Research on Second-Hand Sailboat Prices Based on Machine Learning

Tingting Chen*, #, Chengrui Yuan#, Minghao Wang#

College of Applied Mathematics, Chengdu University of Information Technology, Chengdu, China, 610225

#These authors contributed equally to this work.

*Corresponding author: yilantingting@qq.com

Abstract: With the development of the international shipping market and changes in second-hand ship prices, the operation and trade of second-hand ships are quite active. To accurately evaluate the price of second-hand ships, it is particularly important to establish a pricing model for second-hand ships. The first part of this article uses X-GBoost and PCA to rank and reduce the importance of features, and obtains four principal components. Based on 5 indicators and their importance ratios, establish an optimized X-GBoost regression model for Sine Cosine Algorithm (SCA) to predict the price of second-hand ships. In order to further investigate the impact of regional characteristics on prices, local population density, per capita GDP, and cargo throughput were selected to further classify regional characteristics. A linear regression model was established to investigate the impact of population density, per capita GDP, and cargo throughput on regional error prices, and to explore the impact of various regional characteristics on error prices. Divide sailboats into three categories – low-end sailboats, mid-range sailboats, and high-end sailboats, and discuss the impact of regional characteristics on prices. The study found that there is a significant difference in the prices of low-end sailboats and mid-range sailboats between regions, indicating consistency in regional effects. Studying the pricing of second-hand sailboats can help to correctly select trading markets, develop marketing strategies, purchase ships with higher economic priority, and promote the development and prosperity of the sailing industry.

Keywords: X-GBoost, Sine Cosine Algorithm, Multiple Linear Regression, Principal Component Analysis.

1. Introduction

In recent years, the second-hand ship trading market has been booming. In 2021, the global second-hand ship trading volume reached the highest record in Clarkson Research, reaching 2479 ships, totaling 140 million deadweight tons. Since 2022, the second-hand ship price index has risen to 213.3, an increase of 115% compared to the beginning of 2021 [1].

To accurately evaluate the price of second-hand ships, Li Jing et al. (2021) applied a grey neural network to the pricing of second-hand sailboats [2]. Peng et al. (2022) proposed a second-hand sailboat pricing method based on interval regression model and adaptive neural fuzzy inference system [3].

In most previous studies, although neural network-based models were used to ensure the accuracy of model research, the robustness of the model and the accuracy of feature selection were ignored. Therefore, this article made improvements based on the above shortcomings, optimized the feature extraction method, weakened the model's dependence on indicators, selected a model with high stability and consistency, and used SCA sine cosine optimization to adjust the parameters of the model, selecting the parameters with the highest portability and accuracy.

Our work is shown in Figure 1.

