Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2332257

Used Sailboat Market Dynamic State:

Price Analysis and Prediction Model of Comprehensive Factors

Summary

The sailboat industry is capital-intensive, and sailboats are the primary production material with high economic value. The acquisition of used sailboats can avoid construction risks and has the characteristics of low investment and quick results, which has ample market space. In this work, we use various regression and prediction models to analyze used sailboat price data and its related background information and provide effective price evaluation solutions for brokerage companies.

Firstly, develop a model of the factors influencing the price of sailboats. We analyze the provided data, more data about additional attributes, geographical factors, and economic factors, including quantitative analysis of *dummy variables.* Then we obtain *the degree of influence of each element* using *Random Forest*, and evaluation of the effect through data set division tests. It shows that Displacement, Length, and Age had the highest degree of influence, while there were some differences in the percentage of influence in the two types of sailboats.

Secondly, we further investigate the geographic effects of used sailboat prices. We continue to add geographic and economic factors related to geography and analyze six factors through *Pearson Correlation Coefficient*, concluding that *"geographic factors are weaker than manufacturing factors, but can assist in price determination"*. Thus, with the nine factors combined, *Ridge Regression* is used for price forecasting, with a final *R2_score* of *67.56%* (Monohulled Sailboats) and *75.39%* (Catamarans).

Thirdly, Hong Kong (SAR) is used as a sample for the application of the model. We collect *30 detailed data sets* of used sailboats in Hong Kong (SAR) and analyze their price data characteristics. Then a subset of the given dataset was divided by combining the price characteristics. Regression analysis is performed using this dataset according to the previous model to reflect the accuracy of the price trends under manufacturing and economic effects.

Finally, continue the data mining work. There are a number of valuable *intrinsic regulations*: (a) Monohulled Sailboats prices can be better fitted using only two core factors; (b) there is a clear turning point in the effect of Displacement on Price; (c) as the age of the sailboat increases, there is a clear trend of decreasing in price, also including some special time point.We also provide a report for sailboat broker.

Keywords: Random Forest, Multiple Linear Regression, Ridge Regression, Pearson Co rrelation Analysis, Multivariate Integrated Analysis

Contents

1 Introduction

1.1 Problem background

As one of the luxury goods, the value of used sailboats decays due to various factors, which include internal and external factors. Therefore, using mathematical models to evaluate used sailboat prices is beneficial for decision-makers to understand the trend of sailboat prices， thus promoting the development of services such as sailboat trading and sailboat financing.

We build the model with the help of data provided by a boating enthusiast and other publicly available data sources [1] to enable Hong Kong (SAR) sailing economy companies to have this information. To achieve this goal, the following three questions need to be solved and answered：

(P1) In order to describe the sailboat condition for different variants specifically, more additional parameters from the official website [1] need to be considered based on their physical significance. With this price model, the prices of the different variants and their precision can be predicted and calculated.

(P2) Use the model to explain the region's effect on prices and find generalized patterns across all variants, exploring their practical and statistical significance.

(P3) Using a specific region (Hong Kong (SAR)) as a reference, simulate the regional effect of this region and compare the prediction with the real data. In addition, explore the correlation of regular between the two types of sailboats.

(P4) Combine rich data and continue to search for valuable conclusions and discover intrinsic rules.

(P5) Write a brief report about the price model for the Hong Kong (SAR) sailboat broker.

