Annals of Operations Research (2022) 308:321–338
https://doi.org/10.1007/s10479-020-03804-4

S.I.: ARTIFICIAL INTELLIGENCE IN OPERATIONS MANAGEMENT

Early box office prediction in China's film market based on a stacking fusion model

Yi Liao[1] · Yuxuan Peng[1] · Songlin Shi[1] · Victor Shi[2] · Xiaohong Yu[3]

Accepted: 15 September 2020 / Published online: 6 October 2020

Abstract
Artificial intelligence has been increasingly employed to improve operations for various firms and industries. In this study, we construct a box office revenue prediction system for a film at its early stage of production, which can help management overcome resource allocation challenges considering the significant investment and risk for the whole film production. In this research, we focus on China's film market, the second-largest box office in the world. Our model is based on data regarding the nature of a film itself without word-of-mouth data from social platforms. Combining extreme gradient boosting, random forest, light gradient boosting machine, *k*-nearest neighbor algorithm, and stacking model fusion theory, we establish a stacking model for film box office prediction. Our empirical results show that the model exhibits good prediction accuracy, with its 1-Away accuracy being 86.46%. Moreover, our results show that star influence has the strongest predictive power in this model.

Keywords Artificial intelligence · Film industry · Predictive model · Machine learning · Box office forecast · Stacking fusion model

1 Introduction

In the past few years, China's film market has developed rapidly. According to a 2019 annual film market report, the box office revenue (henceforth called 'box office') of China's film market in 2019 was nearly $9 billion, the second-highest in the world (Leung and Lee 2019). Though China's film market's growth is significant, the continued expansion may discontinue. In 2019, the number of people who watched films in urban cinemas across the country was 1.727 billion, a growth rate of only 0.6%, the lowest level in the past decade,

✉ Xiaohong Yu
21090015@sbs.edu.cn

1 Southwestern University of Finance and Economics, Chengdu 610074, People's Republic of China

2 Wilfrid Laurier University, Waterloo N2L3C5, Canada

3 Shanghai Business School, Shanghai 200235, People's Republic of China

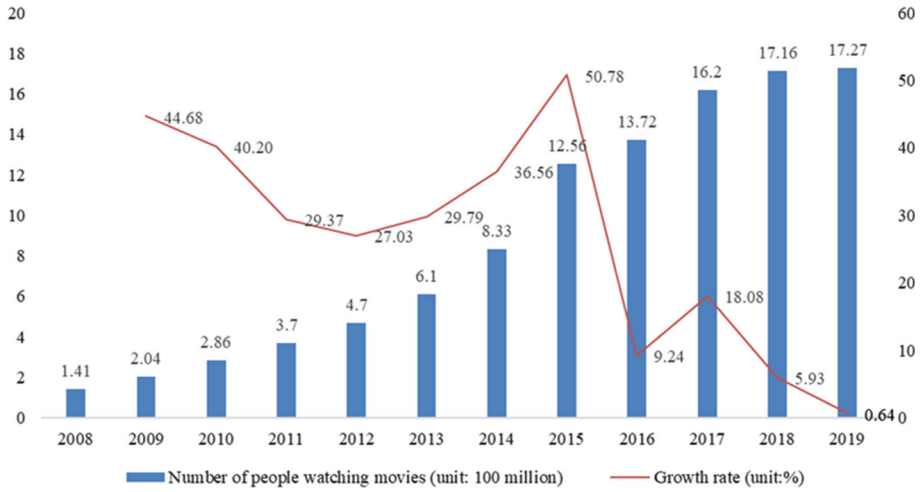

Fig. 1 The number of movie-goers in urban cinemas in China (2008–2019)

as shown in Fig. 1. More importantly, the COVID-19 outbreak has brought severe economic impacts worldwide (Ivanov 2020; Queiroz et al. 2020), especially for the film industry since cinemas are traditionally places for population gathering. As many countries and communities keep banning large public gatherings in response to the pandemic, not only countless films in production have been canceled, but also any new film investment would be examined extremely carefully.

On the other hand, filmmaking itself is highly risky (Ahmed et al. 2019; De Vany and Walls 1999). Before any expected revenue, a significant amount of capital investment is needed. For example, the box office revenue of cinemas usually depends on several popular movies, with around half of the rest losing money (Ghiassi et al. 2015). Furthermore, the movie production cycle is usually long, but market competition is fierce. Thus, it is critical to predicting the box office of a movie at an early stage to avoid potentially huge losses down the road (Ahmed et al. 2019). In addition, such prediction can help film production companies minimize the opportunity cost of wasting limited development resources on mediocre or failed projects (Ghiassi et al. 2015).

A few models have been developed for movie box office prediction based on word-of-mouth (WOM) data, which are normally generated after a movie's official announcement. However, at this stage, the movie project has been largely developed, and its box office revenue prediction has a rather limited impact on the movie's investment risk (Ahmed et al. 2019). Thus, in this research, we ask the following question: How to predict a film's box office revenue at its early stage without using post-release or post-production data? To answer this question, we develop a model to predict a movie's box office revenue at its early stage based on artificial intelligence techniques. Our research is of practical significance as it can help avoid investment failure and save millions of dollars (Sharda and Delen 2006; Delen and Sharda 2010; Ghiassi et al. 2015).

Specifically, combining Extreme Gradient Boosting (XGBoost), Random Forest (RF), Light Gradient Boosting Machine (LightGBM) and k-Nearest Neighbor (KNN) algorithms, we establish a stacking model for box office prediction during a film's early stage of production (shooting period). The model requires data related to the nature of the film itself instead of word-of-mouth or post-film data from social platforms. Extensive numerical experiments

show that the proposed model achieves 69.16% of Bingo accuracy and 86.46% of 1-Away accuracy.

