Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2330185

Broke The waves : Exploring SailBoats' Prices

Summary

Like most luxury goods, the value of SailBoats varies with age and market conditions. In order to better understand the second-hand **SailBoats** market, this article intends to build a versatile model for the pricing rules of second-hand SailBoats, enabling it to be applied to the pricing of SailBoats in different regions.

Analyzing a problem can make it clear that what is needed is a regression model. For regression models, we need more features with high influencing factors. Collect data to find a total of 5 characteristics and attributes related to ships, such as LWL, beam, draft, displacement, and sail area. In addition, consider the regional impact to find the economic characteristics and attributes related to different regions, such as freight throughput, GDP, and GDP per capita.

In order to select the basic model, this article analyzes a total of six machine learning models, namely, LinearRegression, DecisionTree, RandomForest, LGBM, XGBoost, and GDBT, using two evaluation indicators, MAPE and R^2, and selects LGBM in combination with Monohulled SailBoats and Catamarans. On the basic model, the values of R^2 are 0.7676 and 0.8291, respectively. Combining LGBM with Simulated Annealing, this article establishes a **SA-LightGBM** model. Before combining the Simulated Annealing Method, this article adopted the method of **Data Shuffle** to effectively reduce the over fitting phenomenon existing in the data. In order to select the LGBM with the most parameters, this article **Grid Search** is the most widely used hyperparametric search algorithm, which determines the optimal value by searching all points within the search range. Finally, the optimized LGBM was used as the generator of the new solution, and the results obtained by SA-LightGBM were evaluated using R^2. The results showed that Monohulled Sailboats and Catamarans were 0.8130 and 0.9022, respectively, with an accuracy improvement of 5-6%, verifying the effectiveness of the model.

In order to examine the importance of the impact of regional indicators on the model, we conducted an in-depth positive and negative impact analysis of the use of SHAP for each feature, and found that the impact of regional data on the model was not significant. In addition, **Two-factor analysis of variance** was used to analyze whether different levels of the type and region of SailBoats have a significant impact on the final result.

Finally, based on the collected regional data of **Hong Kong market**, we used the SA-LightGBM model to predict, and then used the method of **Paired Sample T-test** to conduct paired sample T test on the above estimated results of the Hong Kong SAR in the data of Monohulled SailBoats and Catamarans, respectively, with our real data. It has been observed that the Hong Kong SAR has a negative impact on both Monohulled SailBoats and Catamarans, but the regional impact on both is different, with the impact on Catamarans being more significant than Monohulled SailBoats.

Keywords:SailBoats,SA-LightGBM,Grid Search,Paired Sample T-test

Contents

1 Introduction

1.1 Problem Background

SailBoats are ships that use the wind to forge ahead. Ranking among multifarious luxury goods, SailBoats enjoy features that values vary as time goes by and as market conditions change.

SailBoats are usually monohull, but there are also Catamarans with strong resistance to wind and waves. A boating enthusiast provided data for both two types, including columns labeled as Make, Variant, Length (in feet), Geographic Region, Country/Reygion/State, Listing Price (in US dollars), and Year (of manufacture).

1.2 Restatement of the Problem

The frequent trading method of SailBoats is through brokers. To better understand the mechanisms that affect prices, we are supposed to solve these problems:

(a) Develop a mathematical model that explains the listing price of each of the SailBoats in the provided spreadsheet with a discussion of the precision of the estimate.

(b) Use the model above to explain the effect, if any, of region on listing prices and discuss whether any regional effect is consistent across all SailSoats variants.

(c) Model the regional effect of Hong Kong (SAR) market on each of the SailBoats prices for the SailBoats in a chosen informative subset. Is the effect the same for both Catamarans and Monohulled SailBoats?

(d) Identify and discuss other interesting and informative inferences or conclusions drawing from the data.

2 Assumptions and Notations

2.1 Assumptions

In order to understand the content of the question more clearly and concisely, we have made a hypothesis here and explained it later.

(a) Inferring a single data from multiple data will be a regression model. Try to find the most suitable method and improve it by combining optimization algorithms.

(b) Find an analytical method to compare the differences between two categorical data and one quantitative data to obtain the degree of impact.

(c) To use regression models to understand the Listing Price in Hong Kong, you should understand other data from Hong Kong.

2.2 Notations

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

Symbol	Description
$F_obs(x)$	Empirical Fistribution Function
D_n	K-S Statistic
μ	Expectations
σ	Variances
y_{base}	Target Sample Mean
SSF	Sum of Squares for Factor
SSE	Sum of Squares for Error

Table 1: Symbol Retrieval Table

3 Data Analysis

Among the data file given, '2023_MCM_Problem_Y_Boats.xlsx', there are two tables that record information about Monohulled SailBoats and Catamarans, respectively. And the column structures of the two tables are consistent. The following data are given in the table: 'Make, Variable, Length (in feed), Geographic Region, Country/Region/State, Listing Price (in US dollars), and Year (of manufacture)'. Where 'Length (in feed), Listing Price (in US dollars), and Year (of manufacture)' are numeric types.

3.1 Data Pre-processing

Our analysis have found that there are three rows of missing values in the Monohulled SailBoats table. Due to the data volume of over 2000 rows, which is far greater than the three missing data, this part of the missing values is directly removed here.

3.1.1 Normality Test

Here, **K-S(Kolmogorov-Smirnov test)** is used to test whether a sequence obeys a normal distribution.

K-S is a nonparametric test used to test whether a sample comes from a specific distribution, or to compare whether two samples come from the same distribution. Assuming there are n mutually independent random samples[1]:

$$x_1, ..., x_n \tag{1}$$

We believe that these values come from a distribution P. Among them, H_0 comes from P and H_1 does not come from P. The cumulative distribution function $F(x)$ is:

$$F(x) = P(X < x) \tag{2}$$

The cumulative distribution function uniquely describes a probability distribution. For the above n mutually independent random samples, their empirical distribution function is:

$$F_{obs}(x) = \frac{nums\ of\ observations\ below\ x}{observations} \tag{3}$$

Sort all observations to obtain:

$$y_1, ..., y_n \tag{4}$$

So:

$$F_{obs}(y_i) = \frac{i}{n} \tag{5}$$

Finally, we get that K-S statistic is:

$$D_n = \max \left| F_{\exp}(x) - F_{obs} \right| \tag{6}$$

After K-S test, it is found that the given data obey normal distribution.

3.1.2 3-Sigma Discrimination

The initial box diagrams of two types of SailBoats based on Listing Price data show in Figure 1 and 2:

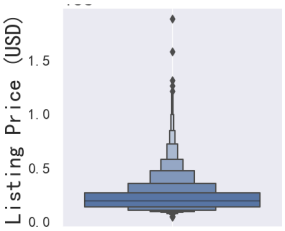


Figure 1: Initial Box Plot for Monohulled SailBoats

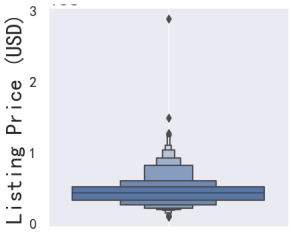


Figure 2: Initial Box Plot for Catamarans

We found that there are partial outliers in both cases, so we consider using the 3-Sigma criterion to remove partial outliers. Due to the probability that distribution of values lying between ($\mu - 3\sigma$, $\mu + 3\sigma$) is 0.9974. Eliminate values outside the range according to the 3-Sigma criterion. 45 data were excluded from Monohulled SailBoats and 15 datafrom Catamarans. The box graphs after processing are removed of outliers, show in Figure 3 and 4:

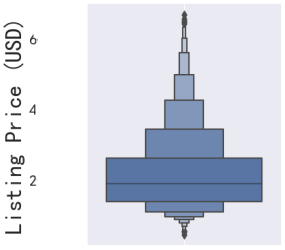


Figure 3: Final Box Plot for Monohulled SailBoats

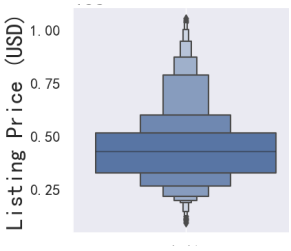


Figure 4: Final Box Plot For Catamarans

The histogram of Listing Price is nearly conforming to a normal distribution, see Figure 5:

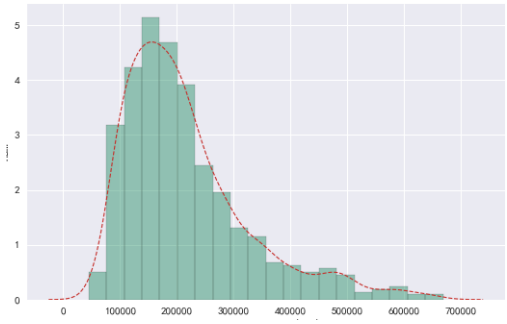


Figure 5: Listing Price (USD) in Millions

3.2 Supplement Data

3.2.1 Data Collection

The additional features'data of SailBoats we need to collect includes **beam**, **draft**, **displacement** and **sail area**, sourcing [2] [3], see in reference. Furthermore, we collect economic data by year and region from the World Bank, International Freight, Trade Association and the World Economic Forum, including annual economics and throughput, and merge them with the original data.

