Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.071001
ISSN 2371-8838 Vol. 7 Num. 10

A Study on the Transaction Price of Second-hand Sailboats Based on the Random Forest Regression Model

Xinyue Zhang*, Xuanyi Xiang

Sichuan University-Pittsburgh Institute, Chengdu, 610044, China
**Corresponding author: 2021141520222@stu.scu.edu.cn*

Keywords: Second-hand Sailboat Market, Price Prediction, Random Forest Regression Model, Decision Tree Regression Model, Supporting Vector Machine Regression Model

Abstract: The second-hand sailboat market is booming but the prices are uncertain, which poses a significant challenge for sellers to determine the optimal selling price. To address this issue, this study employed three regression models, namely Random Forest Regression Model, Decision Tree Regression Model, and Supporting Vector Machine Regression Model, to explore the main factors affecting the pricing of second-hand sailboats, and predict the prices of second-hand sailboats. The result shows that the length of the second-hand sailboats impacts the most and the Random Forest Regression Model has the highest accuracy in predicting the transaction prices of second-hand sailboats. This prediction method can help sellers better price their boats and promote the development of the second-hand sailboat market.

1. Introduction

In recent years, sailing has become an increasingly popular recreational activity, with thousands of individuals and families taking to the water each weekend. As a result, demand for quality second-hand sailboats has increased, as budget-conscious buyers seek out affordable options[1].

However, the pricing of these sailboats is highly complex and uncertain, with numerous factors affecting their value. In order to solve this problem, a lot of researchers in the world have carried out research. Eleftherios Ioannis Thalassinos and Evangelos Politis provide an evaluation model for the pricing of used bulk carriers. Their valuation process relies on cash flow analysis and methodology[2]. Besides, Andreas estimated the shadow price of the most relevant determinant in the second-hand ship hedonic price model through the analysis of the second-hand ship buying and selling fixtures[3]. Despite the efforts of several researchers, a comprehensive understanding of the pricing trend and variation in the used sailing boat market has remained elusive.

To address this problem, the paper aims to develop a prediction model for second-hand sailboat prices through regression analysis to improve the accuracy of price prediction. Three regression models, namely Random Forest Regression Model[4], Decision Tree Regression Model[5], and Supporting Vector Machine Regression Model[6], will be employed to predict the prices of second-hand sailboats and compare them with the actual transaction prices. These three models are used to identify the main factors affecting the pricing of second-hand sailing boats and determine the superiority of one regression model over the others in predicting the transaction prices of second-

hand sailboats. It is expected that this paper will contribute to helping sellers better price their boats and promote the development of the second-hand sailboat market.

2. Sailboat Price Model Construction

This paper extracts transaction data for second-hand sailboats in 2019 from the authoritative boat website boats and categorizes the data by Make and Geographic Region. Every transaction message records information about the sailboat, such as Make, Variant, Length (ft), Geographic Region, Country/Region/State, Listing Price (USD), Year, Make Variant, LWL (ft), Beam (ft), Draft (ft), Displacement (lbs), Sail Area (sq ft).

Considering the uncertainty of second-hand sailboat prices, this paper adopts three methods to predict the price of sailboats, hoping to improve the accuracy of the prediction results as much as possible and provide better pricing suggestions for sellers.

2.1 Data Preprocessing

For the data-analysis problem, it is found that a large amount of raw data contained some incomplete and abnormal values, which could significantly affect the efficiency of modeling and the accuracy of conclusions. Therefore, it is crucial to preprocess the data.

Three methods are used to process the data loss and abnormal values. First, missing data are identified. The ways missing data is processed are: (1) For variables with a large amount of missing data, they are directly deleted. (2) For variables with a small amount of missing data, an interpolation method is used to compensate for the data. After screening out the 3500 sets of data, 3017 sets of valid data are left.

2.2 Data Analysis

2.2.1 Variance Test

The variance test is used to select features by calculating the variance of the features, usually by setting a variance threshold, and considering deleting features that do not reach the variance threshold[7].