1.2 Outline of our work

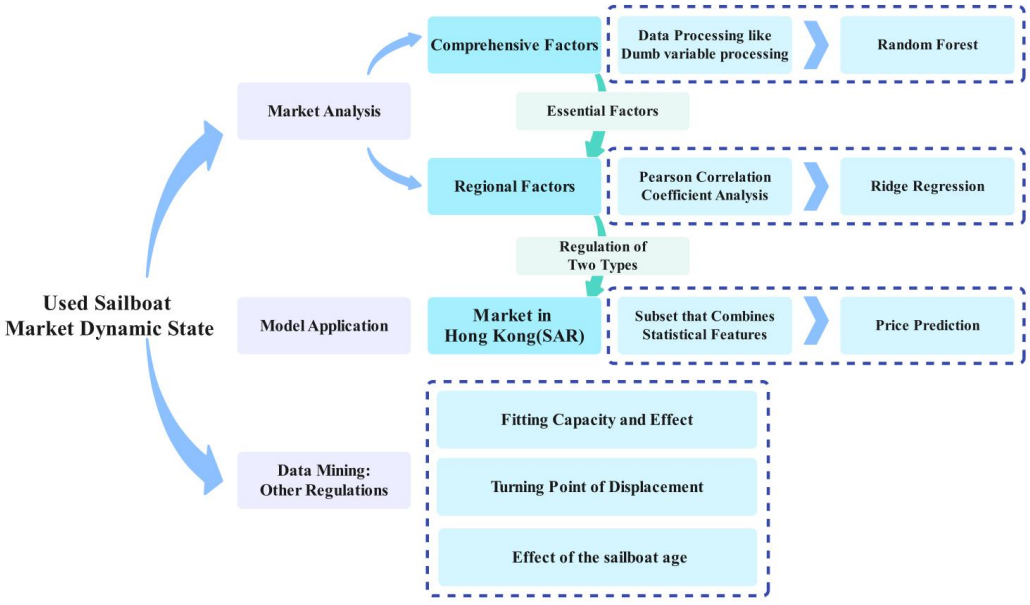


Figure 1: Our work

2 Preliminaries

In this section, we give some preliminaries to simplify the solving of the problems. We first make the following proper assumptions and give the corresponding justifications.

Assumption 1: The price data provided by the sailboat enthusiasts is highly informative and is the market average.

Justification: The design, construction and operation of sailboats are highly technical and specialized in engineering [2], and sailboat prices are volatile and influenced by suppliers and other factors, we consider the data provided by Sailing Enthusiast to be an average of the relevant variant sailboat.

Assumption 2: Ignore the special wear and tear situation of the sailboat

Justification: In the actual scenario of used boat trading, it is necessary to consider not only the internal and external factors of the sailboat, but also the specific condition of the sailboat, such as physical wear and tear, functional wear and tear, etc., without considering the impact of these special factors on the sample.

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Some useful notations in this paper

Symbol	Description
y	Price
x_1	Length
x_2	Displacement
x_3	Age(2020-year)
x_4	GDP
x_5	GDP per capita
x_6	Coastline Length
x_7	Coastline Length and Area Ratio(Coastline/Area)
x_8	Trade Freedom
x_9	Financial Freedom
x_{10}	Draft
x_{11}	Hull Materials
x_{11}	Sail Area

Other variables will be explained if first used.

3 Used sailboat price model

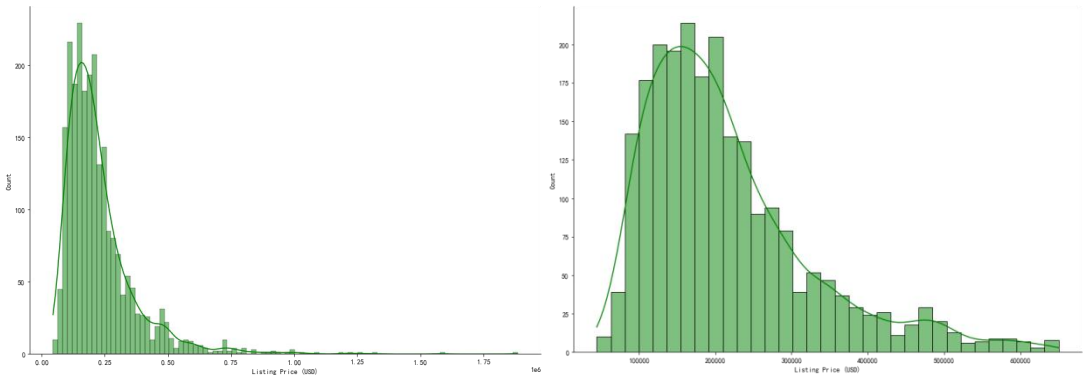
3.1 Data pre-processing

3.1.1 Data cleaning

Based on the data provided by sailing enthusiasts for Comap (including 2347 groups of

Monohulled Sailboats and 1146 groups of Catamarans), specific data on the manufacturer, length, region, year, and price of the different variants of sailboats can be obtained. In order to facilitate subsequent model building, we do two pre-processing aspects.

First, the sample data containing missing values are removed; second, a price distribution plot (Figure 2(a)) is made, from which it can be seen that there are scattered extreme values shifted in the data as a whole, and for the 2% of data with extreme prices, we also eliminate them. After such processing, a new price distribution plot (Figure 2(b)) can be obtained, which is better concentrated.



(a)before **(b)after**

Figure 2: Price distribution statistics

To facilitate the analysis of different variants and ensure the statistical significance of the data, we excluded all variant types with data less than 6. Subsequently, we divided the data set and evaluated the model accuracy on this basis.