Because Make and Variant jointly represent a model, we connect the dataset based on this common model and build a new column of data **MakeVariant**. And then link Monohulled SailBoats and Catamarans to generate a new dataset.

3.2.2 Data Description

The newly generated dataset has about 200 rows of empty data, which has no significant impact compared to the total data of more than 3300 rows. Therefore, rows with empty data are directly deleted. Finally, we obtained a new data with columns containing ['Make', 'Variant', 'Length n (ft)', 'Geographic Region', 'Country/Region/State', 'Listing Price (USD)', 'Year', 'Make Variant', 'LWL (ft)', 'Beam (ft)', 'Draft (ft)', 'Displacement (lbs)', 'Sail Area (sq ft)', 'Average cargo throughput (tons)', and 'GDP (USD in billons)', 'GDP per capital (USD)', 'Average ratio of total logistics costs to GDP'].

Due to the similar nature, we will only take Monohulled SailBoats as an example to describe the data below.

By counting the top ten manufacturers and brands, and geographic regions in the top three, we get bar charts show in Figure 6, 7 and 8 respectively:

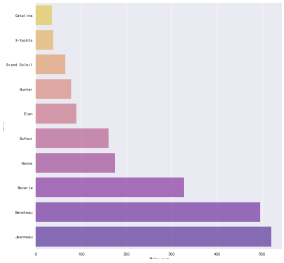


Figure 6: Top 10 Manufacturers

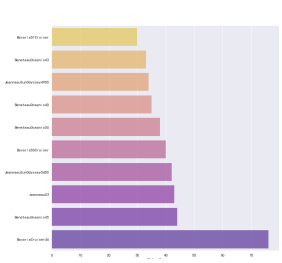


Figure 7: Top 10 Brands

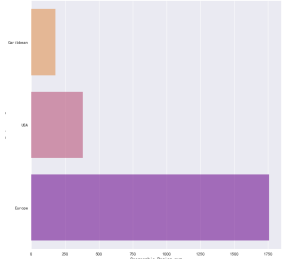


Figure 8: The Geographic Regions in Top 3

In addition, we studied whether the newly added data obeyed a normal distribution, and drew the following image:

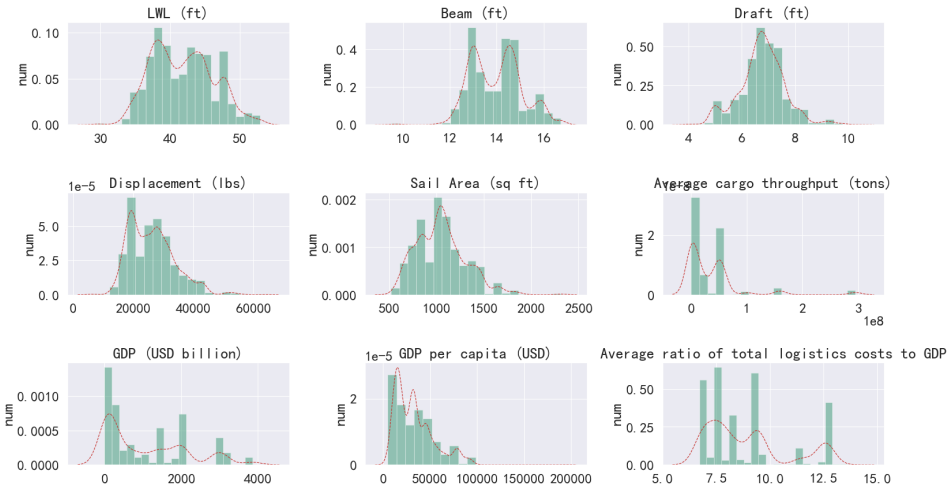


Figure 9: Normality of Different Characteristics

We found that SailBoats' related data are more consistent with the normal distribution than economic data. We simply observed various of data here, and the impact of each characteristic on the Listing Price will be discussed later.

4 Establishment of SA-LightGBM Model

4.1 Model Evaluation Indicators

We use two model evaluation indicators, including MAPE and R^2. Model evaluation refers to the use of indicators and methods to evaluate the generalization ability of the final model output from a specific method. This step typically occurs after model training and selection, and before the formal deployment of the model. The model evaluation method is not specific to the model itself, but only to issues and data. Therefore, it can be used to evaluate the generalization ability of models from different methods and select the final model for deployment.

MAPE : Average absolute percentage error

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{7}$$

Here, MAPE is used to confirm the average percentage error for the training set and the test set, respectively. The closer the trend is to zero, the more perfect the model is.

R^2 : Coefficients of determination

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2} \tag{8}$$

Here, we need to analyze the regional impact on the listing price and discuss whether the regional effects of all SailBoats' variants are consistent.

The closer the R^2 to 1, the better the model is, where the molecular part represents the sum of the square differences between the actual value and the predicted value; The denominator portion represents the sum of the square differences between the true value and the mean value.

4.2 Comparison of Basic Models

To analyze the problem, we need to conduct regression analysis through multiple characteristic factors to obtain the listing price of SailBoats. For regression problems, we use multiple models to evaluate existing data on the spot, including statistical models and machine learning models. We have searched for data on GDP per capita, national GDP, cargo throughput, and logistics as a percentage of GDP for different countries and regions to model. We use all attributes except price as training characteristics and use price as our label. Divide the training set and the test set in a ratio of 1:9, import each column of numerical type vectors as the feature vectors of the model for regression analysis, and compare the predicted values with the test set to obtain the following table.

4.2.1 Fitting Monohulled SailBoats Data

Here we list the fitting of six different machine learning models to our Monohulled SailBoats data, see in Figure 10:

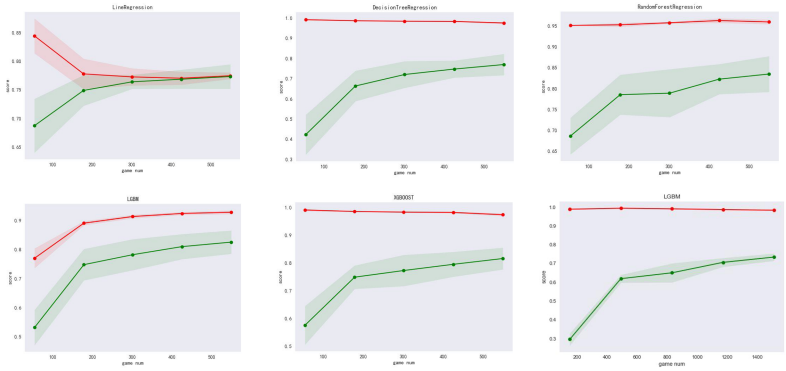


Figure 10: Fitting Graghs

For Monohulled SailBoats, we gave the following scores (see in Table 2), and found that the R^2 of LGBM and RandomForest were relatively high.

	Train MAPE	Test MAPE	R^2 Score
LinearRegression	0.261	0.254	0.5899
DecisionTree	0.135	0.206	0.6681
RandomForest	0.081	0.169	0.7546
LGBMRegression	0.050	0.174	0.7676
XGBoostRegression	0.062	0.183	0.7378
GDBTRegression	0.143	0.180	0.7227

Table 2: Basic Model Evaluation Score

4.2.2 Fitting Catamarans Data

Similarly, as with the above analysis method for Monohulled SailBoats, we also used six different machine learning algorithms for fitting and regression analysis of Catamarans data, learning curves see in Figure 11:

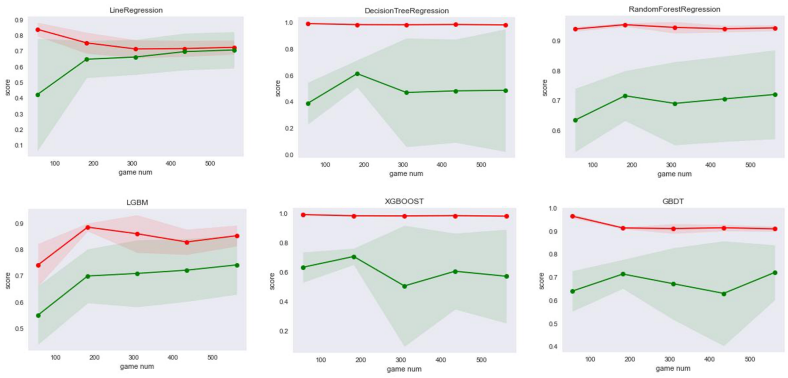


Figure 11: Fitting Graghs

For Catamarans, we gave the following scores (see in Table 3), and found that LGBM and RandomForest also had higher R^2.