The rest of the paper is organized as follows. Section 2 presents a literature review highlighting the existing research to date on the prediction of domestic box office revenue. In Sect. 3, data used in this research and its processing are described, including data acquisition, cleaning, feature extraction, and test data preparation. Next, Sect. 4 provides details of the proposed model. Through numerical experiments, Sect. 5 demonstrates the model's prediction performance results and compares it to those of other models. Lastly, the contributions of this study and potential future research are presented in Sect. 6.

2 Literature review

Litman and Kohl (1989) proposed a box office prediction model based on linear regression, where film rental is used to predict film revenue. Following this influential work, several improved box office prediction methods have been proposed. More recently, various prediction models, such as the Bayesian model (Neelamegham and Chintagunta 1999), support vector machine and neural networks (Ghiassi et al. 2015) have been developed. According to Kim et al. (2015), there are four major types of box office prediction models (Kim et al. 2015): statistical models such as linear regression models (Litman and Kohl 1989; Eliashberg and Shugan 1997), probabilistic models (Neelamegham and Chintagunta 1999; Sawhney and Eliashberg 1996; Eliashberg et al. 2000), time series models, and machine learning models such as deep neural network models (Sharda and Delen 2006; Delen and Sharda 2010; Ghiassi et al. 2015). In the following, we would like to review existing studies that are closely related to our work.

The statistical models based on linear regression have been popular because they can explain the influence of each variable on box office prediction (Kim et al. 2015). Mestyán et al. (2013) defined different prediction features and developed a linear regression model to predict the box office revenue of 312 American films on the first weekend. They assessed the popularity of movies based on the effects of external events on the activity of Wikipedia editors and the number of page views. In Litman and Kohl (1989), linear regression was used as the basic model. Then the adjusted production cost, major distributors, awards, and other seven film characteristics were taken as input variables for regression analysis. Eliashberg and Shugan (1997) also used a linear regression prediction model to analyze the success of movies. However, they focused on the influence of screens, positive reviews, negative reviews, and mixed reviews on the box office.

The probabilistic models can be considered as an alternative to linear regression models (Kim et al. 2015). Neelamgham and Chintagunta (1999) proposed a box office prediction system based on a Bayesian model, which can predict the box office of new films at different stages. Their research found that the number of screens has the greatest impact on the number of viewers. Sawhney and Eliashberg (1996) developed a parsimonious box office prediction system named BOXMOD-I based on queuing theory. Eliashberg et al. (2000) proposed an improved box office prediction model named MOVIEMOD based on Markov chain model. This model can explain the effects of different information sources on consumers' purchase intentions.

The prediction models based on time series is to predict future box office performance according to the past patterns. These models are often used to predict the box office after a film is released based on the box office performance during the previous few weeks. A

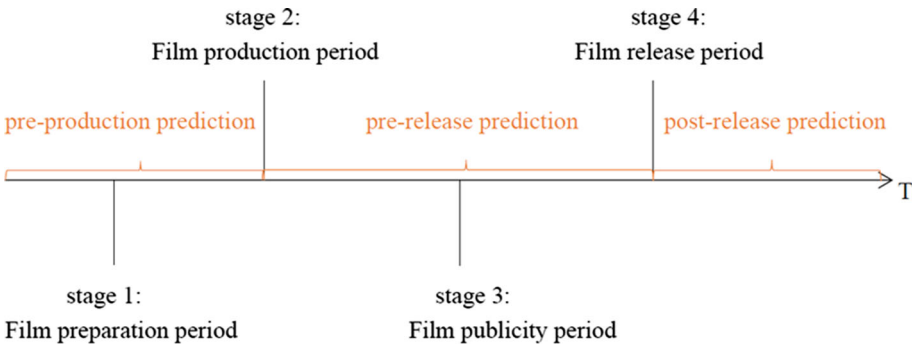

Fig. 2 Film cycle and prediction type

significant limitation of these models is that it only relies on the historical performance of the film but ignores many valuable exogenous factors (Kim et al. 2015).

Dogru and Keskin (2020) define artificial intelligence as an interdisciplinary subject of computer science, statistics, operational management, mathematics, humanities, social science, and philosophy. Its goal is to develop non-biological systems to perform tasks that usually require human intelligence. Artificial intelligence and big data technology have shaped various aspects of business and management (Grover et al. 2020; Kyriakou et al. 2019). For example, Akter et al. (2020) introduced a Netflix's recommendation system, which first mines user data and then adopts different models to determine the most suitable recommendation system. However, the research on applying artificial intelligence in a film's box office prediction, especially at its early stage, is still limited. Hur et al. (2016) proposed a box office prediction model based on film reviews and an independent subspace method, which uses audience review text and nonlinear machine learning to improve the prediction accuracy. Sharda and Delen (2006) set up an effective box office prediction model using artificial neural networks, logistic regression, discriminant analysis, and classification and regression tree algorithm. Ghiassi et al. (2015) added MPAA rating, competition, star value, sequels, and the number of screens to the prediction variables and proposed a pre-release box office prediction model based on a dynamic artificial neural network algorithm.

Overall, for box office prediction, models based on artificial intelligence and machine learning have gradually replaced the traditional one. In this paper, we contribute to this research stream by building an improved box office prediction model before a film's production. Our model is based on the stacking framework with data related to the nature of the film itself.

3 Data

In this section, we describe different aspects of the data, including data acquisition, cleaning, feature extraction, and test data preparation. Before describing the data acquisition stage, we first briefly explain the timeline of film production since the timing is critical for explanatory variable selection, which determines data acquisition eventually. According to the schedule of film production, the box office prediction can be divided into pre-production prediction, pre-release prediction, and post-release prediction, as illustrated in Fig. 2.

