Price analysis of used sailboats based on LightGBM model

Xinyu Gao *, Jingjing Zhu, Rui Yang

Department of Automation, Xiamen University, Xiamen, China

* Corresponding Author Email: 34520202200280@stu.xmu.edu.cn

Abstract. The pricing of second-hand sailboats is a very important issue, because it is directly related to the investment and profits of sailboat operators. This paper mainly conducts further research on the pricing of second-hand sailboats. First, we establish a LightGBM regression mathematical model through data. Second, we identify the two regional factors with the highest correlation with price and conduct qualitative analysis on them. Then, we took Hong Kong as the research object, obtained the regional effect of its prediction results, and verified the accuracy by t-test.

Keywords: LightGBM; paired t-test; linear regression.

1. Introduction

Like many luxury goods, sailboats vary in value as their age and market conditions change. Due to the different performance indicators, regions and the year in which they were sold, their prices often fluctuate up and down. For teams selling used sailboats, it is essential to understand the role of various factors on the price of the sailboat in order to achieve a reasonable estimate of its value, so as to better understand the sailing market.

2. Model building and solution

2.1. the Neural Network Model

2.1.1 Model Overview

The neural network model we establish adopts the fastest descent method, it obtains the error between the output and the expectation through forward propagation, and backpropagates the them. In the process of iterative training, the connection weights between each node of each layer are continuously adjusted until the stop condition is met. In the process, the number of hidden layer nodes we select is 5 according to the cross-validation.

The neural network model topology includes three layers: input layer, hidden layer and output layer. The specific structure is shown in Figure 1.

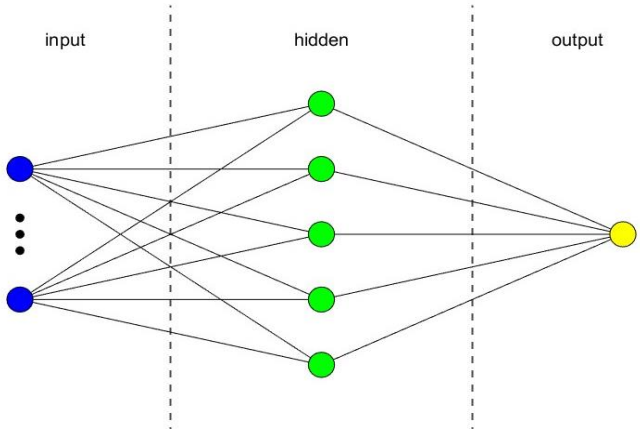


Fig 1. Neural network structure diagram

2.1.2 Model Principle

When forward propagation occurs, the input signal acts on the output node through the hidden layer, and the output signal is generated after the nonlinear transformation. If the actual output does not match the desired output, the error backpropagation process is carried out.

The process of forward propagation in a neural network can be expressed as:

$$a^l = f(w^l a^{l-1} + b^l) \tag{1}$$

Error back transmission is to transmit the output error to the input layer by layer through the hidden layer, and distribute the error to all units in each layer as the basis for adjusting the weight of each unit. By adjusting the threshold and the connection weight between nodes, the error decreases in the gradient direction, so the network parameters (weights and thresholds) corresponding to the minimum error are determined after repeated learning and training. At this time, the trained neural network can process the nonlinear transformation information with the smallest output error for the input information of similar samples.

If a perturbation is applied to the jth node of the lth layer, the error can be expressed as:

$$\delta_j^l = \frac{\partial L(a^l, y)}{\partial z_j^l} \tag{2}$$

The rules for updating parameters can be represented as:

$$b^l = b^l - \alpha \frac{\partial L}{\partial b^l} \tag{3}$$

$$w^l = w^l - \alpha \frac{\partial L}{\partial w^l} \tag{4}$$

2.2. XGBoost

2.2.1 Model Principle

We split the processed dataset into training set, validation set and test set, adjust the parameters to obtain different trained models, and observe their performance on the validation set to determine the corresponding parameters of the optimal model.