	Train MAPE	Test MAPE	R^2Score
LinearRegression	0.142	0.171	0.7157
DecisionTree	0.026	0.120	0.6547
RandomForest	0.047	0.114	0.8509
LGBMRegression	0.066	0.129	0.8291
XGBoostRegression	0.031	0.119	0.7318
GDBTRegression	0.100	0.110	0.8066

Table 3: Optimization Model Evaluation Score

4.3 Model Optimization

Based on the above analysis, we finally selected LGBM as our basic model. On this basis, we further optimized the existing basic model and tried to increase the score of R^2 to a higher level to improve the accuracy of the model.

4.3.1 Data Shuffle

According to the learning curve graph of the machine learning model we drew, we found that the LGBM model has some overfitting phenomenon, which may be caused by the large data gap between the training set data and the test set data. In this case, we use the data shuffle operation to readjust the data, disrupt the original data order, and then divide the data again according to the previous proportion to generate new training and test set data, which can effectively reduce overfitting.

The shuffling algorithm has a good data scrambling effect and can meet the preconditions for big data sampling. In order to prove that the internal rules of the scrambled

dataset are not damaged after sampling, association rule analysis is performed on the pre and post sampling data using data mining methods. By comparing the support and confidence levels of the obtained association rules, as well as the frequency of transactions, it was found that the changes in the association rules obtained from the shuffled data before and after sampling were relatively stable. By comparing the time efficiency and overall sampling error of existing algorithms, it was further theoretically concluded that big data sampling is effective, that is, the overall situation of the data can be inferred through sampling samples.[4]

Based on the data, we have drawn a fitting curve between the actual value and the predicted value. The lower left figure shows the data before the data shuffle, and the right figure shows the data after the data shuffle. After the data shuffle, the fitting rate has significantly improved.

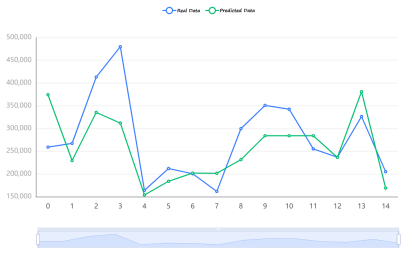


Figure 12: Before Shuffle

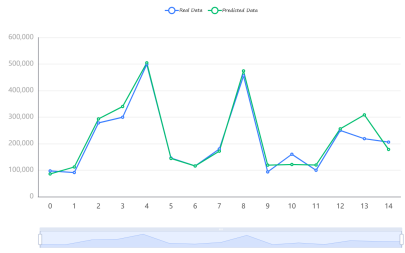


Figure 13: After Shuffle

4.3.2 Hyperparametric Optimization for Grid Search

In machine learning, hyperparametric optimization aims to find the best hyperparameters that enable machine learning algorithms to perform on validation datasets. These hyperparameters return an optimization model that reduces predefined loss functions, thereby improving the prediction or classification accuracy of given independent data. So how to optimize superparameters is crucial.

In order to conduct hyperparametric optimization, we generally use grid search, random search, and Bayesian optimization algorithms. Random search requires a sufficiently large set of sample points, and Bayesian optimization algorithms are prone to falling into local optima. Grid search is the most widely used hyperparametric search algorithm, which determines the optimal value by finding all points within the search range. If a larger search range and a smaller step size are used, WebEx searches with a high probability of finding the global optimal value.[5]

After super parameter optimization, the optimal parameters are 'max depth ': 9,' min data in leaf ': 13,' n estimators': 118. Modify the original parameters to the optimal parameters, use the LGBM model again, and perform model evaluation scoring. The following table is obtained:

	Train MAPE	Test MAPE	R^2Score
Monohulled SailBoats	0.065	0.157	0.8130
Catamarans	0.069	0.092	0.9022

Table 4: Best Model Evaluation Score

Compared to the previous LGBM, the score of R^2 has significantly improved, which can confirm the effectiveness of the grid search method.

4.3.3 Combination of Simulated Annealing and LSBM Model

In order to find a more optimal solution, a heuristic algorithm, simulated annealing, was attempted to optimize the LSBM solution value.

Let $S = s_1, s_2, s_3, ..., s_n$ be the solution space composed of all possible solutions, and C: S $\rightarrow$ R be a nonnegative objective function, reflecting the state S_I is the cost of the solution, the combinatorial optimization problem can be expressed as finding $S_* \in S$, making:

$$C(s_*) = MinC(s_I), \forall s_I \in S \tag{9}$$

The basic idea of simulated annealing for solving combinatorial optimization problems is to convert each combinatorial state s_I into microscopic state of a material system, while $C(s_I)$is regarded as the state s of the material. Using the control parameter T to simulate the temperature of a substance, the internal energy slowly decreases from a sufficiently high value. For each T, the Metropolis sampling method is used to simulate the thermal equilibrium process of the material system under this Ton a computer, that is, a random perturbation is made to the current state s to generate a new state $s^{'}$, the increment $\Delta C^{'} = C(s^{'}) - C(s)$ is calculated, and the probability $exp(-DeltaC^{'}/kT)$ is accepted as the new current state. Theoretically, when the number of times such a disturbance is repeated is sufficient, each probability of state S_I will appear in the final state following the Boltzmann distribution, i.e:

$$f(s_I) = Z(T) * \exp[-C(s_I)/kT] \tag{10}$$

In the above equation,

$$Z(T) = 1/\sum_I \exp[-C(s_I)/kT] \tag{11}$$

$f(S_I)$represents $S_Theprobabilitythat I$will appear in the final state.

If T drops slowly enough and Tß0, it can be seen that the final state will be the global optimal solution.[6]

Due to the small amount of data, in order to achieve better results through simulated annealing, we set the initial state temperature of T_{start} to 200 and end temperature T_{end} to 2 and iterate 200 times on each temperature, $T_{I+1} = \alpha * T_1$. Take α = 0.99 $\rightarrow$ 1 to fully iterate within the value range.

A more fitting image is obtained here by combining calculations, see Figure 14:

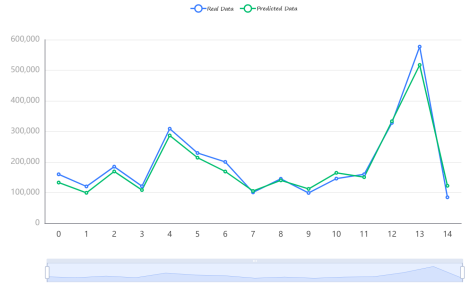


Figure 14: The Result of SA-LightGBM Model

Using this algorithm has two characteristics: first, the final result of this algorithm does not change with the initial solution; Secondly, using LSBM as a new solution generator for the simulated annealing process improves the efficiency of the simulated annealing process and keeps the total calculation time within a reasonable range, which is a practical method.[6]

5 Regional Benefits of the Model

5.1 Importance of Model Indicators

In order to examine the importance of the impact of regional indicators on the model, we conducted in-depth positive and negative impact analysis using SHAP for each feature. Due to the black box attribute of machine learning models, the model lacks transparency, and the **"Shapley Additive Interpretation" (SHAP)** in game theory can provide interpretability analysis for machine learning models[7].

