

Movie Score Predication Model Based on Multiple Nonlinear Regression

Xuemei YOU*, Yongdong LIU, Mingming ZHANG, Man ZHANG

Abstract: In the movie industry, the ability to predict a movie's score before its theatrical release can decrease its financial risk. However, accurate predicctions are not easily obtained. To improve the accuracy and scientificity of movie score prediction, this paper proposes a multiple nonlinear regression movie score prediction model (MSPM) in exponential form. Firstly, the influencing factors of film scoring are analyzed. A single problem is selected for the variables of the existing prediction model. This model combines the metadata variables of the film itself and the characteristic variables of film members to conduct quantitative and qualitative analysis on the factors affecting film scoring. Secondly, MSPM is established and the concept of index is introduced. In order to avoid the redundancy of explanatory variables in the MSPM model, the AIC values of the MSPM model and its five sub-models are also calculated to ensure the necessity of selecting explanatory variables. Douban data set is selected to predict movie scores. Finally, compared with linear regression model (Y_s) and equal scale model (Y_{li}), the actual movie score and predicted value were compared. The results showed that MSPM had the highest prediction effect. Experiments show that the model is effective and robust, and reveals the relationship between film scores and related variables. Real-world data confirms that the MSPM model is a timely and appropriate framework for measuring movie scores.

Keywords: data mining; model prediction; movie score; multiple nonlinear regression

1 INTRODUCTION

"When Robbers Meet the Monster", a 200 million-yuan blockbuster starring Louis Koo, was released in 2019, but the result surprised everyone, scoring only 3.5 points on Douban. "Better Days", directed by young director Zeng Guoxiang, grossed more than 800 million yuan in five days and scored 8.3 points on Douban. Film marketing executives are often overwhelmed by such results [1]. A related question arises: "What are movie scores related to?" Maybe social media data can help predict movie scores [2]. The rapid development of big data technology provides a scientific basis for the prediction of online movie scores and box office [3].

In previous studies, many researchers have used movie-related variables, such as multiple linear regression, to predict movie scores. Palomba [4] mined relevant information of movies and movie consumers to build a movie consumption prediction model. Zhou [5] extracted information from film posters to enrich relevant variables of film prediction. Ivasic [6] can provide more information for predicting box office revenue by combining the features of movie posters with other film-related data. However, feature extraction in these methods is usually done manually, does not have good generalization, and relies on prior knowledge. In conclusion, we found that the variables used in the above studies were single, and the variables related to movies were not completely linear with movie ratings. Due to the lack of objective indicators to measure a certain film score, the choice of relevant variables is often controversial. Therefore, it is of great value to establish a quantitative model to evaluate a film, whether for the market or in mass activities. In addition, there are studies from machine learning algorithms; Wang [7] used the multi-layer perceptron neural network to establish a film prediction model. Ghiassi [8] proposed a model based on dynamic artificial neural network (DAN2) for film revenue prediction during production. Hur [9] used three machine learning algorithms to predict movie scores. PLucey [10] builds models to predict film scores. The body and facial expressions of the audience during watching the movie were also analyzed. Jiang [11] collected data on 1266

online movies from 2013 to 2015 and divided them into three categories: high, medium and low, counted by view. After feature selection and feature creation, ordered support vector machine (SVM) model is used to predict movies. Sometimes it is not very efficient. At home and abroad, a large number of research institutions and researchers have made a lot of efforts in this regard [5], but accurate prediction is not easy to obtain [12]. There are mainly two reasons: (1) insufficient variable diversity [13-17]; (2) simple prediction algorithm [18].

Therefore, this paper takes Douban website as the research object, combines the metadata variable of the film itself and relevant characteristic variable of the film staff, and constructs a nonlinear exponential model -- MSPM, to predict the popularity of the film. By verifying the data set and comparing with other models, it is concluded that the prediction effect is better, which can play a certain scientific guiding role for film investors and cinemas [19-21].

2 RESEARCH DATA AND METHODS

2.1 Data Sources and Preprocessing

This paper selects a total of 9220 movies from 2000 to 2019 from movie platform of Douban website. In the process of data cleaning and preprocessing, we deleted repeat data and missing data, and finally obtained a total of 5062 valid data from 2000 to 2019.

The author took into consideration the following two factors when selecting 10 variables related to movies from Douban website. Some selected metadata variables are directly related to the movie itself, such as title, year, genre, country, release time and movie duration, etc. The corresponding characteristic variables related to movie members are selected in the other part, such as the popularity of director, writer and leading actors.

In consideration of people's diversified life preferences, people may also differ in fancying different types of movies, as shown in Fig. 1. On the whole, the number of drama movies is the largest, followed by comedy movies and horror movies in sequence.

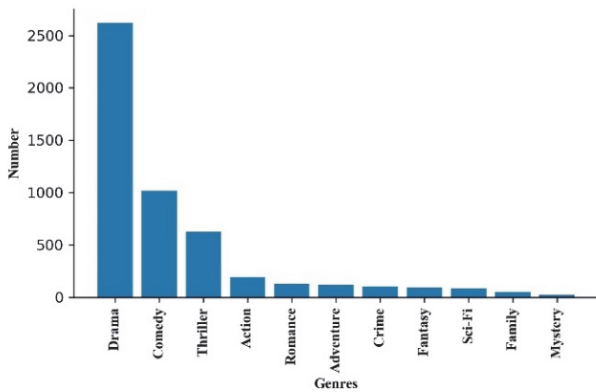


Figure 1 Quantity Analysis of Movies of Various Genres

We took the movie score as an indicator to measure the popularity of the movie, and as the explained variable. The relevant influencing factor indicators, namely, the movie duration, the number of raters, launch time, the director, the writer and the actors, are taken as explanatory variables. The name and symbols of specific variables are shown in Tab. 1.