Table 1: Variance of each catamaran variable

Variable Name	Variance	Variable Name	Variance
LWL	16.0172	Sail Area	8.3085×10^{15}
Length	20.8511	Year	18.8319
Beam	23.6266	ACT	3.3448×10^{15}
Displacement	9.8438×10^{7}	Draft	1.9844

A feature can be considered as contributing significantly to the differentiation of the sample if it has a largely different value across the data set, as shown in Table 1.

From the calculation results, the variable with the smallest variance is Draft, with a variance of 1.9844. The variance is not completely close to 0, which means that the sample value still contributes.

2.2.2 Spearman Test

The Spearman Test is used to calculate the correlation between the independent variables with the correlation coefficients (P_s)[8]. The formula of the Spearman Test is in the formula (1):

$$P_s = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2 \sum_{i=1}^{n}(S_i - \bar{S})^2}} \quad (1)$$

These show the correlation strength between R and S, where P_s is the correlation coefficient of Spearman. And the closer to 1, the correlation strength between R and S is stronger.

Next, the Spearman test is used to calculate the correlation between the independent variables. Then, the testing result is in the thermodynamic diagram in Figure 1.

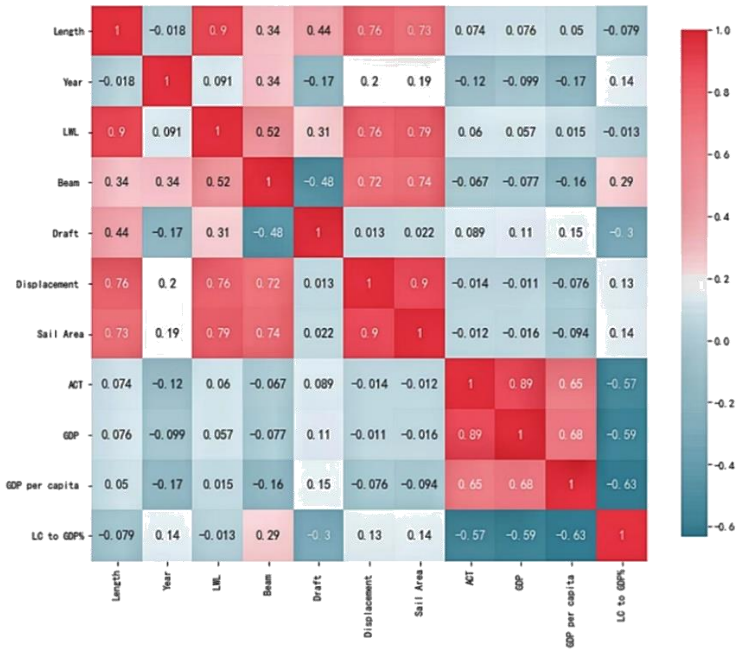

Figure 1: Heat map of correlation coefficient of independent variables

From Figure 1, it can be seen that the correlation coefficient between length and LWL is 0.9, which has a high linear correlation between the two. In fact, a longer sailboat has a longer length of waterline, and the length of the waterline is a representation of length. Hence, in order to avoid the adverse effects of multicollinearity on the model, the variable LWL is eliminated. Finally, the remained features are shown in Table 2.

Table 2: Characteristic variable remained

Variable Name	Units	Variable Name	Units
Length	ft	Sail Area	ft^2
Beam	ft	ACT	tons
Draft	ft	Year	yr
Displacement	lbs		

2.3 Sailboats Price of Random Forest Regression Model (SPRF Model)

The machine forest algorithm is a combination of the Bagging algorithm method and Random Subspace algorithm, and the basic building block is a combination of decision trees (either binomial or multinomial tree)[4].

The initial values are:

$$P = \{(x_{i1}, x_{i2}, \ldots, x_{iM}, y_i)\}_{i=1}^{n} \quad (2)$$

Where n is the number of samples, M is the number of features.

The random forest algorithm consists of two “random” processes: the first “random” process is to randomly generate a training set, with the aim of using the training set to complete the training of the model; The second "random" process is the random selection of a subset of features, which are calculated to select the best-split feature attributes[9].

