Predicting the Popularity of Movies with Social Media

Liangye He
School of Data Science and Engineering, Guangdong Polytechnic Normal University, Heyuan, Guangdong, China

Keywords: Movie Popularity, Correlation Analysis, Descriptive Statistics, Machine Learning.

Abstract: The film industry occupies an important position in the field of culture and entertainment, and the issue of how to avoid the risk of investing in films has received widespread attention, but the existing prediction methods still have certain shortcomings in terms of accuracy and practicality. This paper thoroughly analyses the various dimensions that influence movie popularity, including movie genre, release time, heat, and other factors, and explores the relationship between these factors and movie popularity. This paper analyses and concludes that factors such as the combination of actors and directors, the running time of the film, and the release time have a significant impact on the popularity of the film and thus establish a prediction model for the popularity of the film. Based on this, this paper makes the following suggestions: if film producers want to increase the popularity of their films, they should pay attention to the release time of their films and the combination of directors and actors. At the same time, industry insiders can use the prediction model in this paper to make decisions about film projects.

1 INTRODUCTION

With the rapid development of information technology, modern society has fully entered the information age. Therefore, in the context of this era, information has become a key element in promoting social progress and development in various fields, and the Internet, as the main carrier of information dissemination, brings together a huge amount of data resources and provides sufficient information support for all walks of life. Social media is an important product of the Internet era and has become the most widely used tool in people's daily lives, and the amount of data contained in it is self-evident. According to Digital 2024, published by We Are Social and Meltwater, the number of active social media user identities has surpassed 5 billion, representing 62.3% of the world's population (Social, 2024). Last year, the global total increased by 266 million, with an annual growth rate of 5.6 percent. This data not only demonstrates the widespread popularity of social media but also shows that social media has great data potential as a major platform for information dissemination, social interaction, and cultural exchange.

Film investors want to reduce the risk of investment, so if they can use scientific methods to predict the popularity of the film and adopt a reasonable marketing strategy, film production companies and distribution companies can avoid the risk and reduce losses (Li, 2021). For the audience, what they want more is to be able to watch a good film, and rating prediction can help them decide whether to watch the film or not to enhance the viewing experience. Douban, a famous film website in China, was mined by scholars for rating prediction a few years ago (Geng & Guo, 2021; Zou, 2019). IMDb and TMDB are also popular film websites that contain a lot of information about the film and audience comments, and because of their large number of users, the scores given on the site can be a good reference option.

There have been many scholarly studies on film scoring. Among the more recent findings, Liu et al. were inspired by brain mechanisms to independently encode static features as control signals and modulate time-varying contextual features to predict individual decisions in complex contexts (Liu et al.). Xu et al. constructed a movie information network from the relationship between users, movies, and attributes and used the metapath2vec algorithm to map various attributes to the corresponding feature space in the form of network nodes to obtain feature vectors for rating prediction (Xu et al., 2024). Liu et al. used an attention network to distinguish the importance of interactive features and then used DNN to process higher-order feature combinations to build a predictive teacher model while using knowledge

He, L.
Predicting the Popularity of Movies with Social Media.
DOI: 10.5220/0014150300004942

In *Proceedings of the 2nd International Conference on Applied Psychology and Marketing Management (APMM 2025)*, pages 662-667
ISBN: 978-989-758-791-7

distillation to build a streamlined model of the student model to ensure the accuracy of the teacher model (Liu et al., 2024). Previously, Michael T. Lash and Kang Zhao used social network analysis and text mining techniques to delve deeper into the film's cast and release season, ultimately speculating on the film's profitability (Lash & Zhao, 2016). N. Darapaneni et al. chose XGBoost as the final film success prediction algorithm after working with film outline plots and testing it with several models (Darapaneni et al., 2020). Differently, Rijul Dhir and Anand Raj chose released data from the film as their data set, rather than listening to what critics and others had to say about whether the film would be a success or not, but the final model's predictive accuracy wasn't very high, and they're also looking ahead to future work where they want to include analysis of user comments from social media (Dhir & Rah, 2018). Rather than simply choosing the optimal model, Vedika Gupta et al. integrated and compared different models before using them to predict the success of the film (Gupta et al., 2023).

