

# Predicting the Box Office Model: A Study Based on Network Data

Jie Chen<sup>a</sup>, Shanshan Chen<sup>b</sup>

School of Nanjing University of Science and Technology, Nanjing 210000, China

<sup>a</sup>jie\_chen163@163.com, <sup>b</sup>shanshan\_chen333@163.com

**Keywords:** Film, Box Office Forecast, Multiple Linear Regression

**Abstract:** Box office prediction is one of the important but insufficiently researched fields in the film industry. In this study, a multiple regression prediction model is constructed based on the data of network platform. By collecting the data of 60 films and using the data generated by consumers on the online platform as the index to measure the expectation and consumption intention of watching movies, the factors affecting the box office were discussed. The study found that movie ratings, number of evaluators, attention in the first week of release, and number of "want-to-watch" were significantly correlated with movie box office.

## 1. Introduction

According to data released at the 8th Beijing international film festival, China has become the world's second largest film market after the United States, with the total number of screens leading the world. China's box office grew from 17.07 billion yuan in 2012 to 64.27 billion yuan in 2019, with an annual compound growth rate of 20.85 percent. Behind the rapid growth of China's film market is the rapid expansion of the film consumption market and the film industry. At the same time, the rapid development of the Internet has made great changes in the way films are promoted and distributed and the way they are sold. Before the movie is released, consumers can make consumption decisions through the movie information provided by various channels.

Great changes have taken place in consumers' living environment and movie-watching environment. Film platforms, e-ticket purchasing platforms and new ways of publicity and distribution have brought greater uncertainty to the release and box office performance of films, which makes the prediction of box office more difficult. Therefore, it is necessary to take network factors into account in the new environment to predict the film box office. Various emerging social media and mainstream film websites provide new platforms for film publicity and distribution, such as maoyan film, Mtime, weibo, etc., have accumulated a large number of the most professional and accurate data on film evaluation, film box office, ticket sales, publicity and distribution, and the above data is a reliable data source for predicting film box office.

## 2. Literature Review

By reviewing the previous studies, we can find that researchers have been exploring and verifying the factors related to the film's box office, and incorporating these factors into the box office prediction model, and have achieved fruitful research results. Litman [1], an American film economist, first conducted research on the methods and models of film box office prediction. He chose film idea, launch time and marketing plan as the basic elements, and used the method of hierarchical regression to predict film box office, which laid a foundation for the research of scholars. The rapid development of the Internet makes it easier to obtain data, which promotes the development of the research on film box office prediction. Sharda [2] et al. regard the prediction problem as a classification problem and used neural network to divide films into different categories, but the prediction accuracy of this model was not ideal. The existing research includes many factors that affect the box office of films. Hua [3] et al. verify the influence of factors such as public praise, publicity mode of attention, director, leading actor and actress, and film schedule on the box office of Chinese films. Wang [4] select actors, directors, release time, production company, distribution

company, film review, special effects and first-week box office as independent variables, and found that the first-week box office, director and production company were the factors that greatly affected the box office. Zheng Jian et al. [5] establish the box office analysis model by taking into account various factors such as the director, the plot, the type and the release time, and obtained good prediction results. Taking small and medium-sized films as the research object, Guo Xinru et al. [6] establish a multi-step regression model to predict the box office according to the overall cycle factors such as the comprehensive popularity of the directors and leading actors, the theme and schedule of the films, the network response of the films, the cost input, the continuous popularity after the release, and the comprehensive evaluation after the release.

