Proceedings of the 7th International Conference on Economic Management and Green Development
DOI: 10.54254/2754-1169/42/20232103

Prediction and Analysis of Shoes Popularity Based on Random Forest Regression Model

Botao Cui[1,a,*]

[1]*School of Mathematics and Statistics, Zhengzhou University*
a. email, 20204661@stu.cqu.edu.cn
**corresponding author*

Abstract: Since the era of big data, e-commerce has become a popular industry. People are more inclined to online shopping websites and apps, which are more convenient ways to buy commodities. In order to increase the popularity and fame of websites and apps to attract more customers, e-commerce often analyzes and models customer feedback and purchase types in order to get customers' higher preferences for the performance of products. This paper collected a variety of popular shoes' scoring situation, current price, sales volume, wanted quantity, releasing date, releasing price and other relevant quantities. The HM aggregation operator is used to define the popularity (the popularity of the shoe). In this paper, the random forest algorithm is used to conduct regression modeling analysis with the pre-processed data, and the importance estimation of each criterion and a regression prediction model are obtained according to the established model. Then this paper analyzes the shortcomings of the model, and adds data normalization processing to the steps of data preprocessing to make the model more powerful in estimating outliers. According to the results, some suggestions are made for the research and development priorities of footwear companies and it is verified that the model can basically complete the prediction of popularity.

Keywords: footwear, popularity, random forest, forecast

1. Introduction

As the era of big data occurs, human lifestyle has undergone great changes due to the impact of big data which has become the label of human beings, reflecting human preferences, living habits and so on to a certain extent. In the age of big data, the rapid development of artificial intelligence has fostered change in the way of businesses function and sell. Many e-commerce enterprises, such as Taobao, Jingdong, Pinduoduo, Suning Shopping, Tiktok, etc., have introduced consumer behavior to financial activities, so as to guess customers' preferences and needs, increasing customers' purchase times and the popularity of e-commerce websites and apps.

As early as the 18th century, western scholars began to pay attention to consumer behavior. In 1980s, the research scope of consumer behavior was continuously expanded, the research methods were gradually improved, and its content and depth were further promoted and deepened [1]. With the rise of the Internet, e-commerce and online marketing are developing at an alarming rate. Network consumption, a new consumption behavior, has become an indispensable part of contemporary people's daily life. Among online shopping goods, the transaction scale of clothing products has been rising continuously [2]. In order to meet the needs of online shopping, more and more scholars began

DOI: 10.54254/2754-1169/42/20232103

to study related fields. Dodds and other scholars argue that by shaping the brand image, consumers' perception of brand risk can be effectively reduced [3]. Reichheld emphasized the importance of retaining customers [4]. Koufaris pointed out that consumers' online shopping behavior is influenced by many factors, including but not limited to personal reasons, network environment and online experience [5]. According to Corritore and other academics, trust is a crucial component of internet shopping. [6]. Ahmed believes that perceived risk, psychological factors and perceived interests have a significant impact on consumers' online shopping behavior [7]. In order to solve more quantitative and explanatory results, scholars began to add mathematical algorithms and establish models to quantitatively score and predict known products. Cao Chen et al. summarized the popularity prediction research based on deep learning in recent years, divided it into popularity prediction methods based on depth representation and depth fusion, and analyzed and prospected the development status and future trend of this research direction [8]. In 2014, Kong Qingchao and others put forward an algorithm for predicting the popularity of discussion posts, which combined local characteristics and multiple dynamic factors [9]. In 2018, Xu Huayun and others used FastText and GBDT structure to predict whether the topic of Weibo could become a hot topic, and the accuracy of prediction in the experiment reached 85.27% [10]. In the same year, Regression and classification techniques were employed by Zohourian, Alireza, and others to forecast the quantity of being received as a measure of popularity [11]. The popularity of users on Iranian social media is predicted by the academics. In 2019, GongWeizhi proposed a prediction model for topic popularity that incorporates sentiment and leverages semantic information to better precisely measure and predict popularity [12]. In 2020, Zhang Yixuan and others proposed a model to predict app popularity [13]. In 2018, Wang Jianrong and other scholars put forward a Weibo heat prediction algorithm based on XGBoost and Random Forest [14]. The author developed a machine learning-based model to predict popularity as a result of the aforementioned research.

In the past few years, people can clearly feel that there is a whirlwind of "shoe hot" among teenagers. Especially between 2017 and 2020, the price of many fashionable shoes has been raised by some vendors to 2-3 times the original price and even more than 10 times the original price. According to some simple economic principles, when prices are rising rapidly, sales should decline. The fact is quite the opposite, which a few of shoes whose price is larger increasing are by countless people flocked to, once sold out. This makes people wonder, what is the factor that makes the sales and prices of fashionable shoes crazy rise? Is it feasible to create a regression model that forecasts shoe popularity? If an accurate model can be established, it can not only effectively reduce the production cost of manufacturers, but also improve the satisfaction of customers. In the process of model building, it is also possible to determine which factors have a greater impact on the final result.

This study collected the relevant data and comments (including quality, appearance, foot feeling, price, releasing time, the number of sales, the number of the wanted, releasing price, price difference) of more than 100 AJ (a shoe with the fastest price increase) sold in Dewu APP, and used it as a set of features of this AJ to establish a regression model using a variety of machine learning algorithms. The author analyzes the predicted results after observation and try to build a reasonable prediction model.