At the same time, with the help of social media analysis, the accuracy of the prediction model will further improve, as a result of which provides more scientific decision-making support for film investors. Social media is an important platform for modern information dissemination, contains a large amount of user-generated content, including film reviews, topic discussions, etc., and this data reflects the audience's real-time feedback on the film, as well as their emotional inclination, which is of high research value (Castillo et al., 2021).

Based on this, this paper will examine the correlation between various factors that influence a film's popularity and its rating and will develop a prediction model.

2 METHOD

2.1 Research Design

The purpose of this paper is to explore the potential value and laws of film-related data. Through the comprehensive analysis of multi-dimensional data such as film release time and audience rating, this study reveals the internal relationship between film popularity and these factors and their influence degree, that is, the weight, to establish a film rating prediction model. Meanwhile, social media analysis can be used to capture the expectation value of the audience, i.e., the social media buzz, at this point before the film's release, thus predicting the film's rating and popularity.

Due to the complexity of the factors that influence a film's popularity, it is difficult for a simple prediction model to produce accurate results (Geng & Guo, 2021; He & Yuan, 2021). First of all, this paper obtains the film data set through Kaggle, the reason for choosing Kaggle is because it is a more authoritative and data-rich website, where there is no lack of scholars in various fields to select the data inside for academic research, and then remove null values as well as outliers in the data set. To explore the factors that affect the popularity of films, this paper will carry out statistical analysis and correlation analysis on the popularity of films, which is difficult to express in numerical terms. Therefore, the audience rating of films is selected as an indicator to measure the popularity of films, which is used as a dependent variable for correlation analysis, and the influencing factors are obtained by combining statistical analysis. The first step of this paper is to conduct a descriptive statistical analysis based on the data results to explore the correlation between nonnumerical items, namely, actors, directors, actors, and directors' specific combinations and film scores. The second step is to conduct a correlation analysis between numerical items and film scores and use the analysis results to establish a prediction model of film popularity (Abidi et al., 2020). Because the ultimate goal is to predict the popularity of films, the popularity of films is defined by five levels, namely terrible, poor, satisfactory, good, and excellent. The research design is shown in Fig.1.

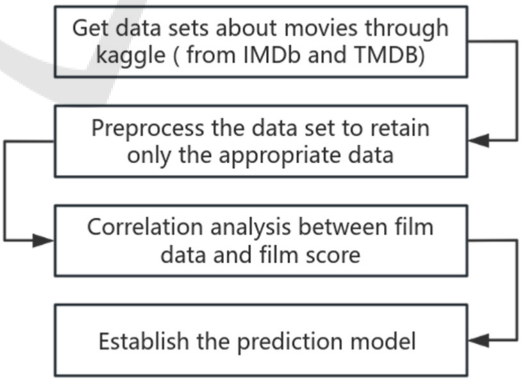
Get data sets about movies through kaggle (from IMDb and TMDB)
Preprocess the data set to retain only the appropriate data
Correlation analysis between film data and film score
Establish the prediction model

Figure. 1. Research design.

2.2 Data Collection

To better predict the popularity of movies, a perfect data set is needed to obtain more accurate results. There are 880,990 original datasets in the data set used in this paper, including 27 dimensions, such as film title, actor, director, film type, score, etc. To ensure consistency, too short and too long movies are removed from the data set, and only movies with a movie duration of 80 to 200 minutes are retained. At the same time, the reference value of film rating obtained when there are few rating users is not significant, so to ensure objectivity, films with rating users less than 100 should be removed. In addition, to avoid randomness, factors affecting the fairness of the research results, such as screenwriters, film travel companies, original titles, etc., are removed. Because Python is selected for prediction in this paper, the prediction of floating-point values may affect the accuracy, so the score is multiplied by 10 manually. Moreover, to realize the prediction, the profit should not appear in the data set because this is the data that can be obtained only after the film is released. Through the above steps, the data set with 12,997 pieces of data is finally obtained, and 11 dimensions are reserved for analysis, including vote_average, vote_count, release_date, revenue, budget, runtime, popularity, genres, cast, director, imdb_rating, imdb_votes。

At the same time, in order to focus on the films with high and low scores, two different data sets are used, top_1000 (score greater than 7.5) and lowest_100 (score less than 4.0), which are mainly used for the statistical analysis of non numerical items, so the dimensions only include date, score, genre, director, crew.

