Possible Forecasting Method for Box Office

Xinyi Zhang
No. 18, Xuezheng Street, Gaojiaoyuan District, Xiasha, Hangzhou City, Zhejiang Province, China
angela@cas-harbour.org

Keywords: Box office, Multiple linear regression, Random forest, Recurrent neural network.

Abstract. The booming development of film industry and the emergence of high-return films have attracted the attention of investors. It is well known that high returns mean high risks. In order to avoid risks, scholars and practitioners have studied various box office prediction models. This paper elaborates the thought about the possible forecasting method for box office. At first, the author selects some features based on the experience and previous research. Next, some traditional features are removed and redefined. After determining the original influencing factors, multiple linear regression and random forest selection are used to select the factors with strong significance. Finally, the author chooses films which are not in the sample range and takes their data into recurrent neural network to get the predicted results. Getting the consequence, the author compares with the actual box office to confirm the effectiveness of the method. This paper is expected to provide readers with a research idea.

1. Introduction

With the continuous development of economy and society, people have paid more and more attention to the spiritual life, and film has gradually entered people's life. The 2018 global box office hits a new milestone, reaching $41.7 billion. China's film market is also diversifying. It is evident that some conclusions or rules of traditional experience about Chinese films and audiences may be failing. This paper first reviews the research results of domestic and foreign film box office prediction. In addition to traditional features, scholars will consider streaming media factors, such as website reviews, film scores and so on. One thing is assumed that investors cannot obtain consumer evaluation, in the early stage of film investment, so these data are replaced by the previous evaluation of similar films. Through the empirical research method, the author builds a film box office prediction model and determines the relationship between features and box office. Finally, the author uses the best way to predict the box office.

2. Literature Review

As a project with high investment risk, the concern of its economic return is one of the core issues in the film industry. With the maturity of films in various stages, such as production and publicity, it is gradually possible to predict the box office of films. Practitioners and scholars are constantly exploring the prediction mechanism, hoping to take a more active position in the market.

Foreign film markets have a relatively early start and Hollywood has a complete set of business models, so the research on film box office is relatively mature. There are several classic prediction models in the study of American box office[1]. Litman[2] believed that the originality of the film, the release time and marketing methods affected the success of a film. Scott Sochay[3] added the number of screening weeks in the dependent varied and increased market concentration in the influencing factors. So far, the research model of film box office has developed from static state to dynamic state. Yong Liu[4] studied the impact of Internet reputation on the box office, taking the number of posts and positive and negative comments on yahoo's film website as factors. Eliashberg[5] said the number of pre-release reviews and post-release reputation were also important.

In early China, researchers analyzed box office qualitatively, and even thought about it from the perspective of art. Only after 2009 did many scholars carry out empirical studies on the influencing factors of film box office. Yusong Zhang et al.[6] established a linear regression model with the box office by selecting film investment, film quality, directors, actors, film sequels and piracy. Minxue Huang et al.[7] predicted the final box office of the film was based on the number of online reviews in the first week of the film's showing. Zheng Wang[8] believed that the box office distribution of Chinese films was skewed, so he established Logit regression after eliminating outliers. Yanru Ma[9] analyzed the box office based on the theoretical basis of online rating influencing consumer decisions and combined it with the other six factors affecting the box office.

Different scholars have put forward different influencing factors, and even the same factors, such as the influence of movie stars, inspired different views. The interpretation of each feature is complex and changing. From the perspective of research methods, most scholars adopt multiple linear regression model for prediction, which provides idea worthy of reference for this paper.

3.Method and Analysis

3.1 Data

This paper assumes to collect the box office data from 2013 to 2018, remove the films with missing information, and delete the data of films with box office less than 10 million in order to avoid the influence of outliers. The author captures the data of network platforms, such as douban, microblog, baidu and other platforms with a large number of users, as the basis for obtaining partial variable values about user emotion.

