Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research on the Influencing Factors of Movie Popularity based on Random Forest

Zhuwen Wei [1, *] and Yifan Zhang [2]

[1] Alen Institute, School of Shanghai Ocean University, Shanghai, 200000, China

[2] Stony Brook College, Anhui University, Hefei, 230000, China

*Corresponding author: 2392401@st.shou.edu.cn

Abstract. Movie, also known as moving or moving pictures, or "reflections," is a form of visual art that communicates ideas, stories, cognitions, emotions, values, or various atmospheric simulated experiences through the use of moving images. As the mainstay of cultural industry, film has brought huge economic and social benefits, and the box office is the most important index to measure the economic benefits of film. In this study, the prediction results of the random forest analysis model are acceptable, it is well in line with the structure of the data, which is conducive to the analysis of the data, and the data is displayed with charts and graphs at the same time, so as to show the readers the indicators and data used in this topic more clearly and clearly, which increases the scientific nature of this paper. This study provides more scientific and accurate strategies for decision-makers and creators, and contributes to the high-quality dissemination of short videos.

Keywords: Movie popularity; multiple linear regression; prediction model..

1. Introduction

As an important form of cultural entertainment, film carries people's yearning for a better life and emotional demands, and is also an important way to reflect social reality and human inner world. With the continuous development and globalization of the film industry, it is of great significance to study the popular factors of the film to understand the cultural trend and promote the development of the cultural industry. Does budget relate to the movie popularity [1]. Head blockbusters stare at the Spring Festival file, National Day file, to create a "weekend file", need more waist and waist below the film support, especially small and medium cost films. However, according to the estimation of Zhang, assistant general manager of China Film Corporation Limited, 16 high-budget films accounted for 92% of the total box office of the 171 domestic films released by June 10, and the remaining 155 small and medium-sized films could only share 8% of the market, which is seriously unreasonable in structure and ecology [2].

The high risk of the film industry has led companies to search for ways to accurately predict revenue. However, because films are affected by very complex social factors, even the most experienced filmmakers often fail to grasp them accurately. In recent years, rising costs, saturation releases in the first two weeks of major box office revenue, declining home video sales, and intensifying media competition have made films increasingly risky. However, with the deepening of people's Internet use, the possibility of accurately predicting the box office through information technology continues to improve. Computer science majors were early to explore the laws of box office forecasting with different combinations of algorithms and variables [3]. Word-of-mouth rating directly determines the quality of the movie. Consumers tend to maximize the utility within a limited time, so they often decide whether to pay to watch the movie according to the rating of the movie [4]. According to the data of the Film Bureau of the State Administration of Press, Publication, Radio, Film and Television, the annual movie box office reached 10.172 billion yuan in 2010, breaking through the 10 billion yuan mark for the first time; In 2021, the domestic box office reached 39.335 billion yuan [5]. As one of the core cultural and creative industries, the development of the film industry has an increasing influence and role on the overall economic development of society, and plays an important supporting

role in the development of China's strategic emerging industries, while promoting the transformation of economic development mode [6].

In practice for more than a century, theatrical feature-length films have generally been between an hour and a half and two hours long (the director's cut that appears later on videotape or DVD can be longer). There are many factors that affect the formation of the film length regulation, among which the more critical are the following two: The film length is controlled within two hours, because this is the bladder endurance time that most people can bear. For the consideration of the number of daily films, the cinema will also reject the screening of too long films. Because of the normative limitations of cinematic time, every minute of cinematic time is precious [7]. A long short term memory network model (LSTM) was established for prediction, and the prediction results were relatively accurate [8]. The random forest prediction model was established with 85% accuracy by selecting variables such as directors, certificates and first-day box office of domestic films with a box office of over 100 million from 2011 to 2018.

