Movie Data Analysis and a Recommendation Model

Hang Yang[1], Yiqing Pei[2], and Zhihui Wang[3](✉)

[1] Data Science, Lanzhou University, Lanzhou, China
[2] Data Science, Durham University, Durham, UK
[3] Culture and Technology, Sungkyunkwan University, Seoul, Korea
floraw0702@163.com

Abstract. The high-risk nature of the movie industry has kept companies looking for better plans to make and invest in a film. With the deepening of Internet applications, it's possible to clearly visualize and demonstrate the overview of the film market and the performance of different movie genres with their unique features in the market. The box office is an important indicator to measure movie performance, and it to some extent reflects the economic development of modern society. So, explore the rules of box office and its correlation with other features of the film. Moreover, to establish a recommendation model, by vectorizing and attaching values to characteristics of the films, the distance between two movies can be calculated to find similar movies. Thus, suggestions on preparing budget, choosing genre and casting could be given to the film producer based on these historical movie data so as to achieve higher audience reviews and financial returns.

Keywords: Visualization · Box Office · Movie Recommendation

1 Introduction

As a core cultural product, film has the dual attributes of culture and economy. On the one hand, movies meet the spiritual needs of consumers; on the other hand, movies have the attributes of commodities, the box office is an important indicator to measure movie performance. And from a deeper perspective, the box office to some extent reflects the economic development of modern society, because it represents people's demand for their own spiritual consumption apart from the basic needs of life. According to data from the statistical company ComScore, global movie box office revenue has increased from USD 13.1 billion in 2011 to USD 42.5 billion in 2019, an average annual increase of 36.75%. Affected by COVID-19, the box office of global theaters has declined in 2020. It is widely believed that with the advancement of COVID-19 vaccination, the epidemic will be effectively controlled, and the demand for movies from the public will recover in an orderly manner, and the global film industry will maintain a rapid rebound. In the past 20 years, the global film industry has developed rapidly, and academic research

H. Yang, Y. Pei, and Z. Wang—These authors contributed equally.

D. Qiu et al. (Eds.): ICBBEM 2022, AHIS 5, pp. 739–751, 2023.
https://doi.org/10.2991/978-94-6463-030-5_74

based on the film industry has also shown a rapid upward trend, mainly involving different disciplines such as marketing science, organizational behavior, economics, and communication. The influencing factors of movie box office and accurate revenue forecast performance have become the research focus, which is of great significance for investment decision-making [2]. Many scholars analyze the main factors affecting the box office from different perspectives and use various modeling and empirical methods to predict the box office and an overview of their research and models will be briefly explained.

Based on the past research, for the analysis model, linear regression is the most used one. When talking the about box office value, Hu, Li and Wu [4] constructed a multiple linear regression model to study the box office value of films. The dependent variable of the regression model is the box office of films, and the independent variables are actors, directors, sequels, remakes, release dates, film types and countries of production. And their results show that different factors have relatively different effects on film box office.

In addition, the influence of different contents of films and other external factors on film earnings has also been mentioned in past studies. For example, bass diffusion model and its variants were used by Yuan and other writers [14] as an analysis model to analyze the relationship between online film reviews and box office revenue of Korean films. Garcia and Zarco [3] used the evaluation model to study the influence of different film contents on film revenue. Also, Rui and others [10] applied the public data on social networks and the commonly used machine learning methods to find out the influence of tweets on film revenue.

Box office forecasts is also a research hotspot. Hur and others [5] built six different machine learning models to improve the preparation of box office predictions when the different time periods of a movie's release were taken into accounts as the main factor. Kim and others [7] achieved machine learning models from three different perspectives and averaged them to get the optimal algorithm model. Feedback neural network algorithm is used by Barman and others [1] to predict the income and cost of a movie, so as to indirectly predict the box office, but this method only uses the information of the movie type.

