Engineering Solutions to Mechanics, Marine Structures and Infrastructures

Eng. Solut. Mech. Mar. Struct. Infrastruct.
2024.1(3).5
ISSN: 3006-2837
https://doi.org/10.58531/esmmsi/1/3/5

Article

A Study on Chinese Film Market Box Office Prediction Based on Random Forest

Minhua Ye[1,#], Xuewen Chen[2,#], Yonpei Cao[2,#], Ruihan Chen[2], Yueyang Wu[2], Di Ning[3], Shenlin Liu[2], Jiawei Luo[2], Jingxuan Cui[2], Kun Tang[2], Zhi Li[2,*]

[1] College of Ocean Engineering and Energy, Guangdong Ocean University, Zhanjiang 524088, China
[2] Faculty of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang 524088, China
[3] College of Economics, Guangdong Ocean University, Zhanjiang 524088, China
[#] These authors contribute to this article equally.
[*] Corresponding Author: Zhi Li

Academic Editor: Dapeng Zhang <zhangdapeng@gdou.edu.cn>

Received: 24 July 2024; Revised: 20 August 2024; Accepted: 21 August 2024; Published: 21 August 2024

Abstract: With the improvement of people's material living standards, people's demand for cultural life is also increasing. As a cultural form, film has attracted more and more people's attention and love, and the Chinese film market is also growing. As the most intuitive indicator to measure a movie, the box office has been widely concerned. Therefore, establishing a model to accurately predict the movie box office helps promote the benign development of movies and provides opportunities for movie makers to improve the quality of movies. This study selects the relevant data information of 421 films from 2002 to 2022, takes the release year, film type, and clustering type as independent variables, and takes film box office as dependent variables, and establishes the stochastic forest classification model and stochastic forest prediction model. Results On the surface, the movie box office prediction model based on random forest established in this study has a high reliability, and the R square reaches 93.6%, which provides a reliable method for the expected box office of the movie box office and mining the influencing factors of the movie box office.

Keywords: Random forest; Data features; Correlation analysis; Movie box office prediction

1. Introduction

Due to the ongoing growth of the digital economy, the needs for living standards and entertainment consumption among individuals are also improving. The film industry involves social and cultural

Citation: Ye M., Chen X. , Cao Y., Chen R., Wu Y., Ning D., Liu S., Luo J., Cui J., Tang K., Li Z. A Study on Chinese Film Market Box Office Prediction Based on Random Forest. Eng. Solut. Mech. Mar. Struct. Infrastruct., year, volume(issue), doi: 10.58531/esmmsi/1/3/5

components in addition to commercial ones. It is a form of art that has significant social and cultural ramifications, and a key metric to assess this sector is box office revenue from movies. Due to the impact of the epidemic, China's overall economic growth has declined, and the film industry has also been affected to some extent. However, the huge benefits it brings to the economic market cannot be ignored, so the development of the film industry has also received more and more attention. While the film industry has high profits, its huge risks also exist. In the market, only a few films can make profits, and most films are in a state of loss. If this research can systematically analyze the multi-dimensional factors that affect film consumption, build an evaluation system of factors that affect film box office income, conduct empirical research based on film big data to predict the commercial value of films in the market in advance and put forward suggestions for the investment and decision-making of film distributors, this study can reduce industry risk, which is of great significance for risk control.

For the research on box office influencing factors, in 2021, Chenhe Pan established a model involving multiple linear regression and adopted one-way ANOVA to find that scores, number of potential audiences, schedule, and source have a significant impact on box office revenue, and the average box office of domestic films is higher than that of foreign films [1]. Through the relevant testing and regression analysis of box office influencing factors, Libo Jin et al. found that word-of-mouth significantly influences a movie's box office, even exceeding the impact of actors on the box office. In 2018, Yao Hua, Li Bo, and others established a semiparametric regression model and found a strong, positive linear association between film type, attention, IP influence, and film box office [2].

For the study of movie predictive models, Ramesh Sharda uses neural networks and 10x cross-validation to build predictive models. In 2015, Xie Tian proposed a new least squares model average estimator for calculating model weights based on PMA criteria and found that using PMA to explain model uncertainty greatly improved the accuracy of box office predictions[3]. In 2022, Tolga Kaya et al. used machine learning technology to establish a Turkish film box office revenue prediction model, including multiple regression analysis, ridge regression, and SVM method[4].

However, the accuracy rate of box-office prediction can not reach a satisfactory result, so how to establish an efficient and accurate box-office prediction model is an urgent problem to be solved. This study mainly accomplishes the following three tasks.