3 RESULT

3.1 Specific Combinations and Film Score

The combination of specific directors and actors may form a brand effect, which can attract their loyal fans to a certain extent, thus affecting the film score. In the most popular top1000 data set, based on its data characteristics for statistical analysis, this study will arrange and combine the directors and leading actors, and evaluate them by combination and score.

As shown in Table 1, when these directors and actors cooperate, the average film score is high, so it can be considered that when the above combination appears, the film score will be high. At the same time, the replacement of the combination will also have an impact on the film score, just like the combination of Akira Kurosawa, the same director, and cooperating with different actors, Toshir ô Mifune and Tatsuya Nakadai. Then, do the same work through the lowest data set, and the rules are similar. Therefore, if a director and an actor appear in a film at the same time, it may cause brand effect and get a higher score. On the contrary, the "chemical reaction" will lead to a lower film score. Therefore, it can be considered that the specific combination of director and actor has an impact on the film score to a certain extent.

Table 1. The top five director and actor groups and the average score of their films.

Director	Actor	Cooperative times	Average film score
Akira Kurosawa	Toshirô Mifune	7	8.24
Richard Linklater	Ethan Hawke	5	7.98
Akira Kurosawa	Tatsuya Nakadai	5	8.10
Peter Jackson	Ian McKellen	5	7.80
John Ford	John Wayne	4	7.80

3.2 Main Factors

In the first section of this chapter, actors and directors are studied. Therefore, in this section, the relevance between them and film scoring will not be explored, but other characteristics will be studied. Through the exploration of correlation, it is found that several factors are significantly correlated with the film score, namely the release time and the playing time, and the corresponding p-value is less than 0.05, indicating that they have a significant correlation with the film score. In addition, there is only a weak correlation between film popularity and film score.

Table 2. Release month and IMDb score.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
rating	62.0	63.0	64.0	63.0	65.0	65.0	64.0	63.0	66.0	65.0	66.0	67.0

The first is to explore the correlation between the release time and the score. To examine the correlation between the two, their Spearman's Rho and Kendall's tau were calculated to be 0.091 and 0.064 respectively, so it can be assumed that the release month has a significant correlation with the film score. As shown in Table 2, the average film score showed an upward trend from January to March and a downward trend from June to August, while the average film score was higher in September, November and December.

What's more, the data shows that when the playing time is too short or too long, the film score will decline. Therefore, most of the films now control the film time between 85 and 110 minutes to reduce the impact of the playing time on the film score. In the descriptive statistical analysis of the data set, the average playing time of the film is 107 minutes. This study chose to use the regression model to find the correlation between the playback duration and the film score. At the same time, to avoid the influence of extreme circumstances, the IQR was used to remove the abnormal value of the playback duration. The filtered results are shown in Fig. 2.

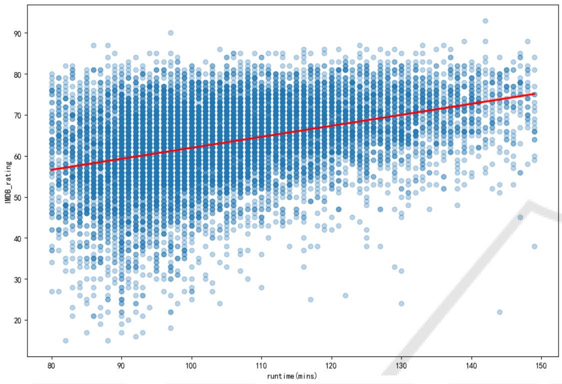

Figure. 2. IMDb_rating by runtime (filtered).