2. Preliminaries

This chapter will detailly introduce the methods and knowledge involved in model building, including HM aggregation operators, random forest algorithms and other machine learning algorithms.

DOI: 10.54254/2754-1169/42/20232103

2.1. HM Aggregation Operation

In 1998, Hara et al. [15]proposed the Heronian mean aggregation operator (HM aggregation operator) as defined in the following way.

Definition 1: Supposed c_i(i=1,2, …, n) is a set of non-negative real numbers, and r, t⩾0. The HM operator can be defined in the following form,

$$\mathrm{H}M^{r,t}(c_1, c_2, \ldots, c_n) = \left(\frac{2}{n(n+1)}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i^r c_j^t\right)^{\frac{1}{r+t}} \quad (1)$$

2.2. Random Forest Algorithm

Random forest technique further introduces random attribute selection in the training of decision trees based on Bagging integration produced by decision tree-based learners [16]. Random forest is regarded as "the method representing the level of integration technology" since it is straightforward, simple to use, and low computationally intensive. Random forest also performs well in many actual tasks.

2.2.1. Decision Tree

The decision tree method is to classify recursively the samples that share the same attributes according to the attribute performance of each sample. According to the result of recursion, determine whether to continue with recursive partitioning or recursive return. A recursive return operation is performed when the following conditions occur. After the division, the samples in each set belong to one class, that is, the required classification work is perfectly complete. The current set is empty, that is, no element can be used for the next operation. All attributes of the samples in the current set are consistent, that is, there is no attribute basis for further division of the remaining samples.

To make it easier to understand, let's take a specific example:

For a sample set, first of all, it is judged by the standard of quality. The part with more than nine points is judged by the score of foot feeling, and the part with less than nine points is classified by the release time. For the part with a score greater than nine, when the foot feeling score is also greater than nine, the appearance score is used for classification. The part less than nine points is further divided by the sale time. When all three scores are greater than nine, the result set and shoes sold before 2020 will be recorded as popular. The rest are recorded as unwelcome. At this time, the construction of a decision tree is completed. The decision tree constructed is as follows in Figure 1:

DOI: 10.54254/2754-1169/42/20232103

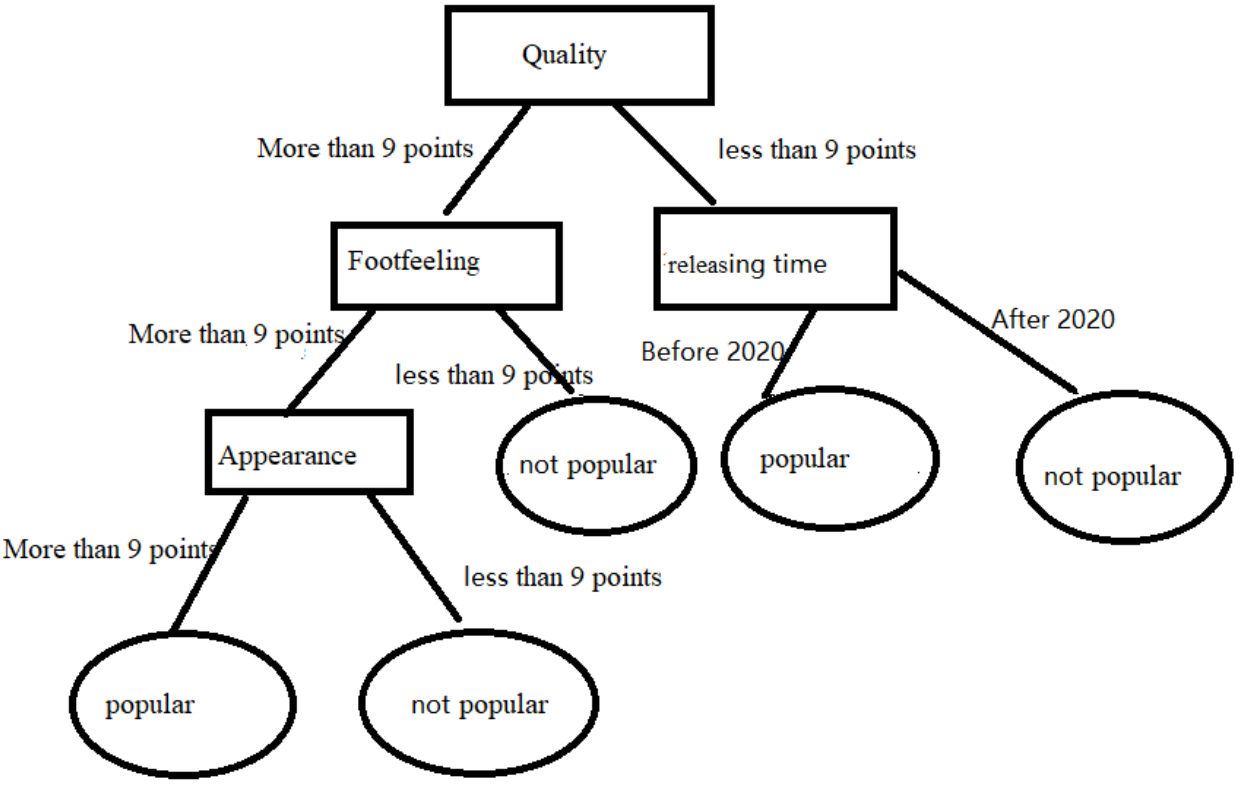

Figure 1: An example of the decision tree.