- Data collection and pre-processing. The data of this study is mainly from Douban.com, the largest fan community and movie database on the Chinese Internet. After collecting the data, the crawled abnormal data was manually screened, the error data and missing data were cleaned up, and finally, a total of 410 movies were collected.
- Analysis of influencing factors: analyze the correlation degree of each index value of the collected film, and use Pearson correlation analysis to explore the correlation degree between various indicators.
- Random forest model building. Due to the high efficiency and accuracy of the random forest algorithm, and the more stable prediction effect after classification, the method of first constructing classification trees and then regression prediction is used to construct a box office prediction model.

2. Data exploratory analysis

Exploratory analysis of the data can help the team to have a preliminary understanding of some characteristics of the data before data analysis so that the data set can be further processed [5]. With the help of the statistical software SPSSPRO, the index correlation test was carried out on the data.

A significant number of individuals utilize the Pearson correlation coefficient to measure the correlation between two variables, and the team used the Pearson correlation coefficient to measure the degree of correlation between the eight characteristic values of the movie box office. Finally, the Pearson correlation coefficient heat map is shown in Figure 1.

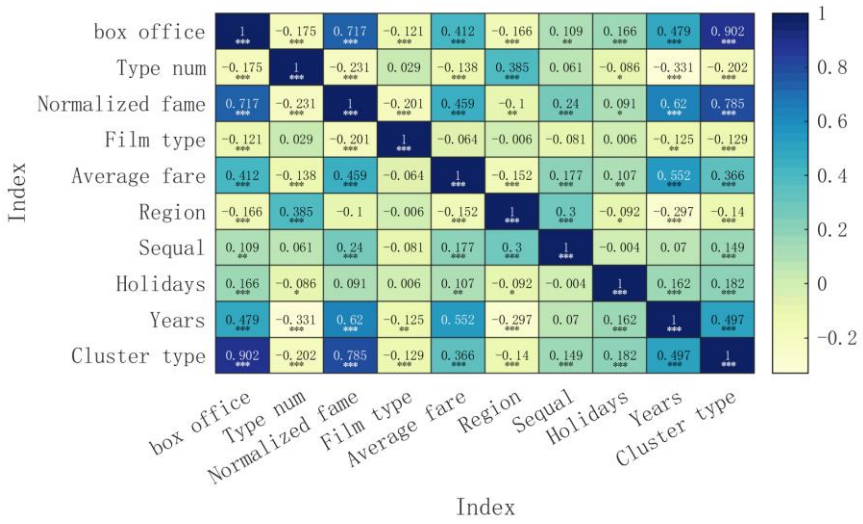

Figure. 1. Pearson correlation coefficient heat map

Pearson correlation coefficient values between -1 and 1, 0 indicate that the two variables are not correlated with each other, and negative values indicate a negative correlation between the two variables, whereas positive values indicate a positive correlation between the two variables. Also, the larger the absolute value, the greater the relationship. A significance level P less than 0.01 indicates a very high significance level, and a significance level greater than 0.05 indicates that the correlation coefficient is meaningless. In the heat map, the color depth shows how relevant the metric is.

Finally, it was found that the box office of the film had a strong correlation with the normalization of fame, the type of clustering, and the release year.

3. The Movie Box Office Prediction System's Evaluation Index System was established

First, feature quantization and feature dimensionality reduction are required. Before classifying and predicting movies, it is observed that the director, actor, company, and release date in the data are all non-digital formats, which is not conducive to subsequent processing and model solving, so this paper needs to perform feature metrics to quantify these important characteristics.

3.1 The indicator of Famous_degree

The first is the four indicators of directors, actors, publishing companies, and distribution companies. Since each of these four indicators contains Chinese names and is relatively scattered and cannot be directly translated, and these four indicators are positively correlated with the box office, the corresponding influence of directors, actors, and companies is calculated and synthesized into one

indicator: famous_degree, this move is also conducive to feature dimensionality reduction and reduces the classification error and regression error of the model. The formula for calculating fame is as follows:

$$\text{famous\_degree}_i = \text{Dir}_i + \text{Act}_i + \text{Com}_i \quad (1)$$

Among them, $\text{Dir}_i$is the director's influence, Act_i is the comprehensive influence of multiple actors Com_i is the influence of the film company.

To make the data more intuitive and easy to calculate,this article normalizes this feature：

$$\text{Famous\_degree}_i = \frac{\text{F\_d}_i - \text{F\_d}_{min}}{\text{F\_d}_{max} - \text{F\_d}_{min}} + 0.01 \quad (2)$$

3.2 The indicator of movie release data characteristics

For the indicator of the release date of a movie, if you only look at the date, you cannot find its relationship with the box office. Wang Zheng and Xu Min found in their research that the box office during holidays, especially during the Spring Festival, is much higher than the usual box office. Therefore, the following Table 1 of date characteristics is established in this paper.

Table 1. A table of movie release date characteristics.