Assume the i-th sample is x_i, and there are total n features, the predicted value of the model for the i-th sample is y_i, the baseline of the entire model is the mean of all sample target variables, defined as y_{base}, the calculation formula for Shapley value is as follows:

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + ... + f(x_{in}) \tag{12}$$

where $f(x_{i1})$ is the Shapley value of the first feature of the sample i. When $f(x_{i1})$is greater than θ, the feature has a positive effect on the predicted value, that is, a positive effect; on the contrary, when $f(x_{i1})$is less than θ, this feature has a downward effect on the predicted value, i.e., a reverse effect. Although the Shapley value calculation formula for a single sample is similar to linear regression, through global analysis of the Shapley values for all samples, the SHAP method can obtain the impact of each feature on the linearity and nonlinearity of the predicted value. In addition, the SHAP interpretation method can also visualize the results, making the impact of features more intuitive.[8]

For Monohulled SailBoats and Catamarans, we draw the following figures:

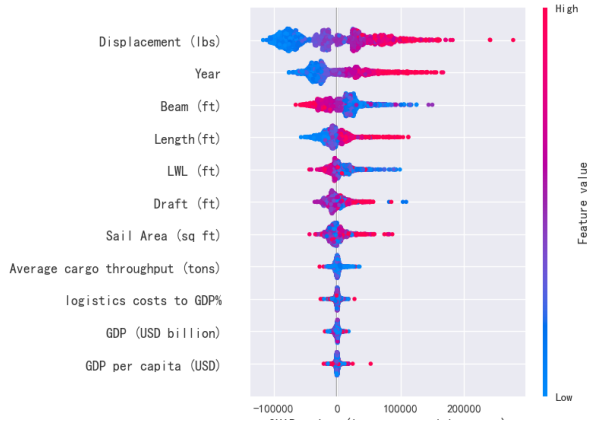


Figure 15: SHAP Value of Monohulled SailBoats

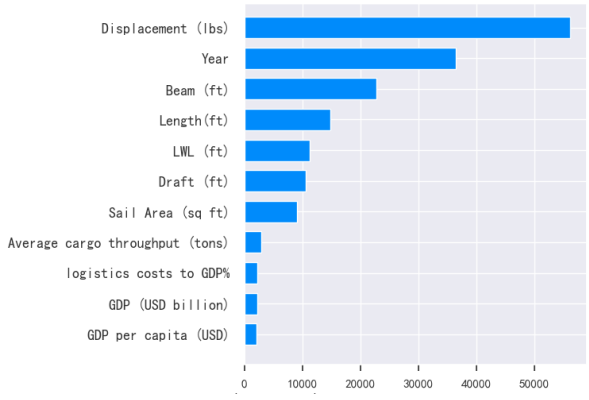


Figure 16: Mean |SHAP Value| of Monohulled SailBoats

From the above figures, it can be found that for Monohulled SailBoats, **Displacement** poses the greatest positive impact while **Beam** exerts largest negative impact.

The rest can be analyzed in turn. And all economic attributes have a certain degree of positive correlation.

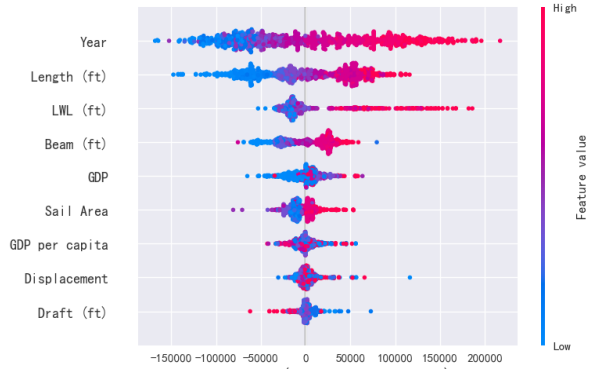


Figure 17: SHAP Value of Catamarans

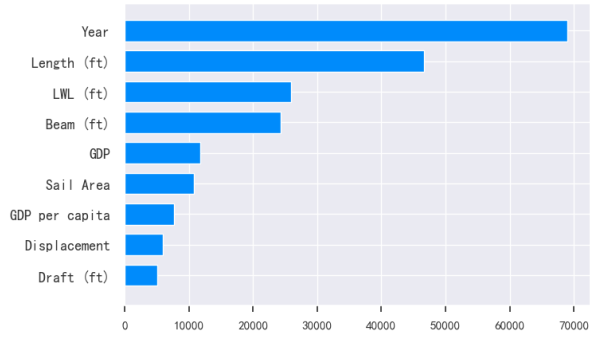


Figure 18: Mean |SHAP Value| of Catamarans

From the above figures, it can be found that for Catamarans, **Year** has the hugest impact, and the rest of the data can be analyzed similarly. For economic attributes, except for **GDP per capital**, which has a certain positive correlation, other attributes have no significant correlation or impact.

The characteristic factors that we divided by region, such as' Average Cargo Throughput ',' GDP ',' GDP per capita ' and' logistics costs to GDP ', all have little impacts on both types of SailBoats. The most significant impacts reside with attributes of the ship itself. It is peculiar that the degree of impact of **Displacement** on Catamarans is far less than that of Monohulled SailBoats, which indicates that the effects of various characteristics on different ships are not consistent.

5.2 Two-factor Analysis of Variance

Two-factor ANOVA is a statistical analysis method that can be used to analyze whether the different levels of two factors have a significant impact on the final result.[9] There are two types of variance analysis. One is a two factor analysis of variance without interaction, which assumes that the effects of factor A and factor B are mutually independent. The other is an interactive two factor analysis of variance, which assumes that the combination of factor A and factor B will produce a new effect. Here we use a **two factor analysis of variance with interactive analysis**.

The error expressions of the two-factor ANOVA are as follows:

Inter group error(sum of squares for factor), denotes as SSF:

$$SSF = \sum_{j=1}^{c} n_j (\bar{x}_j - \bar{\bar{x}})^2 \tag{13}$$

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{c} \bar{x}_j}{n} \tag{14}$$

$$n = \sum_{j=1}^{c} n_j \tag{15}$$

Intra group error(sum of squares for error), denotes as SSE:

$$SSE = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \tag{16}$$

Summary formula for errors:

$$Sum\ of\ Squares\ Total(SST) = SSE + 1stSSF + 2ndSSF \tag{17}$$

We first merged the **Make** column and the **Variant** to form the unique models of SailBoats, which is our **MakeVariant** column.

Then we used the **MakeVariant** column and the **Control/Region/State** column as grouping variables, and the **Listing Price (USD)** as the dependent variable, to perform a bivariate analysis of variance. Results show in the following tables:

	MakeVariant	Country/Region/State	Error
Sum of squares	$4.03 * 10^{15}$	$1.08 * 10^{12}$	$4.39 * 10^{12}$
Free degree	500	71	1772
Mean square	$8.06 * 10^{10}$	$1.52 * 10^{10}$	$2.48 * 10^{9}$
F	32.516	6.156	
P	0.000***	0.000***	NaN
R^2		0.914	
Adjusted R^2		0.886	

Table 5: The ANOVA Results of Monohulled SailBoats

	MakeVariant	Country/Region/State	Error
Sum of squares	$3.03 * 10^{13}$	$6.30 * 10^{13}$	$2.4 * 10^{12}$
Free degree	134	48	700
Mean square	$2.26 * 10^{11}$	$1.31 * 10^{11}$	$3.5 * 10^{9}$
F	65.208	37.916	NaN
P	0.000***	0.000***	NaN
R^2		0.948	
Adjusted R^2		0.915	

Table 6: The ANOVA Results of Catamarans

From the above table, we can see that from the analysis of the results of the F test, the significance P value under **MakeVariant** is 0.000***, showing a significant level, indicating that there are significant differences among different models of SailBoats. The conclusion is the same as **country/region/state** column, that is, significant differences exist on listing prices of SailBoats in different regions.

Figure 19 and 20 show the results of the mean values of the bivariate analysis of variance. By comparing the mean values of different grouping variables and the crossover situation (usually means interaction), we can explore their differences.

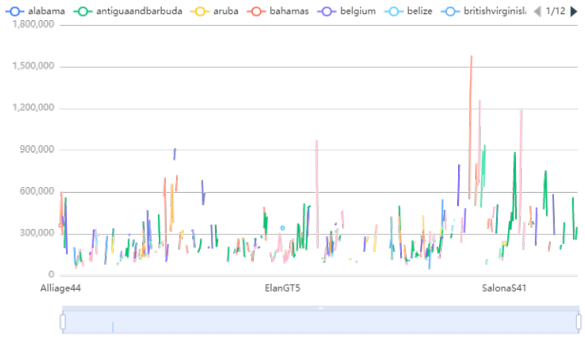


Figure 19: Mean Values of Monohulled SailBoats

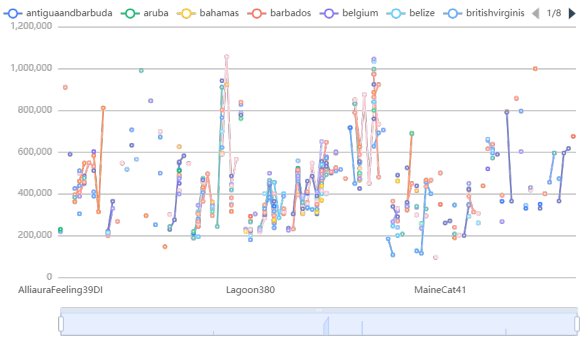


Figure 20: Mean Values of Catamarans

Due to the differences between the regional indicators we are currently considering, including city names and some economic data, we have tried to explore the impact of these indicators on the price of SailBoats, but the actual effect is not significant. Logically, apart from 'GDP per capita', it is true that these economic data are not significantly related to SailBoats, but rather to cargo ships. Therefore, we will seek other data that can refer to regions, such as national coastline, whether coastal countries or not, longitude, latitude and so on, which will be discussed in detail later.