Table 1 The contributing factors and their predictive effectiveness for box office at different stages

Time	Features	Predictive effectiveness
Pre-production prediction	Based on the nature of the film itself, it uses features including release date, type, content, star value, sequel, and duration	With fewer features, it has lower prediction accuracy, but the earliest prediction period, and therefore, the highest practical application value of the prediction results
Pre-release prediction	In addition to the characteristics of the film itself, it also includes social media, search platform data, etc	The prediction accuracy is higher than that of pre-production prediction, and it can guide operational decision-making for cinemas but has little value for early investment decision-making
Post-release prediction	In addition to pre-release features, it also includes a large amount of theatre data, heat index, and audience comment information	It contains the most information and the best predictive effectiveness, but the application value of the results is very low

Table 2 Top 10 box office films in the Chinese market in 2019

Rank	English Title	Box office (unit: 100 million RMB)	Country of origin	Genre
1	Ne Zha	49.34	China	Comedy/Cartoon
2	The Wandering Earth	46.18	China	Fantasy
3	Avengers: Endgame	42.05	US	Action/Adventure
4	My People, My Country	31.46	China	Drama
5	The Captain	28.84	China	Drama
6	Crazy Alien	21.83	China	Comedy/Fantasy
7	Pegasus	17.03	China	Comedy/Action
8	The Bravest	16.76	China	Drama
9	Better Days	15.32	China	Drama/Romance
10	Hobbs and Shaw	14.18	US	Action/Crime

Next, we compare the contribute factors and the effectiveness of box office prediction at different stages (Table 1). In general, box office prediction based on post-show data has the most information to draw on and the best prediction accuracy. Unfortunately for producers or investors who have already spent money, such a late forecast is of little value (Ghiassi et al. 2015). There is no doubt that prediction before production is very difficult but has the greatest practical value and significance in guiding decision-making in the film industry chain.

From Table 1, we can easily identify that the pre-production prediction is generally based on the nature of the film itself, including the release date, type, content, star value, sequel, film length, and other factors. As aiming at the pre-production prediction, we narrow our scope to those factors in the data selection phrase. Furthermore, because domestic films have a strong performance in China's film market, whose box office revenues reached 41.175 billion yuan and accounted for 64.07% of the market and eight of the top 10 films (Table 2) at the box office in 2019 were domestic, it is appropriate to focus on domestic films as our first step.

Table 3 Data description

Variable	Type	Description	Data source
Title	Character string	Title of film	Movie Box Office Database
Actor 1/2/3	Character string	Name of top 3 actors/actresses	Movie Box Office Database Douban Movie
Director	Character string	Name of the main director	Movie Box Office Database
Actor 1/2/3 microblog fans	Value	Actor/actress's number of microblog fans	Microblog
Release area	Category	Category including 'Chinese mainland', 'Hong Kong', 'Taiwan'	Movie Box Office Database Douban Movie
Release date	Date	Film release date	Movie Box Office Database Douban Movie
Genre	Category	Including 18 categories	Movie Box Office Database Douban Movie
Actor (director) awards	Character string	Golden Rooster Awards, Golden Horse Awards and Hong Kong Film Awards for actors (directors)	Baidu

3.1 Data acquisition and cleaning

The original data were obtained from the Movie Box Office Database, Douban Movie, Sina Microblog, and other websites. The original data set contains basic information on films, directors, and actors in the Chinese film market, and the data sources are exhibited in Table 3. After data cleaning and feature extraction, the experimental data set was formed. The experimental data include 1182 domestic films in 2010–2019 (domestic films are defined as those with an origin in China, and for which the directors and most of the main actors are from China).

3.2 Feature selection and extraction

Effective features are the core of box office prediction. Considering the availability of data and the predictive power of features, five pre-production factors are selected based on the film itself: genre, star value, release date, release area, and sequels.

1. Genre

A film's story has depth and complexity, so the classification of film types is diverse. Moreover, people's tastes in film types also change with the market. This study measures the dynamic influence of film types according to the average box office performance of the type of film in the past year. Regarding the main story types of films, the genre classification is shown in Table 4.

Table 4 Genre classification

No.	Genre
1	Romance
2	Action
3	Crime
4	Thriller
5	Fantasy
6	Mystery
7	Sport
8	War
9	Literary adaptation
10	Adventure
11	Ancient history
12	History
13	Family
14	Drama
15	Comedy
16	Music
17	Cartoon
18	Documentary

2. Star Value

Acting staff (mainly actors and directors) can significantly affect the film's quality and the audience's expectations for the film. High-quality actors or directors and high-quality script content are mutually dependent and are also a basis for the film's market value. Delen and Sharda (2010) measured the star value of actors in three categories: high, medium, and low. Without loss of generality, we consider the main director and the top three main actors and use the total box office, average box office, highest box office, lowest box office and the number of acting or directing films of the actor or director within the past 10 years to measure the dynamic influence of the actor or director in that year. The static influence of the actor or director is measured by the number of fans of the actor's microblog, and the number of most popular film awards greater China area, such as Golden Rooster Awards, Golden Horse Awards, and Hong Kong Film Awards won by the actor and director.

3. Release Date and Aera

Concerning the release date, the release schedule plays an especially crucial role in the film's success (Einav 2007), as the release of multiple films at the same time will have a negative impact on each film's revenue due to the smaller market share (Sharda and Delen 2006; Delen and Sharda 2010). According to the film's release month, Delen and Sharda (2010) classified the parameter 'competition' into three categories: high, medium, and low. In terms of the income of Chinese films at different times of the year, the release time is divided into Spring Festival, National Day, summer holidays and normal days depending on the distribution schedule, with codes of 4, 3, 2 and 1, respectively.

The area in which a film is released represents its market. The statistics are all for Chinese films, and are categorized into 'Mainland China', 'Hong Kong' and 'Taiwan'. These markets have obvious differences, where 'mainland China' has a large area as well as a large

population, so it has the largest number of audiences. A one-hot coding method was used to describe the diversity of film release areas. For example, Chinese films released only on the mainland are encoded (1, 0, 0), whilst Chinese films released both on the mainland and in Hong Kong are encoded (1, 1, 0).

