Studies in Economics and Business Administration
https://www.mfacademia.org/
ISSN:0081-7872
Vol. 3, No. 1, 2023

Developing a Predictive Model for Movie Valuation Using Data Mining Techniques

Xiao Ma
Lamar University
Maxiao@lamar.edu

Abstract:

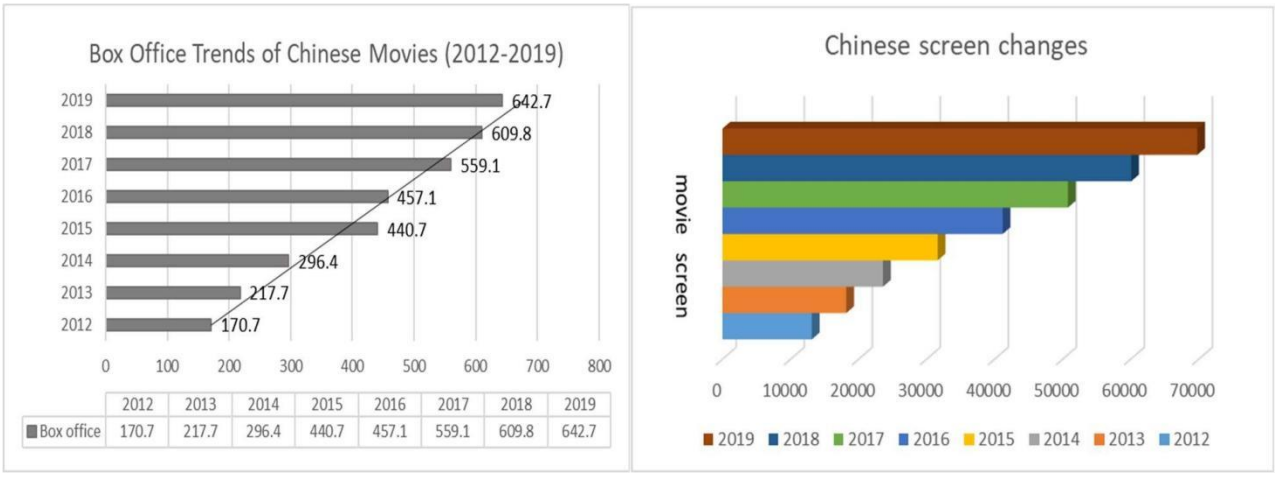
With the rapid advancement of the social economy, film and television culture has become an integral component of public life in China due to its widespread accessibility and appeal. To enhance the overall quality of films and optimize resource allocation within the industry, it is essential to conduct thorough research on the value attributes of films and establish a comprehensive value prediction system. By employing deep learning and data mining technologies for film and television value prediction, relevant information regarding film and television attributes can be extracted. This approach facilitates the concretization and informatization of film and television art, leveraging technology to support the film and television market in nurturing outstanding works.

Keywords:

Deep Learing; Dataming; Movie Value Forecast; Predictive Index System.

1. Introduction

The value of movies is diversified, and it is difficult to directly use scientific calculation methods to evaluate. The following Fig.1 (a) and (b) show the development status of China's film and television market in the past eight years. It can be seen from the figure that China's film and television industry has been developing rapidly from 2012 to 2017, during the period from 2017 to 2019, the development speed is basically stable [1~3]. The fundamental reason is that people's expectations for film and television are higher than the value shown by the film and television industry at this stage.



(a) (b)

Fig 1. Film and television development

The current research on the film and television industry is basically based on the commercial value of film and television as a starting point. The data sample is single, the research method lacks innovation, and it is unable to provide reference value data basis in the early stage of film and television [2]. Data mining and deep learning technology are integrated into film and television research, and the prediction time is extended to the stage of film and television shooting preparation work to minimize Waste of resources.

This article collects 30k film and television data from Douban Movie Data Network (DouBan. com), and applies data mining and deep learning techniques to the film and television industry. Finally, we can complete the evaluation of the film's value before the film is officially produced.