However, the real situation is often not as ideal as the known sample set can be perfectly divided, and then the decision tree model determined by the partition results can be perfectly predicted. Too detailed division will lead to increased algorithm burden and overfitting, while too much of a variance between the projected result and the actual number will make it impossible to accomplish the prediction challenge. In order to address this issue, concepts such as information gain and Gini index have been introduced to determine whether further division is needed.

Definition 2: (information entropy) Assume that the proportion of class k samples in the current sample set D is p_k (k=1, 2, …, |γ|), Then the information entropy of D is defined as

$$\mathrm{Ent(D)} = -\sum_{k=1}^{|\gamma|} p_k \, log_2 p_k \tag{2}$$

Definition 3: (Information gain) Assume that the discrete attribute a has V possible values $\{a^1, a^2, \ldots, a^v\}$。$D^v$ represents the samples in sample set D whose value is a^v when divided according to discrete attribute a, then the information gain obtained after partitioning the sample set D with attribute a is,

$$\mathrm{Gain(D, a)} = \mathrm{Ent(D)} \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{3}$$

DOI: 10.54254/2754-1169/42/20232103

In general, the "purity boost" obtained by utilizing attribute a for partitioning increases in proportion to the information gain. Therefore, when the decision tree is constructed, the information gain can be used to determine whether the next step is divided by this attribute.

2.2.2. Bagging

The Bagging method is a method proposed by Breiman in 1996. The sampling procedure with put back is used to m random samples in a given data collection of m samples. A basis learner is taught for each of the m samples in each of the T sample sets after which these basis learners are concatenated. Bagging commonly employs a basic voting approach for the classification task and a simple mean method for the regression task in the development of predictions, finally yielding the relevant regression result.

Only 63.2% of the samples from the initial training set are utilized in the self-sampling step of the Bagging technique, and the remaining 36.8% can be used as validation sets to get "out-of-packet estimates" of generalization performance. From the perspective of deviation-variance decomposition, Bagging method reduces variance, and the utility of this feature is more obvious on learners susceptible to sample perturbation.

3. Model Establishment and Result Analysis

3.1. Data Collection

This paper randomly records the details of 100 AJ series on Dewu, a popular online shopping APP (please see excel table for details), which includes the scores given by Dewu's own scoring system on the quality, appearance and foot feel of each AJ. The scoring system is based on the average of all the users who have purchased the product's evaluation scores on the three aspects of the product. At the same time, in addition to these three criteria, the price is also an important reason for whether the shoes are sought after by the public. After continuous in-depth understanding of the price changes of these products, the author found that whether it is the same model of the star or a joint model with other brands will also have a certain impact on whether it is sold well. After thinking, the author believes that the release time will also become an important factor affecting the sales of this shoe. For example, shoes released at the right time, such as before the COVID-19 pandemic, generally sell better. Since the price of AJ in the APP is constantly changing, the starting price and the amount of price change are also important bases for judgment. After considering the influencing factors clearly, it is necessary to determine how to express the popularity of this shoe. The author found in the APP that there are rough sales statistics for each shoe. But measuring popularity by sales alone is far from enough. Many audiences, for some reasons, even though they like the shoes very much, they do not buy them. The APP also counted the number of people who marked the shoes as wanted. This data becomes an important part of building popularity.

To sum up, this paper selects the data of quality, foot feeling, appearance, price, release time, release price, price difference, sales volume and desired quantity from Dewu APP for regression analysis and establishment of the forecast model.

3.2. Data Processing

The data processing in this paper is mainly concentrated in three parts, one is whether it is the same or joint name of the star. In this attribute, the author uses "one hot" encoding to describe, that is, if the condition is met, it is 1, and if the condition is not met, it is 0. The second is the processing of the release time. According to the collected data, the expression of the release time of each shoe is different, some are more detailed, such as June 13, 2019, but the release date of some styles is not

DOI: 10.54254/2754-1169/42/20232103

clear, such as spring 2020. For the convenience of recording and subsequent operations. The year is divided into four distinct seasons: spring, which lasts from March to May; summer, which lasts from June to August; fall, which lasts from September to November; and winter, which lasts from December to February. For example, spring 2019 is recorded as 2019.25, summer 2019 as 2019.5, autumn 2019 as 2019.75, and winter 2019 as 2020. The third is the determination of the popularity value, which consists of two parts: sales volume and desired quantity. The time elapsed since the release date and the present time. Figure out the average annual sales volume and the average annual desired volume. The two-attribute information is aggregated by HM aggregation operator. The parameters are configured to have a value of 0.5 to make calculations easier. The whole worth of the shoe is its market value.

However, the authors found that the value of popularity is large and there is no specific evaluation method. The result makes it difficult for decision-makers to have a more obvious understanding of the popularity of this shoe. At the same time, because the value is too large, the error value of the model is too large, which will affect the judgment of the effect of the model to a certain extent. Therefore, this paper deals with the value of popularity and compresses it into the interval [0,1], which will make the error of the model smaller and easier to understand. At the same time, this operation will also allow decision-makers to have a more accurate understanding of the popularity of the shoes.

3.3. Specific Steps

This section explains the specific steps of building and training regression prediction model with MATLAB.

The first step: read the data.