Table 1 Explanatory Variables and Symbols

	Name	Symbol
Explanatory variables	duration	x_1
	people	x_2
	date	x_3
	director	x_4
	writer	x_5
	actor	x_6
Explained variable	score	y

2.2 Construction of Model

2.2.1 Regression Analysis for Single Factor

According to the above analysis, there are six variables that affect the movie score, including movie duration, the number of raters, the movie launch time, the director, the writer and the actors. We first studied the influence of a single performance indicator on score, and guessed that there was a unitary linear relationship between each single performance indicator and score:

$$y = Ax_i + C \tag{1}$$

where y represents the movie score and x_i represents 6 variables respectively. The fitting results and t-test of variables are shown in Tab. 2.

Table 2 Unitary Fitting Goodness and t-test of the Six Indicators and Movie Score

	x_1	x_2	$1/x_3$	x_4	x_5	x_6
p	0.001	0.000	0.000	0.000	0.000	0.008

Obviously, the fitting goodness judgment coefficient R^2 of director, writer and movie score is < 0.3 , and that R^2 of other indicators and score is < 0.1 . In the fitting process, it was found that the movie score is negatively correlated with the launch time. In the following study, we used the reciprocal $1/x_3$ of the launch time to find its relationship with movie score.

Therefore, the influence of a single explanatory variable on score is not very obvious. Several explanatory variables should be combined instead of explaining the change of explained variable and score y .

2.2.2 Movie Score Prediction Model-MSPM

Dietz [22] established the STIRPAT (Stochastic by Regression on Population, Affluence and Technology) model and introduced the concept of index when they were studying carbon dioxide emission assessment models. The specific expression is expressed as:

$$I = aP^b A^c T^d e \tag{2}$$

where a is the constant term of the ratio, b , c , and d respectively represent the coefficient of the variable, and e represents the random error. The logarithmic form of this model is more commonly used in the actual analysis, namely:

$$\ln I = \ln a + b \ln P + c \ln A + d \ln T + \ln e \tag{3}$$

Therefore, this paper introduces the concept of index based on the constructing principle of STIRPAT model, and extends the STIRPAT model. The original three explanatory variables increase to six, namely, movie duration, number of raters, movie launch time, director, writer, actor, all of which have different influences on movie.

Table 3 Pearson Correlation Coefficient Matrix for the 6 Variables

	x_1	x_2	x_3	x_4	x_5	x_6
R^2	0.004	0.060	0.053	0.176	0.220	0.018
p	0	0	0	0	0	0

Tab. 3 shows that x_4 and x_5 , which are weakly correlated, have the largest correlation among variables. It means that the linear relationship between any two of the six explanatory variables is weak so the six explanatory variables can be considered independent from each other.

Therefore, we proposed a multiple nonlinear regression model equation in exponential form --- MSPM. The specific model is expressed as:

$$y = a \left(\frac{1}{x_3} \right)^{\alpha_1} x_1^{\alpha_2} x_2^{\alpha_3} x_4^{\alpha_4} x_5^{\alpha_5} x_6^{\alpha_6} \varepsilon \tag{4}$$

where x_1 represents the movie duration, x_2 represents the number of raters, $1/x_3$ is the reciprocal of the launch time, x_4 represents the popularity of the director, x_5 represents the popularity of the writer, and x_6 represents the popularity of the actors. a is a constant and the exponent of explanatory variable, and ε represents the random error of the model. In the actual regression fitting and analysis, the model can be converted into an algorithm form and into a linear form, namely:

$$\ln y = a_0 - \alpha_1 \ln x_3 + \alpha_2 \ln x_1 + \alpha_3 \ln x_2 + \alpha_4 \ln x_4 + \alpha_5 \ln x_5 + \alpha_6 \ln x_6 \tag{5}$$

where, $a_0 = \ln a + \ln \varepsilon$ represents the sum of the logarithm of the constant term and the logarithm of the error term.

2.3 Test of Model

2.3.1 Fitting Goodness Test of MSPM

Next, the movie data from 2000 to 2019 were put into the logarithmic form of the MSPM, and the fitting goodness R^2 is 0.2848 was obtained after regression fitting.

And, the adjusted determination coefficient $\overline{R^2}$ is 0.2840. This indicates that the obtained model has a relatively high fitting goodness and can better relate the six variables to the score.

Meanwhile, the value p of F-test of the statistics in the regression results of this model is 0.000, which is less than 0.05, indicating that the regression coefficient of at least one explanatory variable is not 0. The value p of t-test for each explanatory variable is shown in Tab. 4.

Table 4 t-test of 6 Explanatory Variables in the MSPM

	x_1	x_2	$1/x_3$	x_4	x_5	x_6
x_1	1.000	0.190	-0.021	0.080	0.047	0.025
x_2	0.190	1.000	0.149	0.132	0.079	0.078
$1/x_3$	-0.021	0.149	1.000	0.128	0.121	0.116
x_4	0.080	0.132	0.128	1.000	0.524	0.271
x_5	0.047	0.079	0.121	0.524	1.000	0.236
x_6	0.025	0.078	0.116	0.271	0.236	1.000

It can be seen that the p values of the t-test of the six explanatory variables are all less than 0.01, so the corresponding coefficients are not 0. All the six explanatory variables have passed the verification of the t-test with a significance of 1%.

2.3.2 Akaike Information Criterion (AIC) Test of MSPM

To avoid redundancy of explanatory variables in the MSPM and verify that all six explanatory variables are included in the MSPM, we also calculated the AIC values of the MSPM and its sub-models.