Display	Value
National Day	1
Spring Festival	2
Summer Vacation	3
New Year's Day	4
Others	0

3.3 The indicator of types of movies

The types of movies are divided into nine types: action, comedy, plot, fantasy, science fiction, animation, romance, disaster, and war, which this article found when viewing movie-related information. Fantasy, science fiction, disaster, and war are four types of movies, and most of them are special effects blockbusters, so this article will directly classify these four types of movies into one category, finally, the type representation table of this article is shown in the following Table 2.

Table 2. Film genre characterization table.

Type	Value
Action	1
Comedy	2
Drama	3
Science Fiction	4
War	4
Fantasy	4
Disaster	4
Animation	5
Romance	6

Thriller	7

3.4 The indicator of sequel

Since the box office and word-of-mouth of a child film are heavily influenced by the parent film, must also extract characteristics for the label of whether the film is a sequel or not, Table 3 presents as whether the sequel.

Table 3. Sequel to Whether.

Sequal	Whether
Yes	1
No	0

3.5 The indicator of film box office classification

After feature quantization and feature dimensionality reduction, K-means clustering is used to classify each movie according to the box office.

K-means clustering is performed and the cluster is determined to be 4.

Table 4. Cluster category area value.

	Cluster category（Mean value + Standard deviation）			
	category1(n=246)	category 2(n=134)	category3(n=21)	category 4(n=8)
Box office	27101.37±183 12.148	102288.157±30790.15 6	262959.0±52010.793	492292.625±64895 .614

As shown in Table 4 and Figure.2, Category 1 is low box office, Category 2 is medium box office, Category 3 is high box office, and Category 4 is super high box office.

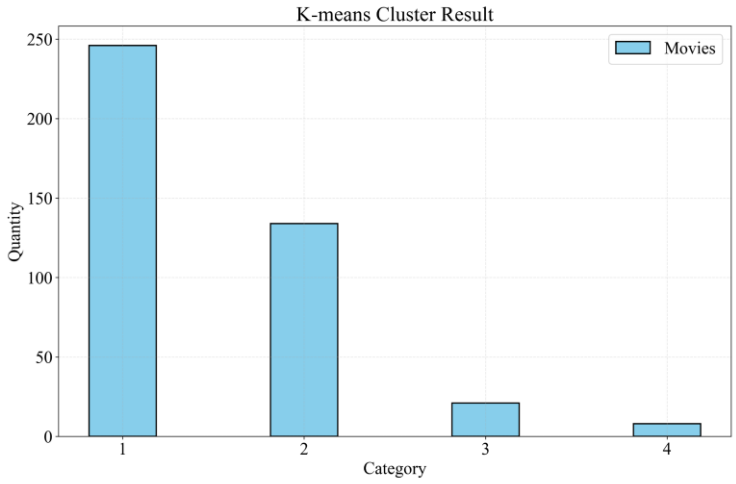

Figure. 2. K-means Cluster Result

4. Introduction to the Principle of the Model

4.1 The K-means clustering model

Introduction to the principles of the K-means model. K-means is a kind of similarity classification of samples clustering algorithm, the algorithm's main goal is to first determine some center points, determine

the category of the rest of the samples, so that the sum of the loss function is the smallest, according to this principle, the sample is initially classified, and then the center point of each class is adjusted to ensure that the sum of the loss functions in each class is the smallest, through continuous iterative adjustment, and finally the classification of the sample is realized. The implementation steps are as follows:

1) Randomly select k center points $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_k}\}$as cluster center

2) Define the loss function, which calculates the Euclidean distance from each sample $\boldsymbol{\{y_1, y_2, \dots, y_n\}}$ to the center of each cluster, as follows:

$$\mathrm{dis}(\mathrm{y_i}, \mathrm{x_i}) = \sqrt{\sum_{\mathrm{j=1}}^{\mathrm{m}} (\mathrm{y_{ij}} - \mathrm{x_{ij}})^2} \tag{3}$$

3) Determine the distance between each object and each cluster center. Assign each object to the class cluster that is closest to the cluster center, and get k class clusters $\{\boldsymbol{S_1}, \boldsymbol{S_2}, \dots, \boldsymbol{S_k}\}$

4) Recalculate the center $\{\boldsymbol{Z_1}, \boldsymbol{Z_2}, \dots, \boldsymbol{Z_k}\}$ of each cluster, that means, find the point that is furthest in distance from the center point in the cluster, and calculate the center of the cluster as follows:

$$\mathrm{x_j} = \frac{\sum \mathrm{y_i}}{\mathrm{s_l}} \tag{4}$$

4.2 Random forest

1) The tree-based models

Since tree-based models serve as the foundation for random forest algorithms, they are discussed first in this work. Recursively separating a given data set into two groups is part of a tree-based model according to specific criteria until a set halt requirement has been met. The leaf nodes or leaves are located toward the base of the decision tree. Figure.3 illustrates the recursive partition of a two-dimensional input space with an axis-aligned boundary -- therefore, every input space is partitioned along an axis-parallel path. Here, $x2 \geq a2$ is where the initial split happens. These two subspaces are then split apart once more: the left branch splits at $x1 \geq a4$. The right branch splits first at $x1 \geq a1$, and one of its subbranches splits at $x2 > a3$. Figure.4 is a graphical representation of the subspace divided in Figure.3.