Step 2: Separate the test set and the training set from samples:

1. The data set is processed using the bagging method, which involves extracting the data set and putting it back.

2. Input columns 2 through 6 of the first 80 rows as the training set, and row 9 as the output of the test set. Take columns 2 through 6 of the following 21 rows as input to the test set, and column 9 as output to the test set.

Step 3: Build the model.

1. Determine the number of decision trees.
2. Determine the minimum number of leaves.
3. Open the error graph (error curve as the number of decision trees increases)
4. Calculate the importance of features.
5. Determine whether the purpose of the model is regression or classification.
6. Model establishment by function.
7. Draw importance chart (Importance indicates the importance of different features to the output result)

Step 4: Simulation test.

Step 5: Calculate the stability of the root-mean-square error analysis model.

Step 6: Drawing the graph.

Step 7: Calculate relevant metrics.

3.4. Result Analysis

3.4.1. Result Display

The result after the code is modeled is shown in the following Figure 2, Figure 3, Figure 4 and Figure 5, respectively:

DOI: 10.54254/2754-1169/42/20232103

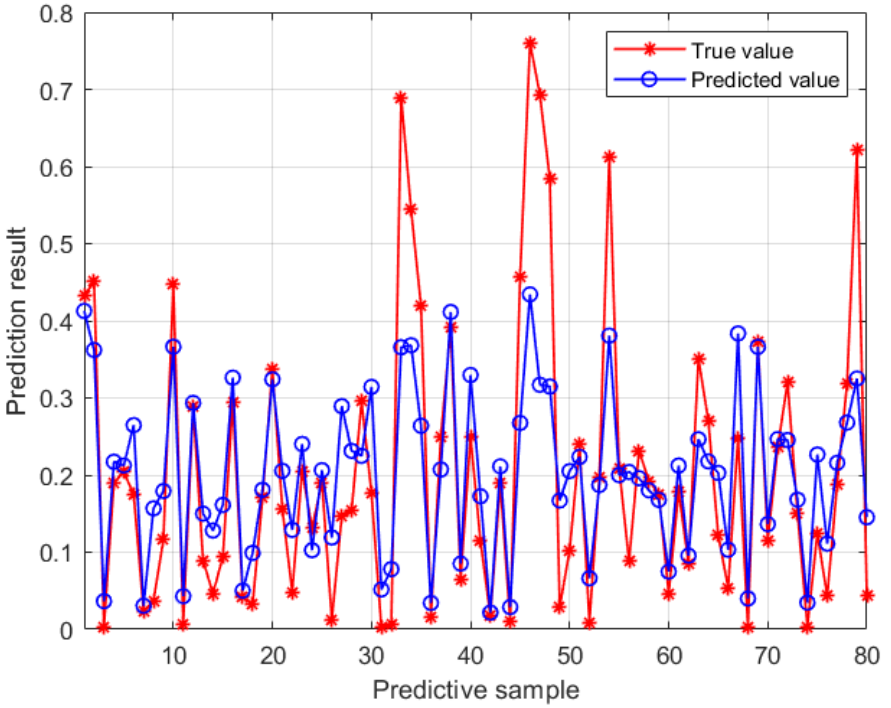

Figure 2: The situation of the training set predicting.

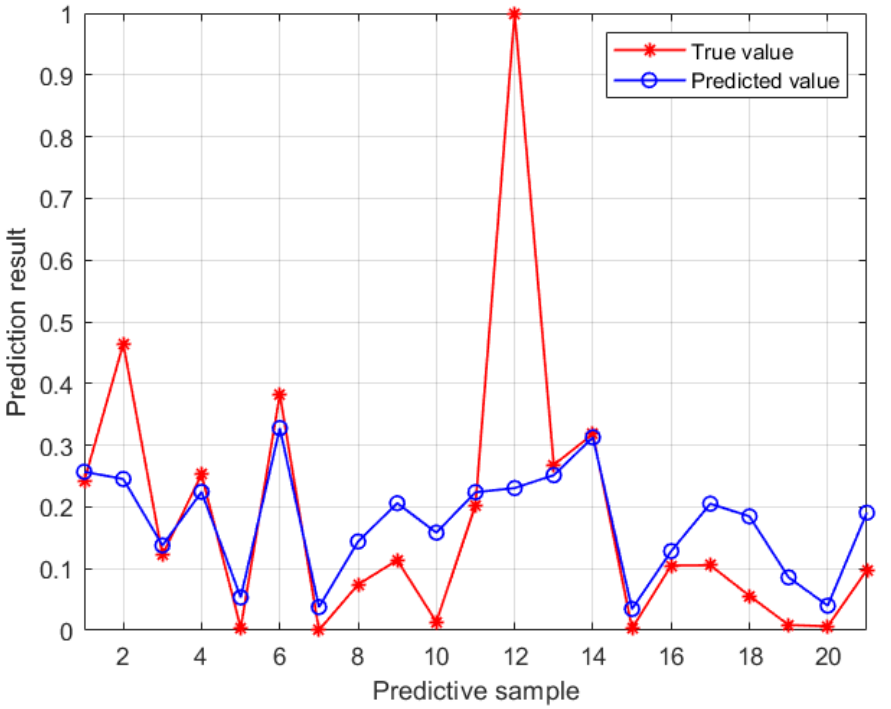

Figure 3: The situation of the testing set predicting.