When scholars continue to explore and verify factors affecting the box office, the trend of "pan-film review" in the film market makes the opinions of the third party become one of the important factors to consider. The information provided by entertainment platforms is valuable for the prediction of movie box office, which is also proved by the research results of many researchers. Sitaram Asur et al. and Bernardo A. Huberman [7] select the evaluation information on Twitter about the movie to be released, numerically processed it, and used emotional scoring to mark the psychological changes of subjective evaluation of different audience groups for the released movie. They used regression analysis to predict the box office of several movies with very different styles, and achieved good prediction results. Zhu [8] takes the film review information on Douban Film and Sina Weibo as the research object, and used the linear regression model to predict the total box office and dynamic box office of the film. Zufryden et al. [9] establish an analysis model on the audiences' operation behavior about the movie. They combine with the audience selection rules of the predicted movie and the competing movie to predict the box office of the new movie, and obtained a good effect. This method did not directly use the evaluation information as the data basis like the traditional method, but an indirect analysis method. Wang [10] establishes a movie box office forecasting model which contains web search data, the research results show that the online search volume and growth trend of the film can predict the box office of the first week of the film. Wu [11] establishes a regression model to find that the deep interaction of users on specific movies was significantly correlated with the box office results. Using Sina Weibo as the platform, Shi Wei et al. [12] study to mine emotional information from microblog and build an auto-regression model to predict the box office of films. The results showed that the model proposed in this study was more effective than the auto-regression model.

Throughout the research findings, scholars have conducted in-depth research on the impact of movie attributes and movie evaluation factors on the box office, but less research has taken into account consumers' preferences and expectations. Before the release of a movie, consumers' purchasing decisions will be influenced by many aspects, including the publicity of the movie, the evaluation of the movie and consumers' expectations. In order to consider the factors of consumers, it is necessary to consider how to quantify the choices of consumers. Therefore, this paper makes box office prediction based on available data on consumer behavior.

### **3. Research Process**

#### **3.1 Selection of Influencing Factors and Data Sources**

This research is carried out from the perspective of consumers. Summarizing previous studies, the author believes that an ordinary consumer's decision on whether to watch a movie is affected by the following three aspects: one is the consumer's attention and expectations of the movie; External evaluation; Third, the "menu" provided by the consumer venues, that is, Film row piece volume. In terms of the first point, this study selects the degree of attention before and after the movie is released, the number of "want to watch", and the amount of publicity materials played as data indicators. In terms of the second point, this paper selects the movie ratings and number of reviewers as data indicators; As far as the third point is concerned, the average number of movies in a movie theater for one week is selected as the data index. The following is a brief description of the influencing factors and data sources.

### (1) Attention

The more attention a film receives, the more attractive it is to consumers and the greater the potential economic benefits. In order to distinguish the influence of attention in different time periods on the box office, this paper selects the attention in the week before the movie's release and the attention in the first week of release as the indicators. The amount of information on baidu search within a specified period of time is used as the measurement index of attention, and the data comes from baidu index platform.

### (2) The number of "want-to-watch"

Before the movie is released, consumers get movie information from various channels and form their own expectations for the movie. In this paper, the number of "want-to-watch" provided by maoyan movie platform is used as the measurement index of consumers' expectations for the movie.

### (3) the amount of promotional materials played

The input of publicity materials reflects the intensity of film publicity, and the amount of film publicity materials can reflect the attention of consumers to the film to some extent. The data are from the amount of promotional materials played on the major video websites in the week before the movie's release provided by maoyan professional edition.

### (4) Word-of-mouth

Starting from the two dimensions of word-of-mouth valence and word-of-mouth quantity, this study explores the predictive effect of word-of-mouth on box office. Word of mouth valence has persuasive effect, the higher the score, the more attractive the film to consumers. Word of mouth has an awareness effect, the more reviewers there are, the more people know about the film. Therefore, in this study, the relevant data of Mtime network were selected as the quantitative data of the third-party emotional evaluation, including the movie score and the number of evaluators.

### (5) Film row piece

Cinemas' expectation of box office and arrangement of films will have certain influence on the box office. Therefore, in this paper, the average number of films row piece volume in the first week of release is selected as the measure of the film arrangement, and the data are from the professional version of maoyan movie platform.

In addition to the above influencing factors, This study randomly selected 60 non-anime domestic films released in 2018 and 2019 and with a box office of over 100 million yuan for empirical study, and the film data came from Mtime.

## 3.2 Descriptive Statistics of Data

In a sample of sixty movies, the average box office of a movie is 790 million, which is a large gap from the maximum box office of a movie. It can be seen from Table 1 that the average of the attention of the week before the movie was 156907.74, and the average of the attention of the first week of the movie was 952184.58, indicating that the contribution of the attention of the first week of the movie to the box office was greater than that before the movie Attention.