2. Data mining association model

2.1. Data mining association model

The construction of the prediction index system needs to be combined with the basic attributesof film and television. The basic attributes of film include: Director, actor, screenwriter, type of film and television, duration of film and television, investment amount of film and television, distribution place of film and television, language type of film and television, and script of film and television.

The number of viewing times of historical works by directors, actors, and screenwriters has a positive relationship with the ratings of film and television works [1~2]. The ratings of films and televisions may affect the number of viewers of the film, but they do not have sufficient For relevance, the following Fig.2 shows the relationship between the ratings of mainstream movies and the number of comments.

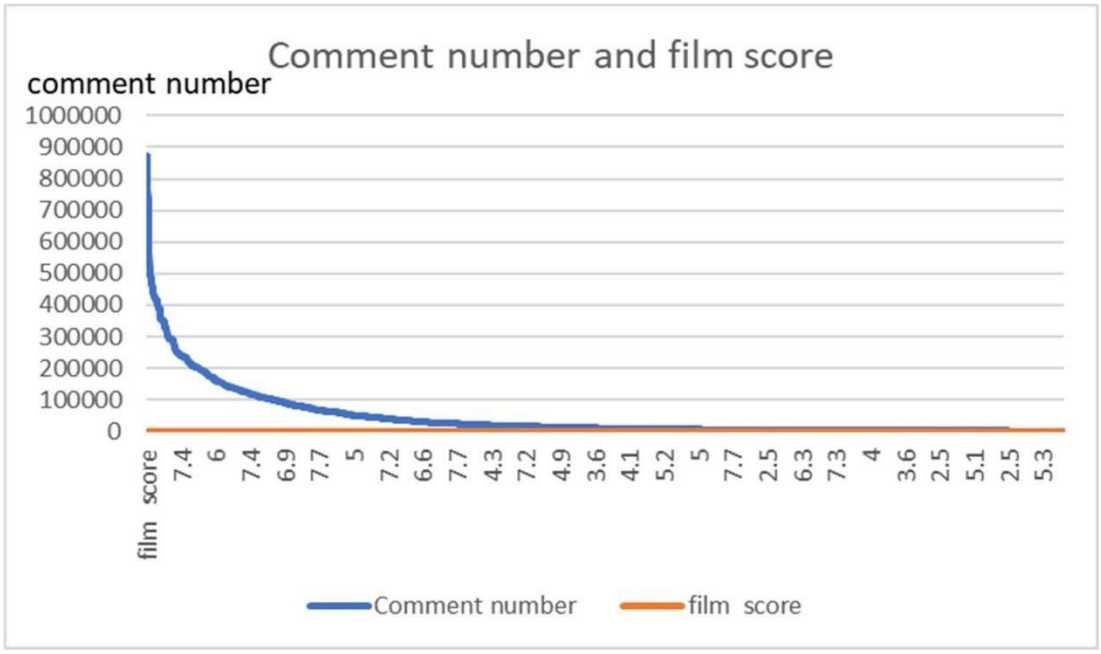


Fig 2. Comment number and film score

After analyzing the correlation between the number of film critics and ratings of historical works of directors, actors, and screenwriters in Figure 2, it is found that the production team isdirectly related to film and television ratings, and the team information belongs to the data available in the early stage. The type of film and television is added to the predictive indicators. Finally, our prediction system includes three first-level indicators and seven second-levelindicators for film and television, as shown in Fig 3.