DOI: 10.54254/2754-1169/42/20232103

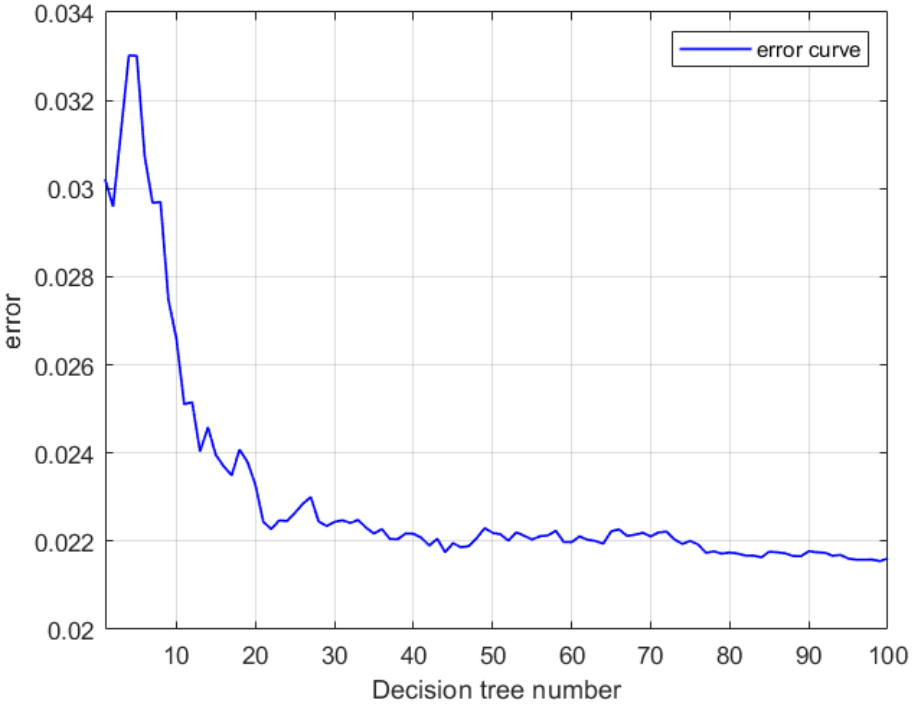

Figure 4: Decision tree number and error.

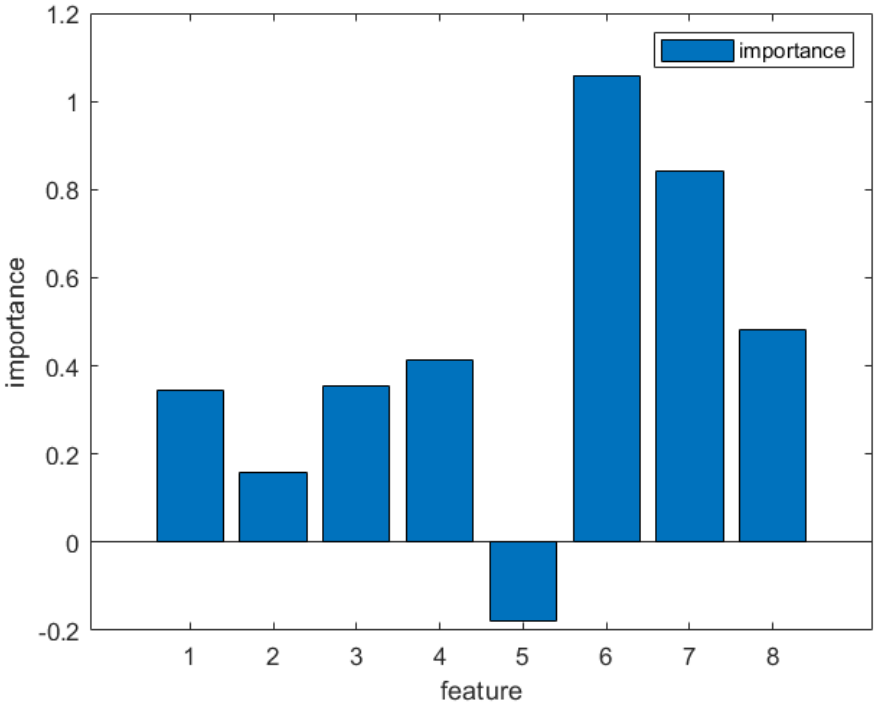

Figure 5: Importance of features.

With the exception of a few outliers, we can observe that the discrepancy between the projected value and the real value is minimal. This is especially true in the training set, where the non-outliers'

DOI: 10.54254/2754-1169/42/20232103

projected values are reasonably close to the actual outcomes. Although for most of the predicted values, their difference from the true value is small, but the number of coincidence (prediction correct) points is very small. At the same time, according to the error analysis diagram, it can be found that with the increase of the number of decision trees, the error is generally decreasing and it decreasing sharply when the number of decision tree arrives at the interval [10.20]. According to the analysis chart of feature importance (FIG. 4), it can be seen that feature No. 6, that is, the release time, and the selling price at the release time of feature No. 7, have a great impact on the final result.

3.4.2. Model Improvement

In order to further improve the model's accuracy, the author decided to add a data normalization method to the existing model to preprocess the data. The role of data normalization is to transform data into a specific range or distribution for better analysis and modeling.

Reasons for using data normalization:

1. As the numerical differences between individual eigenvalues and between eigenvalues and output values in this paper are very large, the model's performance and accuracy will be impacted by this feature. Data normalization can lessen this impact and raise the model's accuracy.