Table 1 Descriptive statistics for variables

	Min	Max	Average	Standard deviation
Box office	1.01	46.40	7.9307	9.64
Number of evaluators	231.00	1321	2116	2290.11
Rating	3.20	8.80	6.4467	1.05
Row piece volume	1216.25	1949831.57	221800.42	343376.65
Want-to-watch playback amount	1.60	154.40	23.69	25.44
	22.70	2745.30	369.11	466.97
Pre-screening attention	2254.00	2620366.00	156907.74	358735.64
Post-screening attention	2695.00	7198093.00	952184.60	1441769.67

## 3.3 Establish the Regression Model

Before the model is established, logarithm the data, whose function is to control the

heteroscedasticity problem that the cross-section data may have, and to eliminate the difficulty in data processing and interpretation caused by the dimension of different variables. Using SPSS for multiple stepwise regression analysis, the following results were obtained.

Table 2 Model summary table

Model	R	R <sup>2</sup>	Adjusted R <sup>2</sup>	Errors in standard estimates
1	.859 <sup>a</sup>	.738	.734	.56442
2	.898 <sup>b</sup>	.806	.799	.49027
3	.909 <sup>c</sup>	.825	.816	.46897
4	.921 <sup>d</sup>	.849	.838	.44040

a. Predictor variable: (constant), ln (number of evaluators)

b. Predictor variable: (constant), ln (number of evaluators),  
ln (Post-screening attention)

c. Predictor variable: (constant), ln (number of evaluators),  
ln (Post-screening attention), ln(score)

d. Predictor variable: (constant), ln (number of evaluators),  
ln (Post-screening attention), ln(score),  
ln(want-to-watch)

Table 3 Coefficients of regression model

Model		Unstandardized coefficients		Standardized coefficients	T	Sig.
		B	Std.err	Beta		
1	(constant)	-5.543	.553		-10.024	.000
	ln (number of evaluators)	.972	.076	.859	12.788	.000
2	(constant)	-7.091	.593		-11.964	.000
	ln (number of evaluators)	.783	.079	.692	9.967	.000
	ln (Post-screening attention)	.226	.051	.309	4.458	.000
3	(constant)	-8.219	.724		-11.357	.000
	ln (number of evaluators)	.734	.078	.649	9.463	.000
	ln (Post-screening attention)	.206	.049	.282	4.192	.000
	Ln (ratings)	.939	.374	.153	2.509	.015
4	(constant)	-8.541	.689		-12.405	.000
	ln (number of evaluators)	.635	.080	.561	7.887	.000
	ln (Post-screening attention)	.146	.050	.200	2.895	.005
	ln (ratings)	1.568	.412	.256	3.802	.000
	ln (want-to-watch)	.237	.081	.209	2.916	.005

a. Dependent variable: box office

From table 2, we can see that with the increase of the variables, the R<sup>2</sup> of the regression model gradually increases, and the R<sup>2</sup> of the final regression model is 0.921, indicating that the model has a better fitting effect. According to table 3, the regression equation is obtained as shown below.

$$\ln(\text{box office}) = -8.541 + 0.635 * \ln(\text{number of evaluators}) + 1.568 * \ln(\text{ratings}) + 0.146 * \ln(\text{Post-screening attention}) + 0.237 * \ln(\text{want-to-watch}) \quad (1)$$

Through regression analysis, we have concluded that the factors that have a significant impact on the box office of a movie are ratings, number of reviewers, attention in the first week of release, and the number of "want-to-watch".

Word-of-mouth effect reflects consumers' personal judgment on movie quality, and influences potential consumers' decision-making. A high rating is the best publicity for a movie, and it will bring more audiences to the movie, a lower rating will do the opposite. The greater the number of word-of-mouth on the Internet, the more people know about the movie, and higher visibility tends to

generate better box office revenue. In terms of attention, the attention of the first week of the release is more important than the attention of the week before the release, which indicates that the publicity of a film needs to consider not only the investment before release, but also how to stimulate the enthusiasm of users and maintain the popularity of the film as a topic after release. The number of people who "want to see" reflects the number of people who have high expectations for the film. This kind of people have a huge possibility to choose to consume, which has a self-evident impact on the box office.

