Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 8, No. 3, 2023

Data Analysis and Research Based on Second-hand Sailboats

Xiaolin Liu[1, *]

[1]School of Modern Posts, Chongqing University of Posts and Telecommunications, Chongqing, China
[*]**Corresponding author:** 2248301625@qq.com

Abstract: With the rise of sailing sports, the second-hand sailing trading market shows a great potential for development. We research a thorough research on the pricing of second-hand sailing. Firstly, the linear regression model and decision tree regression model use machine learning algorithm to accurately evaluate the sailing price; secondly, the two-factor analysis method is used to analyze the influence of manufacturer and geographical region on the market price; and finally, the accuracy of the model is evaluated by paired sample T-test and Pearson correlation analysis methods.

Keywords: Linear regression, Decision tree, Two-factor analysis, Paired sample T-test, Pearson correlation analysis.

1. Introduction

Based on an in-depth survey of the sailing industry, it is found that although there were some fluctuations in the global sailing market when the COVID-19 epidemic first occurred in 2020, so far, the sailing industry has basically escaped the impact of the epidemic and maintained a steady increase in order volume for the third consecutive year. GOB recorded an impressive 1024 projects under construction or ordered, up 24.7% from 821 last year. Therefore, the research on sailing pricing is beneficial to the relevant enterprises to make a forecast on the consumption scale and growth trend of sailing products, and thus help sailing manufacturers to grasp the demand status and demand trend of various user groups for sailing products.

2. Model Building and Solution

2.1. Data acquisition and preprocessing

By finding the relevant data, found that the world bank, international freight and trade association, the world economy BBS website and SailboatData.com-the worlds largest sailboat database website there are a lot of sailing we need its performance related data, using web crawler technology, get web information, the content we have in the web need keyword extraction, get the final data and storage. Through data screening and processing of the data, we finally selected ship width, displacement, sail area, low water level, draft, cargo throughput in different regions, GDP, GDP per capita, and the average proportion of total logistics cost to GDP as other predictive factors that may affect the pricing to further analyze the pricing of sailing boats.

2.2. Build and solve the multivariate linear regression model

The solution of the multivariate linear regression model is basically similar to that of the unary linear regression model, but in this case, there are many independent variables affecting the listing price of sailing boats, so it is difficult to solve the problem by setting up parameters by itself. To further simplify the solution process, we process the data and solve the model with the help of the library and methods brought in Python. Then the linear regression algorithm in machine learning is introduced to process the training values and test values, and find the appropriate parameters, so that the linear weighted values of the respective variables are closer to the real listing price.

Set the assumption function:

$$Y_i = \omega^T X_i + b \tag{1}$$

Where Y_i represents the test data, X_i represents the sample data, and b represents the bias measure used to adjust the data results. According to the law of large numbers and the central limit theorem, the error of the true value and the predicted value of the market price is $\varepsilon \sim N(0, \delta^2)$, and then the loss function is set:

$$loss = \frac{\sum_{i=1}^{n}(\omega^T X_i + b - Y)^2}{n} \tag{2}$$

The loss function value is used as the mean variance to calculate the error of the total sample. Then you can continuously fit the results, strengthen the training set, and obtain the most suitable set of ω and b values, so that the verification set continues to approximate the real value.

To further evaluate the fit of the regression model, we also define the MAPE function to calculate the mean absolute error of the training and validation sets. The formula for setting the average absolute error is as follows:

$$MEA = \frac{1}{n}\sum_{i=1}^{n}\left|pred_{y_i} - y_i\right| \tag{3}$$

After obtaining the final training and validation sets, the coefficient of determination, r-2., can be calculated to determine the fit of the regression equation. Set the formula for calculating the coefficient of determination:

$$r^2 = \frac{\sum_{i=1}^{n}(pred_y - mean_y)^2}{\sum_{i=1}^{n}(y - mean_y)^2} \tag{4}$$