2.4 Sailboat Price of Decision Tree Regression Model (SPDT Model)

Decision Tree is a decision analysis method based on the known probability of occurrence of various situations, to evaluate the project risk and judge its feasibility[10]. To calculate the information gain g(D,A) of feature A on the data training set D:

$$g(D, A) = H(D) - H(D|A) \tag{3}$$

Where the empirical entropy H(D) of data set D is:

$$H(D) = -\sum_{k=1}^{k} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \tag{4}$$

The empirical condition entropy H of feature A on data set H(D|A) is:

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|} \tag{5}$$

In this problem, it is used to decide tree model to analyze the prices of second-hand sailboats, which is applied in sequential decision-making, to take maximum benefit expectation value or minimum expectation cost as the decision criterion.

2.5 Sailboat Price of Supporting Vector Machine Model (SPSVM Model)

Given the linear data (x_i, y_i) to fit the function class (Φ):

$$\Phi = \text{span}\{\phi_1(x), \phi_2(x), \cdots, \phi_n(x)\}$$
$$= \{\phi(x) = a_1\phi_1(x) + a_2\phi_2(x) + \cdots + a_n\phi_n(x), \quad a_i \in R, i = 1,2, \cdots, n\} \tag{6}$$

where $\phi_1(x), \phi_2(x), \cdots, \phi_n(x)$ is linearly independent.

Then, the deviation of $\phi(x)$ on x_i is:

$$\zeta_i = \phi(x_i) - y_i, \quad (i = 1,2,3, \cdots, n) \tag{7}$$

To optimize the linear programming, this paper chooses the support vector machine regression model (SVM). A Support Vector Machine (SVM) is a generalized linear classifier that classifies data binary according to supervised learning. Its decision boundary is the maximum-margin hyperplane for solving the learning sample. It is generally used for classification tasks, and support vector regression (SVR) is a variant of SVM in regression analysis[8].

3. Results

The results show that the three models have made accurate predictions on the second-hand sailboat prices.

In order to evaluate the model prediction effect, there are the model evaluation indicators introduced:

(1) Mean absolute error:

The mean absolute error is the average of the absolute values of the deviations of all individual observations from the arithmetic mean, also known as MAE:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \quad (8)$$

The smaller the MAE, the better the model prediction; conversely, the larger the MAE, the worse the model prediction.

(2) R-squared (correlation coefficient):

$$R^2 = \frac{\sum_{i=1}^{N}\left(\hat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \quad (9)$$

Where y is the actual value, $\overline{y}$ is the mean value, and $\hat{y}_i$ is the fitting value.

$R^2 \in [0,1]$. The closer the value R^2 is to 1, the better the model prediction;

3.1 Result of SPDT Model

The final result of the SPDT Model is in Table 3.

Table 3: Result of SPDT Model

	R-squared	MAE
Training set data	0.87494	0.0920
Testing set data	0.82457	0.1529

The performance of the Decision Tree Regression Model for the training set obtained from feature selection is better, and the test data set is visualized as shown in Figure 2.