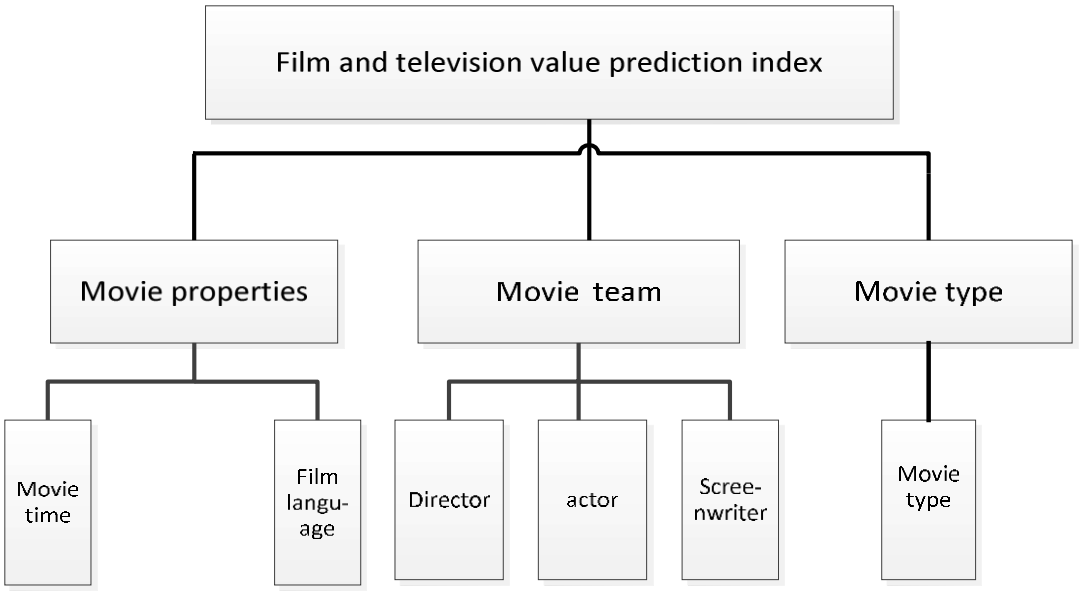


Fig 3. Film and television value prediction index

For the convenience of presentation, in the following, the film attribute is represented by the number 1, the film team is represented by the number 2, and the film type is represented by the number 3.

2.2. Feature extraction algorithm

Based on the feature extraction of the movie team, movie type, and movie attributes, using statistical ideas to calculate the mean value of the three first-level index content of the movie using historical data as the reference point, the specific algorithm is shown in definition. The same attribute of different film works is different. Due to the particularity of the film industry, the method of weighting is used for digitization when processing. The name of the person that appears first has the largest weight, and then gradually decreases, as shown in definition.

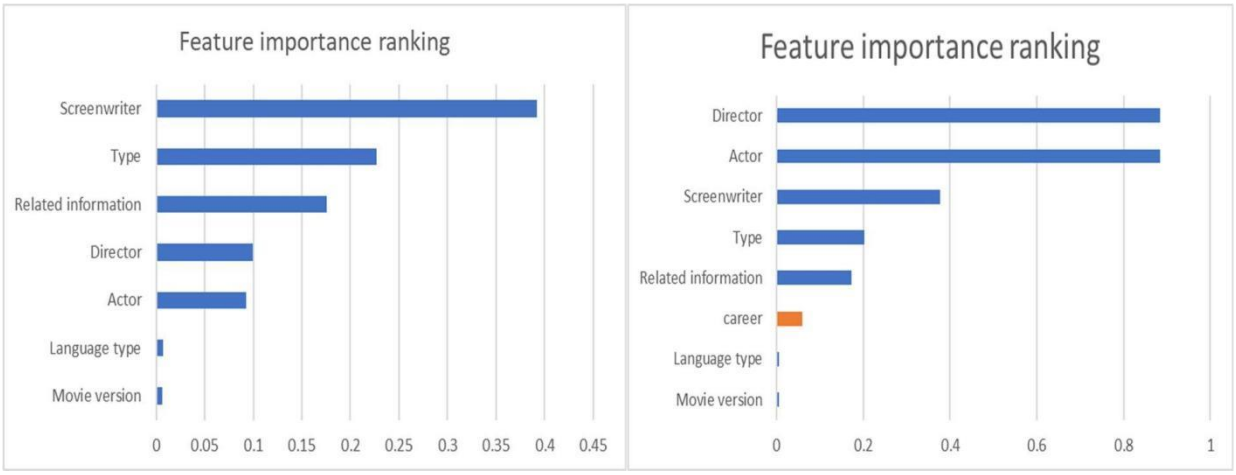
Ht to publish this material in their paper. Use italic for emphasizing a word or phrase. Do not use boldface typing or capital letters except for section headings (cf. remarks on section headings, below).