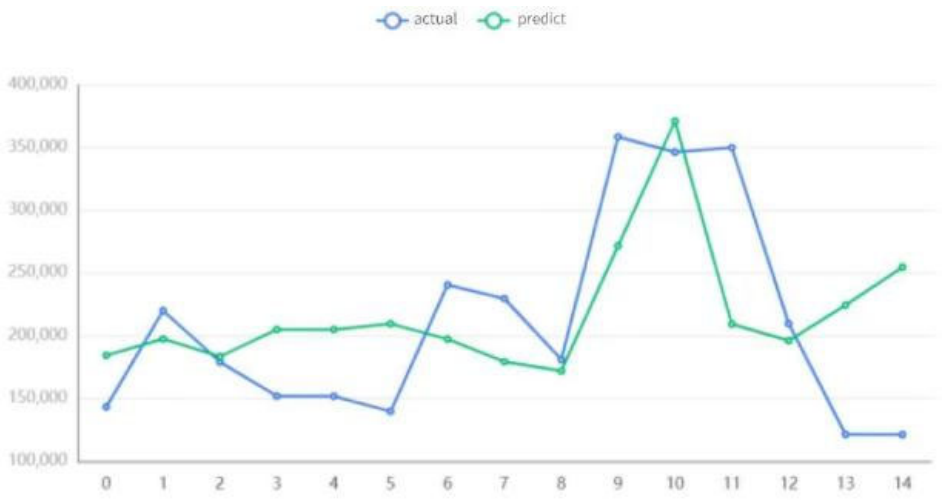

Figure 1. Prediction chart of monohull test data

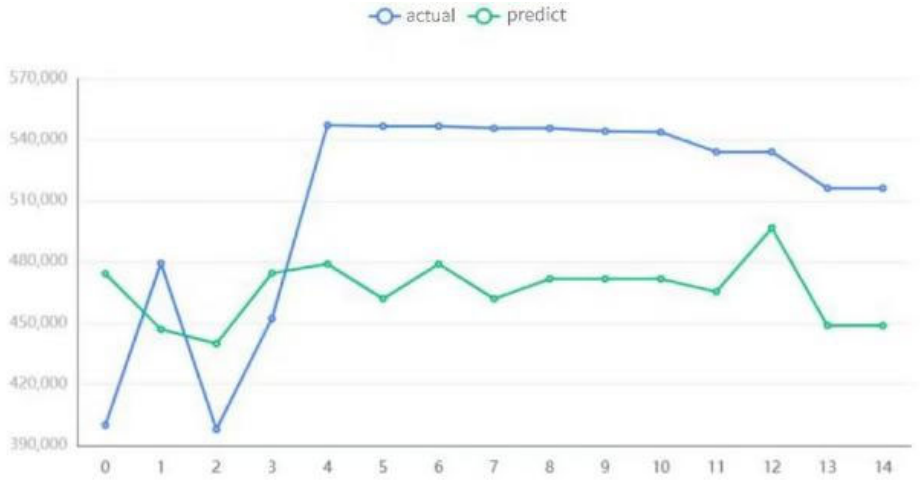

Figure 2. Prediction chart of catamaran test data

From the prediction plots of the test data in Figure 1 and Figure 2, it can be seen that there is a large difference between the real values and the predicted values of monohulls and catamarans, and there is an error in the prediction value of this model, and the training model parameters still need to be adjusted. The accuracy of predictions can be improved by increasing the proportion of training or cross-validation.

3. Model Optimization

Taking a catamaran as an example, the use of linear regression to fit a catamaran is not completely realistic, and the error between the predicted value and the true value of the model is large. Therefore, the catamaran needs to invoke other models to optimize its results. Since there are many independent variables that affect the listing price of catamarans, it can be assumed that the variables that have less weight on the price of these independent variables infect the model and lead to inaccurate results. The above problem is in line with the decision tree model, and in the subtree of the factors influencing the catamaran listing price, we can use pruning treatment to remove the secondary influencing factors and prevent overfitting of the model.