The high-risk nature of the film industry has kept companies looking for ways to accurately predict earnings. However, because movies are affected by very complex social factors, even the most experienced filmmakers often cannot accurately grasp it. Nonetheless, with the deepening of Internet applications, the possibility of accurately predicting the box office through information technology has continued to increase. Explore the rules of box office prediction through computers with different algorithms and variable combinations. The main purpose of this paper is to analyze which movies have better profitability based on historical movie data and future movie trends, and to establish a recommendation model and provide suggestions for movie shooting. The rest of this paper is organized as follows. Section 2 shows the data. Section 3 summarizes the methods. Section 4 shows the results and Sect. 5 concludes the paper.

2 Data

In this paper, the data comes from TMDB [https://www.themoviedb.org/], containing all movie information from 1969 to 2016. It is divided into two parts: information about movie itself and information about credits of movie, and the datatype mainly consist of two parts: number and text.

The acquisition of only text data is not enough for analysis. For data in number there still exist some problems: as the time span is so large and data of different categories of movie varies tremendously, it is unreasonable to measure a film by pure number given in the data set. So, the first step is to standardize the data may be affected by time period and categories. For example, a good way to measure films' profitability is by using earning rate, which is obtained from revenue divided by budget.

Another strategy applied in data preprocessing is to vectorize the data which will be explained in detail in the method part. Here is a brief introduction. The concrete operational approach is to compare the current data with one or a group of value, and then convert textual or numeric data into tabular vectors consists of 0 or 1, based on comparison results. Through this strategy, the waste of textual data in mathematical analysis is greatly avoided, and the data in the form of vector is very favorable in further exploration.

Finally, about data cleaning, recordings contain missing value about their commercial value are removed to make sure the analysis and subsequent recommendation is accurate and effective.

3 Methods

As a preliminary film market research for the film producer and the investment company, this report attempts to demonstrate the overview of film market for decades including the category of each film, its corresponding budget, cast and a series of other possible factors. Thus, making recommendations to the film generator on financing, casting and publicity once a company would like to make a movie with a specific theme. So, the method applied in this research is mainly focused on the movie recommendation system.

3.1 Movie Recommendation Model

Inspired by JIAO, Qing-zheng (2010) [6] and PENG, W. (2013) [8], it's found very reasonable to deal multidimensional data with vectorization. And cosine distance metric learning could be a very straight and efficient way to describe the correlation between different vectorized data.

As a result, the core idea of this recommendation model is that when given a movie that is similar to the one the film producer is preparing, then the top 10 closest movies will be listed along with their related information, so the studios can think of these past films as a guide or a cautionary tale. And to briefly explain the model mathematically, the main approach is to vectorize all the features related to a movie and then use the sum of all these vectors to find the most similar films.

When a movie's profitability is to be measured, the most important thing is to define aspects to be considered. After referring to Wallström, K. (2018) [13], the factors can be determined and given certain weights to fit the target. In addition, as from Walls, W. D. (2005) [12], the revenue and earning rate should be added to make this model more oriented to measure the profitability. So according to the previous data understanding part, there are six features that were generally considered significant when building the model. And from the sorted data structure, these features are the genre of a movie, it's cast, director, keywords and most importantly, the score and earning rate. After identifying the characteristics to be analyzed, what should be noticed is that not all of them are written in the data format of number, so these features need to be quantified at the very beginning.

To introduce the model from the mathematical point of view, each feature of a movie can be changed from its own data format, for example, a list of strings to a vector by the model. In order to do this, logic values "0" and "1" should be assigned to these features through a screening. To be more specific, when a movie covers one of the six features mentioned above, the elements in the feature of a movie can be replaced by "1" in the list of string, otherwise, and elements that are not available can be replaced with 0 in the list of string.

For the numerical variables such as the score and earning rate, the screening should obey a particular standard. To explain in detail, these two features need to be applied with the function 'describe()', first, according to SU, Y. (2018) [11], 75% is a representative number of a standard to distinguish whether a movie is excellent or not. So, 75% is applied to work as the criterion for screening. In other words, these two features are considered valid only when their score and earning rate exceed 75% of the other ones. Among all the movies, the level of 75% are 4.34 for earning rate and 6.8 for the score respectively. So, to conclude, "1" can be set to the earning rate more than 4.34, and to the score more than 6.8. Therefore, vectors can be created and for each feature, vectors among the movies can form a single matrix as the same feature contains the same maximum number of elements, and the model can use the matrix to find the distance between the two vectors (Fig. 1).