The XGBoost algorithm uses a step-by-step forward additive model like the gradient boosting algorithm, and there is no need to calculate the coefficients after each iteration to generate the base learner. The model can be represented as:

$$y_i = \sum_{m=1}^{M} f_m(x_i) \tag{5}$$

The objective loss function can be expressed as:

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{6}$$

The algorithm adds regular terms to the loss function to control the complexity of the trained model, which can reduce the variance of the model and avoid overfitting effectively.

Extend and apply the two formulas above to the first t iterations as follows:

$$y_i^{(0)} = 0 \tag{7}$$

$$y_i^{(1)} = f_1(x_i) = y_i^{(0)} + f_1(x_i) \tag{8}$$

$$y_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = y_i^{(t-1)} + f_t(x_i) \tag{9}$$

Correct the corresponding loss function from the perspective of the sample set of leaf nodes. The formula can be rephrased as:

$$Obj^t \approx \sum_{i=1}^{n} [g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 = \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2 \right] + \gamma T \tag{10}$$

$$G_j = \sum_{i \in I_j} g_i , H_j = \sum_{i \in I_j} h_i \tag{11}$$

The end result can be expressed as:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{12}$$

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{13}$$

2.2.2 Parameter Selection

After the model is initially established, the parameters of the model need to be adjusted and optimized. The best parameters are selected by comparing the performance of the trained model on the validation dataset.

The main parameters that we need to adjust are n_estimators, max_depth, min_child_weight, gamma and subsample, and the rest of the parameters are set by default.

Parameter n_estimators is the number of base learners, usually integers such as 100, 200, etc. Cross-validation yields the highest score when n_estimators = 50.

The learning curve is shown in Figure 2.

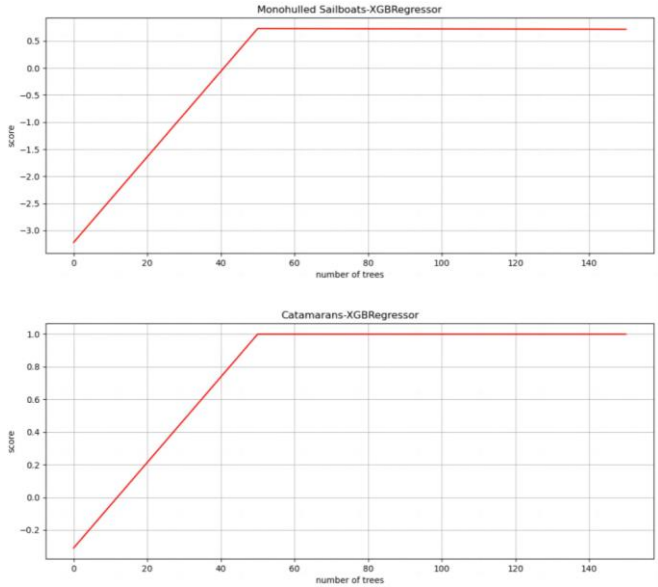


Fig 2. Learning curve of XGB regression model on datas

Gamma determines the descent value of the minimum loss function that the node needs to meet when splitting, which usually takes a value less than 1. The larger its value, the more conservative the algorithm. After several times of debugging, the model works best when the value is 0.

Subsample represents the proportion of each decision tree that randomly takes a sample. If its value is too large, it will cause the model to be overfitting, and if the value is too small, it will produce underfitting, so this parameter needs to have a reasonable value.

Max_depth is the maximum depth of the tree, and min_child_weigh represents the minimum sample weight required for leaf nodes, both of which are used to avoid extreme cases in the fitting of the model.

The value range of parameter max_depth is generally between 5-10, the larger the value, the more specific the data learned by the model, the better the prediction effect, but if the value is too large, it may cause the model to overfit. Parameter min_child_weigh has a value between 0-1, but unlike max_depth, it causes the model to underfit when its parameter value is too large. We use the GridSearchCV function to test any combination of the values of the above two parameters, and finally conclude that when max_depth = 9 and min_child_weight = 6, the score can reach 0.8456.

The options of model parameters are shown in Table 1.

Table 1. Optimal model parameters of XGBoost

n_estimators	max_depth	min_child_weight	gamma	subsample
50	9	6	0.1	0.91

After completing the model parameter optimization, the test set data is put into the trained model to measure the generalization ability of the model, and the R^2 value of the XGBoost model is calculated as 0.9592. This means that 95.92% of the differences between the observations of the explanatory variables can be explained by the established XGBoost regression model.

