An Estimation of the Pricing of Second-Hand Sailboats Based on the Random Forest Algorithm

Chengyuan Yang*[1], Si nan Tang [1], Jia hao Chen[1]

[*] Corresponding author: 1770573981@qq.com

[1] South China University of Technology

Abstract: The sailboat market is a complex market, with prices affected by multiple factors. For buyers, it is important to understand the background information behind the pricing of used sailboats. By understanding the impact of some common factors, buyers can better evaluate the actual value of the sailboat and make better purchasing decisions. Therefore, accurately predicting sailboat prices is crucial for developing pricing strategies and making buying and selling decisions. In response to problem one, the data is processed by using a crawler and collector to expand the dataset to 11 dimensions. The data is then checked for missing values and interpolated. Next, outlier detection is performed using the LOF algorithm, and data cleaning is done using high-dimensional mapping to find outlier data. Text data, such as the brand, is encoded, and logical data, such as the number of sails, is represented by boolean values. For regional factors, four economic indicators related to the region are directly used as replacements.

Keywords: Random Forest, Decision Tree, Lagrange Interpolation, Local Outlier Factor.

1. INTRODUCTION

With the rise of sailing, more and more people are buying their own sailboats, either for leisure or for racing. However, for those who want to buy a used sailboat, they may encounter some problems, such as how to buy a used sailboat cost-effective. The sailing market is a relatively niche market, so price fluctuations can be more dramatic than in other mass markets. The price of a sailboat of the same model can vary greatly from time to time and place to place. In addition, there are many factors to consider in the pricing of used sailboats, such as the boat's age, make, model, size, accessories, maintenance records and so on. Therefore, for buyers, how to predict the price of second-hand sailboats has become its core problem. It is very important for sailing agents to have accurate sailing price prediction results. If the prediction results are not reasonable, sailing ships will be purchased at a price higher than the market price, which will lead to a series of problems[1].

We analyze the relevant data, apply mathematical thought, and establish mathematical model to study the second-hand sailboat pricing.

(1) Relevant models need to be established to describe the relationship between sailing pricing and related factors.

(2) It is necessary to use the model established above to discuss regional effects [2].

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334481

(3) The model is applied to the Hong Kong market to study the regional effect of sailing price and its effect on mono and catamaran vessels.

(4) Provide sailing brokers with our pricing plan [3].

2. SYMBOL EXPLANATION

The primary notations used in this paper are listed in Table 1.

Table 1. Notations

Symbol	Description
MSE_i	The mean square of residual error of the generated data outside B bags
R^2	The goodness of fit of the model reflects how much percentage of the fluctuations of the explained
$score_i$	The importance score of characteristic variables
EVS	The variation of independent variable to explain the variance of dependent variable
SSR	The sum of regression squares
SST P(x)	The total sum of squares Lagrange interpolating polynomial

3. OUR WORK

3.1 Task 1

Task 1 requires us to study the relationship between sailboat listing prices and related factors based on existing data. Estimating the price of sailboats based on brand, model, length, etc [4]. is a regression problem, where the independent variables are factors such as brand, model, and length, and the dependent variable is the sailboat's price. However, we found that discussing only these factors may not fully measure the modeling [5], so we used various crawlers and collectors to process the raw data and data collected from various websites and used a random forest model to process it.

3.2 Task 2

Task 2 requires us to use our model to explain the regional factors and discuss whether the effects of regional factors on the prices of all sailboat models are consistent. Through the analysis of Task 1, we found the correlation between regional variables and pricing, and these regional variables were expressed using four types of regional economic indicators. If the correlation is strong, it indicates that the regional effect has a large impact [6], and if the correlation is weak, it indicates that the regional effect has a small impact.

3.3 Task 3

For the Hong Kong （SAR） region, we discuss Monohulled Sailboats and Catamarans separately. In the process of filling in the economic indices for the Hong Kong region and using our established model for prediction, we explore the regional impact of Hong Kong on the prices of the same type of ships in the subset, and whether the impact of the region on Monohulled Sailboats and Catamarans is the same [7].