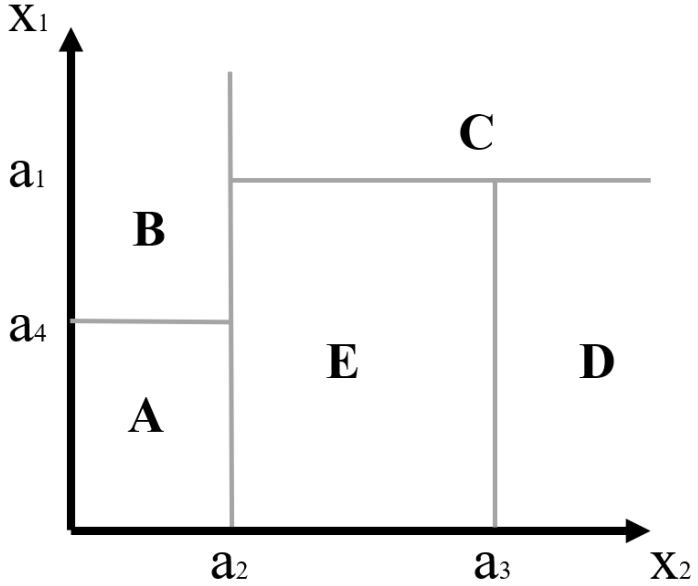

Figure. 3. Recursive binary partition of a two-dimensional subspaces

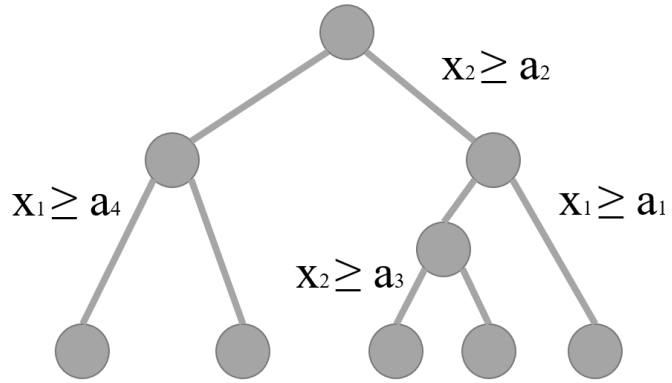

Figure. 4. A graphical representation of the decision tree in Figure.3

Based about how the stopping and partitioning conditions are configured, you can design decision trees for classification tasks (classification results) and regression tasks (continuous results).

The choice of the subset of prediction variables used to divide the internal nodes in classification and regression tasks depends on predetermined split criteria stated as optimization problems. Entropy, a practical implementation of Shannon's source coding theorem that establishes a lower bound on the length of a random variable's bit representation, is a common dividing criterion in classification issues. The following formula provides the entropy through every internal node of the decision tree.

$$E = -\sum_{i=1}^{c} p_i \times \log(p_i) \quad (5)$$

In which c is the total number of distinct classes and pi is the probability for only a particular class. To acquire the most data possible from each decision tree branch, this value is maximized. Regression-related issues, the common split criterion is the mean square error of every individual host.

One drawback of decision trees is how simple it is to overfit them, indicating that now the model is too close to the characteristics of the training set and does not perform well on the test set. Overfitting the decision tree will result in poor global prediction precision, also known as generalization precision.

A way to improve the precision of generalization is to generate several separate trees while just taking into account a small number of observations. The concept of the random subspace techniques was first proposed by Ho [6]. The random forest algorithm system averages estimates for many different trees. Individual trees are constructed using bootstrap samples as opposed to the original samples. It's known as bootstrap aggregation, and it can decrease overfitting.

2)The random forest algorithm

The random forest algorithm is commonly used to classify samples and predict regression, and classify the sample data into the class with the most votes by voting. Here are the steps involved in implementation:

- Choose the training set. K training sets were extracted from the original dataset using the Bootstrap method with random sampling;
- Build a random forest model. For each training, construct a classification tree set with back sampled, generate K decision trees, and form a "forest" without pruning. Among them, the most important thing to construct a classification tree in a random forest is to choose the best splitting method, and this paper proposes to use the Gini coefficient method, the formula of the Gini coefficient is:

$$\text{Gini}(p) = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2 \quad (6)$$

- Create a voting mechanism. K decision trees combined that have been trained similarly, a random forest can be created. The classification outcomes are determined by a straightforward vote for every decision tree's output.

