Research on Mathematical Model of Box Office Forecast through BP Neural Network and Big Data Technology

Debao Dai*, Jibin Chen

School of Management Shanghai University Shanghai, China

*Corresponding author: ddb@shu.edu.cn

Abstract. With the rapid development of China's film industry, the public's enthusiasm for watching films has growing, and the contribution rate of the film industry to the economy has increasing year by year. Based on the research at home and abroad, this paper takes the box office of China's films as the dependent variable, selects a series of indicators of China's film box office for grey correlation analysis, and selects seven related indicators as independent variables, including the number of fans of heroine, the number of fans of hero, the box office of the first week, the number of fans of director, Baidu Index, average ticket price and film type. The influencing factors calculated by grey correlation degree are taken as the input of BP neural network, and the box office of China's films is taken as the output of the network to predict the box office of China's films. The results show that the prediction model based on BP neural network is more accurate than the classical Multiple Linear Regression model. This model can predict to a certain extent, and the prediction effect is good. It can predict the overall trend of film box office and provide reference for investment enterprises.

1. Introduction
With the development of economy, people's cultural consumption also increases. As one of the important entertainment ways, film has become one of the choices of people's cultural consumption. In recent years, China's film market has been booming and its international influence has been increasing. According to statistics, from 2008 to 2018, the total box office of Chinese mainland films has achieved a leap forward development from 4.34 billion yuan to 64.27 billion yuan, becoming the second largest ticket warehouse in the world. In the United States, Japan and other developed countries, film has become a pillar industry, with a relatively mature and perfect film market [1]. However, due to the limitation of film broadcasting cycle and the information asymmetry between film supply and demand, the film industry has certain risks [2].

Therefore, it is necessary to study the box office of China's films for the risk of film investment and the decision-making of film investment. Through the analysis of the factors that affect the box office of the film, investors can be clearer about the arrangement of actors, film arrangement and so on.

2. Literature review
At present, there are many researches on the box office, mainly focusing on the following three aspects: Firstly, research on the factors influencing the box office of films. In recent years, quantitative research has gradually emerged. In foreign countries, Ramesh Sharda (2006) uses cross-sectional survey method
to select some box office influence factors for research, mainly including film title, storyline, star, audition and advertising [3]. Chakra Arty(2010) put forward the film title, content, actors, early promotion and other factors [4]. In China, Hu et al. (2013) analysed 217 films released from 2007 to 2009, and conducted one-way Anova on the variable "remake and sequel", which showed that it had a significant impact on the box office [5]. Taking 50 imported films as the analysis object, Song (2013) divided the attention of films into different types of activities, which showed that publicity and film arrangement had an impact on the box office [6]. Wang (2015) selected 132 films, took film type, online rating, release schedule and other factors as explanatory variables, and analysed them with Eviews. It was found that hot schedule and higher early rating were conducive to the improvement of film box office [7]. Wu et al. (2016) selected 200 domestic films released in 2015 as experimental data, and proposed audience expectation, film self-influence, and competitiveness in the same period as predictive dependent variables [8].

Research on the prediction method of film box office. This research mainly focuses on regression prediction and neural network prediction, and there are few other methods. In terms of regression prediction method, Eliashberg (2006) used star power, budget, sequel and MPAA score as input, and used linear regression model [9]. On the basis of Multiple Linear Regression, Ren (2015) used SSH framework and MVC development mode to design box office prediction system [10]. Ding et al. (2016) took American animated films as the research object, selected 104 animated films released from 2005 to 2014 and predicted their memory by Multiple Linear Regression model, and obtained the similarities and differences between animated films and ordinary films [11]. Luo et al. (2019) used the basic information data and box office data obtained from Douban film and cat's eye film as the data set for linear regression training [12]. In terms of neural network method, Barmand et al. (2012) proposed to use the feedback neural network algorithm to predict the box office [13]. Sharda (2006) proposed to classify the box office based on multi-layer neural network, but when collecting the data, all of them use 0, 1 assignment, which loses the interpretability of the data [14]. Zheng and Zhou (2014) used the feedback neural network to establish a single box office prediction model. The influencing factors are director, actor, schedule, etc., and the effect of the model is good [15]. Yang and Zheng (2014) used the feedback neural network to predict the box office [16, 17]. Other scholars also put forward GBRT model (Han et al.), LARS-SVR model (Chen&Xu), Decision Tree (Wu) and other methods to predict the box office [8, 18, 19].