3.4 Task 4

We summarize the conclusions and rules discovered in the data processing and modeling process, mainly reflected in the characteristics of data features, influencing factors, and the correlation between pricing and regions.

4. MODEL ESTABLISHMENT AND SOLUTION

4.1 A used sailboat pricing model based on Random Forest

Since the factors affecting the price of second-hand sailboats are nonlinear, a multivariate nonlinear model should be built. Random Forest has the characteristics of good nonlinear mapping, and has excellent accuracy among all current algorithms, which has great advantages compared with other algorithms.

4.1.1 Random Forest measures of the importance of characteristic variables

(1) In the used car valuation model based on random forest algorithm, the quantization of the importance of characteristic variables was mainly calculated by permutation, and the main calculation steps are as follows: When establishing the regression tree of random forest, set the mean square of residual error of the generated data outside B bags as： $\mathrm{MSE}_1, \mathrm{MSE}_2, \ldots, \mathrm{MSE}_\mathrm{B}$.

(2) Variables will be randomly replaced in B samples of out-of-pocket data generated, and then a new test sample set of out-of-pocket data will be generated. Then, the established random forest model will be used to regression the newly generated out-of-pocket sample set data [8]. Then, step (1) will be repeated to obtain the OOB residual mean square after random replacement, and the following matrix will be obtained:

$$\begin{bmatrix} MSE_{11} & MSE_{12} & \ldots & MSE_{1B} \\ MSE_{21} & MSE_{22} & \ldots & MSE_{2B} \\ MSE_{31} & MSE_{32} & \ldots & MSE_{3B} \\ \vdots & \vdots & \vdots & \vdots \\ MSE_{\rho 1} & MSE_{\rho 2} & \ldots & MSE_{\rho B} \end{bmatrix} \tag{1}$$

(3) Subtract the column vector$MSE_1, MSE_2, \ldots, MSE_B$ corresponding to the Equation 1 matrix, and divide the average by standard error, that is, the importance score of characteristic variables can be calculated.

$$Score_i = \left(\textstyle\sum_{j=1}^{B} \left(MSE_j - MSE_{ij}\right)/b\right)/S_E, (1 \leq i \leq p) \tag{2}$$

In the process of establishing the random forest, noise interference is added to each characteristic variable, and the accuracy of the random forest model is observed, and the characteristic variables are sorted by observing the change of model accuracy. If the noise of this variable is reduced, the accuracy of random forest is improved, indicating that this feature is of high importance to the prediction results. Random forest model, as a feature of automatic evaluation of the importance of variables, is not only targeted at sample set data with more characteristic variables, but also has a fast-processing speed and a high degree of automation [9].

4.1.2 Test of model

The purpose of establishing the second-hand sailboat model based on random forest algorithm is to carry on the regression analysis to the various influencing factors of second-hand sailboats, and the ultimate purpose of the regression is to forecast the price of second-hand sailboats.

For the test of the model, starting from the evaluation of the prediction effect of the total samples of the model, the better the overall prediction ability of the total samples is, the more reliable the trained model can be used in practical application. Therefore, it is necessary to ensure that the overall error of the samples is within a certain range, not too large.

To test the overall prediction effect of the model, the following indexes are adopted in this paper [10].

(1) R^2check

R^2, the goodness of fit of the model reflects how much percentage of the fluctuations of the explained variables can be described by the fluctuations of the explained variables. The higher the value of, the higher the fitting degree of the regression results of the model to the samples.