In the literature [4], the Robert J research team pointed out that different directors and actors have different influences on their works. The works of film practitioners at different time period also have certain regularities. This article explores the potential relationships among actors, directors, and screenwriters in movies, and encodes the growth of film team members in the industry. The statistical method is as follows: encode the scores of the historical works of the film team members as the digital expression of the potential relationship[5~8]. First, calculate the individual average of the works of each member of the team, and then calculate the overall average, Coding rules for team member growth: based on the principle of nearest neighbor and using different weights to accumulate the work of team members in the film industry, the closer the time point, the greater the weight. The definition of career is as follows. Brackets [3, 4]. The references are to be numbered in the order in which they are cited in the text and are to be listed at the end of the contribution under a heading references, see our example below.

In the film team, due to the influence of personality, appearance, age and other factors, the combination of different members may have a positive or negative impact on the film, encoding the relationship between the film team can be used to maximize the value of the film [15].

Many research articles at this stage have realized and gradually solved the quantification problem of the film team, but many researchers have adopted natural language processing methods for classification. This method is still in the experimental stage, and the commercialization model needs

a certain degree time. The statistical ideas adopted in this article have a certain data basis, are relatively simple, and have good stability. In order to fully explore the characteristics, this article encodes the careers of film team members. This featureis mainly used to distinguish the differences in the arts presented by actors at different stages. According to the research and argumentation in the literature [9], we can see that the actors are The level of acting at different stages is different. Feature Importance Ranking Figure 4, It shows the ranking of actors' career characteristics before and after.



(a) (b)

Fig. 1 Feature ranking

Algorithm: Feature extraction

Begin:

1:Input: Movie directors, actors, screenwriters, etc.
2:While the data is null>=3:
delete the data
3:If the data include Garbled:
re.sub("Garbled"," ")
4:If the team of movie have more works:
$y = w_1 * x_1 + \sum_{i=2}^{n} w_i x_i (w_1 >> \sum_{i=2}^{n} w_i)$
5:Deal the movie type:
Use the one-hot code ⟶ [0,1]
If the film include the type:
Film_type=1
Else:
Film_type=0
6:Get movie works and online time,
if A(time)-B(time)>0:
A(W)>B(W)

End

Fig. 2 Calculation process

As shown in Figure 4, before the feature of career was added, the weight of film language and film version percentage is smallest, and the screenwriter accounted for the largest weight. After the addition of the new feature, the overall ranking did not change much. This shows that the new feature was added this time. It has a certain positive influence on the value of movies [10~12]. The above is the core idea of this feature extraction algorithm. The code writing rulesthis time are shown in Fig 5:

In the era of new media, the frequency of personnel in the film industry is relatively frequent. In order to reduce the influence of the "newcomers" in the film team on the final result, the director, actor, and screenwriter will be separated and counted for each attribute in the featureextraction work. The number of "new generation" groups is used as a quantification method forthe "new generation"

group. Therefore, this feature extraction work extracts a total of 68 features. This feature extraction work fully explores the relationship between the individual members of the film team and the team.

2.3. Forecast model based on data mining

This work collects a large amount of movie data, and the encoding features are relatively independent and sparse. Based on the above points, the prediction algorithm of the model needs to have the ability to deal with sparse and non-linear data, so this article chooses Deep neural network algorithm (DNN) and LIGHT_gbm are used as prediction algorithms [13].

The perceptron is the smallest unit of the neural network, and DNN can be understood as a neural network with many hidden layers. Multi-layer neural network and deep neural network DNN are basically the same. DNN is also called multi-layer perceptron (MLP). DNN is divided according to the location of different layers. The neural network layer can be divided into three categories, input layer, hidden layer and output layer.