First of all, the data needs to be preprocessed, using the concept of information entropy to exclude redundant information. To set the information entropy formula:

$$H(x) = -\sum_{i=1}^{n} p(x_i)\log(p(x_i)) \quad (5)$$

When processing a large amount of data, in order to prevent important factors from being removed and shorten the training time, we also choose the pre-pruning method, and set the accuracy and recall for the pre-pruning to determine whether the target information is discarded.

Accuracy formula:

$$\text{Precision} = TP \div (FP + TP) \quad (6)$$

Recall formula:

$$\text{Recall} = TP \div (TN + TP) \quad (7)$$

Finally, draw a line chart comparing the true value to the predicted value:

Figure 3. Line chart comparing true and predicted values after optimization of the catamaran model

4. Two-factor Analysis

Two-factor modeling analysis was performed using the variables Make Variant and Country/Region/State

4.1. Single sailing catamaran

For the variable Make Variant (manufacturer), it can be obtained from the analysis of the results of the F test, which is significant at the level, has a significant effect on Listing Price (USD), and has a main effect; For the variable Country/Region/State, it can be obtained from the analysis of the results of the F test that there is no significant effect on the level, no significant effect on the Listing Price (USD), and no main effect.

4.2. Catamarans

Due to the difference between monohulls and catamarans, for catamarans, the variable Make Variant (manufacturer), which can be obtained from the analysis of the results of the F test, is significant in level and has a significant effect on Listing Price (USD), and there is a main effect. However, for the variable Country/Region/State, it can be seen from the analysis of the results of the F test that it is significant in terms of level, has a significant effect on Listing Price (USD), and there is also a main effect.

5. Price Forecast for Used Sailboats in Hong Kong

Since the analysis methods of the two sailing boats are similar, we chose a single sailing catamaran as an example to explain the analysis process in detail. For the processing of data related to monohulls, it is also necessary to consider the impact of different regions on the listing price in order to predict the impact of Hong Kong characteristics on the price. It can be seen from the above that the linear regression model fits the model well, so the linear regression model is continued.

5.1. Data substitution

After multiple trainings, the fitted result plot is plotted:

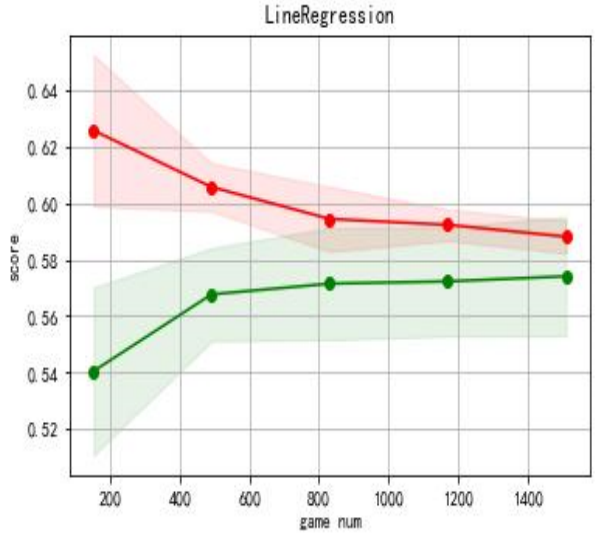

Figure 4. Fitted result plot

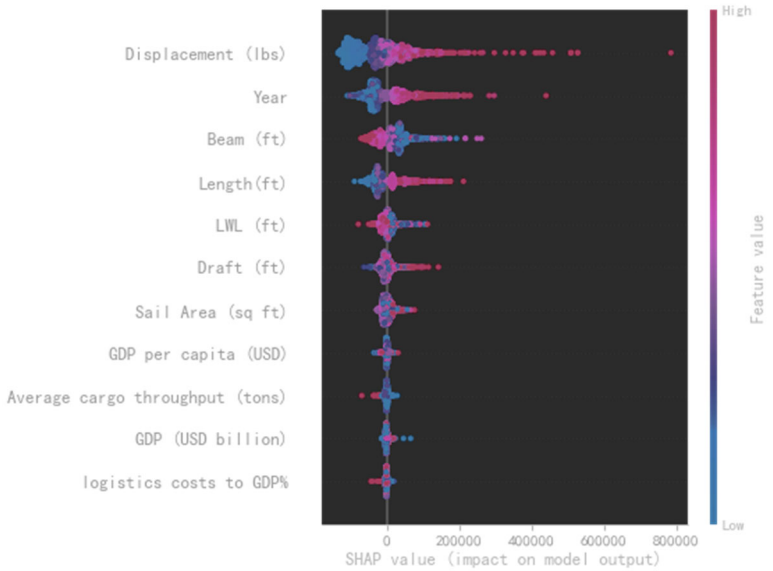

Figure 5. Histogram of feature value weights

Looking at Figure 5, it can be seen that with the increase of training times, both the training set and the test set continue to converge to 60% of the fit degree. And there are no problems such as overfitting, and the degree of fitting is good.

At the same time, various error indicators of linear regression are also calculated: the average absolute percentage error of the training set is 0.438, the average absolute percentage error of the test set is 0.351, the average absolute error is 66834.34542142153, and R2= 0.4777417196056878. All the error indicators showed that the linear regression model had a high fit and better model selection. The weight of the influence of each feature value on the listing price in the linear regression model is obtained, and a histogram is plotted, as shown in Figure 5.

Looking at Figure 5, we can see that the geographical area and the economic indicators within the region have almost no impact on the listing price, so we first find out the relevant economic indicators in Hong Kong, replace all other regions with Hong Kong, and replace all the economic indicator data of each region with Hong Kong data.

5.2. Model solving

After the data is replaced, the linear regression model is trained again, and the listing price in Hong Kong can be predicted after the training is completed. Once the predicted values are obtained, the regional effect of each type of sailing ship is analyzed. To make the results more representative, let's select a subset of sailboats: select a type of sailboat from each real region, and the selection covers all regions and all models. The paired-sample t-test is used below to determine whether there is a significant difference between the true and predicted values. The normality test histogram results are as follows:

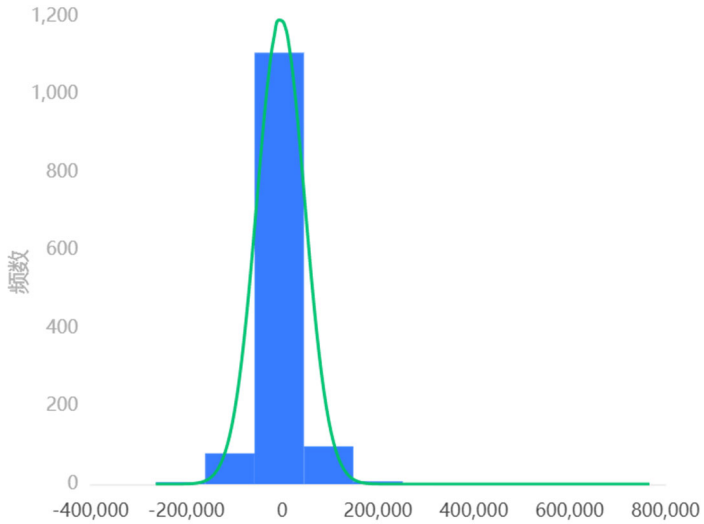

Figure 6. Normality test histogram of monohull sailboats

Looking at Figure 6, it can be seen that the difference between the true value and the predicted value data roughly conforms to the normal distribution model, which can be approximately considered to be in line with the normal distribution.