2. Methods

2.1. Data Source

The prediction data for the movie popularity influencing factors used in this article comes from Kaggle, the raw data is saved in CSV format, and the data for this dataset is scraped from the IMDB website. The top 250 films are selected based on their IMDB scores, and information such as movie titles, directors, actors, ratings, votes, and year of release are collected. No data has been changed or modified in any way, all data is collected in accordance with IMDB's Terms of Use. By analyzing this data set, one can gain insight into the film industry, such as movie ratings and trends in popular genres. The Spring Festival has now become one of the most important schedules in the domestic film market, and it is an experimental field for various types of films to pursue innovation and breakthrough.

2.2. Variable Selection

Indicator description are as follows. Film certificate signifies the official permit required for the theatrical release of the movie in a particular region or country. Year represents the year of the movie's release or production. Run_time specifies the length of the movie in terms of its running time. Budget indicates the amount of money allocated for the production of the movie, including costs for cast, crew, filming, post-production, and marketing. Tagline refers to the catchphrase or tagline used to promote the movie and attract audiences. Director identifies the individual responsible for overseeing the creative and artistic aspects of the movie's production. Cast lists the actors and actresses who portray the characters in the movie. Writer credits the individual or individuals responsible for writing the screenplay or story of the movie. Additionally, genre categorizes the movie into specific types, such as action, comedy, drama, or science fiction, allowing for a more detailed understanding of the movie's content and style. These variables, when analyzed and interpreted together, provide a comprehensive overview of the movies represented in the table, enabling a deeper understanding of their characteristics, performance, and creative teams. (Table 1).

Table 1. Variable introduction

Term	Type	Average Value
certificate	Numeric	8.3072
year	Numeric	1986
run_time	Numeric	60min
budget	Numeric	52458981.53$
tagline	Categorical	-
directors	Categorical	-
casts	Categorical	-
writers	Categorical	-

2.3. Method Introduction

The method chosen for this article is random forests. A random forest is a classifier with many decision trees, which can be used to deal with classification and regression problems, as well as for dimensionality reduction problems. It also has a good tolerance for outliers and noise, and has better prediction and classification performance than decision trees. It can produce highly accurate classifiers for a wide range of data, it can handle a large number of input variables, it can produce unbiased estimates of generalized errors internally when building forests, it contains a good way to estimate lost data, and it can maintain accuracy if a large part of the data is lost.

3. Results and Discussion

3.1. Descriptive Analysis

Figure 1 is a histogram of the movie budget. The budget data on 250-500M is the largest, accounting for almost the entire data.