A study of the history of box office prediction. Wang (2009) divided the development of foreign box office prediction into three stages: basic model, dynamic research and popularity, and proposed the importance of establishing a prediction model suitable for China's development [20]. Wang (2015) divided the whole process into three stages: audience research, prediction model and Internet [21].

To sum up, the research on box office prediction has become systematic and scientific, and scholars at home and abroad have done research from different aspects. However, there are still some deficiencies: the factors that affect the box office have not been defined uniformly, and the factors that affect the box office prediction can continue to be explored. And, the amount of data on the prediction of film box office is relatively limited, which can also increase the amount of data. On the basis of the existing research, this paper selects the relevant indicators, uses BP Neural Network to predict the domestic film box office, and provides reference value for the decision-making of film investors. The innovations of this paper are as follows: firstly, on the basis of scholars' research, the average ticket price index is proposed as one of the influencing factors to measure the box office forecast. The second is to use the grey correlation method to screen the indicators, and the selected indicators are used as the input of neural network.

3. Index selection.

3.1. Variable determination

In this paper, the indicators are determined based on the research situation at home and abroad. Among them, the distribution capacity and marketing capacity data of film companies are not open to the outside
world, so they are not considered when designing indicators. The type of film, the appeal of stars and actors have repeatedly appeared in the research of scholars, which shows its importance. In order to improve the accuracy of the forecast, the first week's box office is also used as one of the indicators. Finally, combined with the characteristics of the era of big data, the web search index of films is also used as one of the predictive indicators. Considering the importance of box office influence factors, the ease of obtaining relevant data and China's national conditions, a series of indicators such as film type, star and actor appeal, first week box office, Baidu index, etc. are finally selected to construct the box office prediction index system of this article. The specific analysis of each index of box office forecast is as follows:

Film type. Refer to the box office division method of Ramesh [22] to classify the box office data, which is divided into 8 levels. The classification is shown in Table 1:

| Film type | Come | Horror | Love | Action | Science |
|-----------|------|--------|------|--------|---------|
| Number    | 1    | 2      | 3    | 4      | 5       |

| Film type | Martial | War | Crime | Plot | Fantasy |
|-----------|---------|-----|-------|------|---------|
| Number    | 6       | 7   | 8     | 9    | 10      |

Charisma of celebrities and actors. Nelson et al. used the number of visits to celebrity homepages on IMDB as a measure of film charisma, and constructed a model of the relationship between celebrities and film box office [23]. The importance of film box office forecasts. Before and during the film's release, the creators will publish the film promotion through Sina Weibo, so the number of fans is an important manifestation of the film promotion. This article selects the number of fans in three dimensions of the film as the box office predictor. They are the number of fans of the hero, the heroine, and the director.

First week’s box office. Dellairocas et al. considered the first week’s box office as one of the influencing factors of the box office revenue, and believed that the first week’s box office is an important factor affecting the film’s future box office [24], the forecast of the first week’s box office for the total box office significant. Therefore, this article selects the first week's box office as a predictive index for predicting the total box office of films.

Baidu Index. Wu Faxiang and others used Baidu Index as one of the indicators to measure the box office of a film to understand the degree of user expectations of the film [8]. Baidu Index is a free data sharing platform based on Baidu's massive user behavior search data. The Baidu Index can reflect the size of the search scale of a certain keyword in the Baidu search engine and the change in a specific period of time. Therefore, this article uses the Baidu index as a predictive index for predicting film box office.

Average ticket price. The ticket price of a film is one of the important factors affecting box office revenue. The average ticket price of a new film in this article is used as a predictor of film box office.