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \sum_{i=0}^{i=n-1} \frac{(y_i - \widehat{y_t})^2}{\sum_{i=0}^{i=n-1}(y_i - \bar{y})^2} \tag{3}$$

(2) Average absolute error MSE

The average of the absolute values of the difference between the predicted and true values. The average absolute error can evaluate the actual prediction error more accurately, and the smaller the MSE value, the better the fitting effect.

$$\frac{1}{n}\sum_{i=1}^{n-1} |(y_i - \hat{y}_i)| \tag{4}$$

(3) Variance score EVS

Reflecting the variation of independent variable to explain the variance of dependent variable, the larger the EVS value, the better the effect.

$$\mathrm{EVS} = 1 - \frac{\mathrm{Var}\{y - \hat{y}\}}{\mathrm{Var}\{y\}} \tag{5}$$

4.1.3 Model result analysis

(1) R^2check.The result ia in figure 1&2.

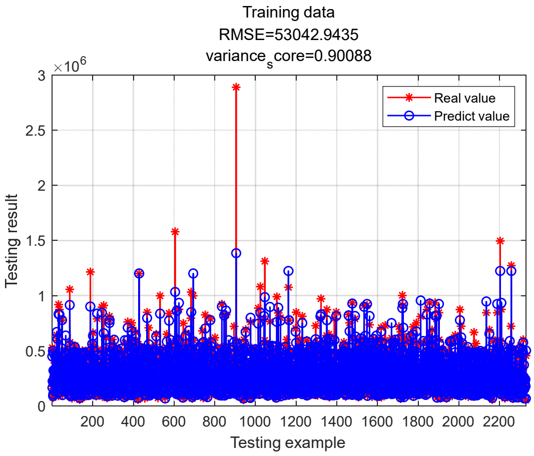

Figure 1. Study results1

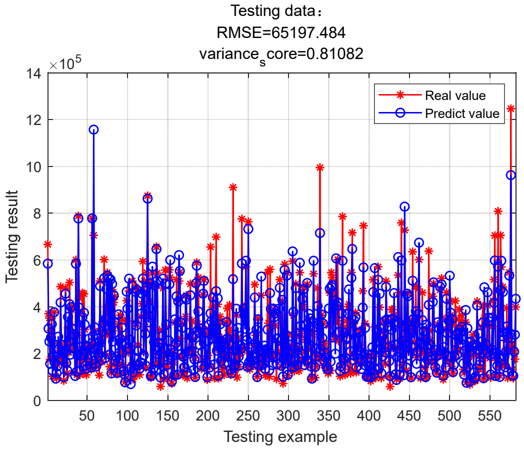

Figure 2. Study results2

The default parameter is set to the number of trees is 100 and the minimum number of cotyledons is 4. It can be concluded that the MSE in the test set is 104555.6237, the variance score is 0.71355, and the variance score and coefficient are close to 1, indicating a good fitting effect.

(2) Variance score EVS. The result ia in figure 3&4.

