figure 1. The methodology consists of several steps outlined below:

3.1. Data Refinement and Preprocessing

We utilized the IMDB dataset for this research, which initially had some missing elements that could potentially compromise the accuracy of our predictive model. To address this issue, we conducted a thorough cleaning and supplementation of movie attributes such as genre, director, lead actors, budget, release time, duration, and ratings, among others. We logically filled in missing values and managed outliers to ensure data quality. Additionally, we applied logarithmic transformations to income and budget, two numerical features, to reduce the impact of disparities. These steps resulted in a more complete and accurate movie dataset.

3.2. Model Enhancement

After selecting the database and refining the data, we used Mean Squared Error (MSE), Standard Deviation (STD), and R^2 as evaluation metrics. We implemented 10-fold cross-validation at each stage to impartially select the most effective box office prediction model. Based on the evaluation results, we made adjustments to feature grouping within models and optimized parameters.

The Lasso function was employed to facilitate feature selection by setting certain parameters to zero through the addition of an L1 regularization term. This regularization term limits the parameters and helps simplify the model. The strength of regularization is controlled by the L1 regularization coefficient: a larger coefficient increases the regularization power, thus simplifying the model, and vice versa.

DOI: 10.54254/2755-2721/73/20240419

Elastic Net further refines the parameter range by using Bayesian optimization to find the optimal parameter combination. Bayesian optimization is a global strategy that efficiently identifies ideal hyperparameter settings. The hyperparameter settings for the Elastic Net model are shown in Table 1. Similarly, we obtained hyperparameter settings for the random forest model.

Table 1. Elastic Net Parameter Configuration

parameter	implication	Valid values
alpha	Strength of the Regularization Term	11
l1_ratio	The weight ratio between L1 and L2 penalty terms	0.7
max_iter	Maximum number of iterations for the optimal solution	1050
tol	Conditions for Stopping Iteration	0.0001
normalize	Deciding whether to normalize the training set	True

Table 2. Random Forest Parameter Configuration

parameter	implication	Valid values
n_estimators	Number of Decision Trees	150
max_features	Maximum number of features used for splitting in each decision tree	None
bootstrap	Sampling with or without replacement	True
max_depth	Maximum depth of the decision tree	None
min_samples_split	Minimum number of samples required for node splitting	2
min_samples_leaf	Minimum number of samples required at a leaf node	1

3.3. Single Model Comparison

We tested six popular machine learning models: Elastic Net, XGBoost, Random Forest (KNN), Kernel SVM, and LightGBM. Each model was evaluated individually using Mean Squared Error (MSE) and Standard Deviation (STD).

- Elastic Net combines the strengths of Ridge and Lasso models to handle multicollinearity issues.
- XGBoost is robust against outliers and demonstrates high prediction accuracy for large-scale datasets.
- KNN effectively handles outliers.
- Random Forest manages high-dimensional features without feature selection while effectively preventing overfitting.
- Kernel SVM solves non-linear problems and achieves good classification results.
- LightGBM processes high-dimensional features without requiring feature selection.

Each model has unique attributes that make it suitable for different data tasks.

3.4. Ensemble Learning Models

Based on the performance of the single models, we created eight ensemble models: 'Elastic Net & XGBoost', 'Elastic Net & Random Forest', 'Elastic Net & LightGBM', 'Random Forest & XGBoost', 'Random Forest & KNN', 'XGBoost & KNN', 'LightGBM & Random Forest', and 'LightGBM & XGBoost'. Comparing these combinations provides a comprehensive understanding of the various model performances in box office prediction.

These combinations merge multiple advantages from individual models. For example, Elastic Net uses L1 and L2 regularization to select features, effectively dealing with multicollinearity issues. Both

DOI: 10.54254/2755-2721/73/20240419

XGBoost and LightGBM demonstrate very high prediction accuracies. Additionally, Random Forest effectively handles numerous features and demonstrates strong noise robustness, while KNN is suited for some nonlinear data analysis tasks.

After combining the base models into new stacked ones, we used the lasso model for second-layer modeling. The lasso model better utilizes weak learner information compared to linear regression models. For instance, although Elastic Net is a linear model, it pairs well with nonlinear models like XGBoost or Random Forest, helping the box office prediction models capture more complex data patterns. This leads to more stable and accurate predictions compared to single models.

For example, the MSE of the individual Elastic Net and Random Forest models were 2.55081 and 4.6141, respectively, while the STD was 0.14948 and 0.7572, respectively. However, when these two models were integrated into an ensemble model, the MSE dropped to 1.4717 with an STD of 0.12842. This indicates that ensemble learning improved performance by approximately 42% and 68%, respectively. Although XGBoost had the best predictive effect among the single models, its predictive results did not improve when used as a base model for ensemble learning.