Next, we used any three of the four explanatory variables of the MSPM to establish six nonlinear models, which are special forms of sub-models of the MSPM. The expressions of several model equations can be expressed as:

$$ME1: y_{ME1} = A_{11} \left(\frac{1}{x_3}\right)^{\alpha_{11}} x_1^{\alpha_{21}} x_2^{\alpha_{31}} x_4^{\alpha_{41}} x_5^{\alpha_{51}} \tag{6}$$

$$ME2: y_{ME2} = A_{12} \left(\frac{1}{x_3}\right)^{\alpha_{12}} x_1^{\alpha_{22}} x_2^{\alpha_{32}} x_4^{\alpha_{42}} x_6^{\alpha_{62}} \tag{7}$$

$$ME3: y_{ME3} = A_{13} x_1^{\alpha_{23}} x_2^{\alpha_{33}} x_4^{\alpha_{43}} x_5^{\alpha_{53}} x_6^{\alpha_{63}} \tag{8}$$

$$ME4: y_{ME4} = A_{14} \left(\frac{1}{x_3}\right)^{\alpha_{14}} x_1^{\alpha_{24}} x_4^{\alpha_{44}} x_5^{\alpha_{54}} x_6^{\alpha_{64}} \tag{9}$$

$$ME5: y_{ME5} = A_{15} \left(\frac{1}{x_3}\right)^{\alpha_{15}} x_2^{\alpha_{35}} x_4^{\alpha_{45}} x_5^{\alpha_{55}} x_6^{\alpha_{65}} \tag{10}$$

$$ME6: y_{ME6} = A_{16} \left(\frac{1}{x_3}\right)^{\alpha_{16}} x_1^{\alpha_{26}} x_2^{\alpha_{36}} x_5^{\alpha_{56}} x_6^{\alpha_{66}} \tag{11}$$

Then, fitting regression was conducted in the MSPM and the other six sub-models respectively, and the fitting goodness and the value ΔAIC and ΔBIC of the MSPM and several models were respectively solved for comparative analysis. The calculation results of fitting goodness and AIC value of MSPM and sub-model are shown in Tab. 5.

Table 5 Comparative Analysis of 5 MSPM Sub-models

Model	R^2	ΔAIC_{MSPM}	ΔBIC_{MSPM}
MSPM	0.2848	0	0
ME1	0.2838	5.12	-1.4
ME2	0.2156	465.88	459.63
ME3	0.2717	90.12	83.68
ME4	0.2764	57.33	50.8
ME5	0.2833	9.03	2.51
ME6	0.2703	100.06	93.54

$\Delta AIC_{MSPM} = AIC_{MEi} - AIC_{MSPM}$, $\Delta BIC_{MSPM} = BIC_{MEi} - BIC_{MSPM}$, ($i = 1, 2, 3, 4$) respectively represent the difference between the value AIC of model ME1-ME6 and the value AIC of MSPM, and the difference between the BIC of model ME1-ME6 and the value BIC of MSPM. It can be seen from the table that the AIC value of the other 6 sub-models is greater than that of the SPM model, which reveals that one of the six explanatory variables in MSPM is indispensable. Thus, the scientific nature and accuracy of Douban movie score model are determined.

3 EMPIRICAL ANALYSIS

In this paper, the prediction of movie score is based on the following assumptions.

1. Score on Douban movie website reflects the movie quality. The higher the score, the better the movie.
2. For most people, good movies are preferable to bad ones.

3.1 Movie Score

On the whole, the movie score is a single peak trend where decrease comes after increase, which is in line with the general law of nature. There are few movies with score below 5, while there are many movies scored between 6 and 7.5. After the peak, the number of movies scored above 8 declines rapidly, which means there are fewer excellent movies or bad ones, as shown in Fig. 2.

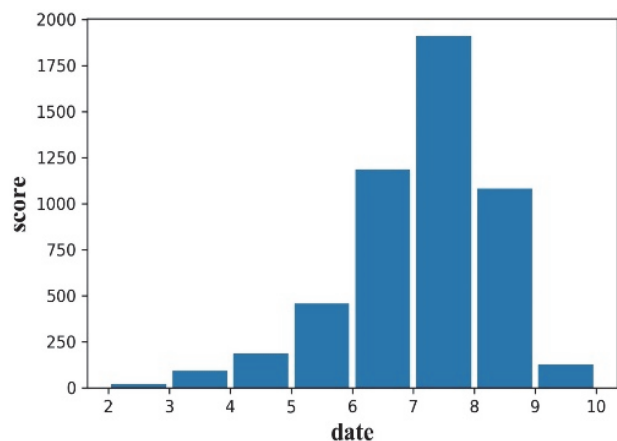


Figure 2 Quantity Analysis of Movies of Various Genres

3.2 Relationship between Movie Score and Other Variables

Movie score is an important indicator to measure the quality of a movie in movie and TV industry. A movie with a high score will attract more people, thus promoting the market supply. Movie score is also influenced by many factors, such as its duration, the number of raters, its launch time and the popularity of director, writer and actor. Therefore, this paper makes an in-depth analysis of the differences in movie score in different dimensions.

3.2.1 Movie Score and Duration

To study the relationship between movie score and duration, scatter plots of movie score and duration are drawn, with the result shown in Fig. 3. The scatter plot of average score and duration is shown in Fig. 4.

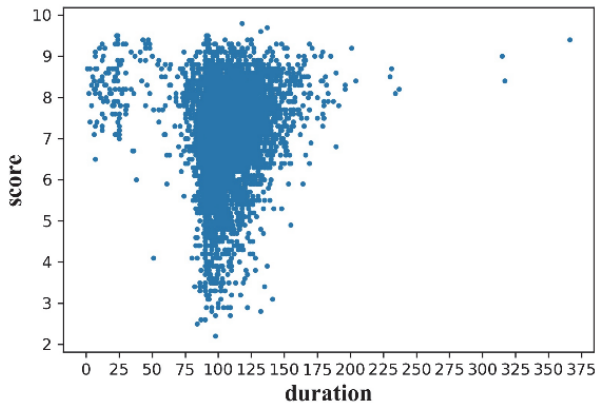


Figure 3 Analysis of Movie Score and Duration

The movie duration is about 80-150 min and the score span is large, with the overall distribution between 4.5 and 9.5 points. But the distribution is more concentrated around 7 to 7.5 points, which shows the general characteristics of the movie duration. It can be seen from Fig. 4 that the score gradually increases within 0-50 min. After the 50 min, the movie score first decreases with the increase of duration and reaches the bottom at about 100 min after which it increases again with the increase of duration.