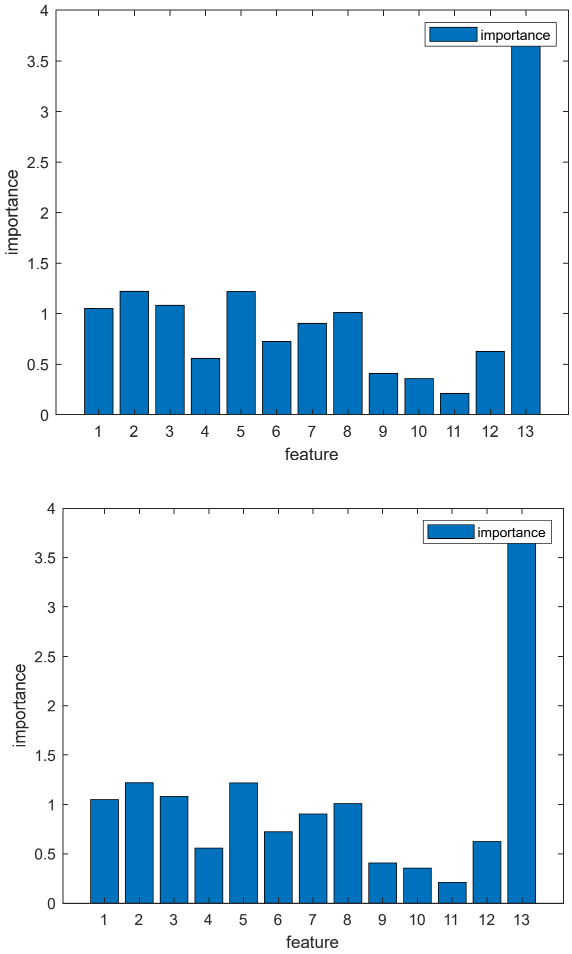

Figure 3. Study results3

A higher importance value means that changes in independent variables are more likely to explain changes in dependent variables. Therefore, it can be seen from Figure 3 that the EVS value of second-hand sailboats in the year is the largest, which has the greatest impact on the price and the greatest importance. We believe that the reason why the year is most important is that sailboats are expendable goods like cars. The longer the year, the aging rate and failure rate of sailboats will also increase, which is closely related to People's Daily use, so the most important degree.

In addition, as can be seen from Figure 3, it can be seen that the influence factor of single and double sailboats on price is the least obvious.

4.2 The effect of region on price

4.2.1 The effect of place on price

Figure 3 shows the importance analysis chart Specific gravity analysis is performed on the importance of the features obtained, and the formula is as follows:

Specific gravity analysis is performed on the importance of the features obtained, and the formula is as follows:

$$\mathrm{importance} = \mathrm{importance}/\mathrm{sum}(\mathrm{importance}, 2) \tag{6}$$

The following table 2 is obtained by the proportion of each data :

Table 2. Notations

feature	Importance
1	0.086477076
2	0.057636756
3	0.085264164
4	0.061211864
5	0.103974502
6	0.052494988
7	0.071256
8	0.061171486
9	0.054993311
10	0.039164626
11	0.05819159
12	0.043161403
13	0.225002274

The features corresponding to regions for 8, 9, and 10 in the table account for a total weight of 0.1553, indicating that the effect of regional factors on price cannot be ignored. Therefore, it can be concluded that regional factors have a certain impact on price.

4.2.2 Regional effects of different types of sailboats

All data for the same type of sailboat were selected, and the random forest algorithm was applied to different types of sailboats to calculate and compare regional weights. Due to space limitations, only part of the sailboats is shown below: The sailboat models are as follows in table 3:

Table 3. Notations

Bavaria Cruiser 46	Jenneau 53	Lagoon 42	Lagoon 400	Lagoon 440	Lagoon 450

The weight graphs for different boat models are shown below,.The result ia in figure 4.:

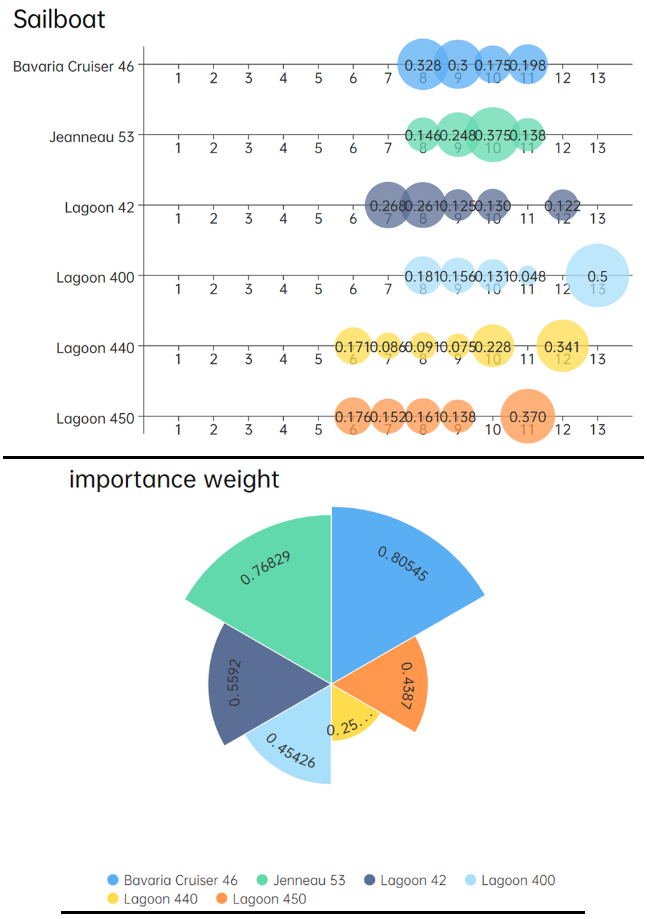

Figure 4. Study results 4

The following table 4 is summarized based on the above table:

Table 4. Notations

Make	**Bavaria Cruiser 46**	**Jenneau 53**	**Lagoon 42**	**Lagoon 400**	**Lagoon 440**	**Lagoon 450**
Proportion	0.80545	0.76829	0.5592	0.45426	0.25867	0.4387

The regional effects on different sailboat types are different.

Index	3,463,000	3,620	48111	8

5. CONCLUSION

Although there are many regression methods for constructing eigen price models, we believe that random forest is one of the most suitable methods, mainly for the following reasons:

(1) Random forest has many advantages that other models do not have. For example, random forest model does not have too many parameters to set, and it takes less time to compute. It can also maintain a high accuracy when processing data with very large sample sets.(2) Among numerous machine learning methods, in a comparative study with support vector machine, boosting and neural network, random forest usually got better results.(3) Able to successfully process categorical variables with multiple categories. For example, in parametric regression or neural networks, many qualitative variables will lead to a large increase in the number of estimated parameters, which usually leads to overfitting of regression results. (4) Because the random forest adopts bagging method to build trees, the influence of outliers on the results will be weakened. This method measures the importance of each variable by comparing the average marginal reduction in the sum of the squares of the residuals for each explanatory variable.

REFERENCES

[1] Ran Lian, Zeng Tianyang. Network clustering behavior in the context of "Internet+Big Data": generation mechanism and governance strategy[J/OL]. Journal of Intelligence:1-8[2023-04-19]. http://kns.cnki.net/kcms/detail/61.1167.g3.20230207.1347.002.html

[2]Chia Chenhang,Chen Gang. The regulation of big data investigation algorithms: the idea of procuratorial supervision [J]. Journal of the Chinese People's Public Security University (Social Science Edition), 2022,38(06):67-73.

[3] Zeng Tuo. Research on big data algorithm for distribution network maintenance based on image data support[J]. Popular Electricity,2022,37(10):57-58.

[4] Li Weihong, Yao Xiaolin, Zhang Ruiqi. Research on optimization of accounts receivable management based on RPA and big data algorithm[J]. China Collective Economy,2022(28):131-133.

[5] Li Xuemei, Ma Wenhui, Zhang Chunqing, Wang Xiuqing. Analysis of cloud computing big data algorithms in network systems[J]. China Science and Technology Information,2022(09):80-82.

[6] Huang Yi, Song Ziyin. The legal regulation of "algorithmic killing" in the context of big data[J]. Zhongzhou Journal, 2022(04):50-54.

[7] Peng Li. Research on monitoring the effectiveness of ideological discourse under the perspective of big data algorithm [J]. Henan Social Science,2022,30(04):17-24.

[8] Volt. Let big data algorithms play more positive energy [J]. China Press,2022(06):7. DOI:10.13854/j.cnki.cni.2022.06.063.

[9] Peng, L. Q., Li, L.. A pilot study on the free relationship between human and algorithm [J]. Journal of Social Sciences, Hunan Normal University, 2022,51(02):20-27.DOI:10.19503/j.cnki.1000-2529.2022.02.003.

[10] Cui Yaomin. Research on the paradigm of sports science research in the era of big data algorithm [C]//Chinese Society of Sports Science. Abstracts of the 12th National Conference on Sports Science - Poster Exchange (School Sports Section). Compendium of abstracts from the 12th National Conference on Sports Science - Poster Exchange (School Sports Section), 2022:4-5. DOI:10.26914/c.cnkihy.2022.010213.