3.1.2 Data Extension

In addition to the data provided by boating enthusiasts to COMAP, we also need other auxiliary materials to enrich the understanding of different variants of sailing.

✧ Additional manufacturing attributes

SAILBOATDATA.COM[1] is the most original and comprehensive database containing information on more than 8800 production and semi-production sailboats dating back to 1900. Most include photographs and drawings from the original plan and brochure library. [1] The site contains a wealth of information and more comprehensive data on the attributes of new and used boats, as well as data for comparison of different boat types.

✧ Economic data

In addition to factors related to the condition of sailing itself, there are other external factors that are worth considering. We used the GDP and GDP per capita of countries around the world (including U.S. states) in 2020 (since the data provided by boating enthusiast are from 2020, only the state of the market economy in that year can be analyzed) published by the United Nations Conference on Trade and Development [3]. Also, we obtained data from the data provided by the Heritage Foundation [4] to obtain data on the freedom of trade,

economic freedom, and monetary freedom for each country in the world in 2020.

3.1.3 Dummy variable handling

In sailboat data, there are not only numeric variables but also literal variables of a certain type. This quantification is achieved by introducing artificial dummy variables, called "dummy variables" or "dummy variables", usually with a value of 0 or 1, to reflect the different properties of a variable, and to serve as a coding tool for the fixed class of variables. This is usually done by introducing artificial dummy variables, which we call "dummy variables" or "dummy variables".

However, for fixed-class variables with n different categories, there will be (n-1) dummy variables and the larger dimensionality has a detrimental effect on the overall factor analysis. In order to show the influence of different types of variables more clearly, we use PCA dimensionality reduction to reduce the dummy variables. In PCA dimensionality reduction, we need to consider the explanatory power of the principal component factors for the original variables. The larger the variance explained, the stronger the explanatory power, and the more it can reflect the most critical factors of the variables, the more effective the extracted principal component factors are, and it is usually considered that reaching 0.8 is more excellent. Therefore, we set the total variance explanation rate to 0.8.

Under such conditions, we process all dummy variables. The dimension reduction of Make is two variables, Variant is three variables, and Country/Region/State is five variables. In addition, Hull Materials involves fewer categories and does not require dimensionality reduction.

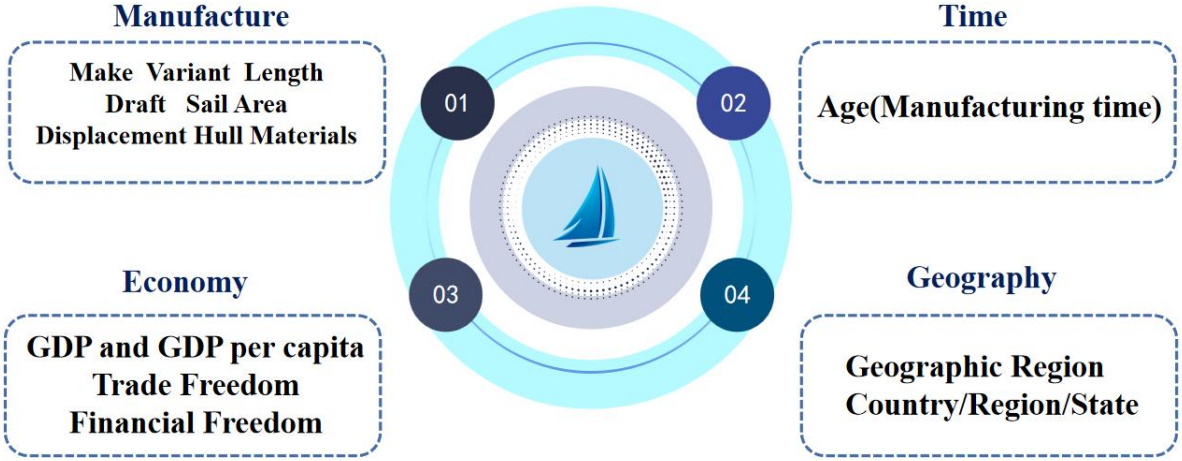


Figure 3: Price Influencing Factors Model

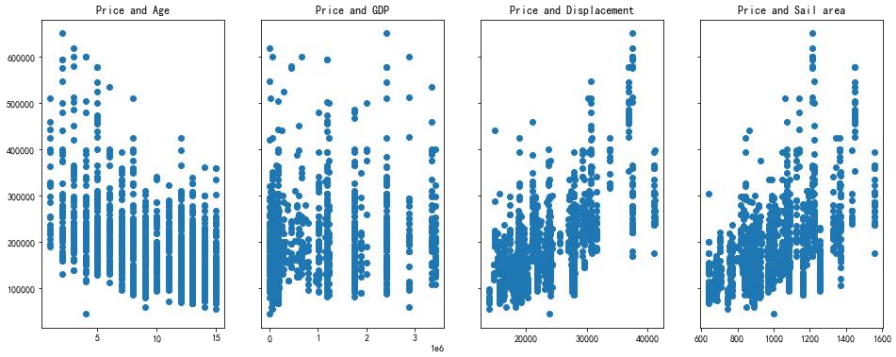
3.2 Price model construction based on Random Forest