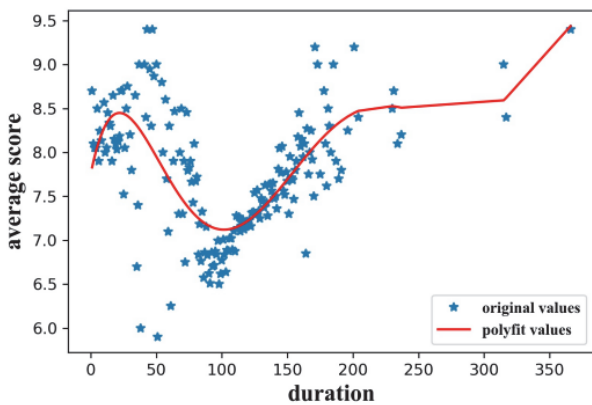


Figure 4 Analysis of Average Score and Duration

3.2.2 Movie Score and the Number of People

To study the relationship between movie score and the number of raters, we made a scatter plot of the two elements, as shown in Fig. 5.

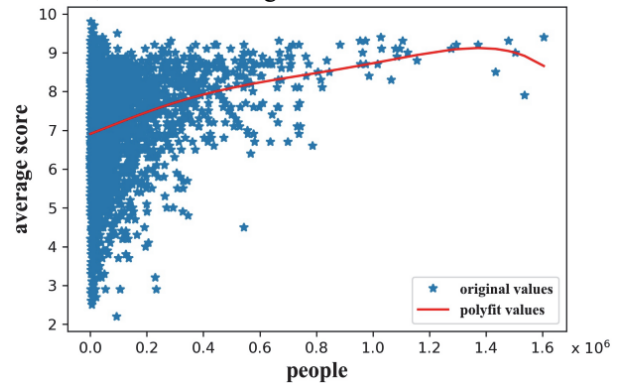


Figure 5 Analysis of Average Score and People

It can be seen from Fig. 5 that there is a positive correlation between the number of raters and the score. The more the raters, the greater the possibility of a high score. At the same time, it also reflects the high popularity of movies and the trend of higher popularity. On the whole, the number of people for most movies is generally under 100,000. The number of raters is an important indicator for users to watch movies and to reflect the excellence of movies.

3.2.3 Movie Score and Date

To study the relationship between movie score and the date, we made a scatter plot of movie score and date, and that of average movie score and the date.

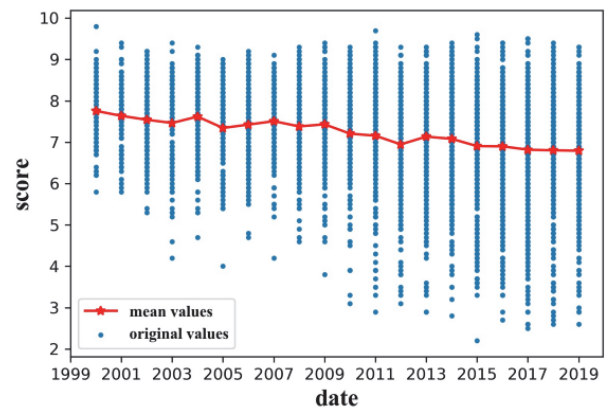


Figure 6 Analysis of Movie Score and Date

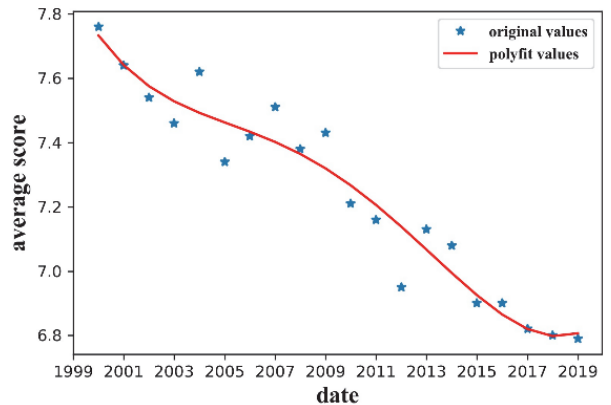


Figure 7 Analysis of Average Movie Score and Date

It can be seen from Fig. 6 that the overall movie score shows a downward trend. Specifically, it can be seen from Fig. 7 that the movie scores from 2010 to 2019 are significantly lower than those from 2000 to 2009. But this does not mean that the movie quality from 2010 to 2019 must be lower than that of previous movies. With the continuous progress of the times, the shooting conditions and transmission channels have been changed substantially.

3.2.4 Movie Score and Director

To study the relationship between movie score and director, we made a scatter plot of movie score and director characteristic variable, and that of average movie score and director characteristic variable.