2.3. LightGBM Model

2.3.1 Model Principle

LightGBM is a histogram-based decision tree algorithm. It uses the GOSS algorithm to exclude most of the samples with small gradients, and only uses the remaining samples to calculate the information gain, so as to achieve the purpose of reducing the amount of data and ensuring the balance of accuracy.

Firstly, the continuous floating-point feature values are discretized and converted into k integers, and then constructed as a histogram with width values equal to k. When traversing the data, the model uses the discretized values as indexes and accumulates the statistics in the histogram.

After a round of data traversal, the required statistics have been accumulated in the histogram. Then iterate in combination with the index of the histogram to find the best segmentation point. Therefore, there is no need to use the sorting algorithm when performing node splitting, and the complexity is optimized from O (#data * #features) to O (k* #features).

2.3.2 Parameter Selection

The optimal parameters after grid parameter finding of the LightBGM algorithm are shown in Table 2.

Table 2. Optimal model parameters of LightBGM

parameter	value	parameter	Value
Max_depth	7	Learning_rate	0.1
Num_leaves	65	N_estimators	40
Min_chid_weight	0.001	Feature_fraction	0.7
Min_child_samples	20	Reg_alpha	0.4

3. Results and Solutions

3.1. Accuracy

We firstly analyze the correlation between each variable and Price.

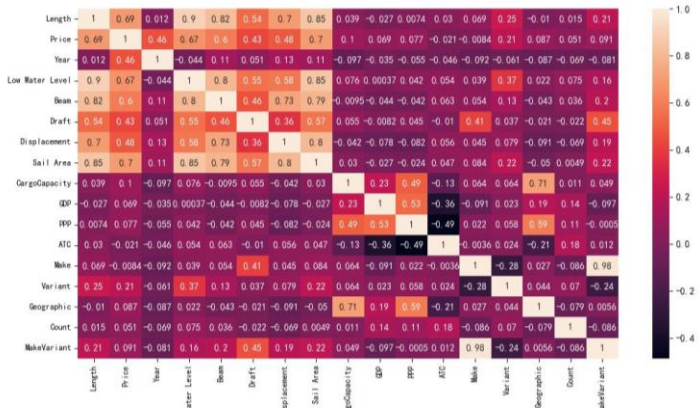


Fig 3. Correlation Analysis Chart

Based on the image, we can see that the price is positively correlated with the individual feature vectors, and we can initially know the initial value of the parameter adjustment.

The prediction accuracy of the model (normalizd) for sailing ship pricing is shown in Table 3.

Table 3. Regression accuracy of different models for different sailboats

Model	Monohulled Sailboats	Catamarans
BP	0.10191	0.03918
XGBoost	0.01860	0.01147
LightGBM	0.00762	0.00324

It can be summarized from the table that LightGBM model has the smallest estimation error on sailing ship pricing, so this model is chosen as the predictor for subsequent questions.

3.2. Analysis of Regional Effects

In this section, we use the XGBoost model to further explore regional effects, explain the impact of regions on listing prices, and discuss whether the sregional effects are consistent across all sailboat variations. From an initial look at the data, we can see that the United States has the highest number of Monohulled sailboats, while the Caribbean has the highest number of catamarans. In addition, the distribution of Monohulled sailboats and Catamarans is not the same in various regions.

To analyze the impact of geography on listing price, we can use a regression model with regional factors (GDP, PPP, ATC, Cargo Handling Capacity) as categorical predictors.

3.2.1 Identify Regional Factors

We first use linear regression model to make a analysis of the correlation between Price and the above four geographical factors.

Take Monohulled Sailboats for example, the regression results is shown in Figure 4.