When the model predicts, the test data is fed into the model, those of the outputs Voting decision trees can be created, and the final forecast result with the highest vote is selected.

5. Forecast Accuracy

5.1 Training results of random forest classification model

To begin the random forest classfication model training process, this research arranges the dots are arranged randomly. Whenever the data is split into training data and test data, the random sort order ensures that the training data is also random. To get reproducible results, this research This study established a seed value. This paper then split the data set into two subsets: 70% of the data for training and 30% for testing. The aforementioned randomization procedure makes sure that the training data comprises observations that fall within each of the categories that are available, as far as there is no appreciable imbalance in the category probabilities. In addition, it eliminates the model's potential dependence on how the observations are arranged in relation to the test data.

After using the training set data to establish a random forest classification model, the Gini coefficient is used as the evaluation value to measure the contribution degree of each indicator feature on the tree, the greater the Gini coefficient, the greater the degree of contribution, and several index features with high contribution degree are screened, and finally 5 indicators: Famous_degree, Years, Holiday, Sequal, Type_num, these five features are input into the random forest, and the model is constructed with the classification category as the output. Figure. 5 displays the characteristics of the random forest classification model and their relative importance.

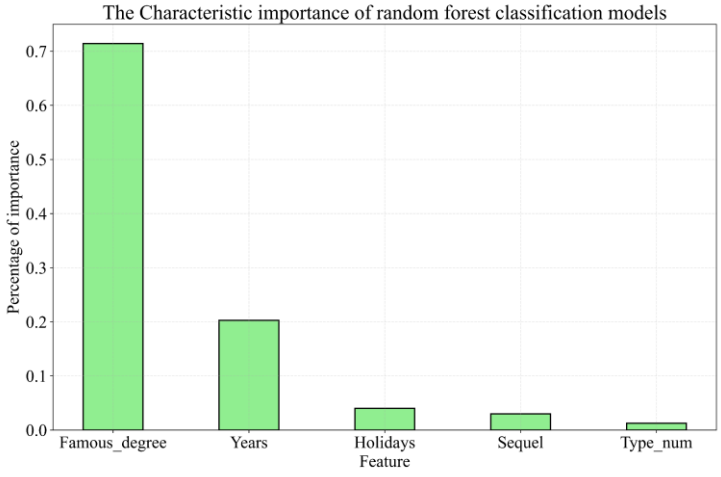

Figure. 5. The Characteristic importance of random forest classification models

Next, this paper adjusts to identify the model with the best test accuracy, the hyperparameters. The parameters of the stochastic forest classification model the following in this essay: n_estimators=200, bootstrap=True, oob_score=True, max_depth=10, min_samples_split=2, min_samples_leaf=2, max_leaf_nodes=50, n_jobs=-1.

The partial decision tree visualization of the random forest classification model is shown in the following Figure.6.