2, because there are individual outliers in this model (some styles are popular, the price is very high and sometimes even different from other styles). The performance of the model is impacted by these outliers. In this paper, the data is changed to the same range by data normalization method, which can make the model more robust and better deal with outliers.

The result after the normalization of the added data is shown in the Figure 6, Figure, Figure 8 and Figure 9, respectively.

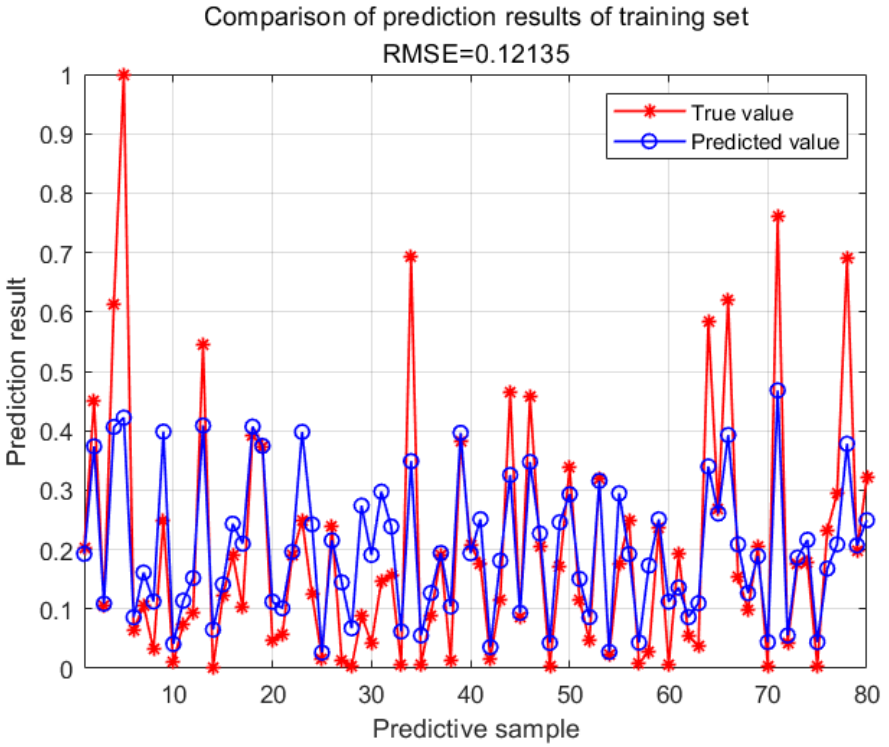

Figure 6: Results of the training set prediction after adding data normalization.

DOI: 10.54254/2754-1169/42/20232103

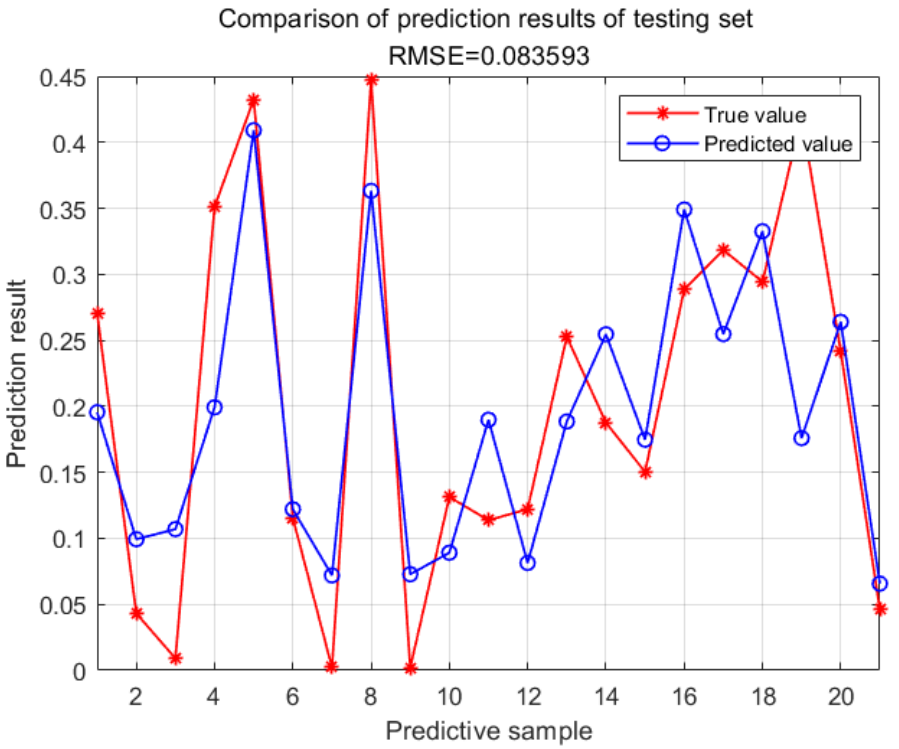

Figure 7: Test set prediction results after addition of normalization.