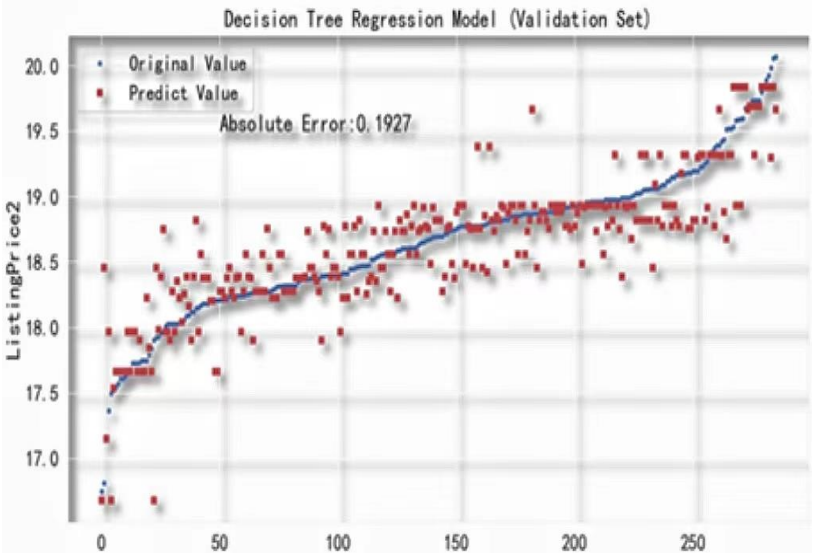

Figure 2: The validation set of the decision tree regression model

Figure 2 presents the validation set of the Decision Tree Regression Model which is applied to predict the listing price of used sailboats. The model is trained and tuned to minimize the mean absolute error (MAE) of the predicted prices, which is achieved through a series of iterations. The low MAE value of 0.1927 indicates that the model is able to predict the listing price of sailboats with reasonable accuracy, i.e., it computes the difference between the actual and predicted price to be within an acceptable range. The low error rate also suggests that the model is not overfitted, which means that it can generalize well to unseen data. The model's good performance on the validation set further supports its ability to accurately predict the listing price of sailboats.

3.2 Result of the SPSVM Model

R-squared and MAE of the SPSVM Model are calculated. The results are shown in Table 4.

Table 4: Result of the SPSVM Model

	R-squared	MAE
Training set data	0.81946	0.1595
Testing set data	0.80454	0.1927

By machine learning, the Support Vector Machine Regression Model is visualized graphically in Figure 3.

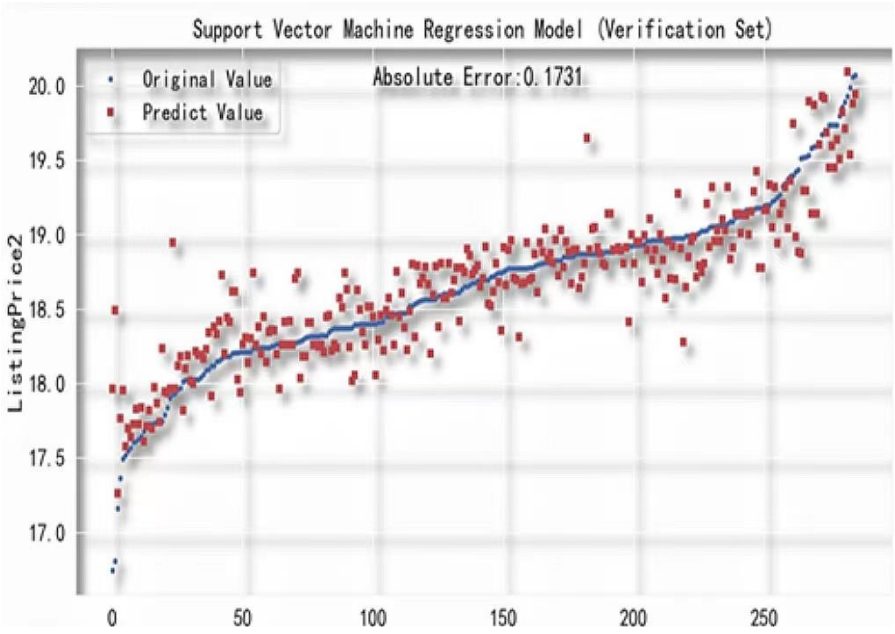

Figure 3: The validation set of the support vector machine regression model

Figure 3 shows the validation set of the Support Vector Machine Regression Model in the market of second-hand sailboats. With the low mean absolute error (MAE) value, 0.1731, which is smaller than that used in the Decision Tree Regression Model, it shows that the difference between the actual and predicted value is also within a reasonable range. Through a series of iterations, the results show that the model can be used to predict the price of second-hand sailboats, comparing the actual and predicted listing prices. The reasonable assessing fit means that the Support Vector Machine Regression Model could predict the unknown data well.

3.3 Result of SPRF Model

Also, the R-squared and MAE of the SPRF Model are calculated. The results are shown in Table 5.

Table 5: Results of the SPRF Model

	R-squared	MAE
Training set data	0.82184	0.1451
Testing set data	0.81672	0.1731

The Random Forest Regression Model is visualized graphically in Figure 4.