6 Specify An Area for Model Verification

In this question, we need to migrate our model for a given geographical area to the Hong Kong (SAR) market to play a role. Here, we believe that the price of SailBoats may be related to the following factors:

1. **Per capita GDP of a certain region**: The price of SailBoats is relatively expensive, and talents with certain economic strength may choose to purchase SailBoats to improve their quality of life.

2. **Cargo throughput in a certain region**: If a region can transport more cargo, it may also transport more SailBoats.

3. **Logistics as a percentage of GDP**: If a region's logistics as a percentage of GDP is relatively high, it will be more likely to purchase more goods, or more likely to purchase SailBoats.

Since we have not found accurate price data for different SailBoats in the Hong Kong market, we want to simulate the characteristics of Hong Kong through the characteristics of per capita production value, cargo throughput, and the proportion of logistics to GDP in Hong Kong, and thereby predict the corresponding SailBoats prices in Hong Kong.

6.1 Forecast the Selling Price of SailBoats in Hong Kong

The following is the relevant data we collected from Hong Kong:

After completing the above model training, we replaced all the economic data in the corresponding records of different types of SailBoats with the economic data of

City	Average exports of goods	GDP ($100 million)	GDP Per Capita ($)	Percentage
Hong Kong	399200000	341	45638	3.3

Table 7: Hong Kong Related Data

Hong Kong, and again made predictions for our new model. From the above training results, we selected the most effective LGBM to predict our data, and finally obtained the following results.

Box diagram showing the distribution of selling prices for Monohulled SailBoats see in Figure 21:

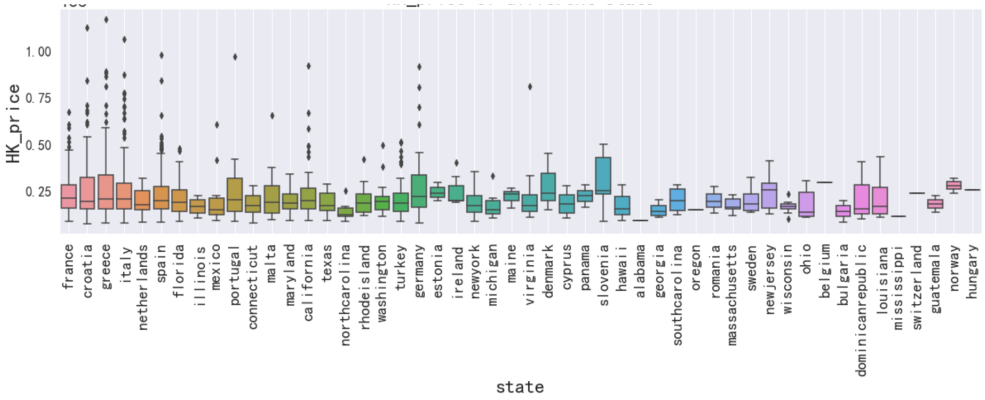


Figure 21: Hong Kong Price of Different States for Monohulled SailBoats (in millions)

From the above figure, we can observe how the sales prices of used SailBoats originally manufactured in different geographical regions will be distributed in Hong Kong. We have noticed that the price of most second-hand SailBoats will be around 0.25, but there are many outliers such as France, croatias, greece, Italy, Spain, California, turkey, and Germany. These outliers represent relatively high prices for SailBoats sold in Hong Kong.

The following is a box diagram of Catamarans' sales price distribution:

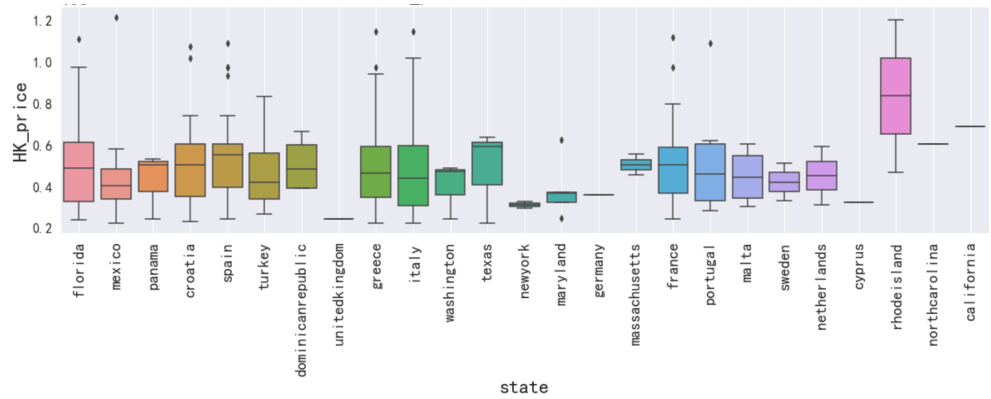


Figure 22: Hong Kong Price of Different States for Catamarans (in millions)

We observe that compared to Monohulled SailBoats, the distribution of second-hand SailBoats from different countries in Catamarans is relatively more concentrated,

and the distribution is mainly around 0.5. That is, compared to Monohulled SailBoats, the price of Catamarans will be relatively mainly distributed in a more expensive range, and there will be relatively few outliers.

6.2 The Impact of Different Characteristics on the Listing Price of SailBoats

Figure 23 below shows the impact of our different characteristics on the price of the Monohulled SailBoats. The closer the color is to red, the stronger the positive impact is, and the closer the color is to blue, the stronger the negative impact is.

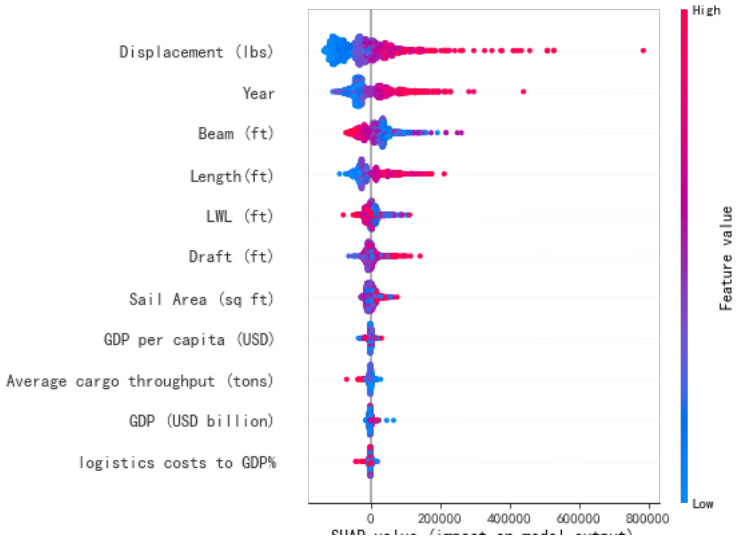


Figure 23: SHAP Value(impact on model output) of Monohulled SailBoats

1. We have observed that **Displacement** has the most influential positive impact on the price of our sales boards. That is, the larger the Displacement of a sail boat, the more expensive its price will be.

2. The closer the **year** is to the present, i.e. the closer the data set is to 2019, the more expensive the price of SailBoats is.

3. Meanwhile, we have noticed that the **Beam** parameter and the **LWL** parameter will have a negative impact on the price of our SailBoats to a certain extent.

Figure 24 below shows the impact of our different characteristics on the price of the Catamarans. The closer the color is to red, the stronger the positive impact is, and the closer the color is to blue, the stronger the negative impact is.

1. Unlike Monohulled SailBoats, the most influential factor for the price of Catamarans is the **Year** factor, which means that the closer the production time of the ship is to the present, the higher the price can be obtained.

2. At the same time, the factors of **Length**, **Sail Area**, and **LWL** will be the secondary influential factors on the price of SailBoats immediately after the Year factor, and these three factors have a positive impact on the price. In contrast to our analysis of Monohulled SailBoats mentioned above, it is possible that a smaller Monohulled SailBoats would result in higher prices, while for Catamarans, larger hulls are more likely to sell at higher prices.