3.2.1 Factor selection and classification

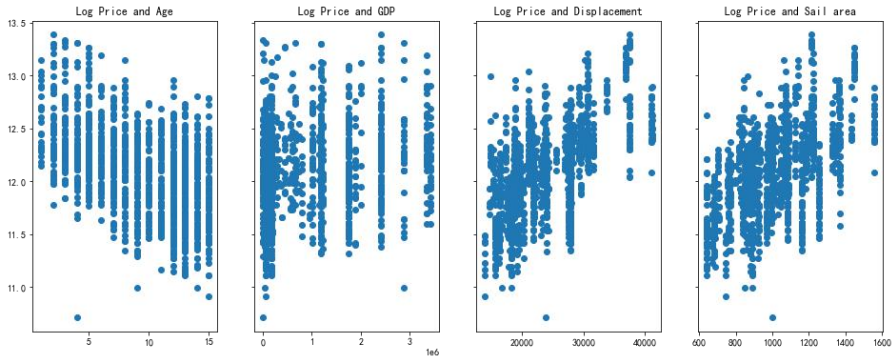
Combining the data we supplemented and processed, the independent variables that can be analyzed include Make、Variant 、Length、Geographic Region、Country/Region/State 、Age（2020-Year）、draft(ft)、displacement(lb)、hull materials、sail area(ft2)、GDP、GDP per capita、 Trade Freedom and Financial Freedom.In order to better describe the factors

affecting the price, we divide the factors into four categories (see Figure3) and explain them as follows.

- **Manufacturing Factors**: Various design parameters should be considered in the sailboat manufacturing process, sailboat variant and manufacturers.In particular, used sailboats are particularly affected by these manufacturing factors.
- **Time Factors**: The sailboat's age determines the remaining service life and operating time of the sailboat and also affects the operating costs such as insurance, maintenance, and fuel consumption of the sailboat[2]. Generally speaking, the older the sailboat is, the shorter the operating time, the less the operating income obtained, and the smaller the sailboat's price.
- **Economic Factors:**The sailboat industry is typically an export industry and therefore the used sailboat price is often influenced by GDP, while different economic and trade freedoms create different spaces for sailboat development, indirectly leading to different sailboat prices.(The focus here is on comprehensive factors, so only GDP and GDP pre capita are selected, Trade Freedom and Financial Freedom will be taken into consideration in Chapter 4)
- **Geographic Factors:**The geographical conditions of different regions are different, and the conditions for the development of the sailboat industry are different, resulting in differences in the scale of the sailboat industry.



(a)before



(b)after

Figure 4: Linear correlation between various factors and price

3.2.2 Relevance Analysis

3.2.2.1 Monohulled Sailboats

Before conducting regression analysis for the multivariate variables, the relationship between price and each factor was observed through scatter plots to verify that the data met the basic assumptions of linear regression, and we found that the logarithm of the price showed a better linear relationship (see Figure 4), increasing the variability between classes.

The Random Forest approach is beneficial for analyzing how influencing factors affect prices and thus building price models.

Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a "forest." We can analyze the extent to which the independent variable affects the dependent variable, thus predicting the dependent variable from the independent variable.As soon as possible the random forest is mainly used for classification, and in a real practical scenario it can be used as a regression model to transform the problem into an equivalent regression problem[5].

Step1.Determination of the root node of the decision tree

80% of all used sailboat data are taken as the training set, and the training sample is used to train a decision tree as the sample at the root node of the decision tree, and the remaining 20% is used as the test set to evaluate the prediction effectiveness of our price model for sailboats with different variants. The sample is randomly sampled to 0, max_depth is unlimited by default, and other metrics are by default. The number of decision trees is adjustable, ranging from 1 to 150, and the step size is 1.