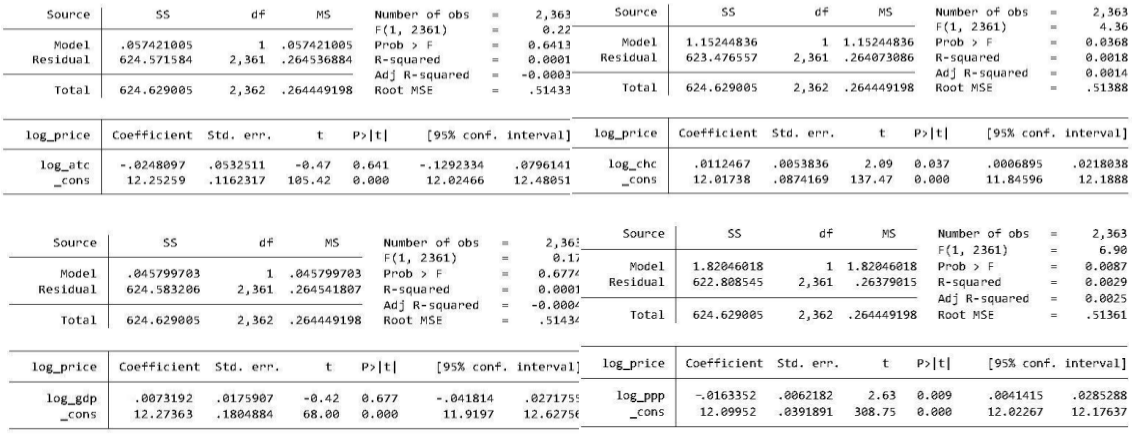


Fig 4. Linear regression results on Monohulled Sailboats

According to the regression results, compared with other factors, the pricing of sailboats has a significant relationship with PPP and CHC, so they are used as regional factors to analyze regional effects. Use PPP and CHC to perform two-factor regression on the pricing of sailboats, and the results of the regression analysis as shown in Figure 5.

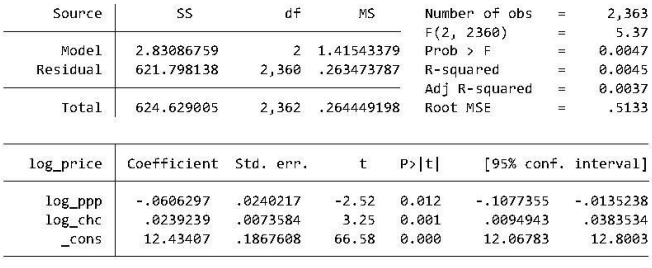


Fig 5. Two-factor regression results

According to the regression results, we can generally know that there is a negative correlation between pricing and PPP and a positive correlation with CHC, which provides a priori knowledge for subsequent quantitative analysis.

3.2.2 Qualitative Analysis

Before using the established model for quantitative analysis, we visualize the pricing of the two types of sailboats separately to qualitatively analyze the regional effects.

The fitted 3D curves, profiles, and contour plots are shown in the Figures below.