### 3.4 Model Validation

In order to verify the reliability of the model, this study randomly selected five films released in 2018 and 2019 as the test. These five films, including comedy, love, art and other genres, all took more than 100 million yuan at the box office, meeting the research requirements of the data. The predicted results are shown in table 4.

Table 4 Movie box office forecast results

Movie name	Actual box office	Forecast box office	Prediction accuracy
Us and Them	13.61	12.24	0.9
Till the End of the World	2.34	2.65	0.87
Looking Up	8.66	7.5	0.87
Kill Mobile	3.66	3.3	0.9
Forever Young	7.54	10.3	0.63

According to the comparison between the predicted box office and the actual box office, the average prediction accuracy is 83.4%, which indicates that the regression model in this study can predict the box office well. However, from the prediction results, there is a large deviation in the prediction results of 《forever young》, which may be attributed to the limitation of film types. The film is an art film rather than a commercial one. Chinese consumers' low interest in literary and art films can be seen by the number of people who 'want to see' them. Compared with other high box office films, the number of people who "want to watch" this film is small, but the good reviews and attention generated by the publicity make the film's box office is overestimated.

### 4. Conclusion

This research takes the perspective of consumers as the entry point and builds a regression model to predict the box office of films based on the relevant data of consumers. Through empirical research, it is found that the number of evaluators, rating, number of “want to watch” and Post-screening attention are significantly related to the box office results. The innovation of the research lies in that it gives up the attributes and creative factors of the film itself (such as director, actor, etc.), and takes consumers as the research object to predict the box office performance of the film based on the degree of attention shown by consumers on relevant platforms. However, there are still some deficiencies in the research. The evaluation data used in this study are from a single platform. Considering the scientificity and universality of the prediction model, multi-platform data can be integrated for modeling in the future. The data in this study may be biased towards certain types of films, such as the box office prediction that is more conducive to commercial films. From the verification results, we can also see that the prediction results of literary films have a large deviation. Due to the limited time, it is not possible to conduct diversified modeling according to the country and market size of the film, so the future research can continue to improve this problem.

### References

- [1] Litman B R, Kohl L S. Predicting financial success of motion pictures: The 80's experience[J]. Journal of Media Economics, 1989, 2(2): 35-50.
- [2] Sharda R, Delen D. Predicting box-office success of motion pictures with neural

networks[J].Expert System with Application,2006,12(30):243-254.

[3] Hua Rui,Wang Senlin,Xu Yang. Research on the influencing factors of Chinese film box office [J]. Statistical observation,2019, (4):97-100.

[4] Wang Jinhui,Pu Yanfang,Li Nian. Empirical analysis of box office prediction model of huallywood films [J], journalism bimonthly,2016,(1):9-14.

[5] Zheng Jian,Zhou Shangbo. Box office prediction modeling based on neural network[J]. Computer application,2014,34(03):742-748.

[6] Guo Xinru,Wang Zhen,Zhang Nan. Research on the factors influencing the box office of small and medium-sized films in China - an empirical analysis based on 100 films[J]. Cultural Industry Research, 2015(03):230-245.

[7] Asur S, Huberman B A. Predicting the Future with Social Media[C]. International Conference on Web Intelligence & Intelligent Agent Technology. IEEE, 2010.

[8] Zhu Mengxian. Word-of-mouth and box office: a study on the commercial value of film reviews on social media[J]. Journal of information resource management,2015:57-64

[9] Zufryden, F. Linking Advertising to Box Office Performance of New Film Releases A Marketing Planning Model[J]. Journal of Advertising Research, 2015, 36(1): 29-41.

[10] Wang Lian,Jia Jian-min.Forecasting box office performance based on online search:Evidence from Chinese movie industry[J]. System engineering theory and practice,2014,34(12):3079-3090.

[11] Wu Yu,Pan Xu. Research on the box office prediction model based on the data of user's content consumption[J]. Global Media Journal,2018,5(3):96-107.

[12] Shi Wei,Wang Hongwei,He Shaoyi.Study on predicting movie box office based on sentiment analysis of micro-blog[J].Journal of Central China Normal University(Nat.Sci.),2015,49(1):66-72.