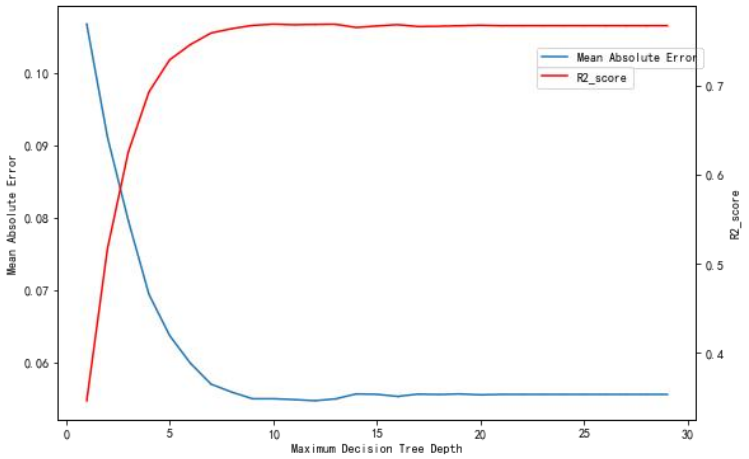


Figure 5: Maximum Decision Tree Depth(Monohulled sailboat)

Step2.Decision tree splitting

The appropriate depth of the decision tree is obtained by the traversal method, and the results are shown in Figure 5.The figure identifies the use of 10 attributes (features) for each sample, and the fitting effect gradually improves as the depth of the decision tree increases. However, if the maximum decision tree depth is greater than 10, it will consume more arithmetic power and may need to be more balanced. When each decision tree node needs to be split, m attributes are randomly selected from these 10 attributes to satisfy the condition m

≪ 10. Then some strategy is used to select one attribute from these m attributes as the splitting attribute of the node.

Step3.Decision tree generation

The decision tree is formed by splitting each node according to Step 2 until it can no longer be split. The whole decision tree formation process without pruning.

Step4.Build a large number of decision trees

Build a large number of decision trees according to steps 1~3.

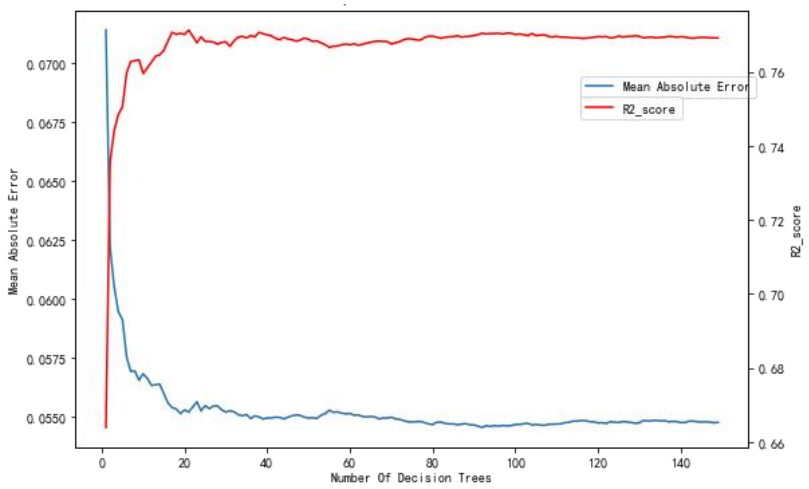


Figure 6: Number of decision trees(Monohulled sailboat)

As is shown in Figure 6,The optimal number of decision trees (depth) for the response with the largest coefficient of determination and the slightest average error is 37. The error data are in Table 2.

Table 2: Error data(Monohulled sailboat)

Term	Result
Mean Absolute Error	0.047
Mean Squared Error	0.004
Root Mean Squared Error	0.065
R^2	0.839

In this decision tree condition, the percentage importance of each factor was analyzed and shown in Figure 7. From Figure 7, we can find that Displacement, Length and Age are the three factors that have the greatest impact on price, followed by Country/Region/State , Draft,Sail Area ,GDP and GDP per capita, while all other factors have less impact on price. In addition, Figure 8 can represent the prediction effect for each variant (20% training set) .