This paper carries out all variables by multiple linear regression. In detail, using stepwise regression is better, and the software helps us find important features. The author mainly adopts the importance score of variables based on permutation, defines the set of the overall training sample set, and expresses the factors affecting the box office by vector. For the overall training sample, Bootstrap sampling is adopted to generate K subtraining sample set and denote the Kth sample subset as D_k. The importance of each influencing factor is measured by construction of the random forest and the importance score of the random forest. In this way, the author compares the importance of various influencing factors and screens indicators to select the box office influencing factors with higher importance for subsequent box office prediction tasks. By observing the results of multiple tests, it can be seen that the value of the importance score of each influencing factor fluctuates within a certain range. Therefore, when measuring the importance of factors, this paper adopts the method of finding the mean value through multiple tests.

In general, this paper selects the final variable by combining the two methods.

3.2 Approach

3.2.1 Multiple Linear Regression

Multiple linear regression[10] attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Since the observed values for y vary about their means u_y, the multiple regression model includes a term for this variation. In words, the model is expressed as DATA = FIT+RESIDUAL. Where the "FIT" term represents the expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_p x_p$, the "RESIDUAL" term represents the deviations of the observed values y from their means u_y, which are normally distributed with mean 0 and variance σ. The notation for the model deviations is ε. Formally, the model for multiple linear regression, given n observations, is

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_p x_p + \varepsilon \quad (1)$$

This model assumes that the explanatory variable x is non-random. When all independent variables are 0, p_0 is the estimated population average value of y. When the independent

variable $x_2,...,x_k$ is fixed, y changes by an average of β_1*1 for every unit of x_1. Parameter $\beta_0,\beta_1,...,\beta_k$ of the multiple linear regression model is estimated by the least square method and satisfies the basic assumption, otherwise the least square estimation is no longer valid.

In order to check whether the above basic assumptions are satisfied, correlation tests are usually carried out on the model, such as LB test, first-order sequence correlation test, multicollinearity test, F test and T test, etc.

3.2.2 Random forest

Random forest[11] is a kind of a combination of multiple independent decision tree integrated classifier. The principle of its decision-making can be described as: a number of experts gathered together identify a particular task according to their own experience and give the correct result. Random Forest through expert voting method, based on the principle of "the minority is subordinate to the majority", finally classifies results. Its generation process can be divided into the following steps: 1) Randomly extract data from the overall training set with Bootstrap method to generate k sub-sample sets and k out-of-pocket data. 2) Select appropriate node splitting algorithm according to the principle and method of decision tree construction to construct k independent decision trees according to the k sub-sample sets generated by random extraction. 3) Integrate the k decision trees generated in Step 2 to build the random forest ensemble classifier. 4) Input the test set into the random forest classifier, and use the random forest classifier constructed in Step 3 to classify it.

3.2.3 Recurrent Neural Network(RNN)

A recurrent neural network [12] is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Unlike feedforward neural networks, RNN can use their internal state to process sequences of inputs.

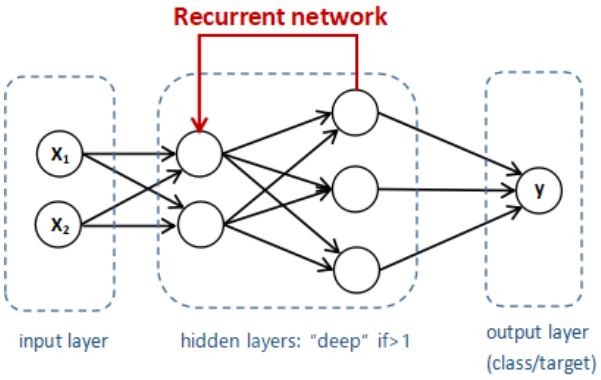

Figure 1. Recurrent Network

Suppose that at time t, the input is x_t, and the state of the hidden layer is h_t. The following functions are commonly used:

$$h_t = f(Uh_{t-1} + Wx_t + b) \quad (2)$$

$$y_t = soft\max(Vh_t) \quad (3)$$

Where U, W and V are all weights, b is the threshold, and the activation function can also be tanh function.

In theory, RNN can process sequence data of any length. However, due to the disappearance or explosion of the gradient of RNN, only short-period dependencies can be learned in practice. In order to reduce complexity, it is often assumed that the current state is only related to a few previous states. Movie box office is closely related to time. Therefore, for the time sequence appearing in the sample, the recurrent neural network can be well solved, which can meet the demand of box office data of the movie.