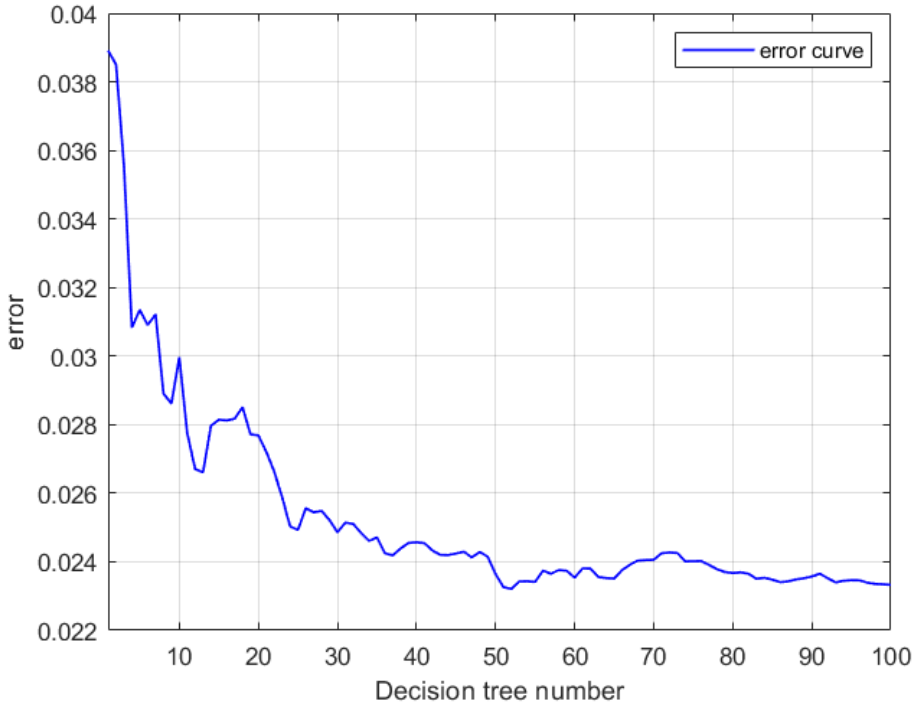

Figure 8: The relationship between the number of decision trees and the error after adding the normalized algorithm.

DOI: 10.54254/2754-1169/42/20232103

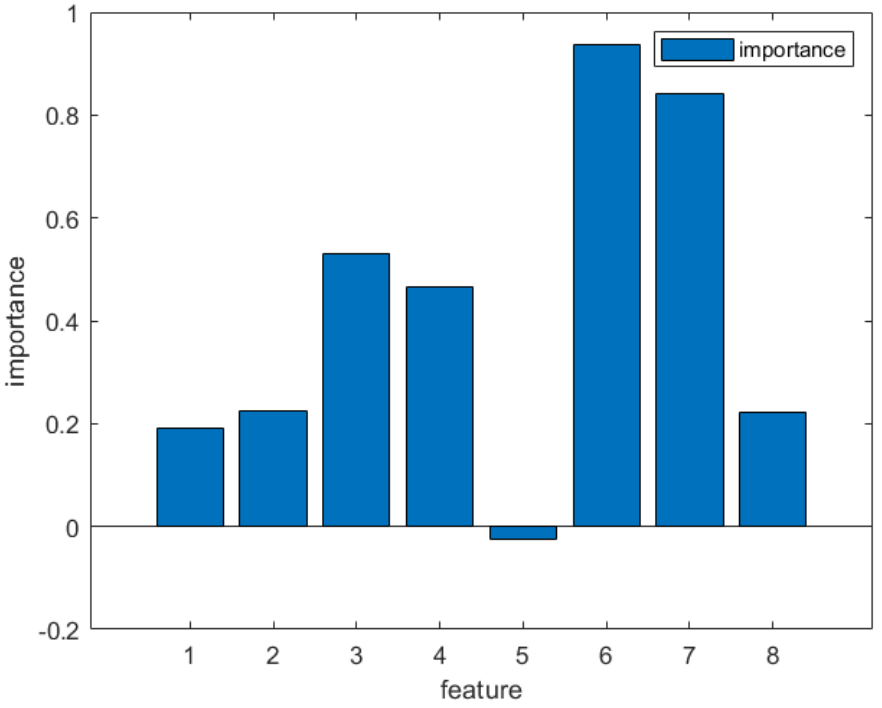

Figure 9: Importance of features after addition of normalization.

The figures above make it abundantly evident that the addition of normalized data processing considerably increased the forecast accuracy. The model has significantly improved its ability to predict outliers, but the prediction results of very popular styles and not very popular styles, that is, the prediction of extreme cases is not accurate. After data normalization, the error decreases faster with the number of decision trees, and gradually becomes stable after cliff-like decline. But the importance of features has changed dramatically. The No. 6 and No. 7 features a cliff-like lead in importance. The importance of number 6 and number 7 has decreased slightly, while the importance of number 3 has increased significantly.

3.5. Analysis

This study analyzes the model's correlation index and chooses R2, MAE, and MBE to further examine the model in order to better understand how well the model performed. The relevant indicators are defined as follows:

The R2 indicator, also known as the coefficient of determination, is a commonly used evaluation model for regression models, a statistic used to measure the model fit of the model and predict how good or bad the data is. The value of R2 ranges from 0 to 1. The accuracy of the model in predicting the data increases as the R2 index approaches 1, but the correlation between what is predicted and the actual value of the model for the data decreases as the R2 index approaches 0.

The calculation formula of R2 index is as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{4}$$

SS_{res} indicates the sum of residuals and SS_{tot} indicates the total sum of squares.

The residual sum of squares is calculated as follows:

DOI: 10.54254/2754-1169/42/20232103

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

The total sum of squares is calculated as follows:

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{6}$$

y_i represents actual value,$\hat{y}_i$ represents the predicted value of the model,$\bar{y}$ represents the actual average value, and n represents the sample size

In order to assess the model's accuracy, the MAE index, also known as mean absolute error, reflects the average absolute error between the model's projected value and the actual observed value.