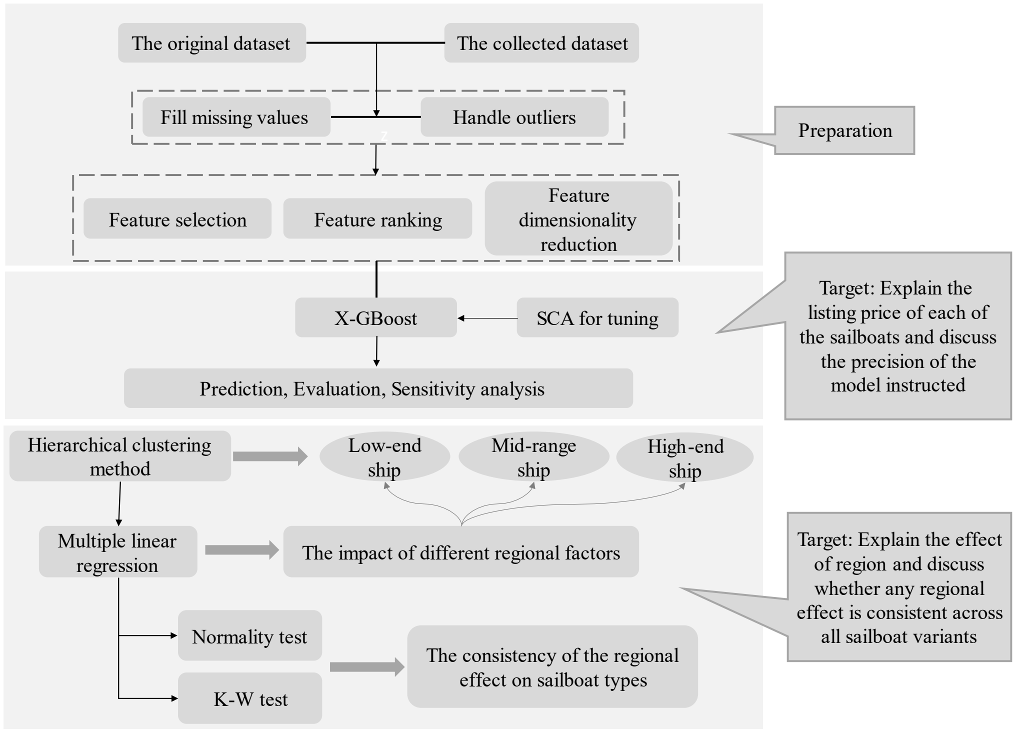


Figure 1. Our work

2. Feature processing

Using the X-GBoost algorithm to obtain the importance of 13 data indicators and sort them, based on feature filtering, we selected 4 dimensionality reduction indicators function, region, structure, and size for PCA dimensionality reduction. The specific steps are as follows: [4, 5]

Step 1 Mean normalization.

Step 2 Calculate the covariance matrix.

Step 3 Calculate eigenvalues and eigenvectors.

Step 4 Select Principal Component.

Step 5 Construct a new feature space.

The four principal components obtained after dimensionality reduction and their weights are shown in Figure 2.

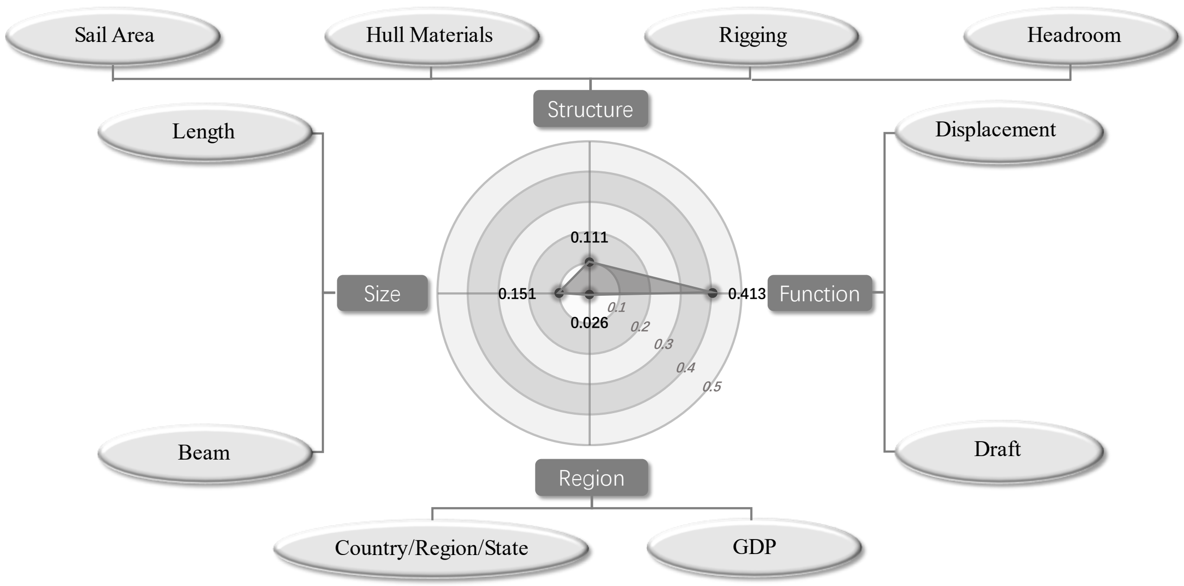


Figure 2. Feature indicators after dimensionality reduction

Based on the results of PCA dimensionality reduction, five indicators were selected for modeling, as shown in Table 1.

Table 1. Parameters and feature importance

Parameter	Feature importance
Function	41.30%
Year	30%
Size	15.10%
Materials	11.10%
Region	2.60%

3. Price prediction based on X-GBoost model optimized by Sine Cosine Algorithm

3.1. The establishment of X-GBoost model

The principle of X-GBoost generating multiple trees is to use the boosting idea in ensemble learning methods to gradually improve the prediction accuracy of the model by iteratively generating decision trees one by one. Specifically, during each iteration, the model will weigh the data based on the prediction results of all previous trees, and then train a new tree to correct the prediction error of the previous model.