3.3 Forecast and comparative analysis of results

This paper supposes to take different types of films released in 2019 and make regression prediction and RNN prediction for their box office. The author compares the results with the actual box office and calculates the error and deviation. In general, RNN results are more accurate after reading several studies.

4 Discussion

4.1 Factors that Affect Box Office

4.1.1Movie

- Film type

Combined with the authoritative classification of film types and the development status of domestic film types, films mainly include drama, love, comedy, action, thriller, fantasy, suspense and other types. When quantifying the type variables, the influence of each type is measured mainly with the help of the historical box office data, and the solution formula is as follows:

$$T_i = \frac{\sum_{j=1}^{N_{ti}} Box_j}{N_{ti}} \tag{4}$$

Where, T_i represents the influence of the i film type, N_{ti} represents the number of films belonging to the i film type in the collected samples, and $\sum_{j=1}^{N_{ti}} Box_j$ represents the box office of the j film type belonging to the i film type. This paper mainly considers the first and second types of films.

- Celebrity effect

Consider the influence of the famous actor persistence as well as the popular actor instantaneity. This paper quantifies actors' influence mainly from two aspects. On the one hand, the author measures his lasting influence from the average box office film in the history. On the other hand, the instantaneous impact is gained with the aid of network platform which extracts relevant actors past average network search volume. The solving formula is:

$$A_i = \alpha * \overline{Box_i} + \beta * \overline{NSV_i} \tag{5}$$

Where, A_i is the influence of actor i, α and β are the coefficients of box office history the and network search volume, $\overline{Box_i}$ represents the average box office of the actor in recent films, $\overline{NSV_i}$ is the actor's average Internet searches in his former movie. Usually a film will have a lot of actors, we only consider the first and the second main actors.

- Directer effect

When quantify the director's influence, the writer takes the influence of his identity as a director and other identities that he has into consideration. This paper mainly measures the influence of Internet search on other professional identities through directors' participation in films as directors and actors' participation in films. If the director wins the award, his value increases points, so the director influence of solving formula is:

$$D_i = \alpha * \overline{Box_i} + \beta * \overline{NSV_i} + \theta * \frac{\sum_{i=1}^{N} w_i award_i}{N} \tag{6}$$

Where, D_i is the influence of director i, α, β and θ are the coefficients of box office

history, the network search volume and the rate of award, $\overline{Box_i}$ represents the average box office for a film he directed or starred in, $\overline{NSV_i}$ is the director's average Internet searches in his former movie. $\frac{\sum_{i=1}^{N} w_i award_i}{N}$ is the rate of award and awards are given different weights at different levels.

- Film schedule

Based on the previous researches on film schedule and practical situation, movies seasons are divided into the following kinds: New Year season (a year before the end of November to the end of February next year), May 1st schedule (every year at the end of April to May 3), summer vacation (every year in early June to August 31st), October 1st season (every year at the end of September to October 7th). In this paper, the author quantifies the scheduling variable and assigns a value of 1 during this time, or 0 if not. At last, binary classification is used.

- Manufacturing technology

There are many movies that use a lot of special effects, such as animated movies and superhero movies. The good or bad production technology also affects whether people will go to see the film, which is measured by a score of 0-5.

- Brand effect

A large number of sequels and adaptations have emerged in the film market, assuming that films adaptations and sequels have good reputation. As its existing audience base will have an impact on the future box office, this kind of works will be given a certain weight.

- Whether to import

Generally, imported films which are introduced in the domestic market have gone through market test, and most of them have good box office performance in the overseas market. Author gives weight to the proportion of imported films' box office in the total box office in the past five years, so the solution formula is

$$IM = \frac{\sum im}{\sum movie} \tag{7}$$

Where, im is the total box office of imported films in the past five years, $movie$ is the total box office in the past five years.

- Box office in the first week

The box office in the first week can basically show the future trend of the film. Since there is no box office in the first week that has not been released, the box office of similar films is taken as a reference, and the solution formula is

$$BF_i = \overline{RMf_i} \tag{8}$$

Where, RMf represents box office of similar films in the first week, similar films mean that index similarity reaches 70% in Movie level.