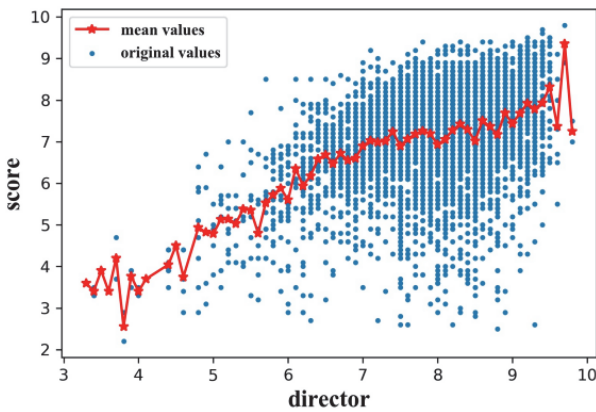


Figure 8 Movie Score and Popularity of Director

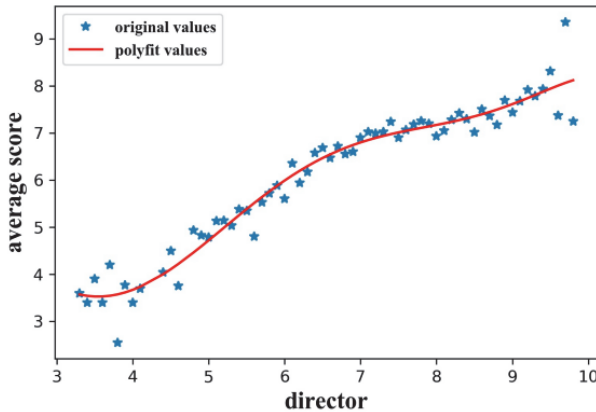


Figure 9 Average Movie Score and Popularity of Director

As shown in Fig. 8, the score of the director is distributed from 6 to 9 points. It can be seen from Fig. 8 and Fig. 9 that the director can greatly influence the movie score which increases along with the popularity of director. The director is just like the supreme commander of the army. The quality of a movie largely depends on the director’s quality and cultivation. The style of a movie often reflects the artistic style and character of the director, and furthermore manifests the director's value of things. Therefore, there is no doubt that a good director exerts dramatic influence on movies.

3.2.5 Movie Score and Writer

To study the relationship between movie score and the writer, we made a scatter plot of movie score and writer characteristic variable, and that of average movie score and writer characteristic variable.

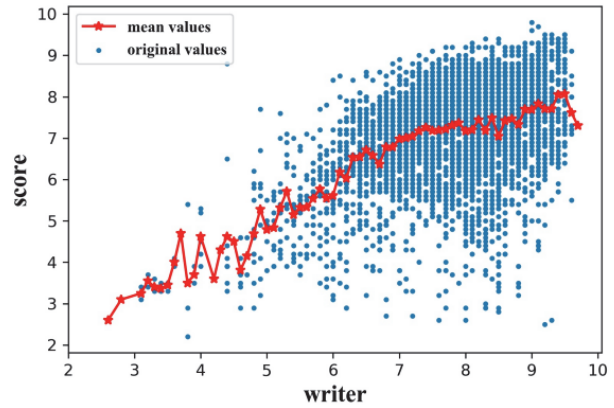


Figure 10 Movie Score and Popularity of Writer

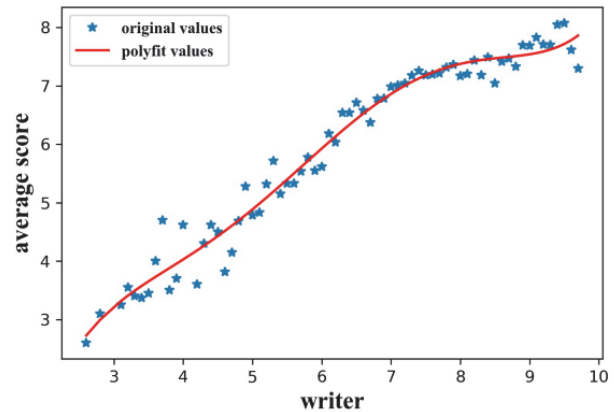


Figure 11 Average Movie Score and Popularity of Writer

It can be seen from Fig. 8 and Fig. 10 that the influence of writer and director on movie scores is similar. That is, the higher the level of writer and director, the higher the movie score. The positioning of a movie and the idea of the script determine the height it can reach in the directing process, the later stage and even when the movie is totally completed.

3.2.6 Movie Score and Popularity of Actors

An excellent movie cannot be separated from the performance of excellent actors. We will study the relationship between movie score and actors in the following part.

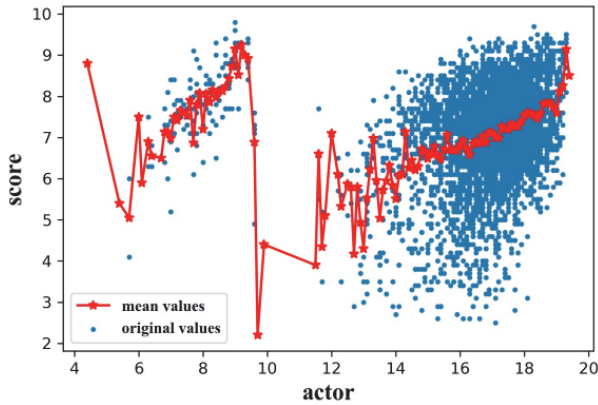


Figure 12 Movie Score and Popularity of Actors

It can be seen from Fig. 12 that the influence of actors on movie score is polarized. According to Fig. 12 and Fig. 13, an actor of average level is also likely to perform a movie with a high score. A good actor has a great possibility of performing a bad movie.

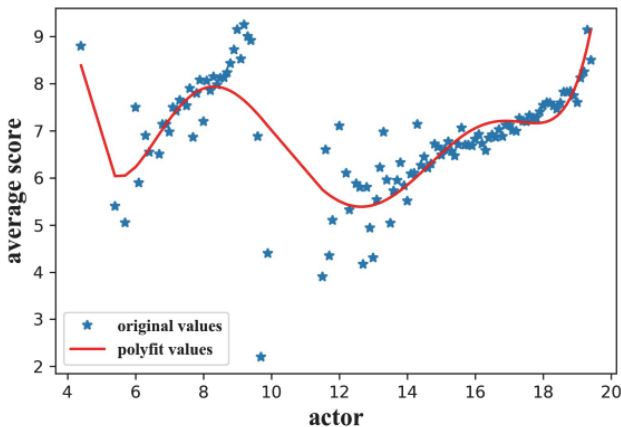


Figure 13 Average Movie Score and Popularity of Actors

4 MODEL COMPARISON AND PREDICTION

4.1 Comparative Analysis with Other Multiple Regression Models

To prove the validity and accuracy of the MSPM, the two constructed multiple regression models are respectively model Y_s and model Y_M , which are compared with MSPM.