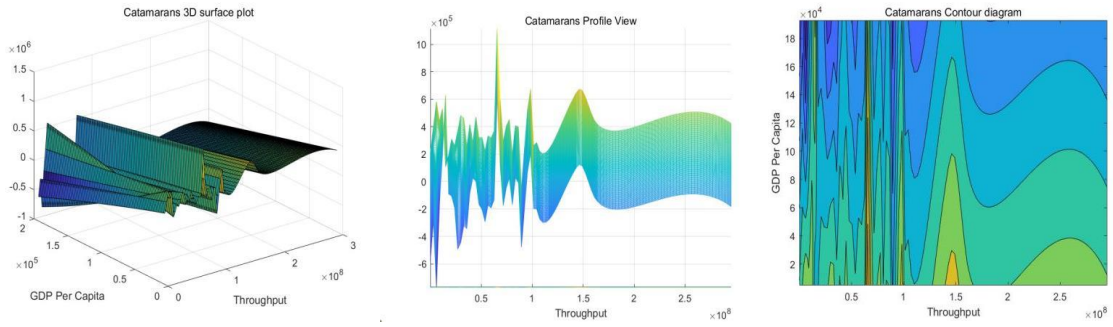


Fig 6. visualization results of Monohulled Sailboats

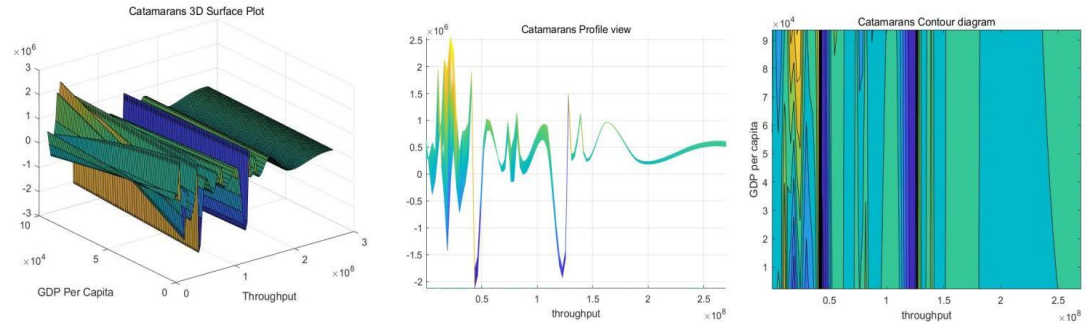


Fig 7. visualization results of Catamarans

A preliminary analysis of regional regional effects based on 3D images shows that the impact of CHC on the pricing of the two types of sailboats is basically the same, while the impact of PPP is slightly different. Under the same conditions (i.e. CHC), the correlation between PPP and sailing pricing is not obvious for Catamarans, and further quantitative analysis is required to determine them. However, for Monohulled Sailboat, PPP shows a clear negative correlation.

To illustrate here, as shown in the scatterplot below, this part of the qualitative analysis does not consider the relationship between pricing and PPP at higher CHC due to the small amount of data.

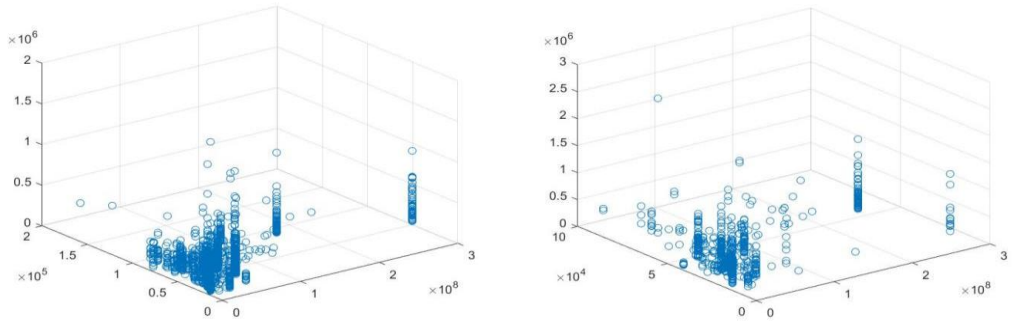


Fig 8. Scatter plot of Monohulled Sailboats(left) and Catamarans(right)

Due to the LightGBM algorithm, the scores of each variable in the figure below can be used as a mapping of the weight of this feature in the model.

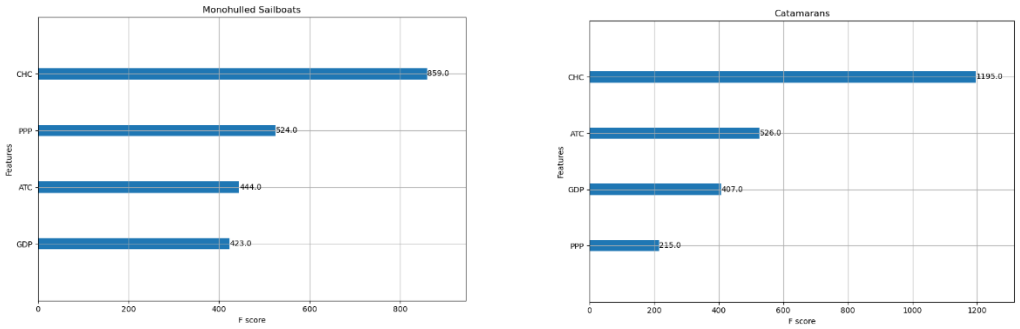


Fig 9. Sort on Monohulled Sailboats(left) and Catamarans(right)

From this, we conclude that the role of regional effects in different sailing variants is not completely consistent, but the factors that have the greatest influence are common. At the same time, the accuracy of the previous regression conclusions are proved.