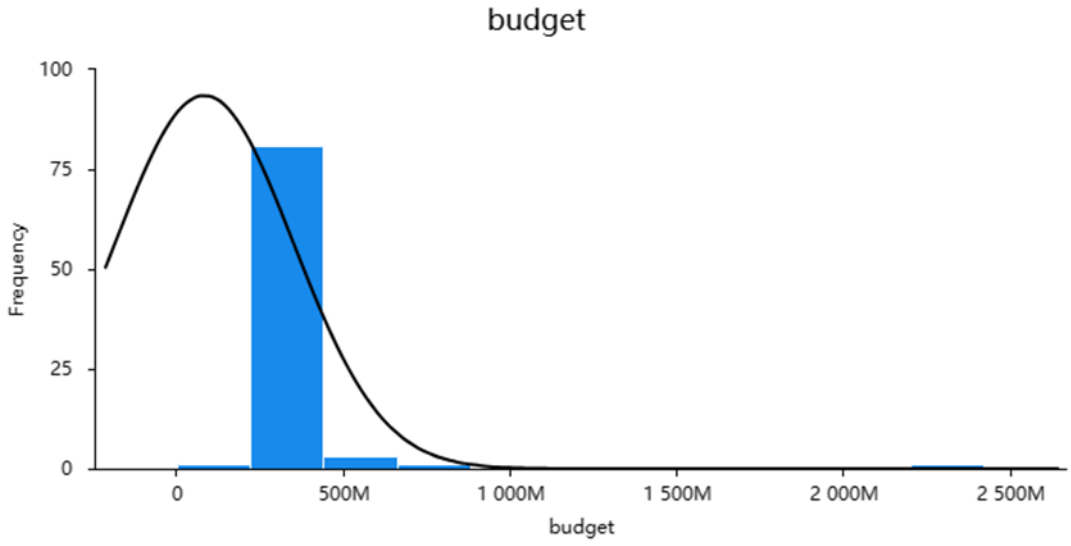

Fig. 1 The histogram of Movie Budget

Figure 2 is a histogram of the year the movie was released. Among them, there are many data released from 2000 to 2020, more than half of which are after 1990, and the most in 2000.

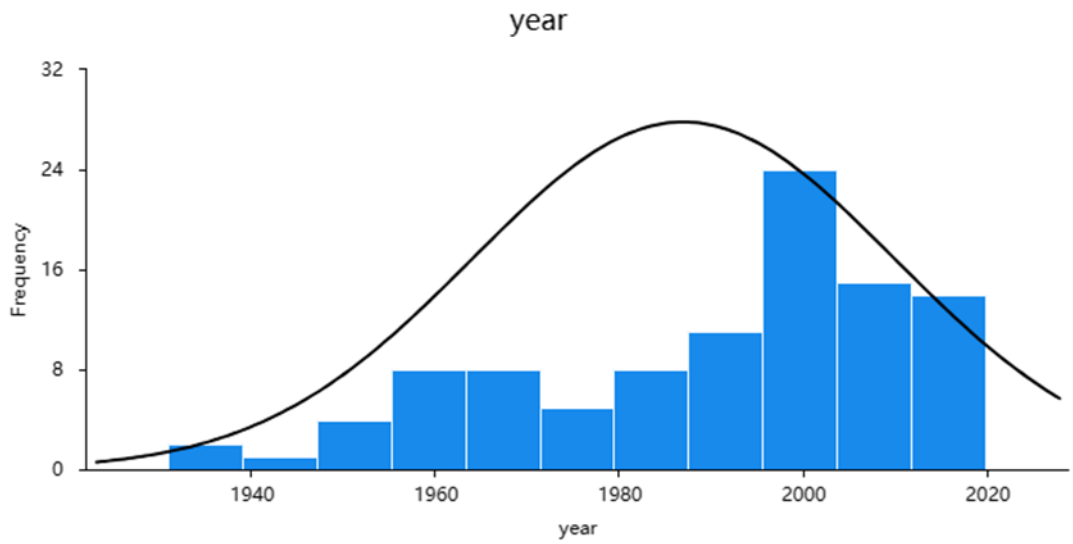

Fig. 2 The histogram of movie year

Figure 3 is a histogram of the length of the film's run. Regarding 20-60 hours, there is a lot of data for 80-90 hours, and the data at 70 hours is significantly lower than the data at 60 hours and 80 hours.