The Y_s model is a multiple linear regression model of six explanatory variables and movie score. The regression equation is expressed as:

$$Y_s = \frac{E_{21}}{x_3} + E_{22}x_1 + E_{23}x_2 + E_{24}x_4 + E_{25}x_5 + E_{26}x_6 + \varepsilon \quad (12)$$

where, Y_s is the movie score, E is the coefficient of explanatory variable, and ε is the error term.

The Y_s model is a multiple linear regression model of six explanatory variables and movie score. The two sides of the equation of this model have equal changes, and the comparison with the MSPM constructed in this paper is also to further confirm that the influence of different factors on the popularity of movies is not equal. It is expressed as

$$Y_M = \frac{E_3}{x_3} x_1 x_2 x_4 x_5 x_6 \quad (13)$$

where, Y_M is the movie score. It was transformed into a linear relation by mathematical transformation and then the following was obtained:

$$\ln Y_M = -\ln x_3 + \ln x_1 + \ln x_2 + \ln x_4 + \ln x_5 + \ln x_6 + \ln E_3 \quad (14)$$

4.2 Comparative Analysis of Multiple Models

The data from 2000 to 2019 were respectively introduced into the MSPM, model Y_s and Y_M model. The fitting goodness and the F-test of the three models are shown in Tab. 6.

Table 6 Fitting Goodness and the F-test of the 3 Models

	MSPM	Y_s	Y_M
R^2	0.2848	0.2745	0.0231
\bar{R}^2	0.2840	0.2012	0.0162
p	0.000	0.000	0.000

4.3 Predictability of MSPM

To evaluate the robustness of MSPM, we substituted training data into the three models MSPM, Y_M , Y_s to test the predictive ability of MSPM. The observed data of 20 years from 2000 to 2019 were respectively used as training data to fit the equation, and the parameter fitting values $a_1, a_2, a_3, a_4, a_5, a_6$ of the MSPM under the training sample sets of different years were obtained. The variation trends of $a_i (i = 1, 2, 3, 4, 5, 6)$ and a_0 are shown in Fig. 14 and 15.

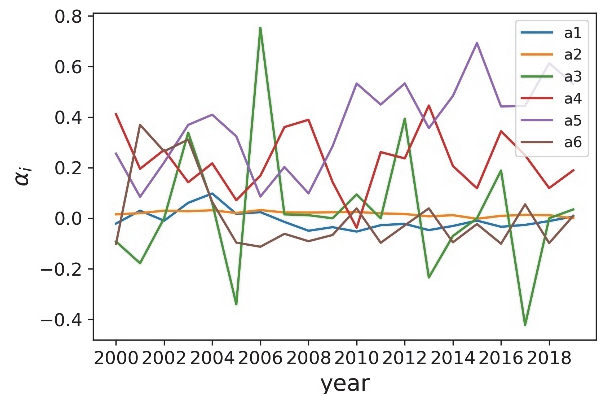


Figure 14 Variation Trend of $a_i (i = 1, 2, 3, 4, 5 \text{ and } 6)$

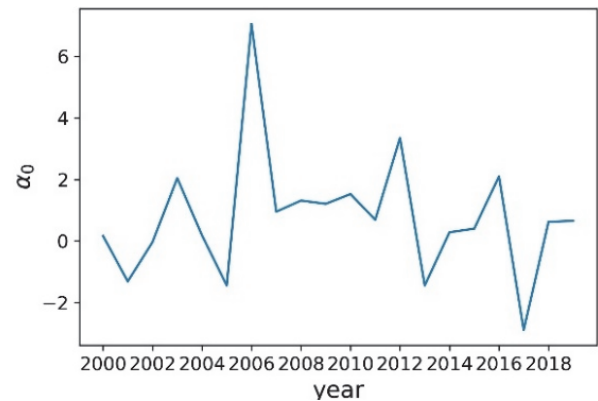


Figure 15 The Variation Trend of a_0

where, $a_0 = \ln a + \ln \varepsilon$. It can be seen from Fig. 14 that the coefficient a_i fluctuates gently over time, and Fig. 15 shows that the $a_0 =$ value fluctuates around $a_0 = 0$. Therefore, we used the observation data in 2019 to fit the MSPM, and then substituted the calculated parameter fitting value $a_0 =$ and $a_i (i = 1, 2, 3, 4, 5, 6)$ into the equation. Then the actual score Y of the movie was compared with the predicted value Y_E . Fig. 16 is the scatter plot of the relationship between the actual score Y and the predicted value Y_E of the movie based on the MSPM.

Fig. 17 is the difference scatter plot between the predicted value and the actual value of the movie score in 2019. where, the red section is the data point for four highly popular movies, which are Better Days, Wandering Earth, Captain Marvel, Frozen 2 and X-Men: Black Phoenix respectively. Fig. 17 shows that most of the data points are distributed near the line $Y = Y_E$, which means that the predicted values are consistent with the observed data set. The red star points in the figure represent the real values and predicted values of the five hot movies, which are distributed near the line $Y = Y_E$.