4. Sequel

Sequels are also an important feature of early film prediction (Delen and Sharda 2010). Due to the brand effect of the parent film, the film audience will have higher and more specific expectations about the film, so sequels perform better than general films (Dhar et al. 2012; Moon et al. 2010). High-quality sequels have many fans and topics while providing potential high-quality content and excellent production teams. Following Ghiassi et al. (2015), we use binary variables to specify whether a film is a sequel (assigned a value of 1) or not (assigned a value of 0). Besides, we use the income of the film's parent film to measure its sequel index. For example, 'War Wolf' is the parent film of 'War Wolf II', with a revenue value of 525 million yuan, so the sequel index of 'War Wolf II' is 52,500. The sequel index of non-sequels is 0.

3.3 Test data preparation

The proposed model was trained on 7 years (from 2010 to 2016) and is being used for the prediction of 2017–2019 films. The whole feature set is shown in Table 5. It is a historical data set based on the film's release year. It uses the features of the film itself without relying on related social media data. The data include two parts: dynamic data and static data. Dynamic data refers to data that changes dynamically each year, such as the dynamic influence of actors, directors, and film types, which is used to measure their influence in the film release year. No previous study of box office prediction research has used dynamic data. After data cleaning and feature processing of the original data, the experimental data feature set includes 37 feature vectors (feature numbers 1 to 37), including five factors: genre, star value, release date, release area, and the sequel, as shown in Table 5. The last column of the data set is the output label (y_0). The category label divides the films into seven categories A–G according to the box office revenue range, representing seven levels from flop to blockbuster. The corresponding box office revenue range of each category is shown in Table 6. A 2010–2016 film data set was used to build the model, and a 2017–2019 film data set was used as the test set to verify the effectiveness of the model. The model is programmed to conduct the calculations in the environment of Python 3.7.

4 Model

In this part, after introducing the Stacking framework first, we then describe the sample division and training method to avoid the repeated learning of the stacking model. The accuracy and characteristic contribution of various models in box office prediction are then analyzed, and the basic models used in this paper are selected based on these two factors. Finally, the whole process of box office prediction based on the stacking model in this paper is described.

In the stacking mechanism, each model class prediction is output, and the output is then taken to predict the final class of model (Doumpos and Zopounidis 2007). The method can be described as follows: take the output of the base model as a new feature, input it into other models, and use this method to stack the models. That is to say, the output of the first level

Table 5 The feature set of domestic box office prediction system

Factor	Feature number	Feature	Data type	Feature description
Genre	1	Dynamic influence of the genre	Continuity	The average box office of this type of film in the past year
Star value	2–22	Dynamic star value	Continuity	The total box office, average box office, highest box office, lowest box office, number of films performed or directed by the director, and the top three major actors in the past 10 years. The sum of the average box office of the top three actors and the director
	23–30	Static star value	Continuity	The number of microblog fans of the three actors, the number of Golden Rooster Awards, Golden Horse Awards and Hong Kong Film Awards won by the director and the top three major actors. The sum of three actors' microblog fans
Release date	31	Film release year	Dispersed	Film release year
	32	Film release schedule	Dispersed	According to the archive data of films released during Spring Festival, on National Day, during summer holidays and on normal days, the codes are 4, 3, 2 and 1, respectively
Release area	33–35	Release area	Dispersed	'Mainland', 'Hong Kong' and 'Taiwan' region codes
Sequel	36	Whether a sequel	Dispersed	Is a sequel (assigned 1) or is not a sequel (assigned 0)
	37	Sequel index	Continuity	Box office income of parent film (unit: 10,000 yuan)

Table 6 Category labels corresponding to film revenue range

Class	Revenue range (unit: 10,000 yuan)
(Blockbuster) G	> 20,000
F	8000–20,000
E	2000–8000
D	800–2000
C	200–800
B	50–200
(Flop) A	< 50

model is taken as the input feature of the second level model, and so on, and the output of the last level model is taken as the result.

For data sets $\mathcal{T} = \{(X_n, y_n), n = 1, 2, \ldots N\}$, X_n are feature vectors of the n th sample, y_n is the real classification value corresponding to the n th sample, and p is the number of features included, that is, each sample feature X contains the feature vector $(x_1, x_2, \ldots, x_p)$. The primary learners are $M_1, M_2, \ldots M_m$, the secondary learner is G_1 and the test data $\mathcal{T}_2 = \left(X', y'\right)$. Formula (1) explains the prediction method of the two-layer stacking model as follows:

$$\hat{y} = G_1(\emptyset) = G\Big(\hat{M}_1\left(X'\right), \hat{M}_2\left(X'\right) \ldots \hat{M}_m\left(X'\right)\Big). \tag{1}$$

4.1 Sample division and training method

The principle of the stacking integration method is to synthesize the learning results of primary learners via secondary learners, adjust the result bias caused by model overfitting in primary learners, and correct the model prediction error. With this method, we can obtain better prediction performance than with a single algorithm. It should be noted that the training set input by the secondary learner is generated by the prediction results output by multiple primary learners. If the secondary trainer is used to train the data set directly, the data may be repeatedly learned, resulting in 'overfitting'. It is therefore necessary to divide the data reasonably.

Taking a fivefold sample data set as an example, the original training data set is divided into five sub-data sets to ensure that the sub-data sets do not overlap. For each primary learner, four data sets are used to train the model, and the remaining data set is used as the validation data set to verify the effect of the learner and to adjust parameters. All five sub-data sets can obtain the prediction results through the learned model. Training each primary learner in turn, the prediction results of multiple primary learners can be combined into a new data set and used as the input of secondary learners. In this way, the data is converted from output to input. For each primary learner, there is a different sub-data set that does not participate in training the learner. This arrangement ensures that the data sets involved in the model training are different to be able to prevent a repetitive training of the data. The model training method with a fivefold data set is shown in Fig. 3.

4.2 Characteristic contribution analysis and model selection

The prediction ability of a single primary learner is the basis of the integrated model's prediction ability. At the same time, there should be a certain degree of difference between primary learners to allow the models to make up for each other. In this part, we comprehensively analyze the prediction performance and feature contribution of various machine learning algorithms and then choose the basic learners of our stacking model on that basis.