Combine the predicted values with the feature data into a larger dataset to feed into the ridge regression model, the resulting prediction coefficients are shown in Tables 4.

Table 4. Predictive features coefficients of Monohulled Sailboats and Catamarants

Coefs	CHC	GDP	PPP	ATC
Monohulled Sailboats	11985.85932794	3814.3423552	-19940.91196292	-1963.86574404
Catamarants	17484.19232426	14854.64234403	-6543.74151404	-2162.64839628

From the perspective of positive and negative correlation, the regional effects of the two sailboats are similar: For Monohulled Sailboats, CHC and GDP are positively correlated with prices, while PPP and ATC are negatively correlated with prices, the same as Catamarans.

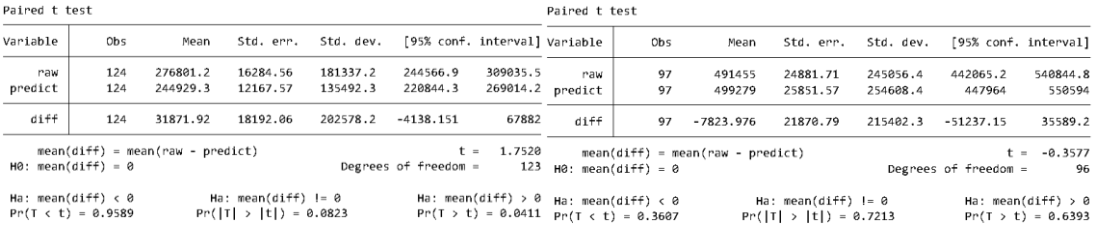
However, from the perspective of importance, we find that the regional effect is different in the two types of sailboats: For Monohulled Sailboats, the most important influencing factors are CHC and PPP, while for Catamarans, the most important influencing factors are CHC and ATC.

3.3. Regional Effect in Hong Kong

We believe that the regional effect in Hong Kong and the global regional effect are slightly different. Therefore, when predicting prices in this question, it is not advisable to use the data from the previous section directly.

In order to simulate the regional effect of Hong Kong, we use PPP and CHC as indicators, and the KNN network are used to screen out similar data to Hong Kong in the existed dataset. Next we replace PPP and CHC indicators of Hong Kong with the corresponding indicators in the original dataset and integrate them into a new data set. Using the new dataset as input, the predicted output can be obtained through the mathematical model established in 5.1, which is the result of the simulation.

We use the paired t-test model to determine the correlation between the predicted prices under the influence of two regional effects. The test results are shown in the following figure.



Paired t test

Variable	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
raw	124	276881.2	16284.56	181337.2	244566.9	309035.5
predict	124	244929.3	12167.57	135492.3	220844.3	269014.2
diff	124	31871.92	18192.06	202578.2	-4138.151	67882

mean(diff) = mean(raw - predict) t = 1.7520
H0: mean(diff) = 0 Degrees of freedom = 123

Ha: mean(diff) < 0 Ha: mean(diff) != 0 Ha: mean(diff) > 0
Pr(T < t) = 0.9589 Pr(|T| > |t|) = 0.0823 Pr(T > t) = 0.0411

Paired t test

Variable	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
raw	97	491455	24881.71	245056.4	442065.2	540844.8
predict	97	499279	25851.57	254608.4	447964	550594
diff	97	-7823.976	21870.79	215402.3	-51237.15	35589.2

mean(diff) = mean(raw - predict) t = -0.3577
H0: mean(diff) = 0 Degrees of freedom = 96

Ha: mean(diff) < 0 Ha: mean(diff) != 0 Ha: mean(diff) > 0
Pr(T < t) = 0.3607 Pr(|T| > |t|) = 0.7213 Pr(T > t) = 0.6393

Fig 10. Paired t-test on Monohulled Sailboats(left) and Catamarans(right)