The setting of the input layer is based on the number of extracted features. This time, a total of 68 dimensions are extracted. Therefore, the input layer is set to 68 neural network nodes. The setting rules of the hidden layer are based on Kolmogorov's theorem. Here, Kolmogorov is Deformation, the final formula used is as follows, the other layers use 1/2 equal ratio formula for dimensionality reduction, and so on to complete the setting of the entire hidden layer network [12~15]. Because the output adopts regression, this time The output layer is set as a node.

$$S = 2N + d(0 < d < 10) \quad (1)$$

$$S = 3N + d(0 < d < 100) \quad (2)$$

Deep neural network is a discriminative model, which can be trained using back propagation algorithm. The weight update can be solved using the following stochastic gradient descent method.The Stochastic Gradient Descent (SGD) algorithm is difficult to choose an appropriate learningrate.The learning rate so small that slow network convergence, too large may affect the convergence, and cause the loss function to fluctuate on the minimum, or even gradient divergence.Adaptive Moment Estimation can calculate the adaptive learning rate of each parameter. It not only stores the exponential decay average of the previous square gradient of AdaDelta, but also maintains the exponential decay average of the previous gradient M(t), which is similar to momentum: M(t) Is the average value of the gradient at the first time, and V(t) is the non-central variance value of the gradient at the second time.

$$M_t = m_t/(1-\beta_1^t) \quad (3)$$

$$V_t = v_t/(1-\beta_2^t) \quad (4)$$

The two formulas are the average value at the first moment and the variance at the second moment of the gradient. The final update formula is as follows:

$$\theta_{t+1} = \theta_t - \eta * m_t/(\sqrt{v_t} + \varepsilon) \quad (5)$$

LightGBM originated from Microsoft Research Asia. Similar to XGBoost, Light_ GBM is still an improved implementation under the framework of the GBDT algorithm. It is a fast, distributed and high-performance GBDT framework based on the decision tree algorithm. It improves the efficiency and scalability of GBDT framework algorithms when facing high-dimensional big data. "Light" is mainly embodied in three aspects, namely, fewer samples, fewer features, fewer internals, Gradient-based One-Side Sampling, Exclusive Feature Bundling, and histogram algorithm (Histogram) three technologies. In addition, in terms of engineering, Light_ GBM has also made many optimizations in parallel computing, supporting feature parallelism and data parallelism, and optimized their respective parallel methods to reduce the amount of communication [13~15].

"Light" is mainly embodied in three aspects: fewer samples, fewer features, fewer internals, Gradient-based One-Side Sampling, Exclusive Feature Bundling, and Histogram. In addition, in terms of engineering [11], LightGBM has also made many optimizations in parallel computing, supporting feature parallelism and data parallelism, and optimized their respective parallel methods to reduce the amount of communication.

2.4. Parameter settings

The parameters of the two algorithms is set according to the rules introduced above. The specific parameter settings are shown in Table 1. This paper uses the function of variance to calculate the respective errors of the two algorithms, and selects the algorithm with relatively small error as the final forecasting algorithm.

Table 1. Parameter algorithm

The parameter of DNN	
The name of the parameter	Parameter
The node of Input layer	68
The number of Hidden layer	7
The node of Output layer	1
Activation function	ReLu
Optimization function	Adam
Error function	Mean squared error
The parameter of Light_GBM	
The name of the parameter	Parameter
Min data in leaf	60
Learning_rate	0.001
Max depth	-1
Min child samples	15
Feature fraction	0.8
Bagging freq	1
Bagging fraction	0.8
Bagging seed	11
Lambda	0.1
Verbosity	-1
Number leaves	600

2.5. Results analysis

Divide the data into two categories based on the amount of data, observe the stability and accuracy of the two algorithms with different amounts of data, and the results are shown in thefollowing table:

Table 2. Algorithm comparison

Error comparison of different data volumes		
The amount of data	Algorithm	Error
150000	DNN	0.09
	LIGHT_GBM	0.07
300000	DNN	0.04
	LIGHT_GBM	0.06

As shown in the above table, the accuracy of the two algorithms is gradually increasing with the increase of the amount of data, but the room for improvement of the deep neural network is greater than that of the Light_ GBM algorithm. This article finally chooses DNN as the final prediction algorithm. In order to verify the effect of the system, my team randomly selected 5 movies and televisions for prediction. The specific effects are shown in the table below.