The filtered results are more intuitive, and the slope of the regression line of the regression model is 0.268, indicating that there is a positive correlation between the playing time and the film score.

This paper uses the Pearson correlation formula as the calculation basis to calculate the correlation between the numerical column in all movies and the movie score, and the correlation matrix is shown in Fig. 3.

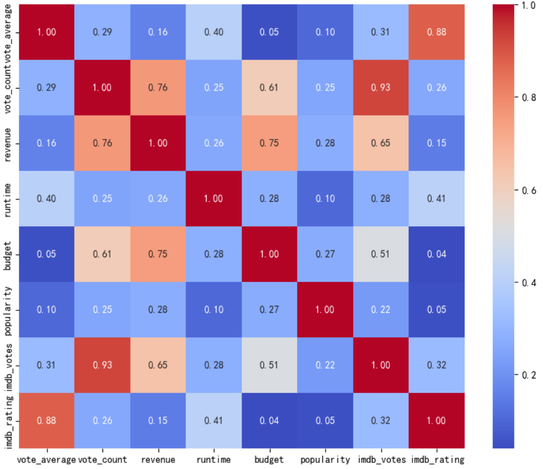

Figure. 3. Correlation matrix.

Figure 3 shows that the correlation between vote_average and imdb_rating and popularity is not great. Therefore, even if a film is very popular on social media, it may only indicate that the marketing strategy before the film's release was relatively successful, that the film's theme is attractive, or that it happens to have a popular actor or director.

3.3 Prediction Model

To avoid the influence of film type on the score, this study will eliminate this influence factor artificially so the model can focus on predicting the popularity of the same type of film. Combined with the above research, the release time, the length of time, and the popularity of the film were selected to establish the model. In this section, classifiers, clustering, and simple linear regression are used. At first, it is divided into nearly 100 categories, that is, each score is a category. At this time, a random forest is used for prediction (Tan, 2021).

To get more accurate prediction results, this study divides every 2 points into an interval of 0,1,2,3,4, corresponding to terrible, poor, satisfactory, good, and excellent. For the most accurate prediction, the score is manually multiplied by 10. This allows research to ignore the inaccuracy caused by floating point values, i.e., every 20 points is taken as an interval. Converting continuous values to discrete values helps to fit the most appropriate values. Taking 80% of all movies as the training set and the remaining 20% as the test set, four models will be selected for implementation, namely random forest, gradient boosting, KNN, and multiple linear regression model. Especially in multiple linear regression, the regression value will be rounded off because the predicted value of linear regression is usually related to the distribution of the target label, so the rounded result can match the classification label in most cases. In particular, when the regression result is 4.6, it will become five according to the rule, but there is no 5 in the classification, so it is necessary to force the value greater than 4 to take four and continue to follow the rounding between 0 and 4. The obtained performance indicators are shown in Table 3

Table 3. Performance index of each machine algorithm.

	Random Forest	Gradient Boosting	KNN	Multiple Linear Regression
Accuracy	86.65%	87.00%	60.46%	84.08%
MSE	0.13	0.13	0.45	0.45
MAE	0.13	0.13	0.41	0.41
R^2	0.59	0.60	-0.39	-0.39

The gradient boosting shows the best prediction results, which shows that the gradient climb algorithm has obvious advantages in dealing with the problem of film popularity prediction. It can effectively extract key features from a large number of data and gradually approach the optimal solution through the iterative learning process to improve the accuracy of prediction and the generalization ability of the model. Therefore, this paper establishes a prediction model by gradient boosting and inputs the budget, popularity, release month, playing time, and user expected score on social media to obtain the corresponding possible popularity level.

4 DISCUSSION

From the above research, it can be concluded that the combination of specific directors and actors may form a brand effect or "chemical reaction", thus affecting the film score. Therefore, when investors choose these specific combinations of directors and actors as investment targets, they can ensure the quality and popularity of the produced films. Similarly, some combinations will lead to the reduction of film scores. Investors can also try to choose some newer portfolios through public feedback so that they can learn from mature and better-performing portfolios so that the new portfolio can better realize innovation for the development of films.