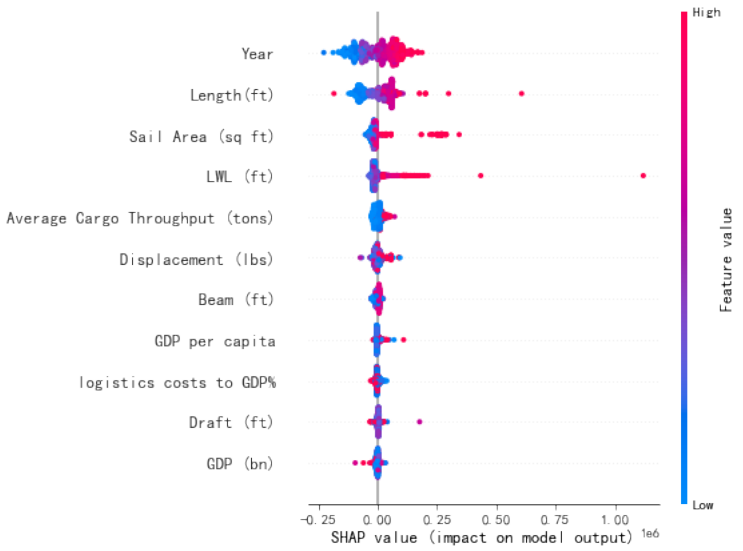


Figure 24: SHAP Value(impact on model output) of Catamarans

6.3 Modelling the Regional Impact of the Hong Kong (SAR) on Sail-Boats Prices

Here, we use the **paired sample T-test method** to compare our above estimated results from the Hong Kong SAR in the data of Monohulled SailBoats and Catamarans with our real data, and use it to compare the difference between our prediction and the real situation of the two fractions.

6.3.1 Paired Sample T-test

If the difference of two paired samples x_{1i} and x_{2i} is independent and conforms to a normal distribution, then whether the parent expected value μ of d_i is equal to μ_0 can utilize the following statistics:

$$t = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}} \tag{18}$$

When $\mu = \mu_0$, it obeys the t-distribution with $n - 1$ degrees of freedom.

6.3.2 Impact of Hong Kong(SAR) Market on Monohulled Sailboats

The result of paired sample T-test shows that based on the variable true Price pairing simulation HK Price, the P value is 0.050**, showing significance at the level, rejecting the original hypothesis, so there exists differences between true Price and simulation HK Price. The Cohen'sd value of the difference is 0.054, which is very small.

6.3.3 Impact of Hong Kong(SAR) market on Catamarans

The result of paired sample T-test shows that based on the variable true Price pairing simulation HK Price, the P value is 0.000***, showing significance at the level, rejecting the original hypothesis, so there exists differences between true Price and simulation HK Price. The Cohen'sd value of the difference is 0.258, which is mild.

Pairing variable	The true Price matches the simulation HK Price
Pair 1	236513.005±157958.795
Pair 2	239242.583±140976.945
Pairing difference	-2729.579±16981.851
t	-1.962
df	1298
P	0.050**
Cohen'sd	0.054

Table 8: T-test Results of Paired Samples

Pairing variable	The true Price matches the simulation HK Price
Pair 1	460987.855±254378.624
Pair 2	495198.76±215184.377
Pairing difference	-34210.905±39194.247
t	-4.056
df	2470
P	0.000***
Cohen'sd	0.258

Table 9: T-test Results of Paired Samples

From the above test data, we can observe that the Hong Kong SAR market has a negative impact for both Monohulled SailBoats and Catamaran. The impact on Catamarans is greater than Monohulled SailBoats.

7 Remaining Characteristics of Data

Average Selling Price of Monohulled SailBoats in Different Regions

We grouped and summed the data from Monohulled SailBoats by country/region/state, averaged them, and presented them on the map, see in Figure 25. The closer the color on the map is to red, the higher the average selling price is, and the closer the color is to blue, the lower the average selling price is.

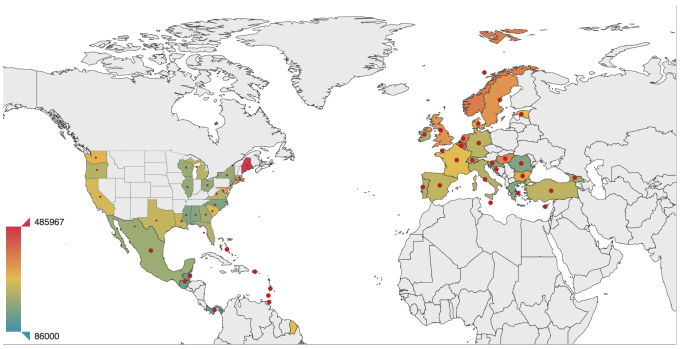


Figure 25: Average Sales of Monohulled SailBoats in Different Regions (US dollar)

It is found that the average total sales price in Maine State is relatively high, followed by some countries or regions in northern Western Europe.

These countries have relatively long coastlines, relatively developed waterways, and numerous islands that are suitable for navigation.

At the same time, we have noticed that the average selling price of Monohulled SailBoats is gradually increasing along the Atlantic coast of Europe as the latitude gradually increases.

For the United States, the eastern coastline of the United States is relatively longer, so the sales distribution of Monohulled SailBoats is also more extensive. At the same time, we note that the states with higher total sales in the United States tend to be those with more economic strength.

From the perspective of wind belts, most of the areas in the picture belong to the parts of the prevailing westerlies, and there are sufficient wind factors to drive the development of SailBoats.

At the same time, we have noticed that several states around the Great Lakes in the United States also have significant average sales of Monohulled SailBoats. Could this possibly mean that people will also engage in sailing in the Great Lakes.

Average Selling Price of Catamarans in Different Regions

Catamarans has a higher average selling price lower limit and a higher average selling price upper limit for Monohulled SailBoats. Unlike Monohulled SailBoats, Catamarans has the highest average sales in Rhode Island. At the same time, there is no longer a significant trend in Western Europe, such as Monohulled SailBoats, where the average selling price gradually increases as the latitude increases.

And we can notice from the colors in the figure that the average sales volume of Catamarans is mainly displayed in blue or light green, which is at the lower middle level in the range of the total sales price distribution. However, for Monohulled SailBoats, the main colors in the graph are orange or yellow-green, which means that their average sales prices are in the middle or upper range of the range distribution, see in Figure 26:

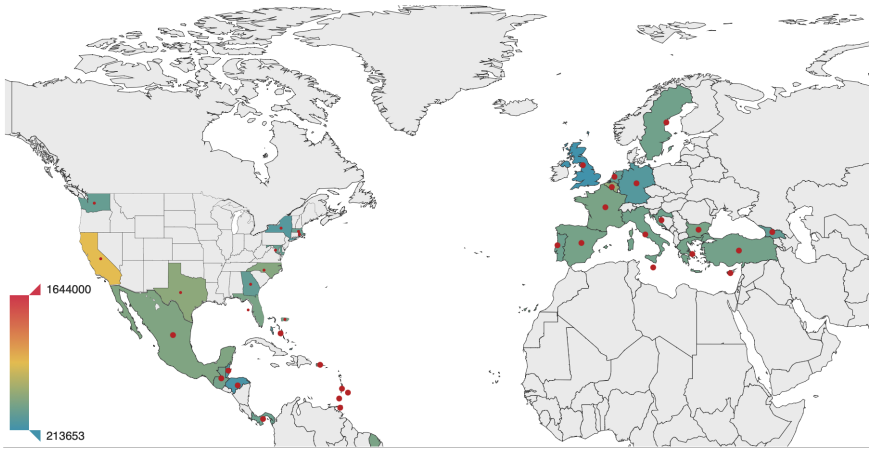


Figure 26: Average Sales of Catamarans in Different Regions (US dollar)

Total Sales Price of Monohulled SailBoats in Different Regions

We can observe from the figure 27 below that the total sales volume of Monohulled SailBoats from the south to northern Western Europe is also gradually decreasing. We speculate that the reason may be that the closer to the Arctic, the more floating ice on the sponge, and the less accessible roads.

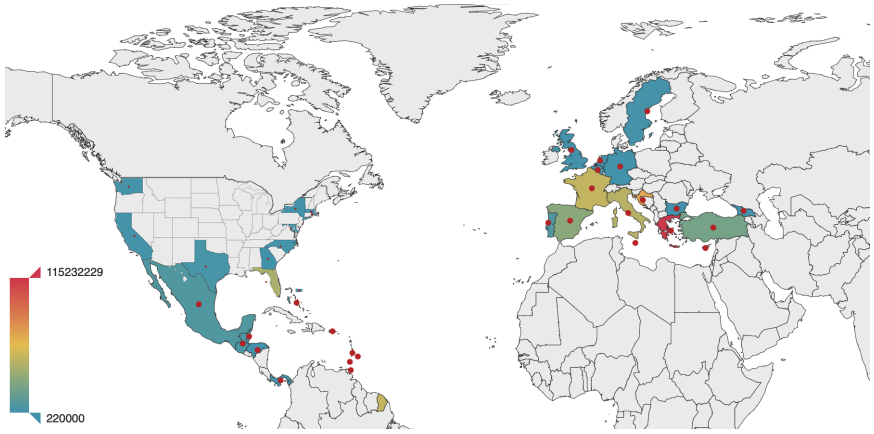


Figure 27: Total Sales of Monohulled SailBoats in Different Regions (US dollar)

At the same time, there are the North Atlantic Warm Current and the Gulf of Mexico Warm Current along the Atlantic coast of Europe and the Atlantic coast of the eastern United States, respectively. The water temperature is relatively warm, and it will also carry more warm water vapor.