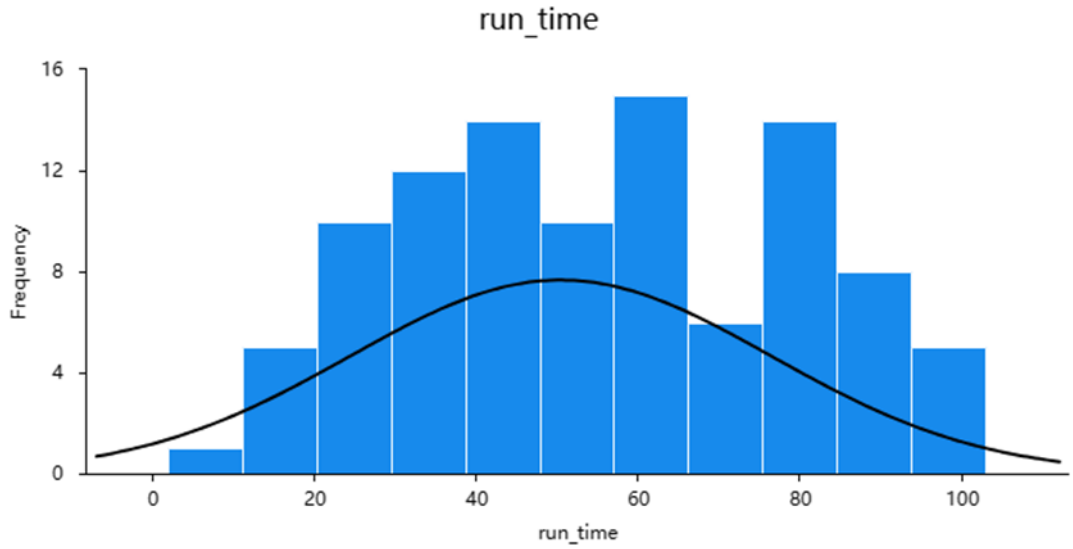

Fig. 3 The histogram of Run_ time

Figure 4 is a histogram of the box office of a movie. The box office data about 250-1250M accounts for almost all of the data, and the box office data about 250-500M is the most.

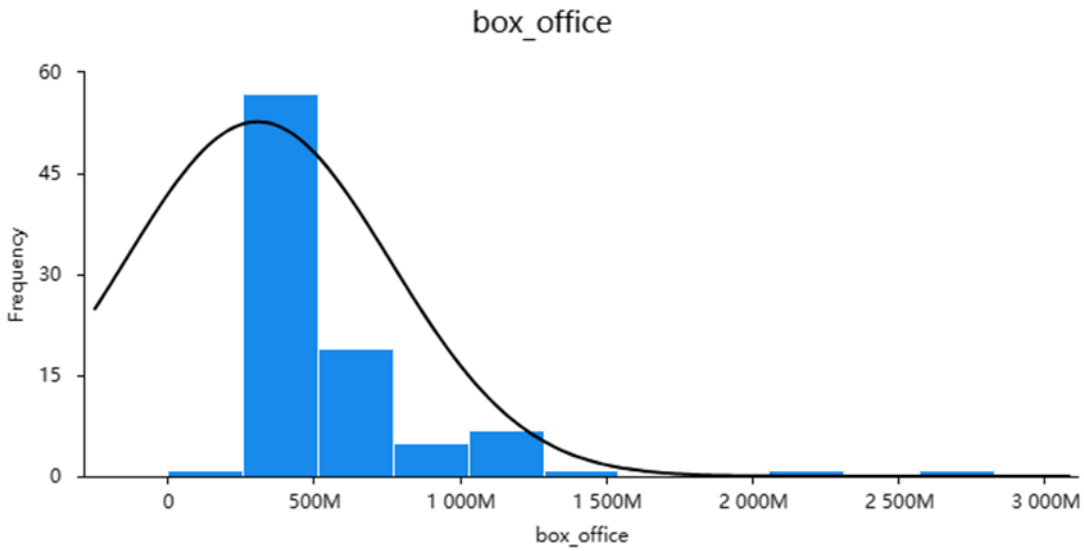

Fig. 4 The histogram of Box_ office

3.2. Correlation Analysis

As can be seen from the Figure 5, use relevant analysis to study certificate, year, directors, casts, writers, run_time, budget and tagline and rank, name, rating, genre respectively. box_office used Pearson correlation coefficient to represent the strength of the correlation between five items.

There is no significant correlation between certificate and rank, name, rating, genre, or box_office, as indicated by correlation values close to 0 (-0.140, -0.009, 0.110, 0.178, 0.024) and P-values greater than 0.05.

Similarly, directors and writers do not exhibit a significant relationship with rank, name, rating, genre, or box_office, with correlation values close to 0 (directors: 0.012, -0.172, -0.004, 0.009, -0.079; writers: -0.067, -0.019, 0.139, 0.047, -0.068) and P-values above 0.05. However, there is a positive correlation between year and box_office, with a significant correlation value of 0.516. No significant correlation is observed between year and rank, name, rating, or genre.