In summary, the film box office forecast indicators selected in this article are shown in Table 2:

| Predictor | Variable | Data Sources |
|-----------|----------|--------------|
| Video type | X1       | Boxofficecn  |
| Actor fans | X2       | Sina Microblog |
| Heroine fans | X3    | Sina Microblog |
| Director fans | X4   | Sina Microblog |
| First week’s box office | X5 | Boxofficecn |
| Baidu Index | X6 | Baidu Index |
| Average ticket price | X7 | Boxofficecn |
3.2. Index correlation

In order to initially measure the relevant procedures of the influence of each attribute on the final film box office, the concept of correlation analysis is introduced here. Correlation is used to indicate how closely two variable factors are related. This paper uses grey relational analysis to get the correlation coefficient between each research index and film box office.

(1) Determine the characteristic sequence and factor sequence. Take the domestic film box office as the characteristic sequence and mark it as $Y_j$ ($j=1, 2...m$). The influencing factors corresponding to domestic films are the sequence of factors, denoted as $X_i$ ($i=1, 2...n$), as shown in Table 2.

(2) Standardization of extreme values. This article uses range standardization to standardize the statistical data to eliminate the influence of the dimensions and magnitude of each variable. Substituting the statistical data into equation (1), the standardized data is obtained to obtain the standardized data, and the number is between $[0, 1]$.

$$x_i^* = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \quad (i = 1, 2, ... n)$$

Where $x_i$ represents the number of the $i$-th index, and $x_i^*$ is the value after the range normalization of the original data.

(3) Calculate the difference sequence.

$$I_i = |Y_j - x_i^*| \quad (i = 1, 2, ... n, j = 1, 2, ... m)$$

(4) Calculate the correlation coefficient. Refer to formula (3) to calculate the correlation coefficient between the feature sequence and the factor sequence, where the resolution coefficient is 0.5.

$$\delta_i(j) = \frac{\min(i) \min(k) I_i + \rho \max(i) \max(k) I_i}{I_i + \rho \max(k) I_i} \quad (i = 1, 2, ... n)$$

(5) Calculate the degree of relevance. Calculate the gray correlation degree between the two sequences according to formula (4), and the calculation result retains three decimal places.

$$Z_i = \frac{1}{n} \sum_{j=1}^{m} \delta_i$$

(6) Sorting Sort the results obtained in step (4) according to the numerical value. The calculation results are shown in Table 3.

| Sorting | Predictor Variable | Correlation |
|---------|-------------------|-------------|
| 1       | Video type X1     | 0.7022      |
| 2       | Actor fans X2     | 0.6902      |
| 3       | Heroine fans X3   | 0.6791      |
| 4       | Director fans X4  | 0.6399      |
| 5       | First week’s box office X5 | 0.6273 |
| 6       | Baidu Index X6    | 0.6129      |
| 7       | Average ticket price X7 | 0.6022      |

To sum up, choose the influencing factors with gray correlation coefficient greater than 0.6, namely $X_3, X_2, X_5, X_4, X_6, X_7, X_1$ as the input of the neural network.
4. BP Neural Network prediction model

4.1. Model construction

BP Neural Network is a commonly used algorithm at present. The basic unit is neurons. The network structure is composed of input layer, hidden layer, and output layer. Neurons in each layer are independent, and neurons in adjacent layers are connected with weights. The basic process of model training is forward transmission and error backward transmission, which is, input variables output results through the action of each layer in the network, and then the actual output is compared with the expected output, and the error between the two is calculated. If the error is within the preset accuracy, output the result; otherwise, modify the connection weights between the neurons in each layer, and repeat the previous step until the preset requirements are met.

This paper uses the neural network toolbox of MATLAB2014 to construct a BP Neural Network to predict the domestic box office. The steps are as follows:

(1) Determine the input and output of the neural network. This article selects X3 heroine fans, X2 male protagonists fans, X5 first week box office, X4 director fans, X6 Baidu index, X7 average ticket price, X1 film type, a total of 7 parameters as the input index system and output index of this model. Y is the box office. The formal input is: Input=[X1, X2, X3, X4, X5, X6, X7], and the output is: Output=[Y]. The training samples are analyzed at the box office of domestic films in the four years from 2015 to 2018. The data set is divided into training set and test set, with a ratio of 8:2. The samples input to the network are first normalized, and the normalized function uses MATLAB’s own function mapminmax ()