Output paired-sample t-test results:

Table 1. Monohull T test results

Paired variable	Mean standard deviation			t	df	P	Cohen's d
	Pairing 1	Pairing 2	Pairing difference (pairing 1- pairing 2)				
True_Price pairing simulation_HK_Price	236513.005±157958.795	239242.583±140976.945	-2729.579±16981.851	-1.962	1298	0.050**	0.054

Note: * * *, * *, * represent the significance level of 1%, 5% and 10% respectively

The following analysis of the catamaran sailboat in the same square, the fitting results after linear regression training are as follows:

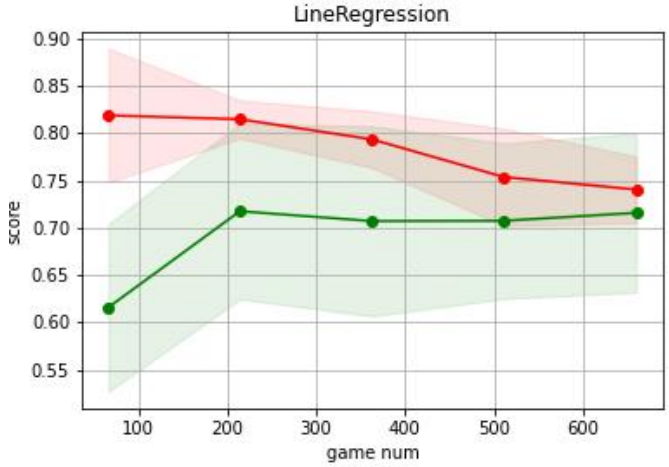

Figure 7. Fitted result plot

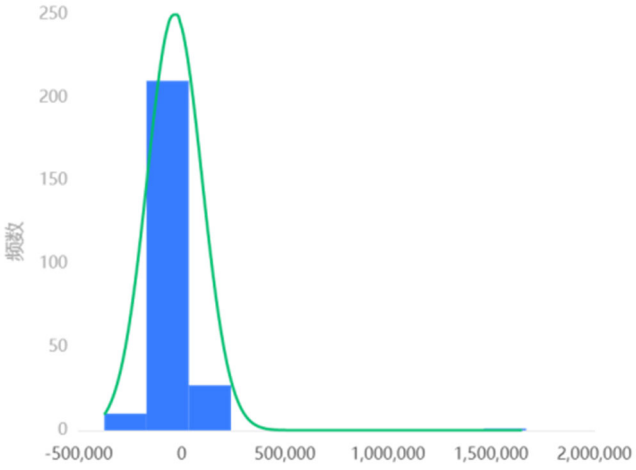

Figure 8. Normality test histogram

The training set and test set continuously converged to 73% of the degree of fit, and the fitting results were good. The normality test histogram results are shown in Figure 8. The image roughly conforms to the normal distribution model, and the two-body can also be approximately considered to be normally distributed.

Output paired-sample t-test results:

Table 2. Catamaran T test results

Paired variable	Mean standard deviation			t	df	P	Cohen's d
	Pairing 1	Pairing 2	Pairing difference (pairing 1- pairing 2)				
True_Price pairing simulation_HK_Price	460987.855±254378.624	495198.76±215184.377	-34210.905±39194.247	-4.056	247	0.000***	0.258

Note: * * *, * *, * represent the significance level of 1%, 5% and 10% respectively

Comparing monohulls and catamarans in the pairing sample T test, the pairings 1 and 2 of catamarans are about twice as high as those of monohulls, but the pairing difference of catamarans is about 12 times that of monohulls. The above conclusion shows that catamarans have a stronger effect than monohulls, and that catamarans are more susceptible to regional influences.

6. Use Pearson Correlation Analysis to Study the Impact of Hull Performance on Pricing

From the above analysis, it can be seen that the listing price of the sailboat is closely related to the performance of the hull itself. Therefore, we will further explore the correlation between different properties and price, and the results can be used as an important basis for brokers when predicting the price of sailboats. Using a monohull as an example, we take the Pearson correlation analysis method to explore the correlation between market price and sailboat performance. Significance P>0.05 indicates a significant relationship. After calculating the correlation coefficient between the two indicators, draw the correlation coefficient heat map:

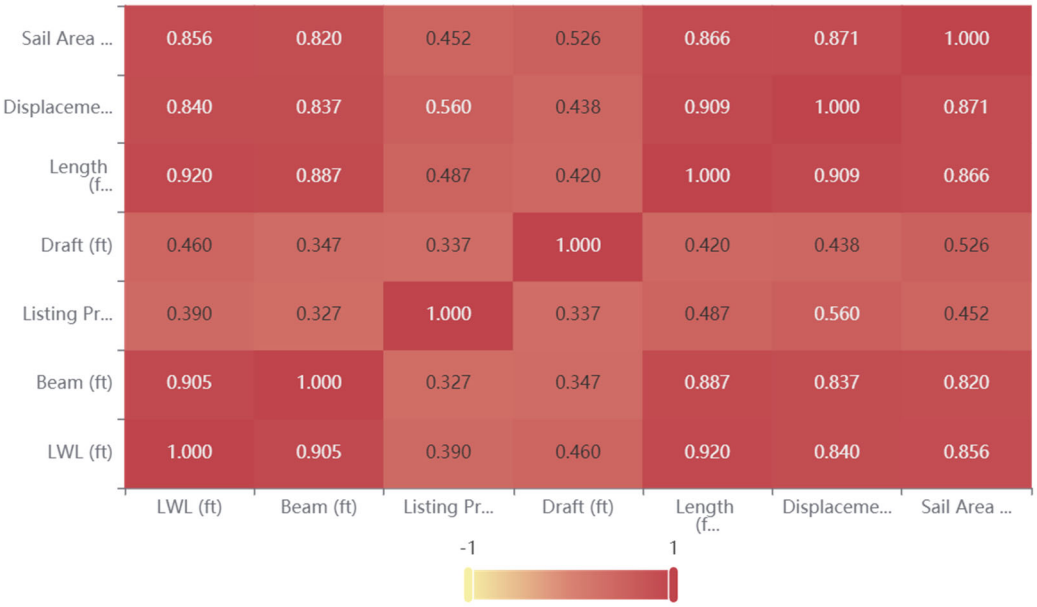

Figure 9. Correlation coefficient heat map of monohulls

Looking at the image, it can be seen that the correlation coefficient between the listing price and the displacement, length and area of the sailboat is large. The correlation coefficient between hull length and displacement is as high as 0.909, which means that length determines the size of displacement. Plot the scatter plot of price and hull length:

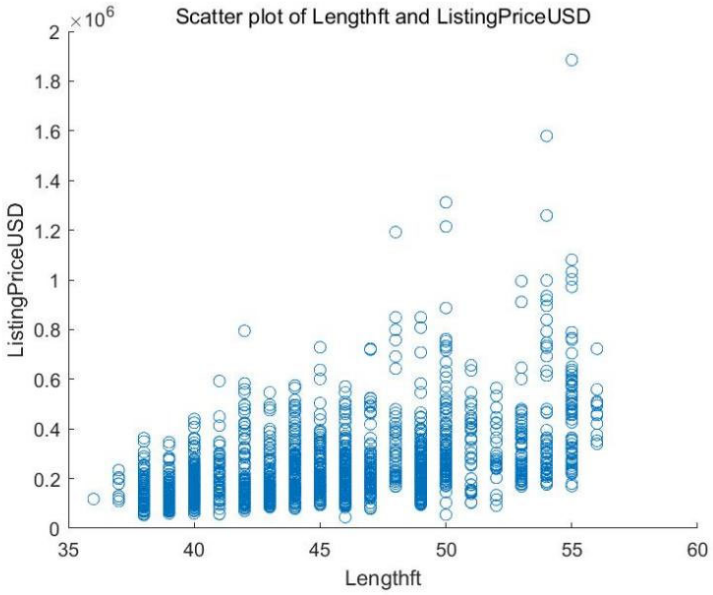

Figure 10. Scatterplot of price distribution with hull length

Looking at the image, it can be seen that as the length of the hull increases, the price shows an upward trend; The price of a monohull sailboat of the same length is not much different.