The box office prediction is a multi-classification problem with no obvious classification boundary. Though the support vector machine (SVM) algorithm has excellent performance in solving small-sample and nonlinear problems, it has difficulty dealing effectively with multi-classification problems. Classification and regression trees (CART), logical regression and decision tree algorithm have bad performance because of the low complexity of the models. However, the ensemble models based on decision tree learning such as XGBoost, LightGBM as well as RF indicate good performance in multi-classification and regression problems due to their strong branching ability. The XGBoost algorithm and LightGBM algorithm are

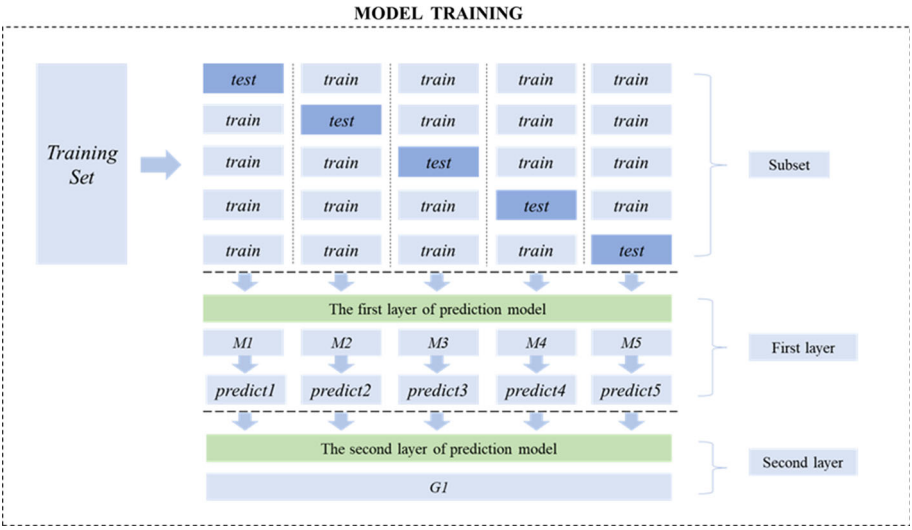

Fig. 3 Training method of the stacking model with fivefold sample data set

boosting integrated algorithms based on GBDT. The former uses a layer-growth strategy, and the latter uses the leaf-node growth strategy. Under the same splitting times, the latter can reduce error, but it is more likely to produce overfitting. In contrast to those two, the RF algorithm is a decision tree algorithm based on bagging integration. To evaluate the effectiveness of each model on the experimental data in this paper, the sub-data set is used to train the model and obtain the characteristic contributions under the training of the XGBoost, LightGBM and RF algorithms, as shown in Fig. 4.

Among all classes of feature factors, the star influence (features No. 2-30) involves the most features and has the strongest predictive power, because not only do stars have an impact on the film's box office in terms of public praise and publicity, but also they have a strong correlation with potential factors such as the film's budget and script quality. Three other factors namely release time, release area and film type, have the same predictive power. The sequel factors are limited by the number of sequel film samples, so their predictive power is weak.

The XGBoost model can fully use all the internal characteristics of input data, thus it has good applicability to both continuous and discrete features and ensures that the contribution of various features in the prediction is relatively uniform. The LightGBM algorithm generates trees in the form of leaf nodes, which cases the main features to have a better degree of fitting and a higher degree of contribution. In the LightGBM prediction, sum of the average box office of directors and actors (feature No. 22) and the sum of three actors' microblog fans (feature No. 30) make significantly higher contributions than other features, while some features' contributions are lost. The RF method is based on bagging integration, which is not sensitive to discrete sparse features. It focuses mainly on a small number of influential features. It performs similarly to the LightGBM algorithm in terms of feature contribution but uses a different integration method. In the prediction, the RF model is better than the other two models in extracting the features of sequels (feature numbers 36 and 37). In sum, the difference between these three primary learners can help the models complement each other's strength and obtain a better integration effect.