The CART decision tree is used to provide the accuracy of the prediction model, and the final prediction value is obtained by summing the prediction results of the decision tree obtained from each round of training. The objective function of its t-th CART decision tree is defined as follows:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, y_i^{'(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{1}$$

Among them:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^{2} \tag{2}$$

3.2. Model optimization - Sine Cosine Algorithm

This article adopts Sine Cosine Algorithm to improve the convergence speed and performance of the model by adjusting the learning rate. These techniques can help avoid overfitting and improve the generalization ability of the model, so as to achieve better prediction results. [6]

$$\mathrm{X_i^{t+1}} = \begin{cases} \mathrm{X_i^t + r_1 * \sin(r_2) * |r_3 P_i^t - X_i^t|} & \mathrm{r_4 < 0.5} \\ \mathrm{X_i^t + r_1 * \cos(r_2) * |r_3 P_i^t - X_i^t|} & \mathrm{r_4 > 0.5} \end{cases} \tag{3}$$

The following are the optimization results of the X-GBoost model using the sine cosine algorithm:

(1) maxiter: The maximum number of iterations per decision tree is 31.

(2)depth_ Max: maximum depth, with a maximum depth of 8 for each tree.

(3)min_ Child: The minimum weight of the leaf node is 1.

In addition, the number of trees was specified as 100. Through sine cosine optimization, the values of these parameters can be automatically adjusted during the training process to improve the performance and convergence speed of the model. The final model will consist of 100 decision trees, each containing a maximum of 8 split nodes, with at least one sample on each leaf node.

3.3. Model evaluation

The model was evaluated using adjusted R-square, and the results are shown in Table 2. [7]

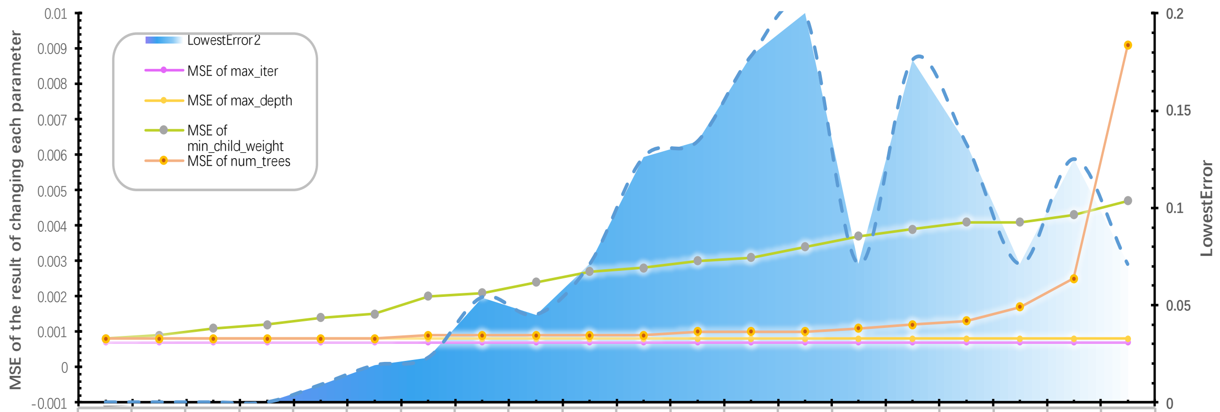
Table 2. Model evaluation

R^2_{adjusted}	MSE	MAE	RMSE
0.9741	0.1	0.01	0.014

It is easy to see that the adjustment determination coefficient R2 of X-GBoost reaches 0.9741, which means that the Goodness of fit of the model is very good; MSE is 0.0, MAE is 0.01, and RMSE is 0.014, which is the most ideal.

3.4. Sensitivity analysis

This article conducts sensitivity analysis by changing the four parameter values of the X-GBoost model. The sensitivity analysis of the X-GBoost is shown as follows in Figure 3.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
max_iter	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
max_depth	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
min_child_weight	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
num_trees	100	95	90	85	80	75	70	65	60	55	50	45	40	35	30	25	20	15	10	5

Figure 3. Sensitivity analysis of X-GBoost parameter tuning

From the figure, it can be seen that when only the value of max_iter is changed or num_trees, while keeping other parameters constant, the error of the model remains unchanged. As is, the accuracy of the train model is changed by max_depth and min_child_weight.

4. Regional effects based on multiple linear regression models

As can be seen from the previous text, the proportion of regional factors to feature importance is 0.026, so the region has an impact on the price of sailboats. To further explore how regional factors affect sailboat prices, this article establishes a multiple linear regression model to explore. Query the data to obtain the 2019 GDP (per) capital and average sailboat price for the three regions of USA, Europe, and Caribbean, as shown in Figure 4.