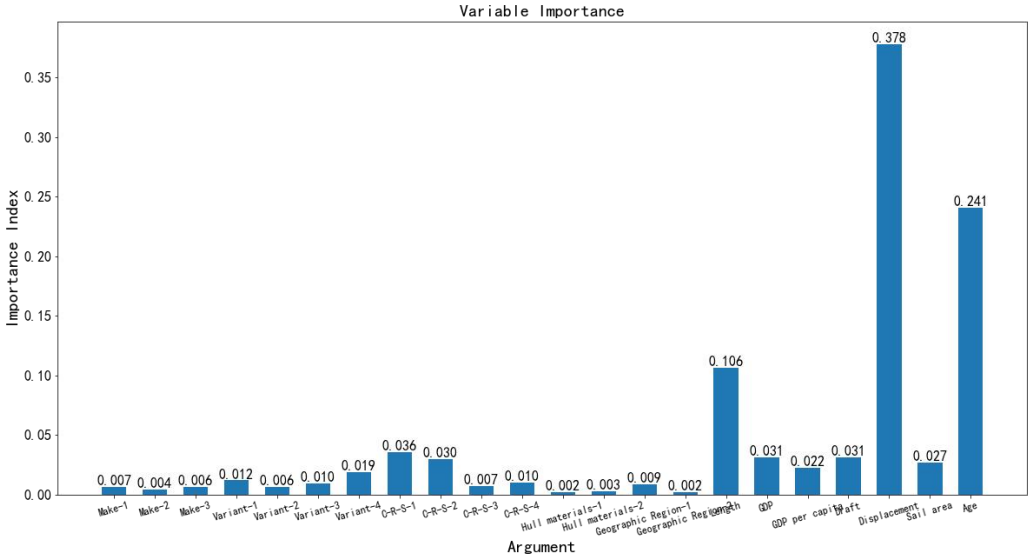


Figure 7: Result of decision trees(Monohulled sailboat)

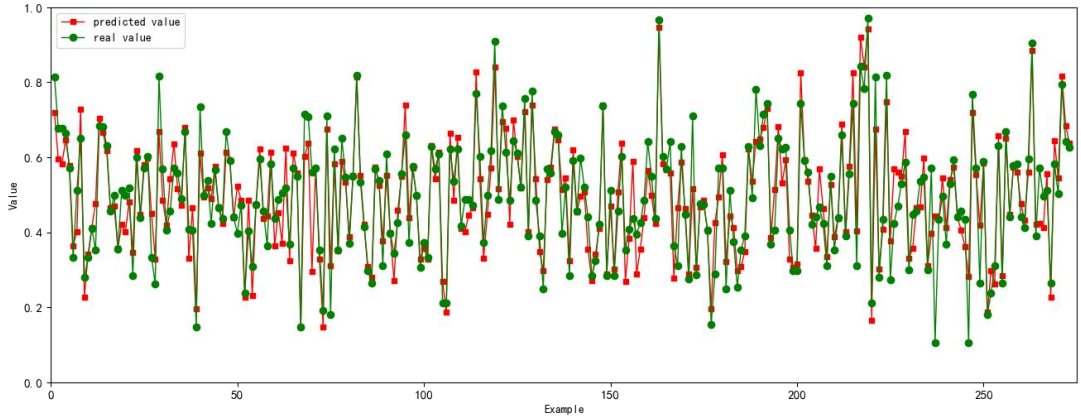
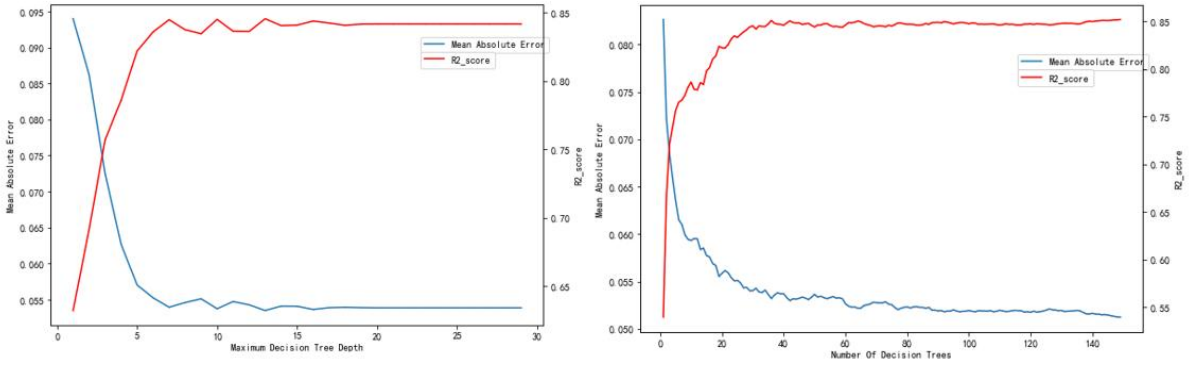


Figure 8: Prediction effect(Monohulled sailboat)

3.2.2.2 Catamarans

Using the same Random Forest processing method as Monohulled Sailboats, only some key results are shown here for comparison. It can be seen from Figure 9 that the maximum decision tree depth is selected as 7 and the number of decision trees is selected as 40 for subsequent fitting analysis.