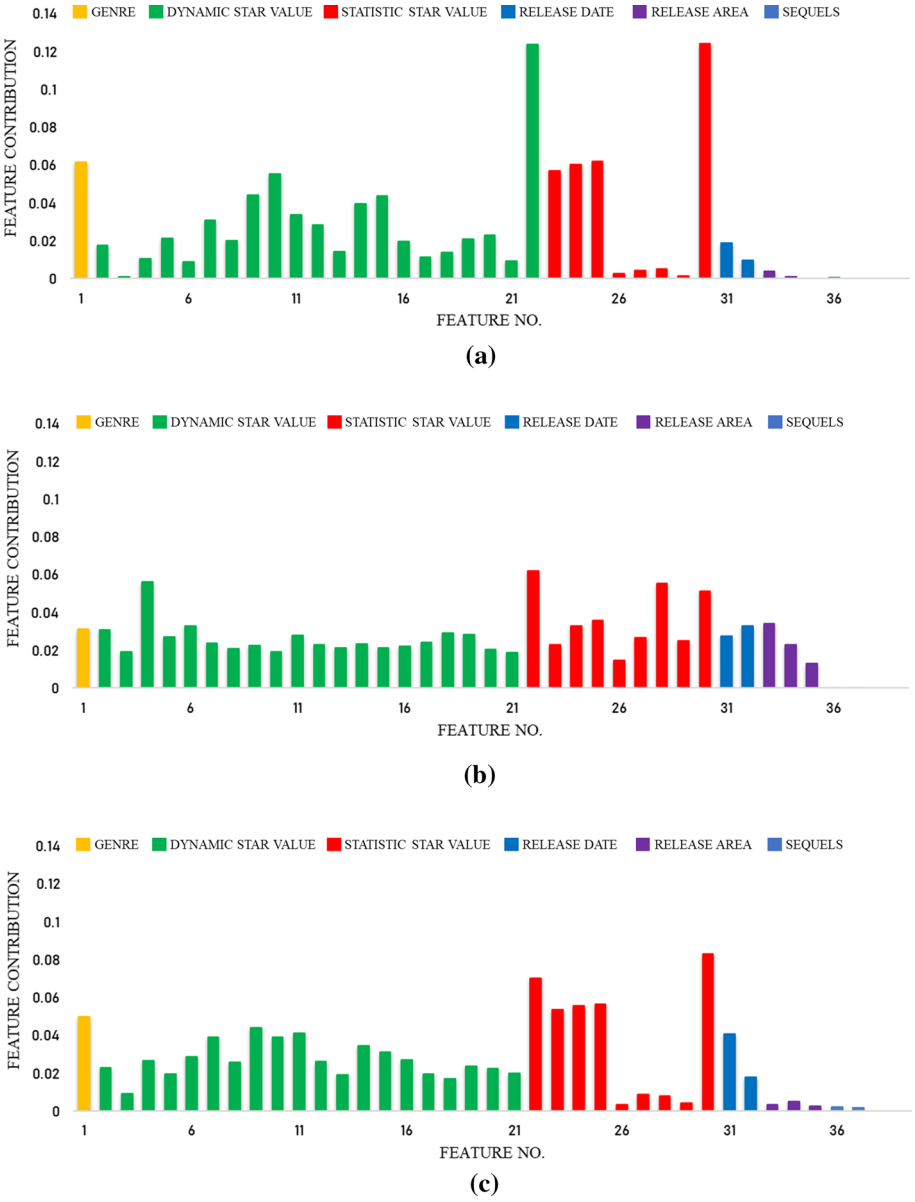

Fig. 4 Characteristic contribution analysis of various algorithms: **a** Feature contribution analysis degree of LightGBM algorithm; **b** Feature contribution analysis of XGBoost algorithm; **c** Feature contribution analysis of RF algorithm

4.3 Model selection and construction

To obtain stable output results, the secondary learners select the models with strong generalization ability, so as to reduce and correct the bias tendency of multiple primary learning algorithms to the training set and to balance the overfitting problem in the decision tree. The

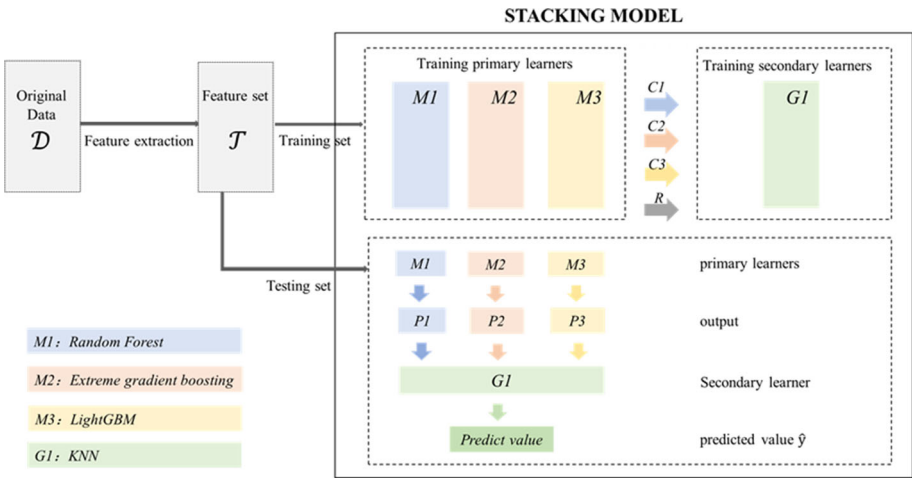

Fig. 5 Box office prediction method based on model fusion in the framework of stacking

k-nearest neighbor (KNN) algorithm is a natural multiclassification model that measures the distance between different eigenvalues. It can effectively integrate the prediction results of multiple models and improve the model's generalization ability.

To summarize, the primary learners used in this paper are XGBoost, RF, and LightGBM, while the secondary learner is KNN. The box office prediction method based on model fusion in the framework of stacking is shown in Fig. 5.

Our box office prediction process based on the stacking model is as follows:

1. Processing data
   - After data cleaning and feature extraction, feature data set $\mathcal{T}$ is generated by original data $\mathcal{D}$.
   - Feature data set $\mathcal{T}$ is divided into a training set $\mathcal{T}_1$ and testing set $\mathcal{T}_2$.
2. Training model
   - Based on the number of samples, the training set is divided into five sub-datasets. According to the training method mentioned above, the primary models are trained, and the appropriate parameters are selected.
   - The prediction class of the training set on multiple primary models are combined into a new data set.
   - At this stage, the new data set is used as input feature to train the secondary learner. So far, the training of stacking model based on multi-model fusion is completed.
3. Output results
   - The test data set is input into the trained primary learners.
   - The generated results are then input into the secondary learner to obtain the model prediction class.
   - The performance of the model is evaluated with respect to the prediction accuracy.

Table 7 Comparison of classification accuracy of different models

Performance evaluation	KNN	Random forest	XGBoost	LightGBM	Stacking model
Count (Bingo)	143	225	226	222	240
Count (1-Away)	224	287	286	290	300
APHR (Bingo)	41.21%	64.84%	65.13%	63.98%	69.16%
APHR (1-Away)	64.55%	82.71%	82.42%	83.57%	86.46%

5 Results

To measure the effectiveness of the model, the average percentage hit rate (APHR) is used as the evaluation index, which represents the percentage of correctly classified samples relative to the total number of samples (Ahmed et al. 2019). In this paper, two types of APHR, Bingo and 1-Away, are used to judge the accuracy of various categories. (1) Absolute accuracy (Bingo): indicates the accurate hit rate; that is, only the classification of the correct class is considered. (2) Relative accuracy (1-Away): In addition to the absolute accuracy (Bingo), the classification results in which the real class is adjacent to the predicted class are also counted. The APHR can be calculated as follows:

$$APHR = \frac{Number\ of\ samples\ correctly\ classified}{Total\ number\ of\ samples} \tag{2}$$

$$APHR_{bingo} = \frac{1}{n}\sum_{i=1}^{K} c_i \tag{3}$$

$$APHR_{1-away} = \frac{1}{n}\sum_{i=1}^{K}(c_i + c_{i-1} + c_{i+1}) \tag{4}$$

where n represents the total number of samples, K represents the total number of categories, and c_i represents the total number of samples correctly classified as category i.