PENG [8] introduced cosine distance to describe the Affinity-disaffinity relationship as it could be easily applied to deal with the multi-dimensional data, and the cosine distance could be expressed as below:

$$dist(A, B) = 1 - cos(A, B) = \frac{\|A\|_2\|B\|_2 - A \cdot B}{\|A\|_2\|B\|_2} \tag{1}$$

Now it's time to consider the weight of different aspects. And the final distance between two movies is shown as below:

$$\begin{aligned} TotalDistance = {} & 0.8 * genereDistance + directDistance + cartDistance \\ & + wordsDistance + 2 * earnDistance + 1.3 * scoreDistance \end{aligned} \tag{2}$$

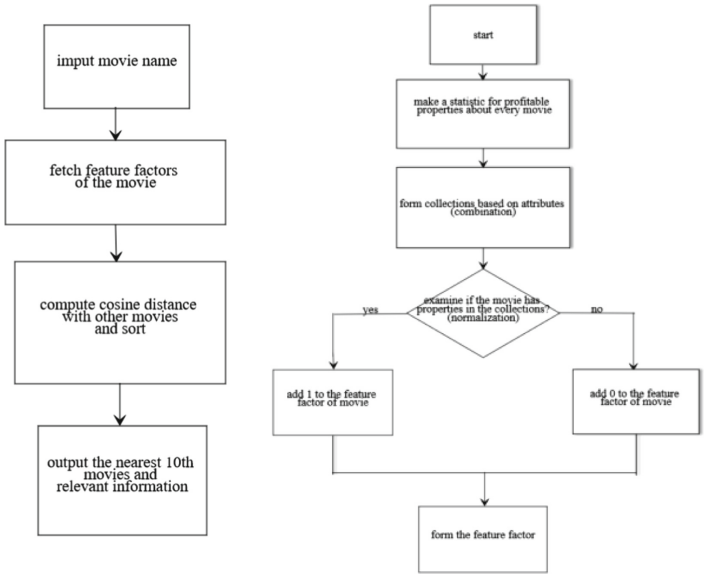

Fig. 1. Specific flow of the algorithm. Photo credit: Original

4 Current Development and Analysis Results

Apart from the movie recommendation mode, since the report is aimed at showing the overview of movie market for the past 10 years as a guide for film generators, the intuitional figures illustrating the relationships between a movie and its corresponding features are necessary. Hence, both the charts and the similar movie recommendations will be covered in this part.

4.1 The Trend of Movie Genres over Time

Begin with the line chart below, it can be noticed that from the distribution of the number of Top 10 movie genres, the top movie genres with the largest number are drama, comedy, thriller and action etc. which are also common movie genres in the theaters at present (Fig. 2).

The multiple line chart is used to represent trends of all kinds of movies changing over time. From the changes in the number of various movie genres year by year, it can be seen that since about 1992, movies entered a period of prosperity and development. Various genres of movies have increased considerably, and the top ones are drama, comedy, thriller and action etc. In addition, the slope between different nodes significantly shows the positive and negative of change. All the movies are classified year by year. And in the collection of movies in one year, the movies are also tagged by their genres. To conclude, the majority of these types of films have a certain relationship with the overall prosperity and development of the art of movie.