(2) Determine the transfer function and learning function. The transfer function from the input layer to the hidden layer uses the unipolar S-type logarithmic function logsig, the transfer function from the hidden layer to the output layer uses the double S-type tangent function tan sig, and the learning function is trainlm ()

(3) Determine the number of hidden layers and nodes. First select a single hidden layer network structure. According to the Empirical Formula of $l = \sqrt{n + m + a}$ (where n is the number of neurons in the input layer, m is the number of neurons in the output layer, and a is a constant between $[1, 10]$), hidden layer neurons the number should be between 3-12. Through trial and error traversing the number of nodes in the range of the empirical formula, it is finally determined that the first hidden layer node is 7, and the second hidden layer node is 6, the error is the smallest. At this time, the neural network structure is shown in Figure 1.

![Neural network structure](image)

Figure 1. Neural network structure.

(4) Set the learning rate, expected error and training times. The learning rate is set to 0.01, the expected error is 0.001, and the maximum number of training times is 5000.

(5) Simulation experiment. The predicted value is denormalized using the function mapminmax (), and the image of the actual result and the predicted result is output.
4.2. Forecast results
After the BP Neural Network model is trained, the comparison between the predicted results of 20 samples and the actual results is shown in Figure 2. The predicted results are relatively close to the actual results.

Use the data of these 20 samples to establish Multiple Linear Regression, and use SPSS software to get:

\[ Y = 139619.741 - 671.322 \times X_1 - 10.843 \times X_2 + 7.278 \times X_3 + 20.291 \times X_4 + 1.104 \times X_5 + 1.298 \times X_6 + 1803.426 \times X_7. \]

Compare the prediction results of the corresponding films with the prediction results of the BP Neural Network model, as shown in Table 4. After comparison, it can be seen that the prediction results of the neural network prediction model established in this article are better than the classic Multiple Linear Regression model, and the prediction results are closer to the actual results.

| Film | Actual Error | Predictive Error | Predictive Error | Predictive Error |
|------|--------------|------------------|------------------|------------------|
| 1    | 5.6793       | 10.25%           | 2.7654           | 51.31%           |
| 2    | 4.6553       | 5.89%            | 4.6668           | 0.25%            |
| 3    | 3.6508       | 10.99%           | 2.4923           | 31.73%           |
| 4    | 3.3921       | 4.61%            | 3.2274           | 4.86%            |
| 5    | 3.3977       | 1.58%            | 2.368            | 30.31%           |
| 6    | 3.0999       | 0.01%            | 3.3092           | 6.75%            |
| 7    | 2.5478       | 0.01%            | 2.9496           | 15.77%           |
| 8    | 2.4399       | 4.38%            | 2.3352           | 4.29%            |
| 9    | 2.2371       | 6.19%            | 3.1553           | 41.04%           |
| 10   | 2.2017       | 0.49%            | 2.2352           | 1.52%            |

5. Conclusion
Based on research at home and abroad, this paper takes the domestic box office as the dependent variable, selects a series of indicators, and analyzes the gray correlation degree of the domestic film box office. Seven related indicators, including index, average ticket price, and film type, are used as independent variables. The influencing factors after the calculation of the gray correlation degree are used as the input of the BP neural network, and the domestic film box office is used as the
The network output is used to predict the box office of domestic films and take advantage of the nonlinear prediction of the neural network. Finally, a more accurate prediction result is obtained, and a model for forecasting the box office of domestic films is established. This model can make predictions to a certain extent, the prediction effect is good, and it can predict the overall trend of the film box office and provide a reference for investment companies. It is worth noting that this article puts forward the average ticket price as one of the indicators to measure the domestic film box office forecast, which performed well in the experiment process. However, the accuracy of the box office prediction model in this paper still needs to be improved due to the limitations of data and discretization. The next step is to focus more on improving the accuracy of the model. On the one hand, we will continue to explore the indicators that affect the domestic film box office forecasts, and use more reasonable and accurate tools to screen out indicators that can represent box office forecasts. The second is to explore other prediction models, such as decision trees, XGB and other research models that are more suitable for film box office prediction.