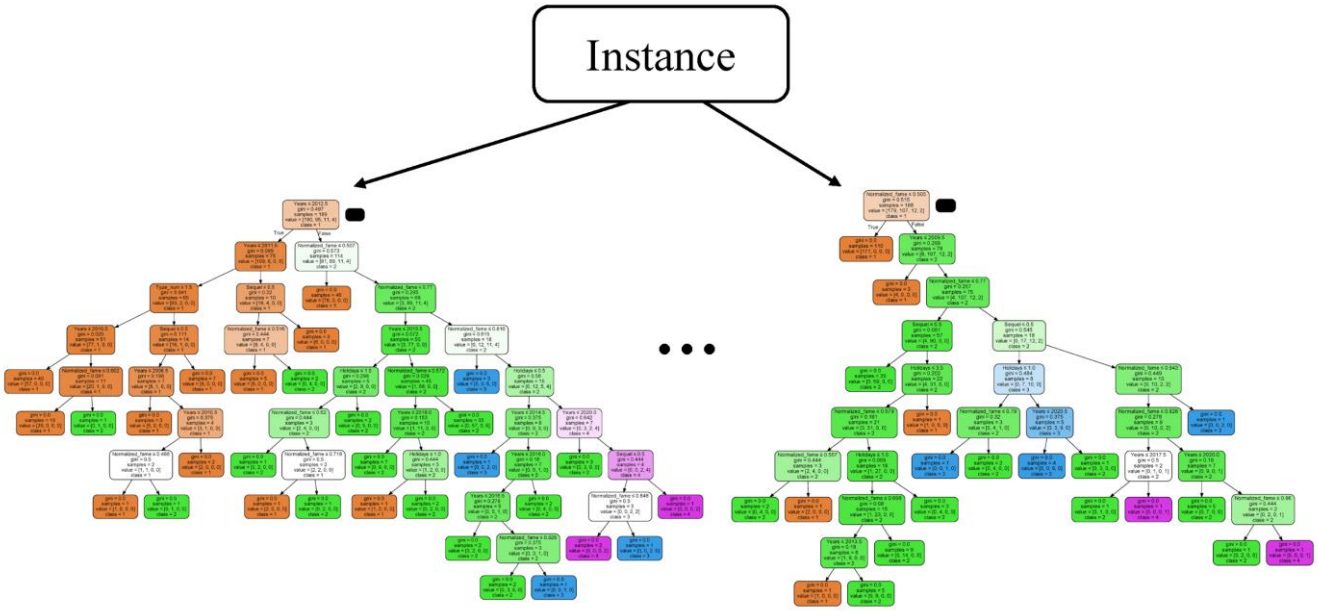

Figure. 6. Visualization of a random forest movie classification model

The classification precision of the training set is 1.00, and the classification precision of the test set is 0.954. The following Table 5 displays the final classification results.

Table 5. Random forest prediction results.

Actual	Predicted
2	2
3	3
1	1
4	4
3	3
⋮	⋮

5.2 Training results of random forest regression model

Just like the random forest classification model. To begin the training process of the random forest regression model, the data points in this study were sorted at random. When training data and test data are separated from the data, the random sort order ensures that the training data is also random. To get reproducible results, this research set a seed value. The dataset is divided into a training set and a test set according to the 7:3 ratio. The previously mentioned randomization process ensures that the training data contains observations that belong to all available categories, as long as the category probabilities are not significantly out of balance.

After using the training set data to establish the random forest regression model, according to the contribution degree of each index obtained by the Gini coefficient: Cluster_type, Famous_degree,Years, Holiday, Sequal, filter the remaining five indicators, input these five characteristic indicators into the random forest, build a random forest model with the box office prediction value as the output. The features of the random forest regression model and the visualization of their importance are depicted in Figure.7.

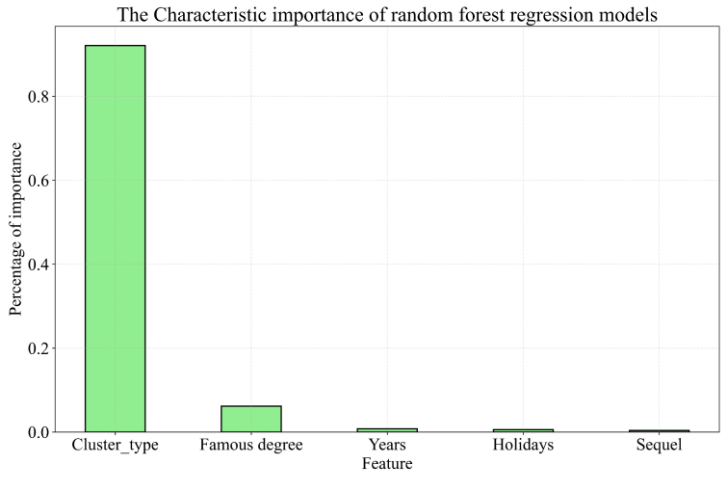

Figure. 7. The Characteristic importance of random forest regression models

Next, this paper adjusts selecting the model with the greatest accuracy level using the hyperparameters. The parameters of the stochastic forest classification model the following in this writings: n_estimators=200,bootstrap=True,oob_score=True,max_depth=20,min_samples_split=2,min_samples_leaf=1,max_leaf_nodes=50,n_jobs=-1.

The partial decision tree random forest categorization visibility model is in Figure 8:

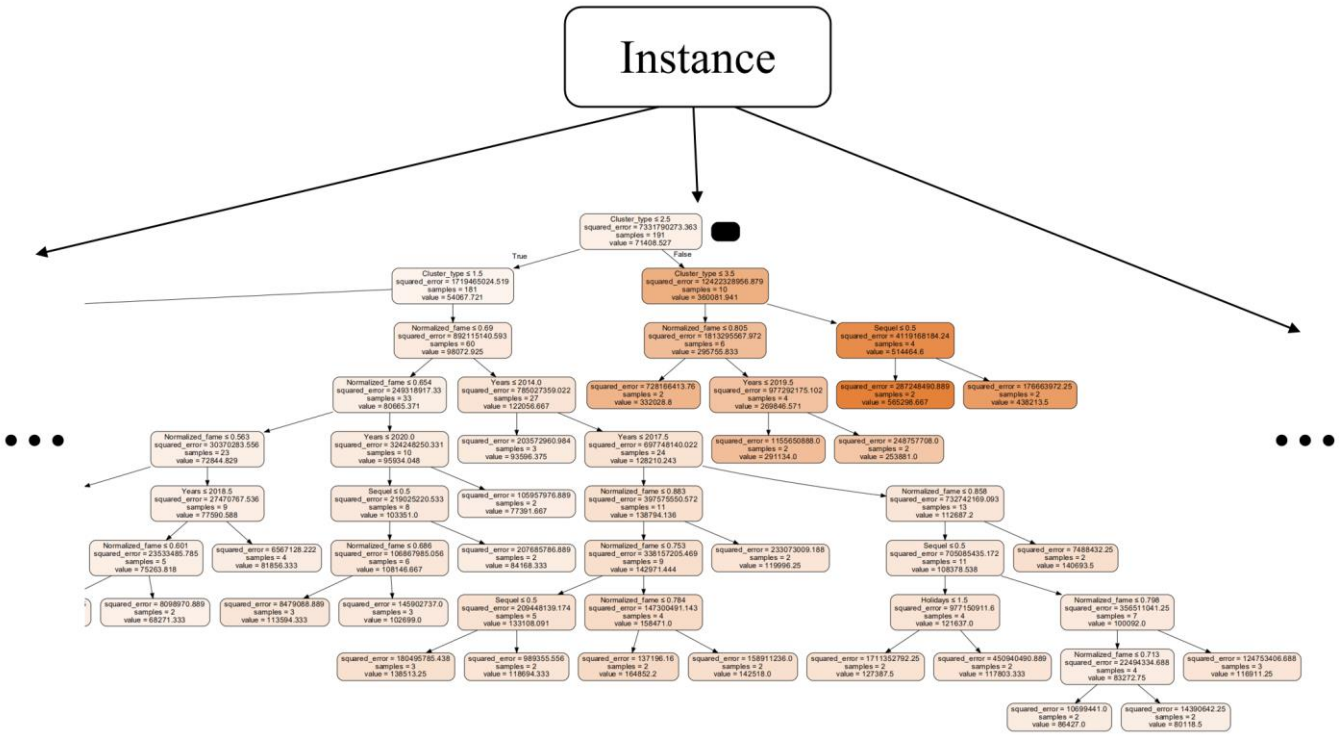

Figure. 8. Visualization of a random forest movie prediction model

Using the random forest prediction model, the training and test set data are predicted respectively, and the actual value of the training set is fitted to the anticipated result, and the actual value and the predicted value of the test set are fitted and plotted in the same way, and the final fitting is shown in the following figure.

The training set prediction and actual value fit plot are in Figure. 9.

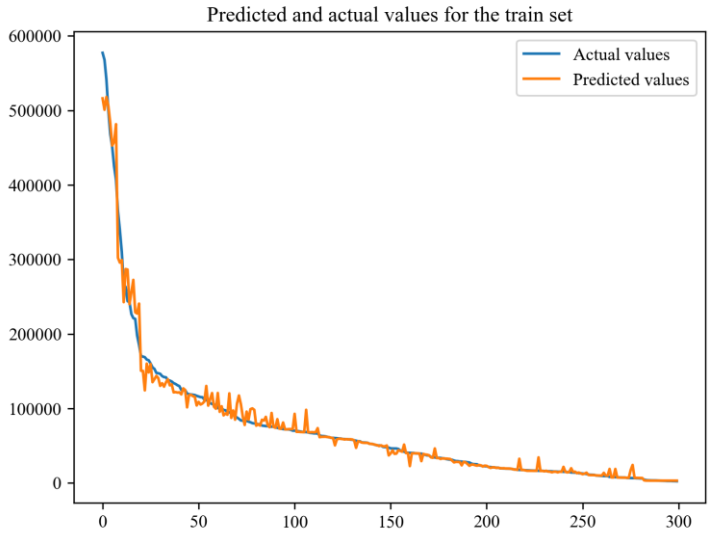

Figure. 9. Predicted and Actual values for the training set

The test set prediction and actual value fit plot are shown in Figure 10.