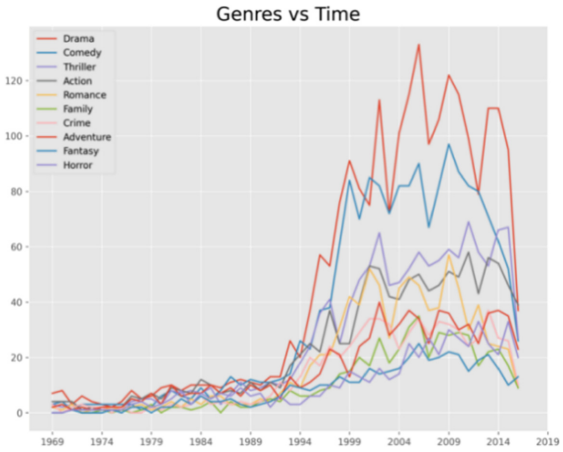

Fig. 2. Line charts showing the movie genres vs time. Photo credit: Original

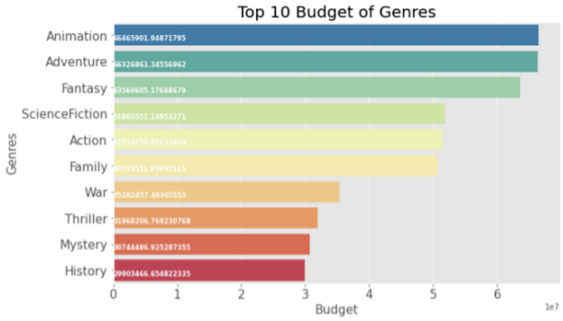

Fig. 3. Bar charts showing the top 10 movie genres vs budgets. Photo credit: Original

4.2 The Relationship Between Movie Genres and Profitability

Then comes to the horizontal bar charts used to present the Top 10 movie genres and the relationship between movie genres and their own budget and profit separately. The category list is broken into individual subjects and calculates quantities of all kinds of movies, listing from many to few and the charts are shown Fig. 3.

As for the relationship between movie genres and budgets, it could be seen clearly, from top to the bottom, the movie genres with higher budget investment include animation, adventure, fantasy, science fiction and action etc. which shows that movies of these genres need more production funds and their average investments are also written on the chart as a guide. When analyzing movie genres and earnings, animation, adventure, and fantasy movies have the strongest profitability (Fig. 4).

Combining these two pictures, the movie genres with the largest budget investment and the ones with the highest revenue are roughly the same. Each choice of the film company has always been verified by the market. But when it comes to the particular

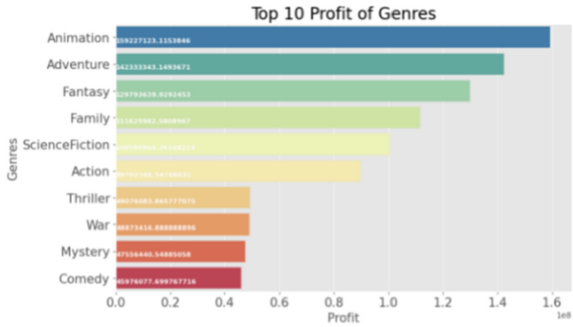

Fig. 4. Bar charts showing the top 10 movie genres vs profits. Photo credit: Original

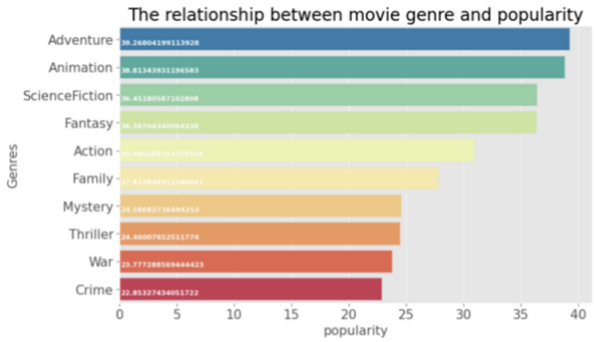

Fig. 5. Bar charts showing the relationship between movie genres and popularity. Photo credit: Original

movie, a high budget does not necessarily mean a high return and other factors need to be considered comprehensively.

4.3 Factors Affecting the Popularity of Movies

How the popularity of a movie is affected will be discussed through the horizontal bar charts. First, the relationship between movie genres and popularity are listed as below. From the top to bottom, the more popular movie genres include adventure, animation, science fiction, fantasy, and action etc. While Fig. 6 demonstrates 20 of the movie genres with their corelated votes from the audience (Fig. 5).