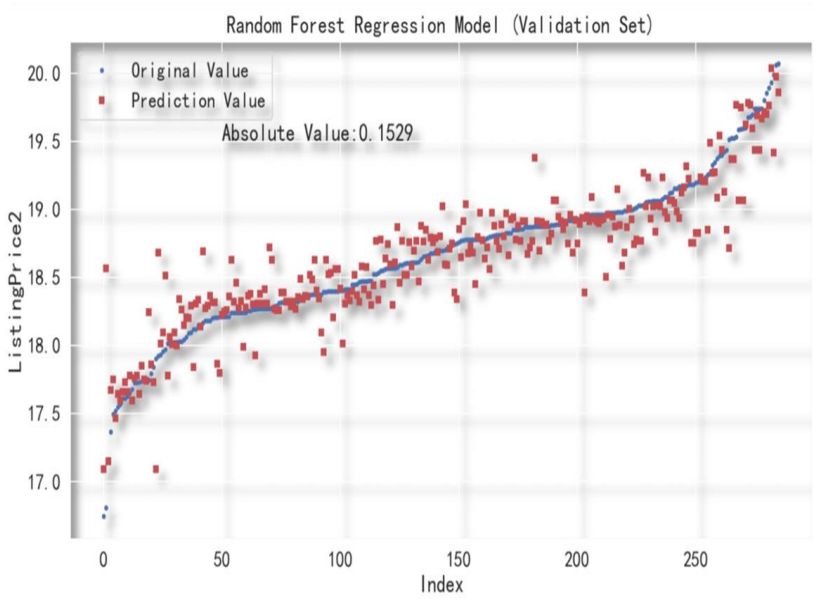

Figure 4: The validation set of the random forest regression model

Figure 4 shows the validation set of the Random Forest Regression Model to predict the listing price of second-hand sailboats. From Figure 4, the mean absolute error (MAE) is 0.1529 which is the lowest value within the models above. In other words, there is not much difference between the predicted value (blue points) and the true value (red points), which predicts well. Compared with the first two models, the MAE of the SPRF model is lower, indicating that this model has a higher degree of fitting and can more accurately shorten the gap between the predicted and the actual listing price of the second-hand sailboats. This helps sellers better predict the listing price of second-hand sailboats.

3.4 Comparative analysis of model results

The prediction results are summarized obtained from the three model training sets and test sets, and obtained in Table 6.

Table 6: The summary of SPRF, SPDT, SPSVM models

	R-squared	MAE
Training SPRF Model	0.87494	0.0920
Testing SPRF Model	0.82457	0.1529
Training SPDT Model	0.81946	0.1595
Testing SPDT Model	0.80454	0.1927
Training SPSVM Model	0.82184	0.1451
Testing SPSVM Model	0.81672	0.1731

Next, obtain the important variable features in the random regression model, and the results through programming analysis in Table 7.

Table 7: Three important variable features in SPRF model

	Feature	Importance
1	Length	0.598819
2	Year	0.290287
3	Make Variant	0.030578

It is not difficult to see from Table 7 that the characteristics that have a greater impact on the pricing

of second-hand sailboats are the length and year of production of second-hand sailboats.

These three main influencing factors are applied to the random forest regression model, and the predicted visual result is shown in Figure 5.

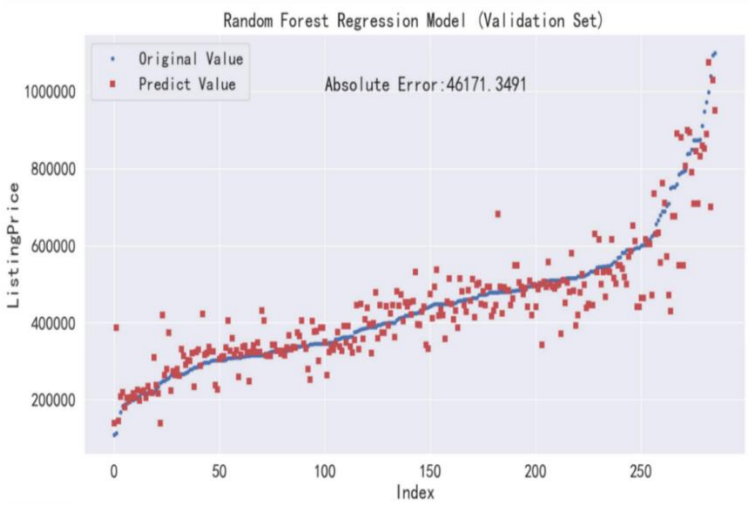

Figure 5: The prediction of the listing price

Finally, the predicted second-hand sailboat prices generated by the random forest regression model are compared with the data obtained from the boats' website. Subsequently, the relative error between each set of predicted data and the actual data will be calculated and plotted as a bar chart in Figure 6.