On the western Pacific coast of the United States, there is a California cold current, where the water temperature will be relatively cold and the air will be relatively drier.

It can be seen from the figure that people in the sea area under the influence of the warm current may be more willing to engage in sailing sports.

Total Sales Price of Catamarans in Different Regions

The region with the highest total sales price for Catamarans is Greece, followed by Croatia, Fcrane, and Italy.

As can be seen in Figure 28, Catamarans are also relatively more purchased in regions along the Mediterranean coast.

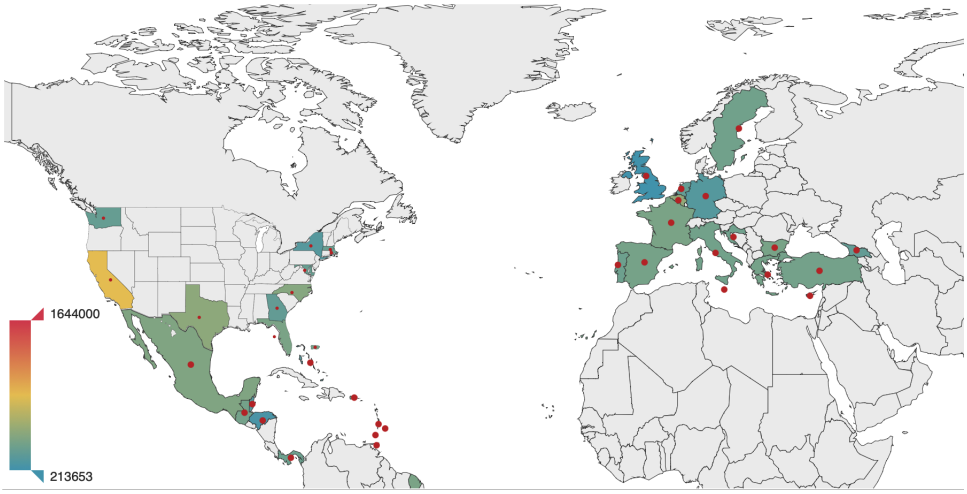


Figure 28: Total sales of Monohulled SailBoats in different regions (US dollar)

Analyze Time Changes

We have summarized the data of Monohulled SailBoats and Catamarans for different years to obtain the number of SailBoats sold in each year. Then we divided it by region and marked it on the map. The following charts show the trend of changes in the number of ships sold from 2005 to 2019.

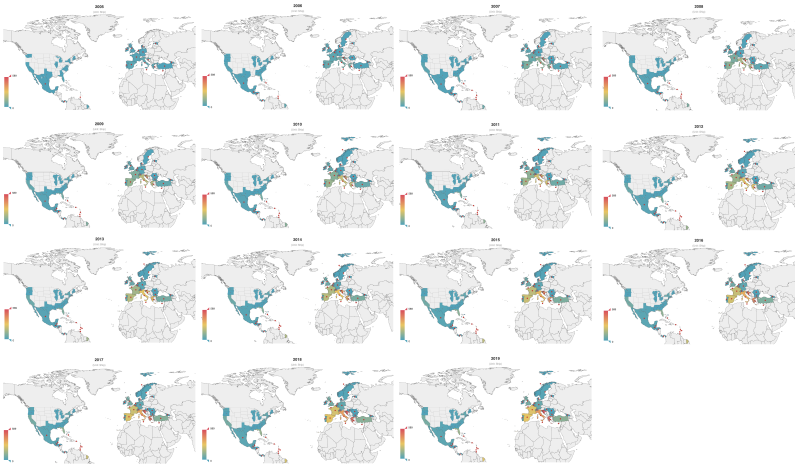


Figure 29: Total Sales of Two SailBoats(US dollar)

For the European region, we can see that over time, there has been a significant increase in the sales of ships from Greece, Croatia, Fcrane, and Italy, especially along the Mediterranean coast.

For the United States, we can see that over time, more regions in the eastern United States have purchased more SailBoats, but the overall growth trend is not as obvious as in Europe.

Analysis of the Proportion of Manufacturers

In our collected data, Jeanneau manufacturers and Beneteau manufacturers account for the highest proportion, followed by Bavaria, Hanse, and Dufour, see in Figure 30:

And from 31 ,we can find that the Bavaria Cruiser is the most popular ship model.

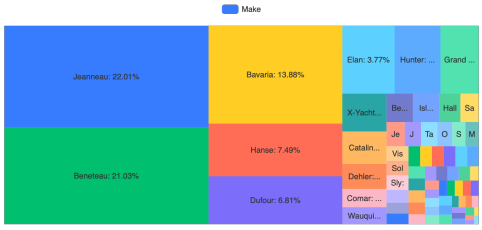


Figure 30: Proportion of Manufacturers

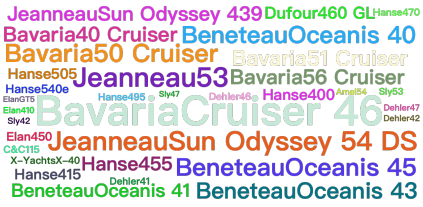


Figure 31: WordCloud

8 Conclusions

The main purpose of this article is to establish a regression model that can evaluate the price of SailBoats, and the model needs to be able to obtain more appropriate prediction results based on different regions. Based on existing data, try to select appropriate features for model training, and be able to obtain the impact of each feature on prediction. Then, appropriate analysis methods are used to obtain the prediction results for the Hong Kong region. In the end, we established the SA LightGBM model, but due to insufficient data acquisition, our model cannot obtain a more suitable explanation in some aspects. Of course, we hope that our model can be helpful for the development direction of this problem.

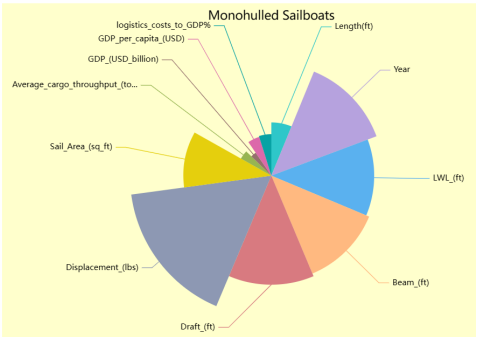
A Report on The Pricing of Used SailBoats

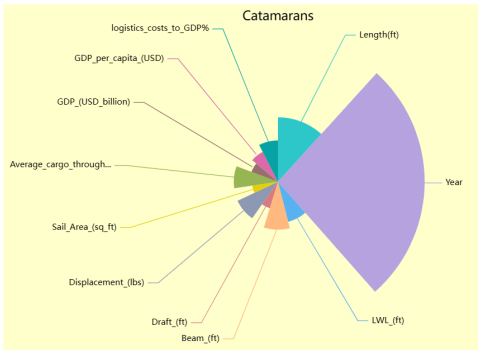
To: SAR Sailing Broker
From: Team # 2330185
Date: April 3rd, 2023
Subject: Hong Kong Market Analysis Report

Dear broker,

It's our pleasure to be commissioned to give advice on the pricing of used SailBoats. After analyzing the data, we have drawn on four conclusions, and let's explore them in more details.

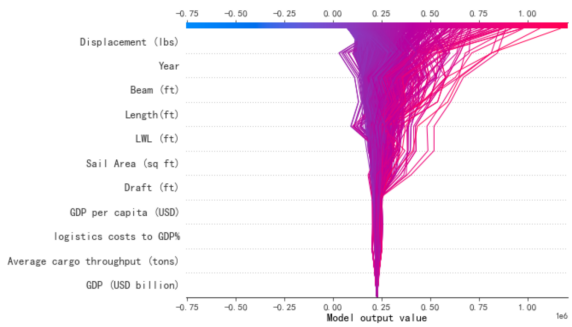
Importance between different characteristics. Different characteristics have different influences on the prediction results of our model. After using regression models to fit the data, we rank the importance of each feature, and the distribution can be seen intuitively in the following figures:

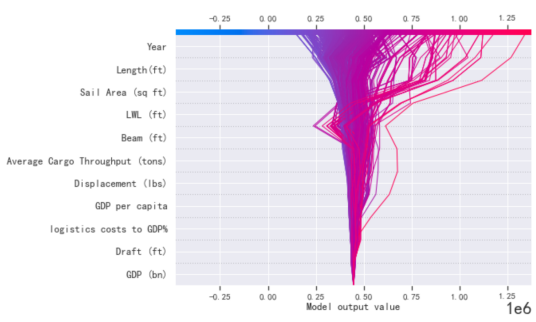




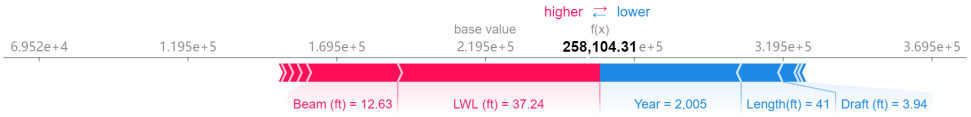
For **Monohulled SailBoats**, **Displacement** is the most important influencing factor for our sales price forecast, followed by **Year**, **Draft**, **Beam** and **LWL**. While for **Catamarans**, **Year** exerts the hugest impact, followed by **Length**. The remaining factors are not particularly important for selling price.