What can be noticed is that the highly rated movies are not the ones people used to be familiar with, they are history, war, drama, music and foreign etc. Then combine these two charts for analysis, it can be concluded that the movies with high ratings are not necessarily highly popular, and there may be some good films with a small audience. While the order of the movie genres with high average scores are not consistent with the high popularities. For example, movies such as history, war and music are not the most popular ones but they get high average scores. In addition, it is worth noting for film production companies that comedy, TV movies and horror movies have low average

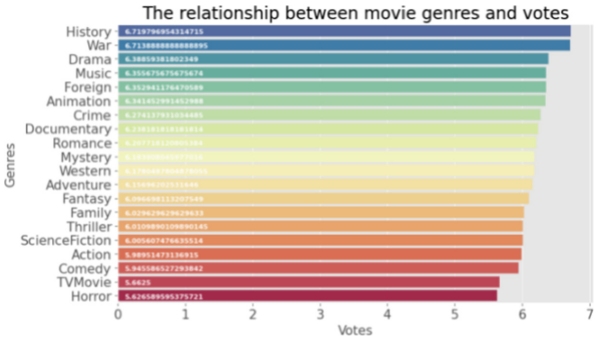

Fig. 6. Bar charts showing the relationship between movie genres and votes. Photo credit: Original

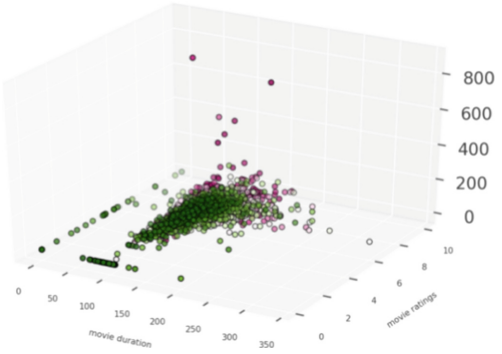

Fig. 7. Relationship between movie duration, movie rating and profitability. Photo credit: Original

scores, meaning that it is not easy for these movies to get a good reputation among the audience. For the science fiction, its single-film revenue is not very high, and the average budget is not the highest, which shows that the film company does not pay much attention to this type and invests relatively little in it. But its popularity is relatively high. In the era of more and more advanced technology, there should be better performance in the future.

4.4 3-Dimentional Scatter Point Chart

It may be not concise and effective to represent relationships between multiple subjects using a bar chart or something analogous. So, the 3-dimension scatter point chart is used to show the relationship between movie duration, movie rating and profitability. Figure 7 is shown.

It can be seen from the figure that X-axis is movie duration and Y-axis is movie rating. And the common ground—profitability, the foreign key to connect duration and rating, is set to be Z-axis. All the movies included in this data set are represented by translucent dots with gradient colors changing with Y-axis and Z-axis to be seen more clearly. One scatter could delegate a movie and its relationship of duration, rating and profitability which is more convenient to be compared.

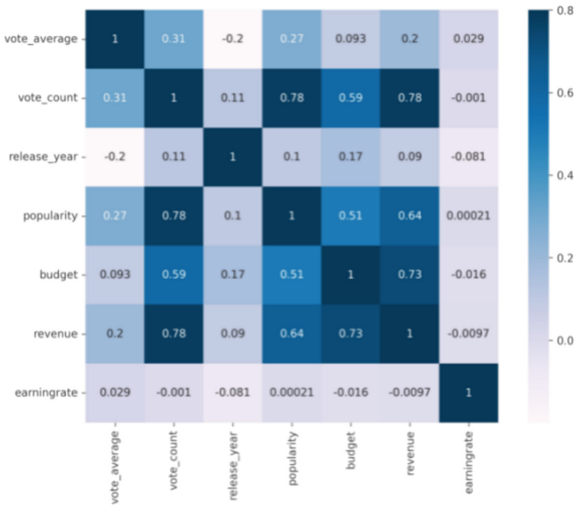

Fig. 8. Heat map showing the factors affecting movie box office revenue. Photo credit: Original

The feature of movie duration hasn't been discussed before, so here, the relationship between movie duration and popularity shows that the duration of the more popular movies is basically between 90–150 min. Movies that are too long or too short will not be welcomed by most audiences.