The calculation formula of MAE index is as follows:

$$\text{MAE} = \frac{1}{n}\sum |y_i - \hat{y}_i| \tag{7}$$

n is the sample number, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

The average degree to which the projected value of the model deviates from the actual value is evaluated using the MBE index, also known as the mean deviation error. Positive MBE indicates that the model's anticipated value is high. MBE indicates the accuracy of the model's prediction when it is negative. If MBE is0, there is no variance between the model's predicted value and the observed value.

The calculation formula of MBE index is as follows:

$$\text{MBE} = \frac{1}{n}\sum (y_i - \hat{y}_i) \tag{8}$$

n is the sample number, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

The relevant indices are as follows in Table 1:

Table 1: Correlation index.

	R2	MAE	MBE
Training set	0.64824	0.079625	0.0016517
Testing set	0.64834	0.066568	-0.0085365

According to the definition of the above indicators and the calculated data, it can be seen that the constructed model has the ability to predict outliers to a certain extent. The model's forecast for the training set is a bit larger than that for the testing set, and it is more accurate when predicting the training set outcomes than the test set results. Simple predictions of popularity can be made, but the model still needs to be further adjusted and trained so that it can make more accurate predictions.

4. Conclusion and Future Study

According to the establishment and running results of the model, it can be determined that the model can perform simple prediction work for related problems. However, the prediction and processing of

DOI: 10.54254/2754-1169/42/20232103

outliers or variance values in this model are not precise enough and need to be strengthened. In the process of model building and training, this paper made statistics on the impact of each feature on the result (that is, the importance of the feature) and found that the shoe's foot feeling, sale time and sale price had a high impact on the popularity of the shoe. Manufacturers can invest in the research and development of shoe technology (such as BOOST, ZOOM, etc.) to increase the comfort of users. At the same time, the release time and pricing are adjusted in order to get a better market response.

Obviously, there are still many theoretical defects in the model, which will be solved in the follow-up research: the credibility of the outlier has a great impact on the model, some traders use some economic means and public heart, theory to hype a certain shoe, making it sought after by the public. For example: it is possible that the user experience of the shoes is not very ideal, but some speculators have monopolized the shoes, and have used means such as reducing the circulation on the market and increasing the unit price to enhance the popularity of the shoes in the crowd. The evaluation criteria are insufficient. Due to reasons such as time and knowledge reserve, this paper only finds 6 characteristics that may be related to popularity. The number of relevant features will be increased in subsequent studies.

References

[1] *Wang, J.P. (2013) Clothing digital media consumption behavior analysis and empirical research. MS thesis. Xi 'an Engineering University.*

[2] *Zou, F.Y. (2019) Research on clothing online consumption behavior of college students in Jiangxi Province. MS thesis. Nanchang University.*

[3] *Dodds, W. B. , Monroe, K. B. , and Grewal, D. (1991) Effects of price, brand, and store information on buyers' product evaluations. Journal of Marketing Research, 28(3), 307-319.*

[4] *Reichheld, F.F., and Phil S. (2000) E-loyalty: your secret weapon on the web. Harvard business review, 78(4), 105-113.*

[5] *Koufaris, M. (2002) Applying the technology acceptance model and flow theory to online consumer behavior. Information systems research, 13(2), 205-223.*

[6] *Corritore, C.L., Beverly K., and Susan W. (2003) On-line trust: concepts, evolving themes, a model. International journal of human-computer studies, 58(6), 737-758.*

[7] *Ahmed, R.R., et al. (2016) Empirical analysis of factors influencing the online shopping phenomenon: evidence from Pakistan. 3rd International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM.*

[8] *Qi, C., et al. (2021) Survey on Deep Learning Based Popularity Prediction. Journal of Chinese Information Processing, 35(2), 1-18,32.*

[9] *Chao, K. Q., and Ji, M. W. (2014) Predicting popularity of forum threads based on dynamic evolution. Journal of Software, 2767-2776.*

[10] *Huayun, X. , Zeyang, G. , Zhengjie, H. , Zheng, Z. , and Jinhao, S. (2018) Deep learning based topic popularity prediction. China Computer & Communication.*

[11] *Zohourian, A., Hedieh S., and Arefeh Yavary. (2018) Popularity prediction of images and videos on Instagram. 4th International Conference on Web Research (ICWR), IEEE.*

[12] *Gong, W. , Zheng, Z. , Gao, X. , and Chen, G. (2019) SATP: Sentiment Augmented Topic Popularity Prediction on Social Media. International conference on service-oriented computing, 575-577*

[13] *Zhang, Y.X., et al. (2020) App popularity prediction by incorporating time-varying hierarchical interactions. IEEE Transactions on Mobile Computing 21(5), 1566-1579.*

[14] *Wang, J.R., et al. (2018) Research on hot micro-blog forecast based on XGBOOST and random forest. Knowledge Science, Engineering and Management: 11th International Conference.*

[15] *Hara, T. , Uchiyama, M. , and Takahasi, S. E. (1998) A refinement of various mean inequalities. Journal of Inequalities & Applications, 1998(4), 387-395.*

[16] *Rigatti, S.J. (2017) Random forest. Journal of Insurance Medicine, 47(1), 31-39.*