5.1 Analysis of model prediction results

To evaluate the model's performance in this paper, the prediction results (Bingo and 1-Away) are measured by the APHR index. We compare the prediction results of all three primary learners and the stacking model, as well as the results of the KNN model, as shown in Table 7.

XGBoost performs best among the single prediction models, with a Bingo accuracy of 65.13%, followed by RF, with a Bingo accuracy of 64.84%, and then the LightGBM model, with a Bingo accuracy of 63.98%; these are all much higher than KNN's accuracy of 41.21%. However, the stacking model performs better than all these single-prediction models, achieving 69.16% Bingo accuracy and 86.46% 1-Away accuracy (Table 8).

The prediction accuracy of the model for low-income (Class A) samples is slightly higher than that for other types of samples, presumably because these low-level films' box-office income is relatively low due to the lack of reputation and influence; and because of the 'long tail effect' of the film market, the number of such samples accounts for the highest proportion in the film market. However, if the high-level films are to obtain the corresponding high returns, they are also subject to many factors, such as market environment, audience preference, and even weather, with greater uncertainty. In general, the box office prediction model is more sensitive to risk than to revenue. When it is applied to budget allocation, it

Table 8 Prediction accuracy matrix of the stacking model for film box office prediction

Predict	A	B	C	D	E	F	G	APHR (Bingo) (%)	APHR (1-Away) (%)
Actual									
A	92	15	4	0	1	1	1	80.70	93.86
B	4	22	6	0	1	0	0	66.67	96.97
C	1	7	48	3	5	4	0	70.59	85.29
D	0	0	2	5	1	1	0	55.56	88.89
E	0	1	1	3	18	6	3	56.25	84.38
F	0	1	0	1	4	9	2	52.94	88.24
G	0	1	3	6	11	7	46	62.16	71.62
AVG								69.16	86.46

should focus on the avoidance of excessive investment in films with low predicted box-office incomes.

From 2010 to 2016, the Chinese film market grew rapidly. During this period, the market continued to expand, and the influence of stars also showed an upward trend. Since 2017, the growth rate has slowed down, and the influence of the market and stars has stabilized. In this paper, the 2017–2019 film period is used as the prediction set. At this stage, the model may be slightly affected, but after the market tends to be stable, the impact of data fluctuations on the model will become less significant, the prediction effect will be gradually stable, and the prediction performance will be better in the future.

5.2 Comparing the numerical results

To further evaluate the effectiveness of this feature model on box office prediction, the prediction results of this model are compared with the results of previous studies in China and abroad. One of the most representative studies in the field of box office prediction is the prediction model advanced by Delen and Sharda (2010). Their model selected seven pre-release factors such as competitiveness and star influence and used five machine learning methods and model fusion methods to predict the box office for 214 Hollywood films before their release. The accuracy of prediction ranged from 40.46 to 56.07%, among which the mixed model had the best predictive effectiveness. Quader et al. (2018) collected 755 films and television works from around the world, extracted 15 dimensional features before and after release, and classified the film box office data into five categories according to size. Seven classic machine learning methods were used for prediction, with an accuracy range of 43.29% to 58.5%, among which the multilayer perceptron (MLP) model had the best predictive effectiveness.

Compared with the movie box office prediction model proposed by Delen and Sharda (2010) and Quader et al. (2018), the prediction model proposed in this paper adopts multi-dimensional data sets and new feature processing methods. The prediction time node is earlier than the above model, and it can predict the movie box office performance before the film production period. The comparison of model prediction results is shown in Table 9. The results show that the features and models proposed in this paper have a good effect on box office prediction. The accuracy of new movie prediction (bingo) has been improved by at least 13.09% and 10.66%, respectively.

Table 9 Comparison of model results

Model	Features	Period	APHR (Bingo) (%)
Stacking model (in this paper)	New features	Pre-production	69.16
Hybrid model (Delen and Sharda 2010)	Old features	Pre-release	56.07
MLP model (Quader et al. 2018)	Old features	Pre-release	58.50

6 Conclusions

In the literature, existing box office prediction models usually depend on data pertaining to the after-film production period, which has a rather limited effect on stakeholders, because it can only affect the later refinement of advertising or distribution strategies at this stage. In this paper, we propose a novel model to predict the box office revenue of a film during its early stage of production. This model uses data about the nature of the film itself but not the word-of-mouth or post-film data from social platforms. By incorporating the XGBoost, RF, LightGBM and KNN algorithms, we establish a stacking model for film box office prediction. Our results show that our model performs well in terms of box office prediction with 69.16% of Bingo accuracy and 86.46% of 1-Away accuracy.

To better understand the impacts of different variables, we further use the trained prediction model for preparing movies. In this experiment, the decision makers of entertainment companies can identify with high accuracy how much a specific actor, a specific genre, a specific release date, or whether or not it is a sequel, can contribute to the success of a movie. Through the analysis of the contributing factors to film performance, management can better plan the film production and distribution. For example, star influence is crucial to films, but famous stars obviously require higher pay. It is unwise to choose most famous movie stars blindly, as it may increase their market demand for top stars, further increasing their pay. De Vany and Walls (1999) and Walls (2005) confirmed the 'curse of the superstar'. If a star gets the expected revenue growth associated with his or her performance, the movie almost always loses money. Fortunately, stars and film production companies are a two-way choice. Film companies can screen out more cost-effective actors through the box office prediction system or negotiate with them about film remuneration and control budget to maximize their profits.