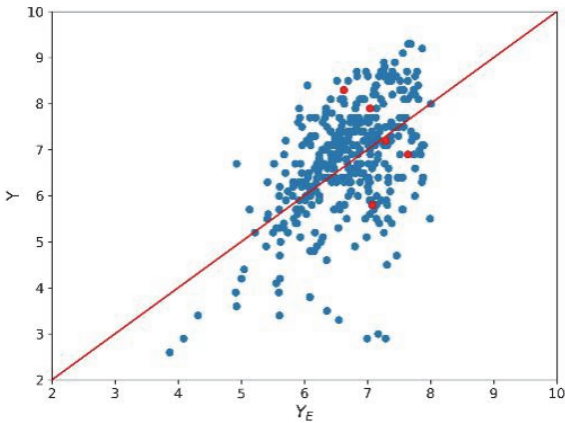


Figure 16 Comparison of Actual Value and Predicated Value of MSPM

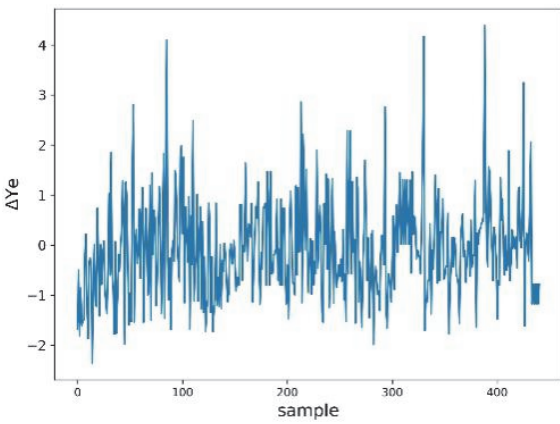


Figure 17 $\Delta Y_e = Y_E - Y$

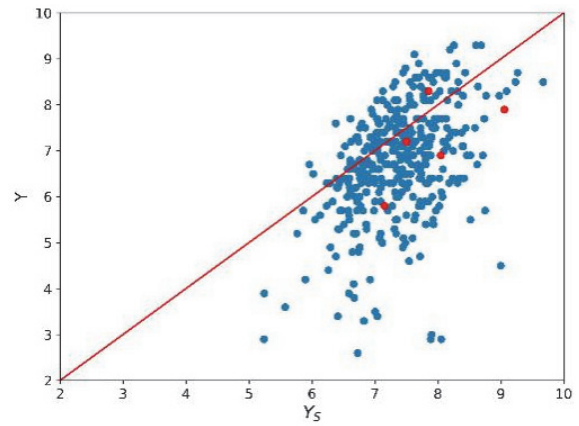


Figure 18 Comparison of Actual Value and Predicated Value of Model Y_s

Fig. 18 and Fig. 19 are the scatter plots of the relationship between the predicted value and the actual score of the model, and the difference diagram. Although the points in Fig. 18 are uniformly distributed around line, the difference between the two in Fig. 19 is mostly above 0 compared with that in Fig. 17. It reveals that the prediction accuracy is higher than that of the model.

Most of the points in Fig. 20 are above, indicating that the predicted value of the model is basically greater than the actual value of the movie score. As shown in Fig. 21, the difference is between -5 and -10 , which is a big error. Therefore, the predictive power of MSPM is better. To sum up, MSPM has more accurate prediction ability.

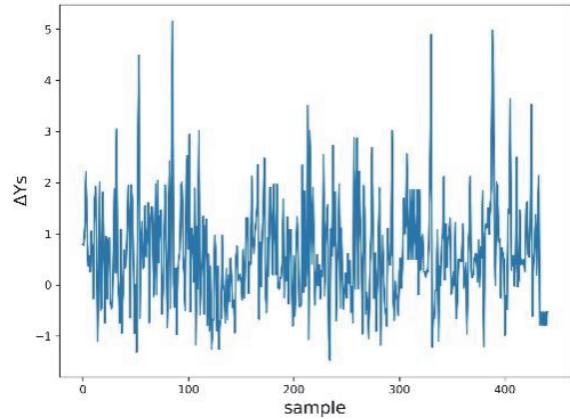


Figure 19 $\Delta Y_s = Y_s - Y$

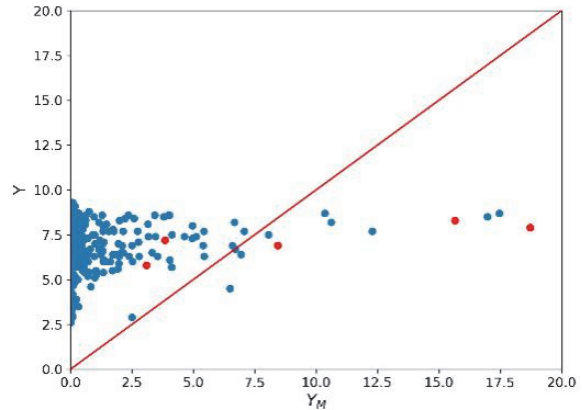


Figure 20 Comparison of Actual Value and Predicated Value of Model Y_M

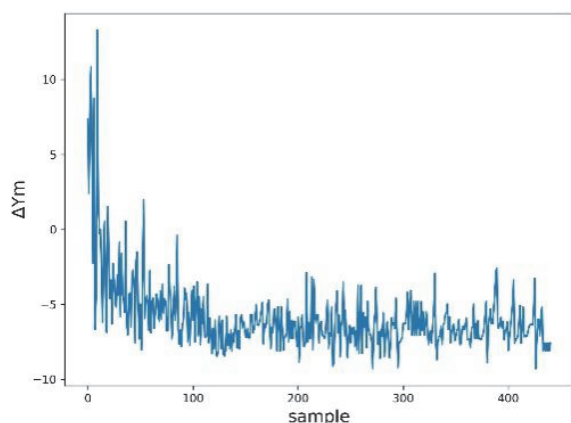


Figure 21 $\Delta Y_m = Y_m - Y$

5 CONCLUSIONS AND FUTURE WORK

This prediction can be used by the audience to choose a movie according to the movie online score, serve as the basis to make the screening plan for the movie theater, and also provide the direction and reference for the peripheral industries. On the basis of the previous studies, this paper explores the factors that influence movie score, and concludes the relationship between the three characteristic factors, namely, the director, the writer and the actors, and movie score. Also, it proposes a prediction model of movie online score on the Douban website -- MSPM. Compared with many other models, the constructed MSPM has a better prediction effect. MSPM can evaluate performance and observation value in other fields that can be independently measured. In other fields, such as scientific research, the reputation of some scholars is closely related to their published records and research findings.

This paper presents a movie online score prediction model based on Douban data. In the follow-up work, experiments are expected to be based on more movie score platforms to continuously optimize the adaptability and prediction accuracy of the movie online score prediction model.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 71701115, Grant 71471106, Grant 71704098, Grant 71904110 and Grant 61903236, the Natural Science Foundation of Shandong Province (No. ZR2017MF058).