Film duration also has a partial impact on film ratings, so it should be limited effectively. Meanwhile, it should not abandon the integrity of the content in order to control the length of the film. At that time, the harm to the film is greater than the advantage, so investors should pay more attention to the overall balance between the content and the time.

Choosing a more appropriate time for the release of the film can enhance the popularity of the film. Different types of films will be released at different times. For example, comedy can be released during the Spring Festival in China, and love films can be released on Valentine's Day, which may make the film more popular.

In any case, making a good film is the key to winning the popularity of the public. Good content plus the above key factors can make the film more popular (Zou, 2019). Absolutely, early publicity work is necessary, otherwise no one knows even though the film is of high quality, which is one of the reasons for the low popularity.

5 CONCLUSION

This study firstly explores the influence of actors and directors' combinations on film ratings through statistical analysis, then uses regression analysis to select the main influencing factors and establish the prediction model of film popularity, and finally selects the gradient climbing algorithm to complete the modeling. This paper has been conducted on a sample of films with a duration ranging from 80 to 200 minutes. It is acknowledged that certain film genres, such as epic films, may extend beyond this range. Consequently, the data about these films is incomplete. It is recommended that subsequent studies extend the scope of research to include these films.

The release month and rating of films in different months have been shown to have a certain impact, but different regions may show different results. For example, the release of comedies during the Spring Festival in China is a significant factor in their popularity, and different regional cultures have been found to affect films. In the future, the exploration of this possibility in a specific region or a region with the same culture will be a valuable avenue for further research.

REFERENCES

A. Castillo, J. Benitez, J. Llorens, X. (R.) Luo, Social media-driven customer engagement and movie performance: Theory and empirical evidence. Decis. Support Syst. 145, 113516 (2021)

J. Geng, M. Guo, Douban top 250 movie data mining and score prediction. Hebei Enterp. (02), 11-13 (2021)

J. Tan, Research on IMDB movie score prediction based on random forest algorithm. Mod. Comput. (30), 24-31 (2021)

M. T. Lash, K. Zhao, Early predictions of movie success: The who, what, and when of profitability. J. Manag. Inf. Syst. 33(3), 874-903 (2016)

N. Darapaneni et al., Movie success prediction using ML, 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 0869-0874 (2020)

N. Li, Research on investment risk and box office revenue prediction of China's film industry, Ph.D. thesis, Dongbei University of Finance and Economics (2021)

Q. He, F. Yuan, Research on the influencing factors of film consumption and box office prediction in the era of digital economy: Based on the perspective of machine learning and model fusion. Price Theory Pract. (09), 163-167+204 (2021)

R. Dhir, A. Raj, Movie success prediction using machine learning algorithms and their comparison, 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 385-390 (2018)

S. M. R. Abidi, Y. Xu, J. Ni, et al., Popularity prediction of movies: from statistical modeling to machine learning techniques. Multimed. Tools Appl. 79, 35583-35617 (2020)

Social, Digital 2024 Global Overview Report (2024). Retrieved from: https://wearesocial.com/cn/blog/2024/01/digital-2024-5-billion-social-media-users/

T. Liu, S. He, Y. Peng, Q. Yuan, D. Zeng, Research on movie score prediction based on strategy coding. Syst. Eng. 1-10 (no date)

T. Liu, S. Yu, W. Ni, Movie score prediction based on attention mechanism and knowledge distillation. Comput. Appl. Softw. (07), 5-60 (2024)

V. Gupta, N. Jain, H. Garg, et al., Predicting attributes based movie success through ensemble machine learning. Multimedia. Tools Appl. 82, 9597-9626 (2023)

X. Xu, M. Zhang, R. Zhao, Y. Zhu, Movie score prediction based on interaction attribute enhancement. Soft. Guide, (01), 182-189 (2024)

Y. Zou, An empirical study on the influencing factors of movie word-of-mouth and the relationship between word-of-mouth and box office: An analysis based on Douban score data from 2011 to 2017. West. Leather (06), 106-107 (2019)