When building the prediction model, we consider the availability and nature of data with five factors, including the release area and star influence. However, due to data unavailability, several influencing factors (e.g., budget) are omitted. Intuitively, movie box office prediction methods rely on reliable data, some of which are confidential. Hence, first, future research can endeavor to cooperate with film studios and obtain additional data that are not available to outsiders. Second, in the current research, we emphasize on the prediction model of the box office in the early stage of film production by incorporating the characteristics of the film only. However, a movie project may be adapted from intellectual property, which has a huge fan base. Therefore, future research can still utilize social media data before film production to build an even better box office prediction model.

Acknowledgements The authors are very thankful to the editor and the referees whose detailed reviews and suggestions helped improve this article. The research is supported by National Natural Science Foundation of China (No. 71871186 and No. 71871184) and the Fundamental Research Funds for the Central Universities (JBK18JYT02, JBK1902009, and JBK190504).

References

Ahmed, U., Waqas, H., & Afzal, M. (2019). Pre-production box-office success quotient forecasting. *Soft Computing*. https://doi.org/10.1007/s00500-019-04303-w.

Akter, S., Michael, K., Uddin, M., McCarthy, G., & Rahman, M. (2020). Transforming business using digital innovations: The application of AI, blockchain, cloud and data analytics. *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03620-w.

De Vany, A., & Walls, W. D. (1999). Uncertainty in the movie industry: Does star power reduce the terror of the box office? *Journal of Cultural Economics, 23*(4), 285–318. https://doi.org/10.1023/A:1007608125988.

Delen, D., & Sharda, R. (2010). Predicting the financial success of hollywood movies using an information fusion approach. *Industrial Engineering Journal, 21,* 30–37.

Dhar, T., Sun, G., & Weinberg, C. (2012). The long-term box office performance of sequel movies. *Marketing Letters, 23*(1), 13–29. https://doi.org/10.1007/s11002-011-9146-1.

Dogru, A., & Keskin, B. (2020). AI in operations management: Applications, challenges and opportunities. *Journal of Data, Information and Management, 2,* 67–74. https://doi.org/10.1007/s42488-020-00023-1.

Doumpos, M., & Zopounidis, C. (2007). Model combination for credit risk assessment: A stacked generalization approach. *Annals of Operations Research, 151*(1), 289–306. https://doi.org/10.1007/s10479-006-0120-x.

Einav, L. (2007). Seasonality in the U.S. Motion picture industry. *The Rand Journal of Economics, 38*(1), 127–145. https://doi.org/10.1111/j.1756-2171.2007.tb00048.x.

Eliashberg, J., Jonker, J.-J., Sawhney, M., & Wierenga, B. (2000). MOVIEMOD: An implementable decision-support system for prerelease market evaluation of motion pictures. *Marketing Science, 19,* 226–243. https://doi.org/10.1287/mksc.19.3.226.11796.

Eliashberg, J., & Shugan, S. M. (1997). Film Critics: Influencers or predictors? *Journal of Marketing, 61*(2), 68–78. https://doi.org/10.1177/002224299706100205.

Ghiassi, M., Lio, D., & Moon, B. (2015). Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Systems with Applications, 42*(6), 3176–3193. https://doi.org/10.1016/j.eswa.2014.11.022.

Grover, P., Kar, A., & Dwivedi, Y. (2020). Understanding artificial intelligence adoption in operations management: Insights from the review of academic literature and social media discussions. *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03683-9.

Hur, M., Kang, P., & Cho, S. (2016). Box-office forecasting based on sentiments of movie reviews and independent subspace method. *Information Sciences*. https://doi.org/10.1016/j.ins.2016.08.027.

Ivanov, D. (2020). Viable supply chain model: integrating agility, resilience and sustainability perspectives—-lessons from and thinking beyond the COVID-19 pandemic. *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03640-6.

Kim, T., Hong, J., & Kang, P. (2015). Box office forecasting using machine learning algorithms based on SNS data. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2014.05.006.

Kyriakou, I., Mousavi, P., Nielsen, J., & Scholz, M. (2019). Forecasting benchmarks of long-term stock returns via machine learning. *Annals of Operations Research*. https://doi.org/10.1007/s10479-019-03338-4.

Leung, W.-F., & Lee, S. (2019). The Chinese film industry: Emerging debates. *Journal of Chinese Cinemas, 13*(3), 199–201. https://doi.org/10.1080/17508061.2019.1678235.

Litman, B. R., & Kohl, L. S. (1989). Predicting financial success of motion pictures: The '80 s experience. *Journal of Media Economics, 2*(2), 35–50. https://doi.org/10.1080/08997768909358184.

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PLoS ONE, 8,* e71226. https://doi.org/10.1371/journal.pone.0071226.

Moon, S., Bergey, P., & Iacobucci, D. (2010). Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. *Journal of Marketing, 74*(1), 108–121. https://doi.org/10.1509/jmkg.74.1.108.

Neelamegham, R., & Chintagunta, P. (1999). A Bayesian model to forecast new product performance in domestic and international markets. *Marketing Science, 18,* 115–136. https://doi.org/10.1287/mksc.18.2.115.

Quader, N., Gani, M., & Chaki, D. (2018). *Performance evaluation of seven machine learning classification techniques for movie box office success prediction.* Paper presented at the 2017 3rd international conference on electrical information and communication technology (EICT), Khulna, Bangladesh, February 2018.

Queiroz, M. M., Ivanov, D., Dolgui, A., & Wamba, S. F. (2020). Impacts of epidemic outbreaks on supply chains: mapping a research agenda amid the COVID-19 pandemic through a structured literature review. *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03685-7.

Sawhney, M., & Eliashberg, J. (1996). A Parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science, 15,* 113–131. https://doi.org/10.1287/mksc.15.2.113.

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications, 30*(2), 243–254. https://doi.org/10.1016/j.eswa.2005.07.018.

Walls, W. (2005). Modeling movie success when 'nobody knows anything': Conditional stable-distribution analysis of film returns. *Journal of Cultural Economics, 29*(3), 177–190. https://doi.org/10.1007/s10824-005-1156-5.