Acknowledgment
The corresponding author of this article is Chen Jibin. The work described in this paper was supported by the Planning Foundation Project for Humanities and Social Sciences Research of Ministry of Education of China (17YJA880014).

References
[1] Yin H, Peng K, Yin Y. Research report on the development trend of the world film industry [J]. Modern Communication(Journal of Communication University of China), 2014, 36(8):1-8.
[2] Zhang J, Rui M. The film industry that economic researchers pay attention to--A review from the perspective of industrial economics [J]. Journal of Social Sciences, 2014(11):39-49.
[3] Sharda R, Delen D. Predicting box-office success of motion pictures with neural networks [J]. Expert Systems with Applications, 2006, 30 (2):243-254.
[4] Chakravarty. The differential effects of online word-of-mouth and critics' reviews on pre-release movie evaluation [J]. Journal of Interactive Marketing, 2010, 24 (3):185-197.
[5] Hu X, Li B, Wu Z. Analysis of influencing factors of movie box office [J]. Journal of Communication University of China(Science and Technology), 2013,(1):62-67+39.
[6] Song Z. Research on the influencing factors of movie box office revenue in the development of cultural industry——Based on an empirical analysis of 50 imported movies [J]. Zhejiang Finance, 2013, (12): 66-68.
[7] Wang Y. Analysis of factors affecting the box office of movies in China [J]. Marketing Research, 2015, (5): 30-32.
[8] Wu F, Qian J, Liu J. A study on box office prediction based on C5.0 decision tree algorithm [J]. Science Mosaic, 2016 (4): 186-192.
[9] Eliashberg J, Elberse A, Leenders M. The motion picture industry: Critical issues in practice, current research, and new research directions[M]. INFORMS, 2006.
[10] Ren D. Design and implementation of movie box office prediction system based on multiple linear regression model [D]. Sun Yat-sen University, 2015.
[11] Ding H, Zheng X, Zhou Z. Analysis of the influencing factors of American animation movies’ box office and construction of forecasting models——A comparative study with real-life movies [J]. Journalism Research, 2016, (1): 76-86+149.
[12] Luo G, Jiang Y, Chen W, et al. Research on movie box office prediction based on linear regression algorithm[J]. Computer Knowledge and Technology, 2019, 15(1): 202-203+213.
[13] Barmand, Chowdhury, Singhha R K. To predict possible profit/loss of a movie to be launched using MLP with back-propagation learning[C]. Proceedings of the 2012 International Conference on Communications, Devices and Intelligent Systems. Piscataway, NJ: IEEE Press, 2012: 322-325.
[14] Shard R, Delen D. Predicting box-office success of motion pictures with Neural Networks
[15] Zheng J, Zhou S. Movie box office prediction modeling based on neural network [J]. Journal of Computer Applications, 2014, 34(3): 742-748.

[16] Yang W. Research on the prediction model of movie box office based on Weibo data [D]. Anhui University, 2014.

[17] Zheng J, Zhou S. Movie box office prediction modeling based on neural network [J]. Journal of Computer Applications, 2014, (3): 742-748.

[18] Han Z, Yuan B, Chen Y, et al. An effective early movie box office prediction model based on GBRT [J]. Application Research of Computers, 2018, 35(2): 410-416.

[19] Chen B, Xu M. Research on the prediction model of movie gross box office based on LARS-SVR [J]. Journal of Shaanxi Normal University (Natural Science Edition), 2018, 46(1): 10-15.

[20] Wang J. The development and evolution of contemporary Western movie box office forecasting research [J]. Film Art, 2009, (1): 45-49.

[21] Wang X. A brief review of the history of movie box office forecasting research [D]. Chongqing University, 2015.

[22] Sharda R, Delen D. Predicting box-office success of motion pictures with neural networks [J]. Expert Systems with Applications, 2006, 30 (2):243-254.

[23] Nelson R A, Glotfelty. Movie stars and box office revenues: an empirical analysis [J]. Journal of Cultural Economics, 2012, 36 (2):141-166.

[24] Dellairocas C, Awad N, Zhang X. Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning [J]. ISCS 2004 Proceedings, 2004, 14 (2):30.