4. Experiments and Results

4.1. Dataset

The TMDB dataset ("The Movie DataBase") is a large database containing movie information, including basic descriptions of movies, cast and crew information, trailers, ratings from audiences and critics, box office revenue, etc. This dataset is rich in content, providing specific information about directors, actors, duration, genre, budget, language, release date, and more for each movie. Due to the richness of its data content, it plays an important role in tasks such as predicting box office performance, analyzing movie genres, or studying viewing trends. This article conducts a study on the performance of a box office prediction model based on this dataset.

4.2. Evaluation Metrics

MSE (Mean Squared Error)

MSE stands for Mean Squared Error, which is a measure used to quantify the difference between predicted values and actual values. The calculation method of MSE is to take the square of the prediction error and then calculate the average. The smaller the error, the smaller the value of MSE, indicating higher accuracy of the prediction model. In machine learning and statistical model predictions, MSE is a commonly used evaluation metric.

The formula (1) for MSE is as follows:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2 \tag{1}$$

Where θ is the vector of observed values of the variable being predicted, with θˆbeing the predicted values.

R^2 (Coefficient of Determination)

R^2, also known as the coefficient of determination, is a statistical indicator used in statistics to evaluate the goodness of fit of a model. Its value ranges from 0 to 1, representing the proportion of the squared correlation between predicted and actual values of the target variable to the total sum of squares. The closer R^2 is to 1, the stronger the explanatory power of the model and the higher the prediction accuracy. Conversely, if R^2 is close to zero, it means that the prediction performance is not satisfactory. The formula (2) for R^2 is as follows.:

$$R^2 = 1 - \left(\frac{\mathrm{SSR}}{\mathrm{SST}}\right) \tag{2}$$

Where:

SSR (Sum of Squared Residuals) represents the deviation between the model prediction and the actual observed values.

DOI: 10.54254/2755-2721/73/20240419

SST (Total Sum of Squares) represents the deviation between the observed values and their average.

In this article's forecast, mean square error (MSE) will be used together with R^2 R^2 to increase prediction accuracy.

4.3. Baseline

Traditional prediction models include the following types.

Elastic Net: Elastic Net is a linear regression model that incorporates L1 and L2 regularization to prevent overfitting. It combines the characteristics of Ridge and Lasso Regression, using an adjustment parameter for balance. Ridge Regression restricts large absolute values of model coefficients by adding their square (L2) to the loss function, while Lasso Regression uses the sum of absolute values (L1) of these coefficients. Elastic Net's feature selection capability makes it suitable for handling high-dimensional data with many features.

XGBoost (Extreme Gradient Boosting): XGBoost is an ensemble model developed by Tianqi Chen and Carlos Guestrin from the University of Washington. It is based on the gradient boosting decision tree algorithm. XGBoost starts with a simple prediction and iteratively improves it by calculating and updating each sample's residuals with new decision trees until a specific performance threshold or maximum iterations are reached. It employs second-order gradient optimization along with L1 and L2 regularization, similar to GBDT (Gradient Boosted Decision Tree). XGBoost is known for its high predictive power and execution speed, especially for large datasets.

KNN (K-Nearest Neighbors): KNN is a supervised learning algorithm used for classification and regression problems. It calculates distances between an unknown sample and all dataset samples, selects K points with the smallest distances, and classifies the unknown sample based on the majority class among these K points. Different K values can significantly impact predictions: small values may lead to overfitting due to complexity, while large values may result in underfitting due to simplicity. KNN is suitable for multi-classification problems but has drawbacks such as imbalanced samples causing prediction bias and high storage and computational costs for large datasets.

Random Forest: Random Forest is an ensemble model consisting of multiple decision trees. Each sub-tree is formed based on randomly selected variables and optimal splitting methods. The results of the individual trees are relatively independent until no more child nodes can be constructed at some node. Averaging the results improves prediction accuracy, but Random Forest is sensitive to outliers.

SVM (Support Vector Machine): SVM is primarily used for classification and regression analysis. Kernel SVM, in particular, excels at handling complex data that involves linearly inseparable problems. By mapping samples from the original space into a higher-dimensional feature space using kernel functions, linear SVM finds an optimal hyperplane to achieve sample classification. This allows SVM to make non-linearly separable issues become linearly separable in high-dimensional spaces. Kernel SVM offers flexibility, sparsity, and good generalization ability by adapting different types of data and problems through the choice of different kernel functions.

LightGBM: LightGBM enhances performance through techniques like GOSS (Gradient-based One-Side Sampling) and EFB (Exclusive Feature Bundling). GOSS retains samples with larger gradients while keeping some smaller gradient samples for data distribution consistency. EFB increases training speed and reduces memory consumption by bundling mutually exclusive features, thereby decreasing the number of features.