6 REFERENCES

- [1] Ahmed, U., Waqas, H., & Afzal, M. T. (2019). Pre-production box-office success quotient forecasting. *Soft Computing*, 24(9), 6635-6653. <https://doi.org/10.1007/s00500-019-04303-w>
- [2] Barman, D., Singha, R. K., & Chowdhury, N. (2013). Prediction of Possible Business of a Newly Launched Film using Ordinal Values of Film-genres. *International Journal of Intelligent Systems and Applications*, 5(6), 53-60. <https://doi.org/10.5815/ijisa.2013.06.07>
- [3] Liu, T., Ding, X., Chen, Y., Chen, H., & Guo, M. (2014). Predicting movie Box-officer venues by exploiting large-scale social media content. *Multimedia Tools and Applications*, 75(3), 1509-1528. <https://doi.org/10.1007/s11042-014-2270-1>
- [4] Palomba, A. (2020). Consumer personality and lifestyles at the box office and beyond: How demographics, lifestyles and personalities predict movie consumption. *Journal of Retailing and Consumer Services*, 55, 102083. <https://doi.org/10.1016/j.jretconser.2020.102083>
- [5] Zhou, Y., Zhang, L., & Yi, Z. (2017). Predicting moviebox-office revenues using deep neural networks. *Neural Computing and Applications*, 31(6), 1855-1865. <https://doi.org/10.1007/s00521-017-3162-x>
- [6] Ivasic-Kos, M., Pobar, M., & Mikec, L. (2014). Movie posters classification into genres based on low-level features. *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. <https://doi.org/10.1109/mipro.2014.6859750>
- [7] Wang, Z., Zhang, J., Ji, S., Meng, C., Li, T., & Zheng, Y. (2020). Predicting and ranking box office revenue of movies based on big data. *Information Fusion*, 60, 25-40. <https://doi.org/10.1016/j.inffus.2020.02.002>
- [8] Ghiassi, M., Lio, D., & Moon, B. (2015). Pre-production for casting 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>
- [9] Hur, M., Kang, P., & Cho, S. (2016). Box-office for a casting based on sentiments of movie reviews and Independent subspace method. *Information Sciences*, 372, 608-624. <https://doi.org/10.1016/j.ins.2016.08.027>
- [10] Navarathna, R., Lucey, P., Carr, P., Carter, E., Sridharan, S., & Matthews, I. (2014). Predicting movie ratings from audience behaviors. *IEEE Winter Conference on Applications of Computer Vision*. <https://doi.org/10.1109/wacv.2014.6835987>
- [11] Jiang, F. & Shang, W. (2018). Prediction of Movie Playback Based on Ordinal Support Vector Machine Classification. *2018 IEEE International Conference on Big Data (Big Data)*. <https://doi.org/10.1109/bigdata.2018.8622468>
- [12] Oh, C., Roumani, Y., Nwankpa, J. K., & Hu, H. (2017). Beyond likes and tweets: Consumer engagement behavior and movie box office in social media. *Information & Management*, 54(1), 25-37. <https://doi.org/10.1016/j.im.2016.03.004>
- [13] Oghina, A., Breuss, M., Tsagkias, M., & Rijke, M. D. (2012). Predicting IMDB Movie Ratings Using Social Media. *Lecture Notes in Computer Science Advances in Information Retrieval*, 503-507. https://doi.org/10.1007/978-3-642-28997-2_51
- [14] Ratnayake, M. P. (2018). Can we predict movie Revenue? *International Journal of Multidisciplinary Studies*, 5(1), 83. <https://doi.org/10.4038/ijms.v5i1.114>
- [15] Tatar, A., Amorim, M. D., Fdida, S., & Antoniadis, P. (2014). A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5(1). <https://doi.org/10.1186/s13174-014-0008-y>
- [16] Ghiassi, M., Lio, D., & Moon, B. (2015). Pre-production for a casting 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>
- [17] Zhang, H., Li, J., Ji, Y., & Yue, H. (2017). Understanding Subtitles by Character-Level Sequence-to-Sequence Learning. *IEEE Transactions on Industrial Informatics*, 13(2), 616-624. <https://doi.org/10.1109/tii.2016.2601521>
- [18] Gopinath, S., Chintagunta, P. K., & Venkataraman, S. (2013). Blogs, Advertising, and Local-Mark et MovieBox Office Performance. *Management Science*, 59(12), 2635-2654. <https://doi.org/10.1287/mnsc.2013.1732>
- [19] Hur, M., Kang, P., & Cho, S. (2016). Box-office for a casting based on sentiments of movie reviews and Independent subspace method. *Information Sciences*, 372, 608-624. <https://doi.org/10.1016/j.ins.2016.08.027>

- [20] Kim, T., Hong, J., & Kang, P. (2015). Box office for a casting using machine learning algorithms based on SNS data. *International Journal of Forecasting*, 31(2), 364-390. <https://doi.org/10.1016/j.ijforecast.2014.05.006>
- [21] Giretti, I., Correani, A., Antognoli, L., Monachesi, C., Marchionni, P., Biagetti, C., Carnielli, V. P., et al. (2021). Blood urea in preterm infants on routine parenteral nutrition: A multiple linear regression analysis. *Clinical Nutrition*, 40(1), 153-156. <https://doi.org/10.1016/j.clnu.2020.04.039>
- [22] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723. <https://doi.org/10.1109/tac.1974.1100705>

Contact information:

Xuemei YOU, Associate professor
(Corresponding author)
Business School, Shandong Normal University,
Jinan 250000, China
E-mail: 412095307@qq.com

Yongdong LIU
Business School, Shandong Normal University,
Jinan 250000, China
E-mail: 1565062265@qq.com

Mingming ZHANG
Business School, Shandong Normal University,
Jinan 250000, China
E-mail: 1359309797@qq.com

Man ZHANG
Business School, Shandong Normal University,
Jinan 250000, China
E-mail: 2017158501@qq.com