(a)Maximum Decision Tree Depth **(b)Number of decision trees**

Figure 9: Hyperparameter selection(Catamarans)

Using these hyperparameters to generate a decision tree and the error data are in Table 3.

Table 3: Error data(Catamarans)

Term	Result
Mean Absolute Error	0.037
Mean Squared Error	0.003
Root Mean Squared Error	0.054
R^2	0.904

The percentage importance of each factor was analyzed and shown in Figure 10.Similarly, Displacement, Length, and Age are the core factors, with Length being more influential than Age (compared to Monohulled sailboat). GDP, GDP per capita, Draft, Sail Area, variant, and Country / Region / State also have some impact. Other factors have less impact on price.

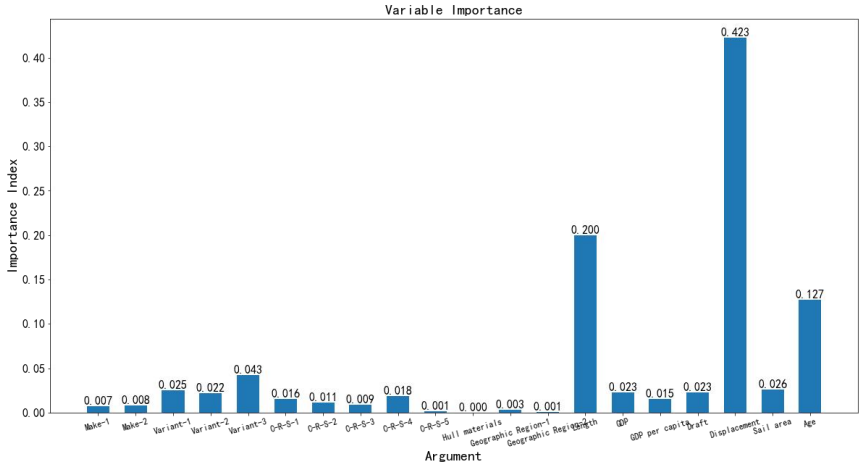


Figure 10: Result of decision trees(Catamarans)

The prediction effect(20% training set) is shown in Figure 11.

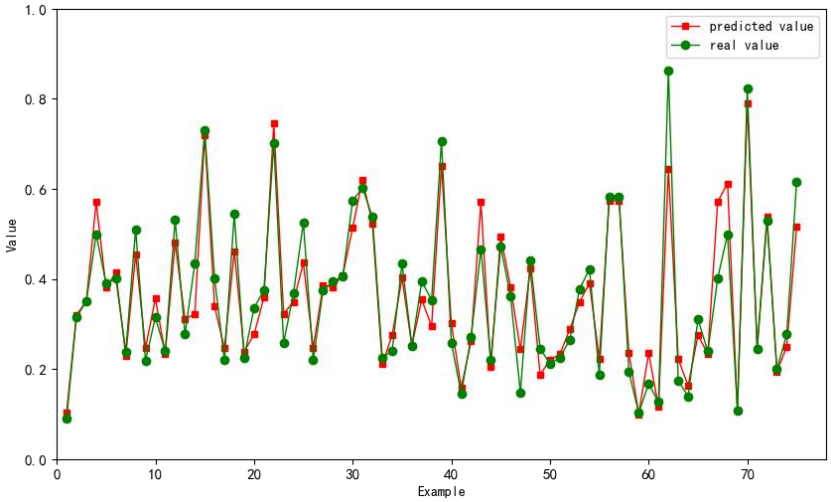


Figure 11: Prediction effect(Catamarans)

4 Regional impact model

4.1 Model background

The beginning point for developing the ocean economy statistics was to determine the geographical scope of the account[6].

If a country has a long coastline and a wide navigable area, the sailboat will be more adaptable to the market and the higher the price of the sailboat. Under the same conditions, a large sea sailboat can be engaged in small sea sailboat transportation, while a small sea sailboat cannot be engaged in large sea sailboat transportation. At the same time, another part of region-related factors is economic factors, which can translate the overall prosperity of trade to a certain extent, which has a greater impact on the price of sailboats.

4.2 Geographical price distribution

The data provided by the sailboat enthusiasts included sailing prices for several small areas in three large regions.

First, we conducted an overall analysis of the geographic region and used one-way ANOVA(Table 4) to explore the distribution patterns that existed.