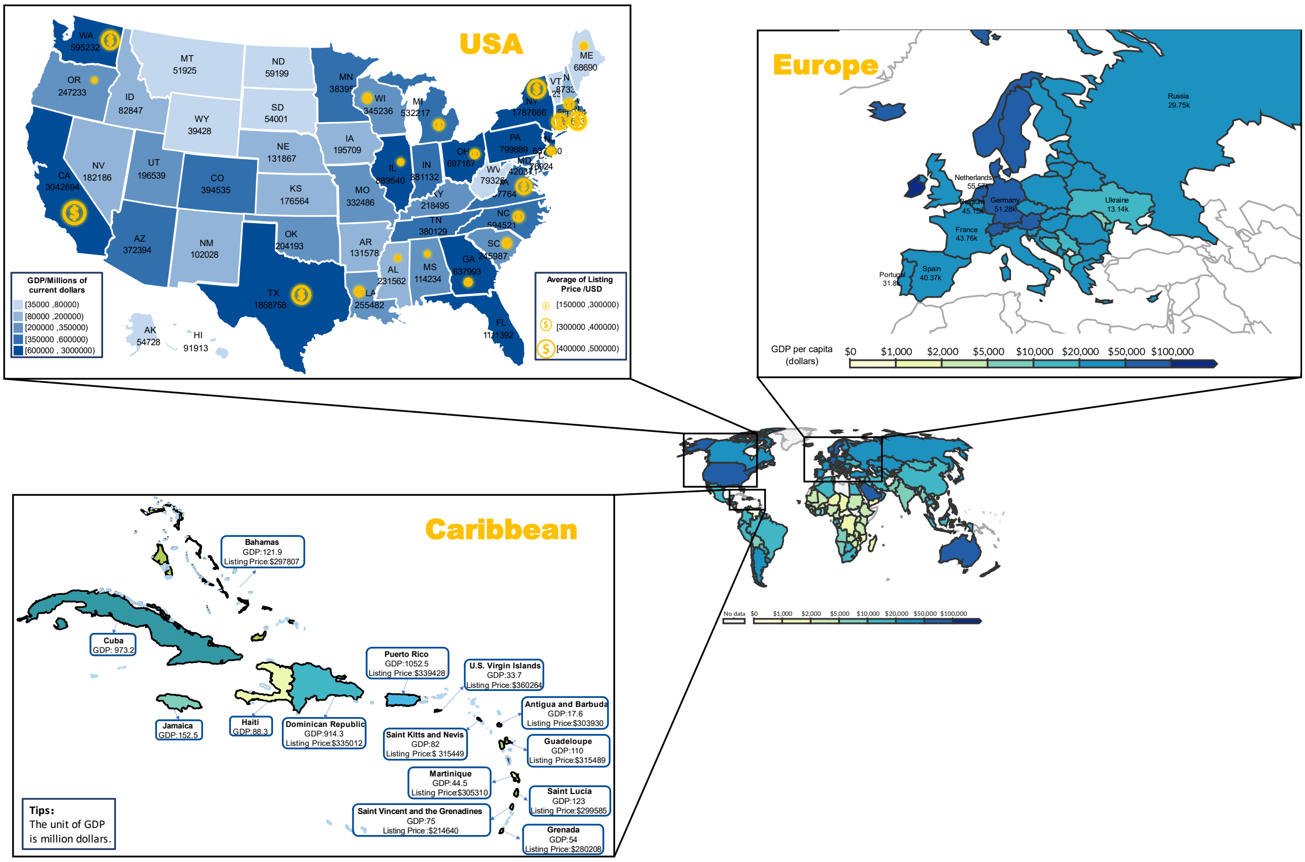


Figure 4. GDP (per) capita and average sailboat price in different regions

4.1. Hierarchical clustering method

To minimize the impact of other indicators on the model, all factors that do not include regional indicators are clustered and the sailboat is divided into three designated clusters: affordable sailboat, mid range sailboat, and luxury sailboat. The following steps are taken: [8]

Step 1 Initialization.

Step 2 Class hierarchical clustering - calculate distance matrix.

Step 3 Merge the nearest cluster.

Step 4 Update the distance matrix.

Visualize the clustering results, as shown in Figure 5.