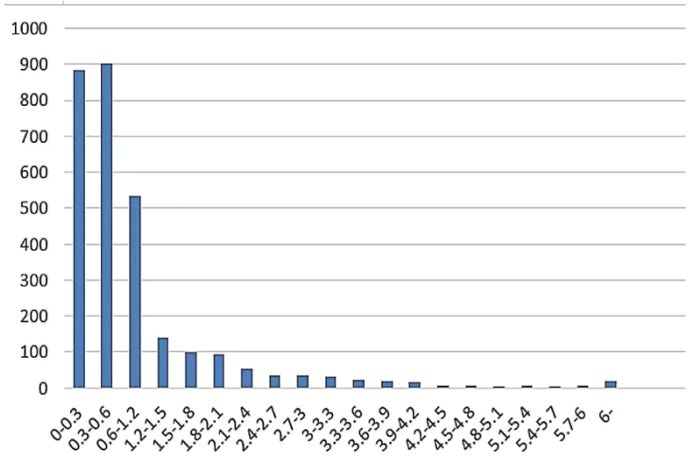

Figure 6: The error of the predicted value

It is not difficult to see from Figure 6 that most of the relative errors of the forecast are within the range of 0~1.2%, which indicates that the accuracy of our model is very high. From the aspect of absolute error, it is found that the average absolute error between the price predicted by the random forest regression model and the real price is 0.1529. These two error values prove that the random forest regression model is highly accurate in predicting the price of second-hand sailboats, thus providing very valuable suggestions for sellers in pricing.

4. Conclusion

By using the random forest regression model, decision tree regression model, and support vector machine regression model to predict the second-hand sailboat price, and comparing it with the actual transaction price, this paper puts forward an effective method to solve the challenge of the uncertain

market price of second-hand sailboat. The results show that the decision tree regression model has the highest accuracy in predicting the second-hand sailboat transaction price. This forecasting method is helpful for sellers to set reasonable prices, improve the success rate of second-hand sailing transactions, and promote the development of the second-hand sailing market.

References

[1] Liu, Xiaolin. Data Analysis and Research Based on Second-hand Sailboats. Frontiers in Business, Economics and Management.8.3 (2023): 234-241.

[2] Thalassinos, Eleftherios Ioannis, Evangelos Politis. Valuation Model for a Second-Hand Vessel: Econometric Analysis of the Dry Bulk Sector. Journal of Global Business & Technology 10(2014).

[3] Peng Z , Huang Q , Han Y .Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm[C]//2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT).IEEE, 2019.DOI:10.1109/ICAIT.2019.8935894.

[4] Abdulkareem, Nasiba Mahdi, Adnan Mohsin Abdulazeez. Machine learning classification based on Radom Forest Algorithm: A review. International journal of science and business 5.2 (2021): 128-142.

[5] Myles, Anthony J., Robert N. Feudale, Yang Liu, Nathaniel A. Woody, Steven D. Brown. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society 18.6 (2004): 275-285.

[6] Widodo, Achmad, Bo-Suk Yang. Support vector machine in machine condition monitoring and fault diagnosis. Mechanical systems and signal processing 21.6 (2007): 2560-2574.

[7] Schluter, Dolph. A variance test for detecting species associations, with some example applications. Ecology 65.3 (1984): 998-1005.

[8] Myers, Leann, Maria J. Sirois. Spearman correlation coefficients, differences between. Encyclopedia of Statistical Sciences 12 (2004).

[9] Xiang Z , Wen Y , Hanqing L I ,et al.Research on the Spatial and Temporal Evolution Characteristics of the Price of Second-hand Housing in Beijing[J].Journal of Geo-Information Science[2023-11-16].

[10] Xu, Min, Pakorn Watanachaturaporn, Pramod K. Varshney. Decision tree regression for soft classification of remote sensing data. Remote Sensing of Environment 97.3 (2005): 322-336.