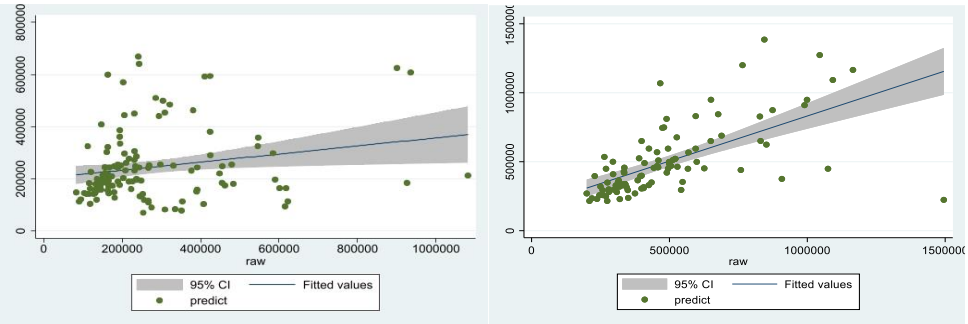


Fig 11. Visualization of t-test on Monohulled Sailboats(left) and Catamarans(right)

The results show that the results pass the paired t-test, so the conclusions we draw above about regional effects also apply to Hong Kong.

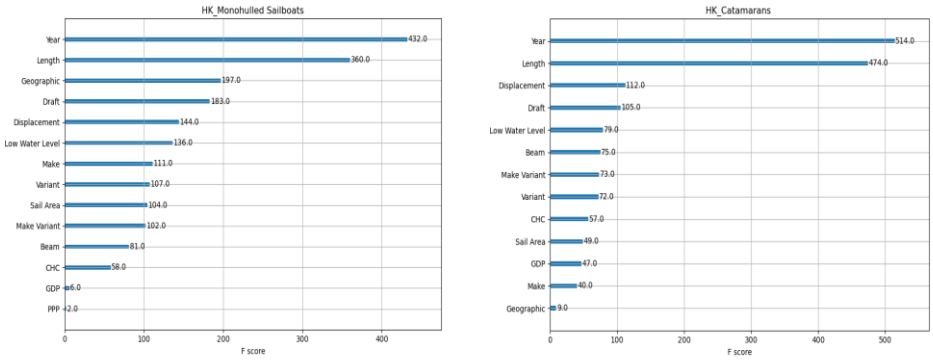


Fig 12. Sort on Monohulled Sailboats(left) and Catamarans(right)

4. Strengths and Weaknesses

4.1. Strengths

(1) We chose the LightBGM regression model for a more comprehensive and accurate assessment of reliability and accuracy.

(2) In the process of analyzing regional effects, we first use linear regression to obtain prior knowledge, and then validate and modify the prior knowledge through a model. This process reduces the error probability and the uncertainty of the analysis, and improves the reliability of the conclusions.

(3) When we apply the global regional effect to the Hong Kong region, we fully consider the particularity of Hong Kong and reuse the model to analyze the factors that have the greatest impact on prices in the specific region, making the conclusion more targeted and practical.

4.2. Weaknesses

(1) In fact, there is a close relationship between the price of second-hand sailboats and the depreciation rate of the engine, but the depreciation rate is not included in the data we are looking for, making the established model unreliable.

(2) When simulating the regional effect of Hong Kong, after we find out that the regions are similar to Hong Kong, we use the ship prices in these regions as the real price of Hong Kong ships, without considering the differences between different regions. Also there is a problem of not being comprehensive enough.

References

[1] Mao Pan. Research on Influencing Factors of used car price evaluation based on BP neural network. XiHua University, 2020.

[2] Jia Pengxiang. The Used Cars' Price Forecast Based On LightGBM. Shandong Normal University, 2021.

[3] Wu Xiqingyong. Research on Used car pricing based on data mining. 2095-6649(2020)08-0006-003, 2020.

[4] Wei Haijiao, A Literature Review of the Inflfl uencing Factors of Used Car Price Assessment in China. Renmin University of China.

[5] Ma Wenyi. Research on value evaluation and investment decision of second-hand dry bulk carriers, Shanghai Jiaotong University, 2020.

[6] Li EnLIn. Study of Second-hand Ship Purchasing Decision-making Model Based on Price Fluctuations. Harbin Engineering, 2021