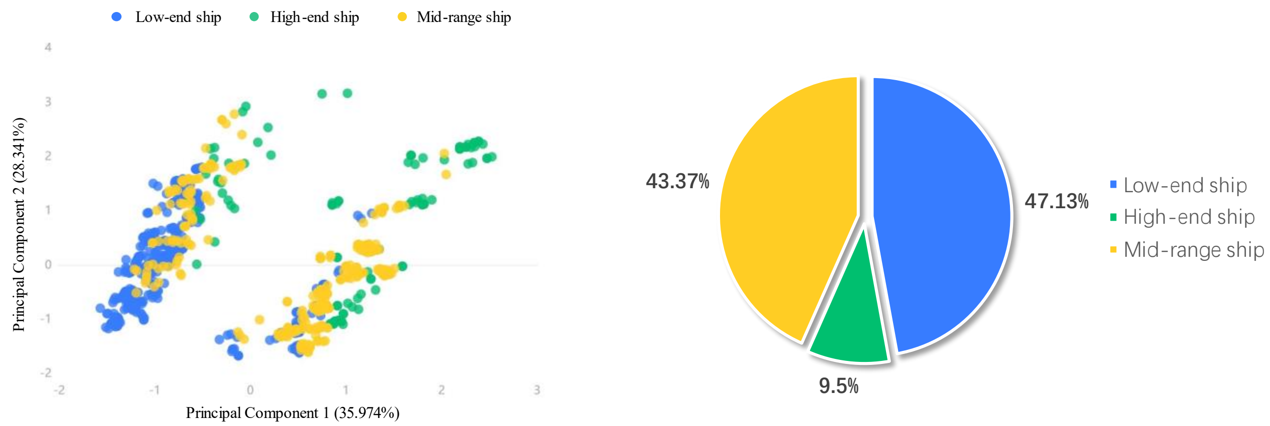


Figure 5. Visualizing clustering results

4.2. Multiple linear regression

Multiple linear regression model is a statistical analysis method used to predict the relationship between a dependent variable and multiple independent variables. It is based on the best fit of known data and can help us predict future data results. [9]

Select three indicators representing the region - population density, cargo throughput, and regional GDP - and conduct multiple regression analysis within each group to study the impact and relationship of different regional factors on the internal data of the group.

Step 1 Collect data to determine the type of linear relationship between the chosen independent and dependent variables.

Step 2 Parameter estimation.

Using the least squares method, estimate the parameters in the multiple linear regression model to obtain expressions for the prices of low-end, mid-range, and high-end sailboats:

$$\begin{aligned} y_l &= 0.471 + 0.069 * structure + 0.189 * size + 0.063 * function \\ &\quad + 0.019 * dens + 0.063 * GDP + 0.001 * CT \\ y_m &= 0.37 + 0.148 * funtion + 0.003 structure + 0.084 * size \\ &\quad + 0.153 * dens + 0.022 * GDP - 0.003 * CT \\ y_h &= 0.14 + 0.0009 * function + 0.01 * size - 0.0025 * structure \\ &\quad - 0.0000147 * dens + 0.000003 * GDP + 0.0025 * CT \end{aligned} \tag{4}$$

Step 3 Model verification.

Use various statistical indicators, such as F-test, t-test, and R-squared values, to evaluate the degree of fit of the model and test whether the hypothesis is valid [10]. The results are shown in Table 3.

Table 3. Model testing

High-end sailboats		Mid-range sailboats		Low-end sailboats	
VIF		VIF		VIF	
1.974	F=3.766 P=0.002***	4.921	F=10.218 P=0.000***	9.726	F=9.656 P=0.000***
1.901		1.419		5.243	
1.079		5.747		7.33	
2.075		1.061		1.129	
2.092		1.148		1.575	
1.162		1.125		1.686	

The result analysis of F-test shows that the significance P value is 0.000 * * *, which is significant at the level and rejects the original hypothesis that the regression coefficient is 0, so the model basically meets the requirements. For variable collinearity performance, VIF is all less than 10, so the model does not have multicollinearity problem, and the model is well constructed.

4.3. Consistency of regional effects on the sailboat varieties

Step 1 Normality test.

First, perform a normality test on the data, and the results are shown in Table 4.

Table 4. The result of the normality test

variable	sample size	average	SD	SN	KT	S-W Test	K-S Test
Price	1800	245992.152	67969.896	0.019	-0.864	0.974 (0.000***)	0.078 (0.004***)

SD: standard deviation SN: skewness KT: kurtosis

It can be seen from the table that the P-value of S-W test is less than 0.05, the original hypothesis is rejected, and the data does not meet the normal distribution, so the Kruskal-Wallis test can be conducted.