4.5 Factors Affecting Movie Box Revenue

As for factors affecting movie box office revenue, to dig out the coefficients of association between different parameters, the heat map is applied to show one to one correspondence. The attributes on axis are symmetrical so that the correlation could be found in the cross of one row and column. The deeper the color lump is, the tighter the relationship, and the Fig. 8 is shown.

According the heat map, there are four sets of characteristics that are very closely related to each other and they will be explained one by one. First is about the popularity and the vote count whose correlation coefficient is especially high to 0.78, meaning that there is a strong interaction between the two. The result is realistic and is easy to understand, because when audience vote a high score to a film, it will definitely engage more people to watch it, thus, gaining higher popularity for the film. Second, the relationship between revenue and audience votes is also considerable, with a coefficient of 0.78, its logic is just like the former one. When a higher rating attracts more people to the cinema, both the cinemas and film producers will receive correspondingly more revenue.

Then is for the revenue and budget of a movie with a coefficient of 0.73. Although there is no definite causal relationship between budget investment and income in a movie market, the heat map above shows that higher budget does bring more substantial revenue. This can be interpreted as the higher quality of films can be presented through better casts, more elaborate special effects and a more logical script when given higher

budgets. Therefore, a well-made film in all aspects is bound to attract more audience to buy tickets for it and get better profitability.

Finally, the relationship between film revenues and popularity cannot be ignored either as they also have a coefficient close to 0.65. Combined with the discussion of the previous three groups of relationships, it is not difficult to explain the connection between them. When a movie becomes more popular, audience will have a better chance to hear about the film from their friends and other media sources. Thus, are more willing to pay to see it, which in turn brings in more revenue.

4.6 Similar Movie Recommendations

After showing and analyzing the overview of movie market for the past 10 years through various kings of charts and figures, the results of similar movie recommendation system will be demonstrated in this part. According to Rhee, T. G. (2016) [9], where Rhee use neural network classification approach to make the predict of profitability, this method here could be seen as a reverse progress to find similar movie for recommendation with

Table 1. Recommendations for similar movies

Input	Avatar	
Name	Genres	
Aliens	'Horror', 'Action', 'Thriller', 'ScienceFiction'	
Rating	Main actors	
7.7	'JamesRemar', 'MichaelBiehn', 'PaulReiser', 'SigourneyWeaver'	
Budget	Revenue	Earning rate
18500000	183316455	9.91
Name	Genres	
The Abyss	'Adventure', 'Action', 'Thriller', 'ScienceFiction'	
Rating	Main actors	
7.1	'EdHarris', 'LeoBurmester', 'MaryElizabethMastrantonio', 'MichaelBiehn'	
Budget	Revenue	Earning rate
70000000	90000098	1.29
Name	Genres	
Terminator 2: Judgment Day	'Action', 'Thriller', 'ScienceFiction'	
Rating	Main actors	
7.7	'ArnoldSchwarzenegger', 'EdwardFurlong', 'LindaHamilton', 'RobertPatrick'	
Budget	Revenue	Earning rate
100000000	520000000	5.2

a simpler process. When a studio is planning to make a movie that closely resembled a known movie in genres or other aspects, it can be entered into the recommendation system as an input, then 10 similar movies will be listed as reference along with their names, genres, ratings, main actors, budgets, revenues and earning rates. An example is shown as Table 1.

Since making a good movie is not a subject that could be realized easily, so when a recommendation is made to make a movie, the information provided should be as complete and effective as possible. The best way is to learn from good examples, then it is very useful to offer movies matching demands with essential information.

So, combined with the above list as an example, in a real case, when a company is preparing an adventure movie, they can be told that the total number of adventure movies in the last decade has not been very large. Then, the investment it requires and the income it brings are both very high compared with other types of films. In addition, although an adventure movie is rated in the middle of the pack among all genres, it remains extremely popular. Finally, refer to other adventure movies for casting, performance in the real market, and earning rates etc., the company can reach the conclusion that, if they can afford a bigger budget, they are likely to reap higher returns with proper casting and directors.