Also using Pearson correlation analysis, a correlation coefficient heat map of a catamaran sailboat is plotted:

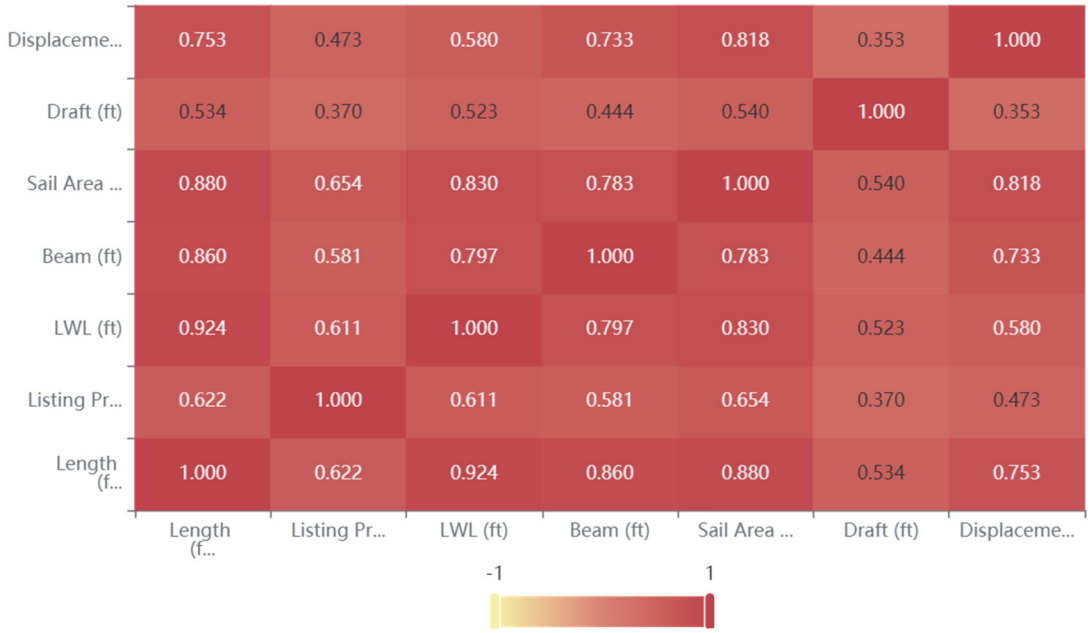

Figure 11. Correlation coefficient heat map of a sailing catamaran

Observing the image, it can be seen that the price of a catamaran is mainly affected by the sail area, length and low water mark, and the length is closely related to the low water mark. Therefore, compared to monohulls, the sail area needs to be considered when predicting the price.

Scatterplot the distribution of price and hull length, and scatterplot of price and sail area:

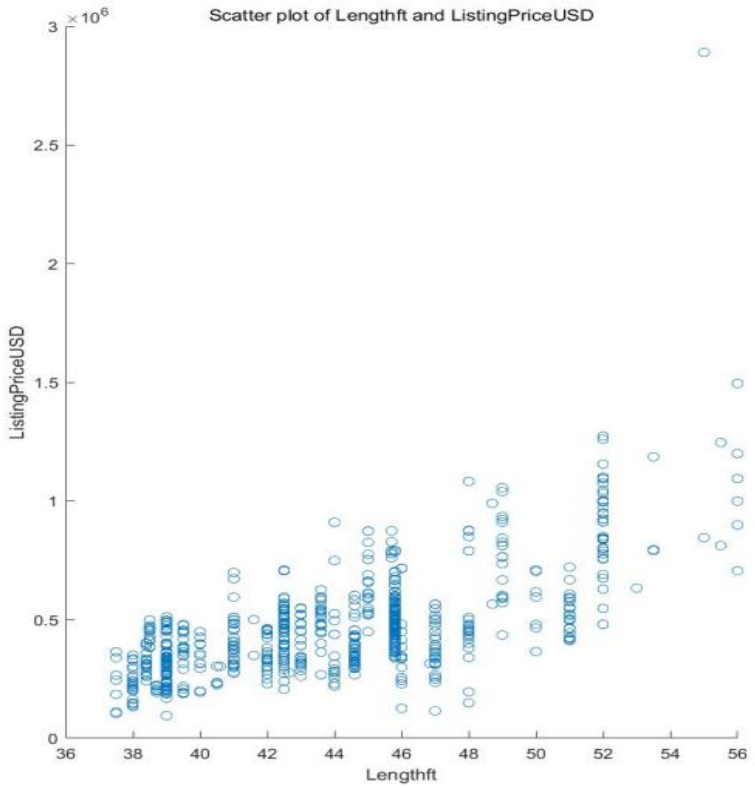

Figure 12. Scatterplot of price distribution with hull length

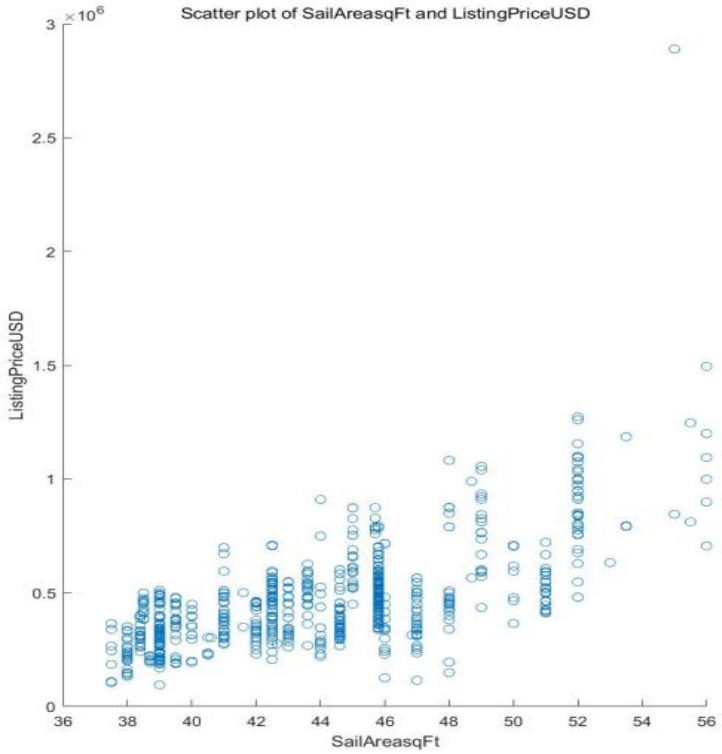

Figure 13. Scatter plot of the distribution of prices and sail area

Looking at the image, it can be seen that as the length of the hull increases, the price shows an upward trend; As the sail area increases, so does the price. With a hull length between 38-43 feet and a sail area between 39-46 square feet, the price is more affordable.

7. Model Evaluation

7.1. Model advantages

Based on the multiple linear regression model, the established model is closely related to the reality, and the proposed problems are solved in combination with the actual situation, so that the model is closer to the reality, and the versatility and generalization are strong.

The model has strong operability and a wide range of applications, and the model based on the fourth question using Pearson correlation analysis is more accurate, and the simulation results are more reasonable.

The model has no strict restrictions on data distribution, sample size and indicators, which is suitable for small sample data and large systems with multiple evaluation units and multiple indicators, which is more flexible and convenient

7.2. Model disadvantages

The fitting result of linear regression is about 70%, and there is a certain deviation between the predicted value and the true value.

In the modeling process, taking a monohull as an example, applying the same method to a catamaran may lead to ignoring the particularities of the catamaran, such as finding that the linear regression model is not suitable for catamarans in the modeling process of problem one.

References

[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com .

[2] Fisher Box, Joan. Guinness, Gosset, Fisher, and Small Samples. Statistical Science. 1987, 2 (1): 45–52.

[3] Draper, N.R. and Smith, H. Applied Regression Analysis. Wiley Series in Probability and Statistics. 1998.

[4] Zhou Zhihua Machine Learning [M].Tsinghua University Press, 2016.