Step 2 K-W test.

The inspection results are shown in Table 5.

Table 5. The result of K-W test

	P	Cohen's f
Low-end sailboats	0.000***	0.214
Mid-range sailboats	0.008***	0.016
High-end sailboats	0.363	0.017

For low-end sailboats and mid-range sailboats: $P<0.05$ showed significant statistical results, indicating a significant relationship between regions, indicating consistent regional effects. Cohen's f of 0.214 is a minor difference. This indicates that factors such as market demand and level of competition for low-end and mid-range sailboats in different regions will have an impact, leading to price differences.

For high-end sailboats: $P>0.05$, there is no significant relationship between regions, and the regional effects are inconsistent. Cohen's f is less than 0.1, indicating no difference. This indicates that the prices of high-end sailboats are relatively stable across different regions and will not differ significantly due to factors such as geographical location.

5. Conclusions

The first part of this article is to model and explain the trading price prediction of second-hand sailboats. First, the data is retrieved and preprocessed to process the missing values and delete outlier. Subsequently, the importance of each feature was obtained through feature extraction and X-GBoost feature selection techniques, with beam (0.38) being the most important. Then, PCA is used to reduce the dimensionality of the features and obtain four principal components. Based on five indicators and their importance ratios: function (0.413), year (0.3), size (0.151) Structure (0.111), and region (0.026), an X-GBoost regression model was established, and the model parameters were optimized using SCA to obtain preliminary results. According to the adjusted R-square, MSE, MAE, and RMSE, the model has a high fit, with an adjusted R-square of 0.9741. Finally, by setting relevant perturbations on the X-Gboost parameters, it was observed that the error variation was relatively small, indicating that the model has good robustness and generalization.

In the second part of this article, in order to delve deeper into regional characteristics, local population density, per capita GDP, and local cargo throughput were selected to further classify regional characteristics. A linear regression model was established to illustrate the impact of regional characteristics on the error price caused by population density, per capita GDP, and cargo throughput. In order to ensure the accuracy of regression prediction, first, hierarchical clustering is carried out to reduce the impact of other indicators, and similar data samples are divided into the same group, which is divided into three groups – low-end ships, mid-range ships and high-end ships. Then, regression is carried out within each group to obtain the fitting results of errors, and at the same time, nonparametric Kruskal Wallis test is carried out internally. There are significant differences between low-end and mid-range sailboats and regions, indicating consistency in regional effects.

References

[1] Global ship trading volume remains strong and transaction amounts hit record highs. Tianjin Navigation, 2022, 164(03): 82.

[2] Li, J., Zhang, H., Huang, Y., & Chen, W. Grey neural network analysis for pricing of second-hand sailboats [J]. Journal of Marine Science and Engineering, 2021, 9(2): 143.

[3] Peng, J., Wang, L., Huang, Y., & Chen, W. A hybrid pricing method for second-hand sailboats based on interval regression and adaptive neuro-fuzzy inference system [J]. Ocean Engineering, 2022, 241, 109442.

[4] Qin, T., Zhang, C., Chen, K., & Zhang, X. XGBoost-Based Causal Inference on Time Series Intervention [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 1-12.

[5] Liu, Y., Wang, C., & Liang, Z. Locality Preserving PCA and Its Applications [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 1-14.

[6] Li, J., Huang, T., & Liu, Q. A self-adaptive Sine Cosine Algorithm Based on Global and Local Search for Function Optimization [J]. Expert Systems with Applications, 2021, 185, 115665.

[7] Lin, L., Wang, P. Adjusted R-squared for Generalized Linear Mixed Models [J]. Statistics in Medicine, 2020, 39(23): 3305-3315.

[8] Qin, Y., Zheng, Z., & Song, H. Hierarchical Clustering of Traffic Flow Patterns Based on Density Peaks for Fault Diagnosis in Complex Transportation System [J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 1-11.

[9] Cai, J., Fan, W. Sparse Group Lasso with Overlapping Group Structure for High-Dimensional Multinomial Logistic Regression [J]. Journal of the American Statistical Association, 2020, 116(536): 764-774.

[10] Yu, X., Li, J., Zhang, Z. PLS Regression with Missing Data for High-Dimensional Multicollinear Data [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 1-13.