The correlation analysis reveals significant relationships between run_time and both rating and box_office, with positive correlation values of 0.262 and 0.232, respectively. However, no significant correlation is observed between run_time and rank, name, or genre.

On the other hand, budget and tagline do not exhibit significant relationships with rank, name, rating, genre, or box_office. The correlation values for budget are close to 0 (0.147, -0.040, -0.075, -0.189, 0.177), and all P-values are above 0.05, indicating no correlation. Similarly, the correlation values for tagline are also close to 0 (-0.141, 0.002, 0.091, 0.040, -0.007), with all P-values exceeding 0.05.

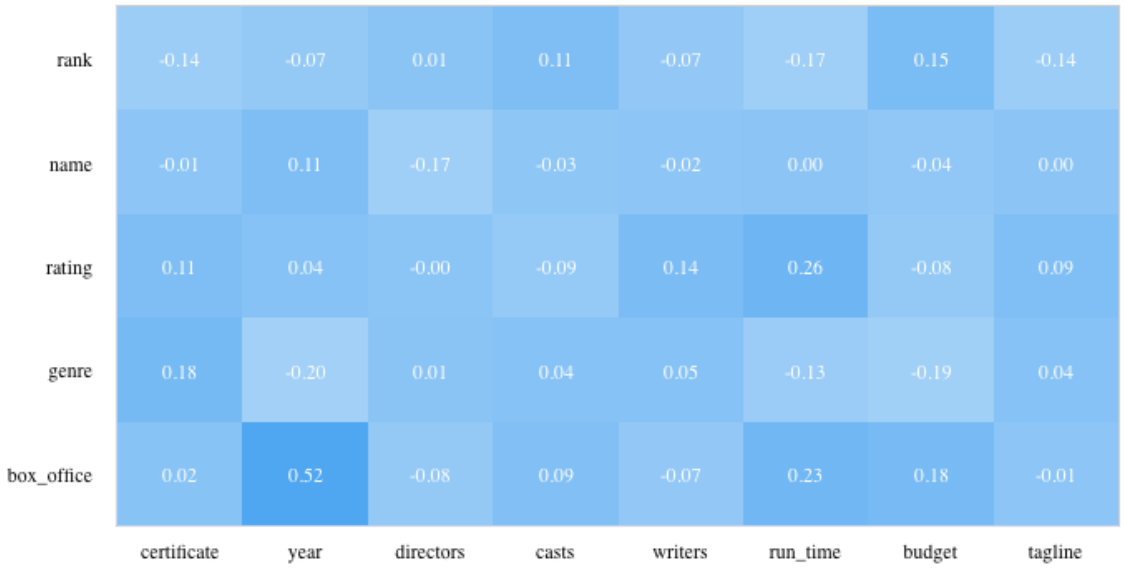

Fig. 5 Pearson Correlation

3.3. Random Forest Results

The feature weights serve as indicators of the significance of each factor's contribution to the model, with their collective sum totaling 1. From the provided table, it's evident that "run_time" holds the highest weight of 37.10%, making it the most critical factor in model construction. This suggests that the duration of a movie plays a pivotal role in the model's predictions.

The second most significant factor is "budget," accounting for 23.59% of the total weight. This underscores the importance of a film's budget in shaping its potential success and influencing the model's outcomes. Closely following is "box_office" with a weight of 21.94%. This feature reflects the market performance of a movie and its impact on the model's predictions. Collectively, these three features: run_time, budget, and box_office-account for 82.64% of the total weight, highlighting their centrality in the model.

Moreover, genre categorizes the movie into specific types, such as action, comedy, drama, or science fiction, giving a deeper understanding of the movie's content and style. The remaining 14 features, including various movie genres such as crime, romance, drama, adventure, biography, sci-fi, and action, as well as different ratings like Approved, PG, G, 18+, PG-13, Not rated, and Passed, possess varying degrees of influence on the model. Each of these factors, though individually carrying lower weights, contributes uniquely to the model's overall functionality, providing a more comprehensive and nuanced understanding of the movie landscape.