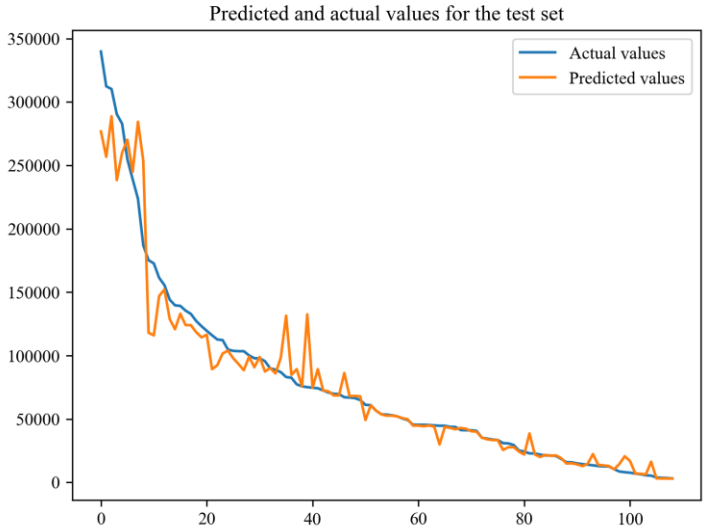

Figure. 10. Predicted and Actual values for the test set

To evaluate the prediction model, this study utilizes three evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2). The MAE metric quantifies the level of agreement between the predicted values and the actual values, with a smaller MAE indicating a better-fitting model. Similarly, the MSE metric also measures the goodness of fit between the predicted values and the actual values, with a smaller MSE suggesting a superior fitting model. The Coefficient of Determination (R^2) is a statistical measure employed in regression analysis to assess the goodness of fit of the model, with values ranging from 0 to 1. A higher R^2 value closer to 1.0 signifies a more robust fit of the model.

Below are the formulas for Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (8)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

Where n is the sample size, y_i represents the true values, $\hat{y}_i$ represents the predicted values by the model, and $\bar{y}$ denotes the mean of the true values.

The performance evaluation outcomes of various carbon emission prediction models are displayed in Table 6.

Table 6. Random Forest Movie Regressor Model Summary.

R^2_{Train}	R^2_{Test}	MAE_{Test}	$RMSE_{Test}$
0.981	0.936	9360.58	18498.328

The value of R^2 in the training set is 0.981 and the test's value in the test set is 0.981, $R^2$is 0.936, so the predictive model constructed in this paper works well, establishing a model to accurately predict the movie box office is helpful to promote the benign development of movies and provide opportunities for movie makers to improve the quality of movies.

6. Discussion

By collecting a large number of official movie box office data, this study establishes a random forest model to predict the movie box office, and the prediction effect reaches a satisfactory result. However, if we want to further improve the accuracy of movie box office prediction, we should explore the deeper factors affecting the movie box office, so as to better help the movie box office prediction. The future work of this study is devoted to exploring the deeper influencing factors of the movie box office, and combined with more advanced intelligent algorithms, is committed to building a more perfect movie box office prediction system. In addition, the study found that cheesy and gory movies are difficult to attract people's attention. Secondly, Chinese films combined with Chinese national culture can attract more moviegoers.

However, if we want to further improve the accuracy of movie box office prediction, we should explore the deeper factors affecting the movie box office, to better help the movie box office prediction.

The future work of this study is devoted to exploring the deeper influencing factors of the movie box office, and combined with more advanced intelligent algorithms, is committed to building a more perfect movie box office prediction system.

7. Conclusion

Using the random forest model, a Chinese film box office prediction model was constructed, but due to the impact of uncontrollable factors such as the epidemic in recent years, the scale of the film market fluctuated downward, so when predicting the future box office, some characteristic indicators can be appropriately added according to the situation to improve the accuracy of prediction.

In light of the research's findings, the suggestions below are made for the growth of the domestic film market: First, filmmakers should focus on the improvement of the quality of their works, as consumers' tolerance for bad films continues to decrease, it will be difficult for tacky and bloody films to occupy a place in the film industry. Secondly, the Chinese film industry should attach importance to the external display of Chinese culture, and film, as a form of culture, also bears the burden of spreading Chinese culture to the world. Therefore, the creation should pay more attention to the display of national spirit, promote excellent works with the imprint of Chinese culture to the world, and drive China from a film country to a film power.

Acknowledgments

I would like to thank the research partners for their cooperation and the support, help, and understanding of friends around for this study.

Funding: This project is funded by the Guangdong Provincial Science and Technology Innovation Strategic Special Fund (No. pdjh2023b0247) and the Guangdong Ocean University Undergraduate Innovation Team Project (No. CXTD2023014).

Conflict of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References:

1. Liao Lin, Huang Tao. The effect of different social media marketing channels and events on movie box office: An elaboration likelihood model perspective. Information & Management. 2021; 58: 7. Doi: 10.1016/j.im.2021.103481
2. Y. Hua, B. Li, J. Ren, and H. Yan, Research on influence Factors of IP movie box office based on semi-parametric model, 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS). IEEE: Beijing, China. 2018.
3. Tian Xie. Prediction model averaging estimator. Economics Letters, 2015; 131: 5-8. Doi: 10.1016/j.econlet.2015.03.027
4. Anil Gürbüz, Ezgi Biçer and Tolga Kaya. Prediction of gross movie revenue in the turkish box office using machine learning techniques. 2022 International Conference on Intelligent and Fuzzy Systems. 2022; 505: 86–92. Doi: 10.1007/978-3-031-09176-6_10
5. Maozhu Jin,Yanan Wang,Yucheng Zeng. Application of data mining technology in financial risk analysis. Wireless Personal Communications. 2018; 102: 3699‑3713. Doi: 10.1007/s11277-018-5402-5
6. Tin Kam Ho. Random decision forests, In Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE: Montreal, QC, Canada. 1995.