In a typical experiment, definitions and environmental interpretations are used. During the preprocessing of the dataset, there is no specific need to group the data. In this study, during data cleaning, factors such as whether the data was 'produced in China', 'English spoken', or 'used young actors extensively' were identified. These factors were found to significantly affect box office predictions. The study draws some results based on these factors:

1. The film's language can influence its global acceptance;
2. The age range of actors may attract specific audience demographics;
3. The film's country of origin fundamentally impacts its box office performance.

DOI: 10.54254/2755-2721/73/20240419

Introducing these features enhances model interpretability and helps understand which factors might impact box office revenue and their relative influences better. Moreover, identifying these features could improve prediction accuracy to some extent since standard machine learning models often struggle to capture such influential factors from raw data. However, explicitly incorporating these elements into predictive data could increase predictive precision slightly.

After applying stacking ensemble learning, some ensemble models showed superior performance on metrics such as Mean Squared Error (MSE), with lower standard deviation (STD) indicating more consistent metric values.

4.4. Results

This study employed a range of individual and multiple machine learning models, including Stacking, to forecast film box office earnings. The stacking model demonstrated superior predictive accuracy when trained and validated on the TMDB dataset. Feature processing played a crucial role in this research, with different feature classification filters significantly enhancing model prediction accuracy and minimizing the risk of overfitting. Key factors influencing box office performance include movie budget, director, lead actors, genre, and release date.

Model performance was evaluated using Mean Squared Error (MSE) and R^2 as benchmarks. Each result underwent 10-fold cross-validation, with MSE indirectly derived from Root Mean Square Error (RMSE). As shown in Table 3, among the single models, XGBoost stood out for its predictive power with an MSE of 2.27929 and an R^2 value of 0.4415 on validation datasets.

As shown in Table 4, a Stacking ensemble model utilizing Elastic Net and Random Forest as base models delivered superior predictive results on validation datasets with an MSE score of 1.4717 and an R^2 value of 0.8429. Elastic Net also performed well among single models after XGBoost, followed by LightGBM and Random Forest. Stacking ensembles based on Elastic Net and XGBoost showed higher prediction accuracies compared to their base models, demonstrating better predictive abilities than standalone models.

However, one drawback of ensemble learning is that some low-accuracy models may noticeably affect high-accuracy ones, thereby reducing overall prediction effectiveness. Moreover, certain Stacking ensembles may struggle to leverage the high-accuracy advantages of their base models, resulting in drops in post-ensemble prediction accuracies, possibly due to overfitting issues.

In conclusion, stacking ensemble models prove to be more accurate in predicting movie box office revenues.

Table 3. Single model prediction results

	Elastic Net	XGBoost	KNN	Random Forest	Kernel SVM	LightGBM
MSE	2.55081	2.27929	7.1020	4.6141	9.5847	4.4300
STD	0.14948	0.14678	1.2773	0.7572	1.2747	0.8389
R^2	0.3027	0.4415	0.2414	0.5068	-0.0248	0.5275
STD	0.0241	0.0534	0.0852	0.0494	0.0222	0.0557

DOI: 10.54254/2755-2721/73/20240419

Table 4. Stacking integrated model prediction results

	Elastic Net and XGBoost	Elastic Net and Random Forest	Elastic Net and LightGBM	Random Forest and XGBoost	Random Forest and KNN	XGBoost and KNN	LightGBM and Random Forest	LightGBM and XGBoost
MSE	2.28016	1.4717	4.6208	8.71324	8.7295	8.7613	8.6873	8.6927
STD	0.13871	0.12842	0.9039	1.72506	1.7447	1.6041	1.7790	1.7706
R^2	0.4419	0.8429	0.5086	0.07	0.0700	0.0657	0.0747	0.0741
STD	0.0400	0.079	0.0472	0.13	0.1366	0.1171	0.1424	0.1410

5. Conclusion

This paper introduces a box office prediction model developed for the film industry, utilizing stacked ensemble learning with Elastic Net and Random Forest algorithms. The study begins by sorting and filtering various factors from IMDB, including movie duration, director details, cast lineup, and movie region. Six individual models are then employed to predict box office performance, and their Mean Squared Error (MSE) is compared to evaluate their accuracy. Among these models, XGBoost performs the best, with an MSE of 2.27929 and an R^2 value of 0.4415 on validation data sets, demonstrating its reliability and precision. Based on the accuracy rankings of the single models, the top performers are selected as base models to create a second comparison stacking model using Lasso as the second layer method. The results show that employing Elastic Net and Random Forest as base models in the stacking ensemble model yields higher accuracy, with an MSE of 1.4717 and an R^2 value of 0.8429 on validation data sets. This outperforms the single models significantly, highlighting the advantages of ensemble modeling. Additionally, the choice of Lasso for second layer modeling is found to have a better predictive effect than simple linear modeling, possibly because simple second-layer modeling may lead to overfitting.