In summary, the feature weights reveal the intricate balance between different factors that shape the model's predictions. Each feature, regardless of its weight, plays a vital role in the model's ability to accurately assess and predict outcomes related to movies.

Table 2. Random forest results

Item	Weighted Value	Item	Weighted Value
Drama	0.018	PG-13	0.006
Crime	0.046	PG	0.007
Action	0.01	G	0.007
Biography	0.013	Approved	0.009
Adverture	0.014	Not rated	0.003
Sci-Fi	0.012	18+	0.007
budget	0.236	Passed	0
box_office	0.219	run_time	0.371
R	0.023	-	-

4. Conclsuion

This study uses relevant analysis to study certificate, year, directors, casts, writers, run_time, budget and tagline and rank, name, rating, genre respectively. Box_office used Pearson correlation coefficient to represent the strength of the correlation between five items. We also use the feature weight to show the importance of each title's contribution to the model. By analyzing a large number of data, this paper obtains a series of statistical results. Then we have a conclusion that the certificate, year, run_time, budget, tagline, directors, casts don't relate to the movie popularity. Since films are characterized by high investment and high risk, studying the influencing factors of film box office and their influencing degree is a means to ensure the return of film distribution and control the risk of distribution, which has important reference value for the investment decision of films. Based on the random forest models, the director can take movies according to this in the future. Therefore, they do not need to consider the above factors when studying the influencing factors of film popularity but to pay more attention to the quality and content of the film itself. At the same time, future research can further explore other factors that may influence popularity to more fully understand the people's prefer to the movie. Chinese moviegoers have undergone great changes compared with the past. Under the impact of new media such as short videos, the number of moviegoers and young moviegoers is decreasing, while the age group of moviegoers is increasing. The mainstream moviegoers of the past have now reached middle age, and this group of people still retains the enthusiasm of the past. But younger audiences, with no movie-going habits, will have a long-term impact on the overall film market. For today's audience, aesthetic attributes give way to social attributes, we should use schedule, marketing and other ways to develop the derivative value of films, fully tap the social attributes of films to attract young audiences, and supplement new forces for our film market.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Qi Wei. Fan Film Preference, Viewing Orientation, and Box Office Balance: A Study on the Current Status of Online Film Rating. Contemporary Film, 2015, 5.

[2] Tan Wei. Analysis of Opinion Leaders on Film Rating Websites. Liaoning University, 2024.

[3] Zhou Wenle, Zhu Ming, Jiang Dan. A method for predicting movie ratings based on comprehensive time and rating factors. Electronic Technology: Shanghai, 2015.

[4] Wang Heng, Tang Xiaoguo, Guo Junliang. Python based web data crawling for movie ratings. Heilongjiang Science, 2022.

[5] Tan Jiazhu. Research on IMDB movie rating prediction based on random forest algorithm. Modern Computer (Professional Edition), 2021, 27(30): 24-31.

[6] Lu Junzhi. A movie rating prediction model based on random forest regression algorithm. Jiangsu Communication, 2018, 34(1): 4.

[7] Tao Yijia, Cao Jing. Interpreting the Trust Crisis in Film Rating in the Era of Information Explosion: An Improved Design Using Douban Film Platform as an Example. Industrial Design Research, 2017, 7.

[8] Jiao Yanan. Analysis of the Relationship between Douban Rating and Film Box Office. Youth Journalist, 2018.

[9] Wen Qu, KaiSong Song, YiFei Zhang, et al. A novel movie recommendation method based on multi visual content analysis and semi supervised enhancement. local of computer science & technology, 2013.

[10] Huang Dongjin, Ji Hao, Geng Xiaoyun, et al. A movie rating prediction model based on text vector features. Modern Film Technology, 2019, 7.