Table 3. Predictive effect

Effect of model			
Movie	Real Score	Prediction score	Error
Venom	7.2	7.1	1.4%
Shazam	5.6	5.6	0
Andhadhun	8.3	8.4	1.2%

Mirage	7.8	7.7	1.3%
Extreme Job	7.8	7.3	6.4%

Through the analysis of movie data, it is found that the reason for the high error is that some members of the movie team are missing data and cannot be obtained, thereby reducing the accuracy of the prediction. It shows that the way this article deals with the new generation of actors needs to be improved. In the literature [13-15], it is proposed to replace the work information with academic information. In the future research work, we can consider researchin this direction.

3. Summary

Scoring is one of the manifestations of film value, The film industry expects to have a scientific forecasting system to reduce investment risks, This paper redefines the film value score, buildsa film value prediction index, and advances the film prediction timeline, so that minimizing therisk of film investment. The article uses data mining technology to associate independent film information, DNN is used as a prediction algorithm to predict the film scores, Finally, we validate the effectiveness of the proposed model through case study. and finally the overall error is controlled within 7%. Future research work can be linked to personal learning experience, and personal personality can be used in this research.

References

[1] Global Views - Communication and Mass Media; New Communication and Mass Media Data Have Been Reported by Researchers at University of California Santa Barbara (A Graph-Learning Approach for Detecting Moral Conflict in Movie Scripts)[J]. Journal of Engineering,2020.

[2] Machine Learning; Islamia College Peshawar Researchers Add New Data to Research in Machine Learning (Summarizing Online Movie Reviews: A Machine Learning Approach to Big Data Analytics) [J]. Information Technology Newsweekly,2020.

[3] Science - Science and Technology; Study Data from University of Porto Provide New Insights into Science and Technology (Hallucinating Goncalo M. Tavares' "Short Movies")[J]. Science Letter,2020.

[4] Science - Social Science; Study Data from State University Rio de Janeiro Provide New Insights into Social Science (The movies I live in: cartogenealogies of the present)[J]. Science Letter,2020.

[5] Global Views - Translating and Interpreting; Data from Katholieke Universiteit Leuven Broaden Understanding of Translating and Interpreting (How to understand the Other. Multilingualism and Translingualism in the Latin American Road Movie)[J]. Politics & Government Week,2020.

[6] Li Zhixun, Jin Xiche. A Study on the Influence of Filmmaking Factors and Promotions on the Intention of Watching Movies[J].Paper of Korea Entertainment Industry Association,2019,13(7).

[7] Jin Zhenghao, Jin Zaicheng. Performance analysis of Directors, Producers, Main Actors in Korean Movie Industry using Deciles Distribution (2004-2017)[J]. Paper of Korea Cultural Information Society,2018,18(10).

[8] Antaigan, sun Jingai. Research For Shaw Brothers[J].Global cultural content,2016(25).

[9] Myeong Hwan Kim. Determinants of revenues in the motion picture industry[J]. Applied Economics Letters,2013,20(11).

[10] Li Xiaoren. A Study of Yun Bong-chun's Diary during 1935~1937 focused on analysing of realrity ofthe Korean movie world[J]. Film Studies,2013(55).

[11] Jooyoung Kwak,Liyue Zhang. An Empirical Study on the Determinants of the Box-Office Performance of the Foreign Films in China *[J]. International Area Studies Review,2011,14(2).

[12] Permanent residents. Study on the factors affecting the box office - Focused on the factor of patriotism - [J]. Research on Cultural Industry,2010,10(2).

[13] Robert J. Cirasa. Movies Were Always Magical: Interviews with 19 Actors, Directors, and Producers from the Hollywood of the 1930s through the 1950s (review)[J]. Film & History: An Interdisciplinary

Journal of Film and Television Studies,2004,34(2).

[14] Jeong Yoon Su. Subnet Generation Scheme based on Deep Learing for Healthcare Information Gathering[J]. Journal of Digital Convergence,2017,15(3).