Significant factors such as movie budget, directorship, main actors, genre, and release date have a substantial impact on box office revenue. Properly handling and processing these features plays an important role in improving prediction accuracy and reducing the risk of overfitting.

However, it is important to note that there are certain drawbacks to ensemble modeling. While it leverages the strengths of multiple models, it may be affected by low-accuracy ones, potentially reducing overall predictive effectiveness. Therefore, when implementing stacking ensembles, appropriate base models should be chosen to ensure they collectively enhance rather than hinder development.

In conclusion, this paper presents an effective and accurate tool for predicting film box office revenues based on the Elastic Net Random Forest Stacked Ensemble Model. This model offers high application value in the field of box office prediction.

References

[1] W. Lu, X. Zhang, and X. Zhan, “Movie Box Office Prediction Based on IFOA-GRNN,” Discret. Dyn. Nat. Soc., vol. 2022, 2022, doi: 10.1155/2022/3690077.

[2] D. Choudhery and C. K. Leung, “Social media mining: Prediction of box office revenue,” ACM Int. Conf. Proceeding Ser., vol. Part F1294, pp. 20–29, 2017, doi: 10.1145/3105831.3105854.

[3] T. Liu, X. Ding, Y. Chen, H. Chen, and M. Guo, “Predicting movie Box-office revenues by exploiting large-scale social media content,” Multimed. Tools Appl., vol. 75, no. 3, pp. 1509–1528, 2016, doi: 10.1007/s11042-014-2270-1.

[4] 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.

[5] 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.

DOI: 10.54254/2755-2721/73/20240419

[6] C. T. Madongo and T. Zhongjun, "A movie box office revenue prediction model based on deep multimodal features," Multimed. Tools Appl., no. 100, 2023, doi: 10.1007/s11042-023-14456-4.
[7] Y. Ru, B. Li, J. Liu, and J. Chai, "An effective daily box office prediction model based on deep neural networks," Cogn. Syst. Res., vol. 52, pp. 182–191, 2018, doi: 10.1016/j.cogsys.2018.06.018.
[8] J.-Y. Chang, "An Experimental Evaluation of Box office Revenue Prediction through Social Bigdata Analysis and Machine Learning," J. Inst. Internet Broadcast. Commun., vol. 17, no. 3, pp. 167–173, 2017, doi: 10.7236/jiibc.2017.17.3.167.
[9] L. Liao and T. Huang, The effect of different social media marketing channels and events on movie box office: An elaboration likelihood model perspective, vol. 58, no. 7. Elsevier B.V., 2021. doi: 10.1016/j.im.2021.103481.
[10] Y. Zhou, L. Zhang, and Z. Yi, "Predicting movie box-office revenues using deep neural networks," Neural Comput. Appl., vol. 31, no. 6, pp. 1855–1865, 2019, doi: 10.1007/s00521-017-3162-x.
[11] L. Jiang and Z. Wang, "Predicting box office and audience rating of Chinese films using machine learning," ACM Int. Conf. Proceeding Ser., pp. 58–62, 2018, doi: 10.1145/3300942.3300951.
[12] M. S, I. S. April, A. Rajasekar, and K. Yan, "Data Mining for Analyzing and Predicting the Success of Movies," 2021.
[13] L. Haiyan, Z. Yang, and Z. Hui, "Film box office prediction based on factor analysis," Proc. - 18th IEEE/ACIS Int. Conf. Comput. Inf. Sci. ICIS 2019, pp. 439–443, 2019, doi: 10.1109/ICIS46139.2019.8940299.
[14] N. D. Roy, Proceedings of International Conference on Data Analytics and Insights , ICDAI. 2023.
[15] V. Gupta et al., "Predicting attributes based movie success through ensemble machine learning," Multimed. Tools Appl., vol. 82, no. 7, pp. 9597–9626, 2023, doi: 10.1007/s11042-021-11553-0.
[16] K. Lee, J. Park, I. Kim, and Y. Choi, "Predicting movie success with machine learning techniques: ways to improve accuracy," Inf. Syst. Front., vol. 20, no. 3, pp. 577–588, 2018, doi: 10.1007/s10796-016-9689-z.
[17] S. Sahu, R. Kumar, H. V. Long, and P. M. Shafi, Early-production stage prediction of movies success using K-fold hybrid deep ensemble learning model, vol. 82, no. 3. Multimedia Tools and Applications, 2023. doi: 10.1007/s11042-022-13448-0.
[18] 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.
[19] D. Carr, "Early prediction of a film ' s box office success using natural language processing techniques and machine learning Sean O ' Driscoll Supervisor :," 2016.
[20] D. Park and J. Paik, "Prediction of movies box-office success using machine learning approaches," in Proceedings of the Korean Society of Computer Information Conference, 2020, pp. 15–18.