Impact on overall data. In the figure left below, we found that for **Monohulled SailBoats**,**Displacement (lbs)** is the most important factor that has a **positive** impact on our prices. However, **Year**, **Beam**, **Length**, and **LWL** have a **negative** impact on the price of SailBoats to a certain extent. This may indicate that when customers choose to purchase Monohulled SailBoats, smaller boats will receive higher prices. As for **Catamarans**, the first three factors that have the most significant impact are the **Year**, **Length (ft)** and **Sail Area (sqft)** in sequence. All have a certain degree of negative impacts on the pricing.



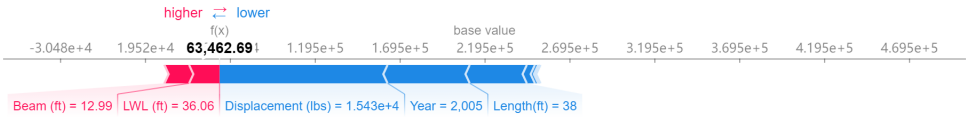


Impact on individual data. For a single sample, the influencing factors of its selling price will be affected by its own different characteristics. As shown in the figure below:

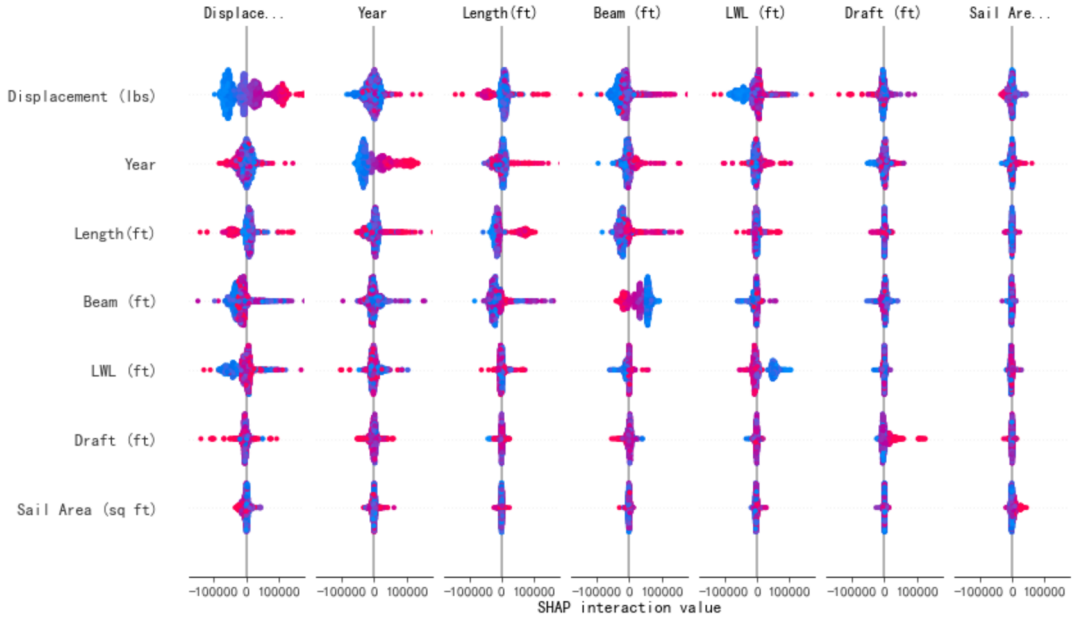


LWL (ft) has the most significant positive impact on the selling price (the longest red stripe), **Beam** secondary. While the **Year** factor has the most significant negative impact (the longest blue stripe), followed by **Beam (ft)** and **Draft (ft)**.

Of course, if we take a different piece of data, we will find that its different characteristics will have different influencing factors on our model. As shown in the figure below, we note that Displacement (lbs) has a significant negative impact on the price of our model, followed by Year and Length (ft), which have a negative impact on our sales price. The positive impact of Beam (ft) and LWL (ft) on our selling price is relatively weak.



This is because different features also have a certain degree of influence relationship, and the increase or decrease of one feature may also lead to the increase or decrease of other features. For example, a change in the size of a Displacement may lead to a change in the size of the hull, and thus drive a change in our final selling price. Below is a distribution diagram of our different features.

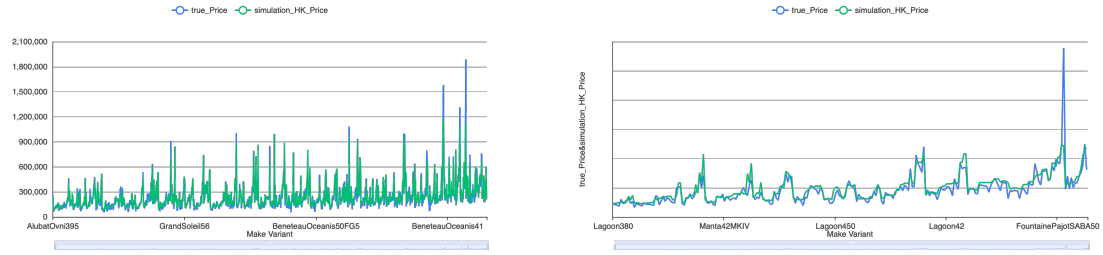


The impact of Hong Kong on the selling price. We have collected some factors that may affect prices by geography to simulate the sales environment in Hong Kong.We first query the above characteristics, and then we replace all the region-related features

in the above data with the relevant characteristics data of Hong Kong.Finally, we use our trained model above to make predictions for our Hong Kong data.

After that, we conducted a paired sample T-test between the predicted data and the actual selling price of the ship, and compared the impact results on Monohulled SailBoats data and Catamarans data. We found that the regional impact factors in Hong Kong will more significantly affect the selling price of Catamarans type ships compared to SailBoats.

The following figure on the left shows the impact that the regional impact of Hong Kong may have on Monohulled SailBoats. We can see that the predicted selling price in Hong Kong may be relatively higher than our previous selling price, but the increase is not particularly significant. The following image on the right shows the impact that the geographical impact of Hong Kong may have on our Catamarans ships. We have noticed that the price change is more significant compared to Monohulled SailBoats.



Therefore, in Hong Kong, the Catamarans ship may have a more significant sales price increase than the Monohulled SailBoats ship to some extent.

Yours sincerely,

Team # 2330185

References

[1] Yuanhui Xiao. A fast algorithm for two-dimensional kolmogorov-smirnov two sample tests. *COMPUTATIONAL STATISTICS & DATA ANALYSIS*, 105:53–58, JAN 2017.

[2] `https://www.sailboatdata.com`.

[3] `https://www.yachtworld.co.uk`.

[4] Zhang Chunsheng Liu Hanyue. Effectiveness analysis of big data sampling based on shuffle algorithm. *Computer Application Research*, (3049-3054), 2021.

[5] Liu Jiaxing. Support vector regression based on grid search hyperparametric optimization. *Scientific and technological innovation*, (71-74), 2022.

[6] Huang Minxiang Ni Qiulong. Application of simulated annealing algorithm based on branch switching in distribution network planning. *Journal of Electric Power Systems and Automation*, (31-35), 2000.

[7] Yuren Yang, Ye Yuan, Zhen Han, and Gang Liu. Interpretability analysis for thermal sensation machine learning models: An exploration based on the shap approach. *INDOOR AIR*, 32(2), FEB 2022.

[8] Wang Lei Liu Tianchang and Zhu Qinghua. Research on user churn prediction of smart elderly care service platform based on shap interpretation method. *Data Analysis and Knowledge Discovery*.

[9] Hancun Dai Jinhui. Comparison of two factor anova methods. *Statistics and decision-making*, 34(4):30–33, 2018.