5 Conclusion

The main objective of our report is to give an overview of film market for decades including the genre of each film, its corresponding audience response, financial expenditure and revenue, and cast as reference factors for analysis. By visualizing the data and referring to the results of similar movie recommendations, film financiers and production companies can get proper advice on preparing budget, choosing genre and casting. Thus, achieving more financial rewards and high reputation in the film market.

During this process, the data set was sorted first to gain clearer and more logical visualization diagrams including bar charts, line charts, 3-dimentional scatter diagrams and the heat map. Combined with these figures, some suggestions can be concluded for the companies. For specific movie genres, the animation, adventure and action movies' profitability and popularity are among the best but they require a lot of budget investment, therefore large companies are recommended to produce them. When talking about comedy and drama whose average budget and revenues are low, but the single-film rate of return is not inferior to the top genres. So small and medium-sized companies may try to produce them. The second method is to vectorize the data, and by calculating the distances between each vector, similar movies with their corelated features can be recommended. According to the performance of these movies in the real market, the movie production companies and investors will be able to choose solutions that are more appropriate for their own program so as to ensure the long-term and stable development in the future.

For this paper, it started with an overview of main aspects of movie industry, which gives fundamental states and offers a direction for further learning --- movie profitability. Then a recommendation system based on cosine distance using vectorized data is realized. Thus, a recommendation list of profitable movies could be generalized to help

film producer to do choice. The strengths of the system are its stability and sufficiency of the information it offers. But there still exists some deficiencies: the model may not be such precise and the result of recommendation is not very concrete.

References

1. Barman, D., Chowdhury, N., & Singha, R. K. (2012, December). To predict possible profit/loss of a movie to be launched using MLP with back-propagation learning. In *2012 International Conference on Communications, Devices and Intelligent Systems (CODIS)* (pp. 322–325). IEEE.
2. Eliashberg, J., Elberse, A., & Leenders, M. A. (2006). The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing science*, *25*(6), 638–661.
3. Garcia-del-Barrio, P., & Zarco, H. (2017). Do movie contents influence box-office revenues?. *Applied Economics*, *49*(17), 1679–1688.
4. HU, X. L., Li, B., & WU, Z. P. (2013). The analysis of the factors which influence film box office. *Journal of Communication University of China (Science and Technology)*, 01.
5. Hur, M., Kang, P., & Cho, S. (2016). Box-office forecasting based on sentiments of movie reviews and Independent subspace method. *Information Sciences*, *372*, 608–624.
6. JIAO, Qing-zheng, WEI, & Cheng-jian. (2010). Text categorization approach based on probability standard deviation with evaluation of distribution information. Journal of Computer Applications, 29(12), 3303–3306.
7. Kim, T., Hong, J., & Kang, P. (2015). Box office forecasting using machine learning algorithms based on SNS data. *International Journal of Forecasting*, *31*(2), 364–390.
8. Peng, K., Wang, W., & Yang, Y. P. (2013). Pseudo-K-nearest neighbor text classification algorithm based on cosine distance metric learning [J]. Computer Engineering and Design, 6.
9. Rhee, T. G., & Zulkernine, F. (2016, December). Predicting movie box office profitability: a neural network approach. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 665–670). IEEE.
10. Rui, H., Liu, Y., & Whinston, A. (2013). Whose and what chatter matters? The effect of tweets on movie sales. *Decision support systems*, *55*(4), 863–870.
11. Tang Ying, Sun Kang-gao, Qin Xu-jia, Zhou Jian-mei (2018). Local Model Weighted Ensemble for Top-N Movie Recommendation. Computer Science, 45(11A), 439–444.
12. Walls, W. D. (2005). Modeling movie success when 'nobody knows anything': Conditional stable-distribution analysis of film returns. Journal of Cultural Economics, 29(3), 177–190.
13. Wallström, K., & Wahlgren, M. (2018). What are the main factors affecting movie profitability?.
14. Yuan, X., Zhang, H., Jiang, X., & Li, Z. (2010). The study of online WOM marketing: an empirical study based on online film review and box office receipts of Korea. *Journal of Marketing Science*, *6*(1), 41–58.