4.1.2 Channel

- Issuing corporation

The film industry has developed for many years and formed a large number of film companies, some of which even have certain brand guarantees in the industry. The ability of their distribution companies is measured by the success rate of their investment, and the return on investment of a film (total box office * 0.4-cost-1) over 50% is considered as a successful investment.

- Publicity

A good film needs not only good content, but also good publicity. And now we often see that a film's publicity expenditure accounts for the majority of the cost, which is measured by 0 to 5 points.

- Film arrangement rate

The film screening rate reflects the optimistic degree of cinemas on the film, which also affects its box office. People have more opportunities to see it due to the high screening rate. Since there is

no actual data before shooting, the average value of similar films in the past is used as an alternative.

4.1.3. Consumer

- Number of film evaluations

The number of comments represents the audience's level of discussion and expectation, which reflects the popularity of the film. We can refer to the discussion degree and preference of major social media for releasing advance notice, and measure it with a score of 0-5.

- Movie rating

The unavailability of data requires us to find alternative data. After much consideration, we refer to the average score of previous similar films, although it has some errors.

4.2 Suggestions

Some of the above factors are traditional and some are combined with modern viewing procedures. Not all factors based on experience and previous research are significantly effective. The paper uses the statistical methods listed above to research and screen the various factors and get the optimization equation. On the choice of method, random forests can simplify subsequent prediction model of input, and the prediction effect of recurrent neural network is very good. However its internal hidden layers belong to the black box and we cannot obtain transfer data from the input layer variable and the training times is also difficult to choose. Because of the randomization of parameters, the prediction results cannot be reproduced. This paper cannot directly show its prediction effect without direct data support. Also we cannot assure what is the best forecasting model. The meaning of paper is to provide some thoughts for readers and expect greater development in this field.

5. Conclusion

This paper mainly studies the influencing factors of box office and tries to predict box office. Based on previous researches, the author assumes that use various hypothesis testings in multiple linear regression and random forest to analyze and screen the influencing factors, and utilize multiple linear regression and recurrent neural network to predict the out-of-sample films. This paper tries to figure out which method is the best estimate. Because the prediction in this paper is the box office prediction made during the investment period, it lacks timeliness compared with other factors selected in other articles. It mainly takes the traditional box office prediction into consideration, and most of the user experience variables have been deleted and replaced. Therefore, compared with the actual box office of the film, the error is relatively big, and the total box office at the later stage is greatly affected by the public opinion.

References

[1] J. L. Wang, Development and Evolution of Contemporary Western Box Office Prediction Research [J]. Film Art, 2009,324 (1): 45-49.

[2] B. R. Litman, L. S. Kohl, Predicting Financial Success of Motion Pictures: the 80s experience[J]. Journal of Media Economics, 1989, 2(2): 35-50.

[3] S. Sochay, Predicting the Performance of Motion Pictures[J].Journal of Media Economics, 1994, 7(4): 1-20.

[4] Y. Liu, Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue[J]. Journal of Marketing, 2006, 70(3): 74-89.

[5] J. Eliashberg, S. K. Hui, Z. J. Zhang, From Story Line to Box Office: a New Approach for Green-lighthg Movie Scripts[J]. Management Science, 2007, 53(6): 881-893.

[6] Y. S. Zhang, X. Zhang, Analysis of Influencing Factors of Box Office[J]. Economics BBS, 2009, (8): 130-132.

[7] P. Pei, Y. L. Jiang, Analysis of Influencing Factors of Domestic and Foreign Box Office[J]. Cooperative Economy and Technology, 2014, (2) : 18-21.

[8] Z. Wang, M. Xu, Analysis of Influencing Factors of Box Office -- Research Based on Logit Model [J]. Exploration of Economic Problems, 376(11) : 96-102.
[9] Y. R. Ma, Research on the Impact of Internet Word-of-Mouth on Box Office[D]. Capital University of Economics and Business, 2014.
[10]Z. H. Tang, Research on Chinese Box Office Based on Multiple Regression and Neural Network[D]. Hunan Normal University, 2018.
[11]Leo Breiman, Random Forests[J] . Machine Learning. 2001, (1) : 5-32.
[12]C. M. Mi, Y. Lu, Q. T. Lin, Box Office Prediction Model Based on Weighted K-means and Local BPNN[J]. Application of Computer System, 2019, 28(02): 15-23.