Table 4: Results of one-way ANOVA test

Type	Geographic region	size	average	Standard deviation	F	P
Monohulled Sailboats	Europe	1048	184720.088	85649.396	18.002	<0.001
	USA	201	225680.418	102294.68		
	Caribbean	119	191864.353	89751.966		
Catamarans	Europe	257	435039.918	284986.649	3.827	0.023
	USA	22	526127.273	205365.303		
	Caribbean	95	406430.211	175787.024		

Second, a quantitative analysis of effects was performed to analyze the differences between the data.Their results are shown in Table 5.

Table 5: Effect quantitative analysis table

Term	Difference between groups	total deviation	Partial η^2	Cohen's f
Monohulled Sailboats	2.83×10^{11}	1.10×10^{13}	0.026	0.162
Catamarans	2.59×10^{11}	1.28×10^{13}	0.02	0.144

According to the ANOVA results, the p-value is less than 0.05, which means that the

statistical results are significant, i.e., there is a significant difference between different Geographic Regions in Listing Price.

4.3 Analysis of regional factors

4.3.1 Data Preparation

The geographical factors discussed in this part are actually the careful consideration of the geographical factors and economic factors selected before.

For geographical factors, the size of the ocean is an essential factor affecting the economic prosperity of the sailing area. Therefore, we find data on the coastline of each region from The World Factbook [7], including the length of the coastline and the ratio of the coastline to the national area. These data are of great reference value for understanding the marine distribution of the region because the longer the coastline, the more light the land area where sea activities can be carried out. The larger the ratio of coastline to the area, the greater the possibility of marine activity in the country.

For the economic factors, in addition to the previous GDP and GDP per capita, we add the previously mentioned Financial Freedom and Trade Freedom to measure the prosperity of the regional economy and trade.

As shown above, we have to consider the impact of six factors on prices in total: coastline length, coastline-to-country area ratio, GDP, GDP per capita, Financial Freedom, and Trade Freedom.

4.3.2 Relative degree by Pearson correlation coefficient

Monohulled Sailboats and Catamarans are first analyzed separately to derive the relationship with the six factors, i.e., to calculate the sample Pearson correlation coefficient. We expect to derive the degree of influence of the six factors on prices through Pearson correlation coefficients. However, it should be noted that from the conclusions in Chapter 3, we can find that the essential geographical factor of GDP is not highly correlated with sailboat prices. We can only try to find other factors or explain the auxiliary effect of geographical factors on prices.

For n sets of data belonging to the same sailboat type considered as n samples, the correlation between components X and Y is compared. That is $X:\{X_1,X_2,X_3,\ldots,X_n\}$ and $Y:\{Y_1,Y_2,Y_3,\ldots,Y_n\}$

Sample means and covariances:

$$\bar{X}=\frac{\sum_{i=1}^{n}X_i}{n},\bar{Y}=\frac{\sum_{i=1}^{n}Y_i}{n},Cov(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1} \tag{1}$$

Sample standard deviation：

$$S_X=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}},S_Y=\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}} \tag{2}$$

Sample Pearson correlation coefficient:

$$r_{XY} = \frac{Cov(X,Y)}{S_X S_Y} \tag{3}$$

A matrix representing the relationships of the six factors was derived from the Pearson correlation coefficients, and due to the large amount of data, a heat map is used here for visual representation(Figure 12 and Figure 13）.

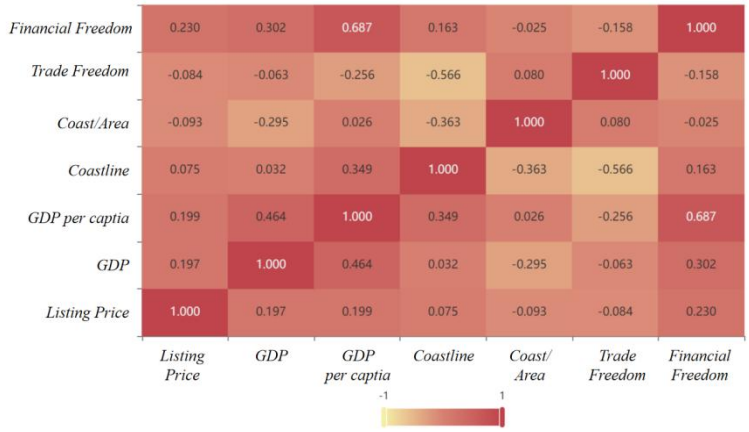


Figure 12: Pearson coefficient matrix（Monohulled Sailboats）

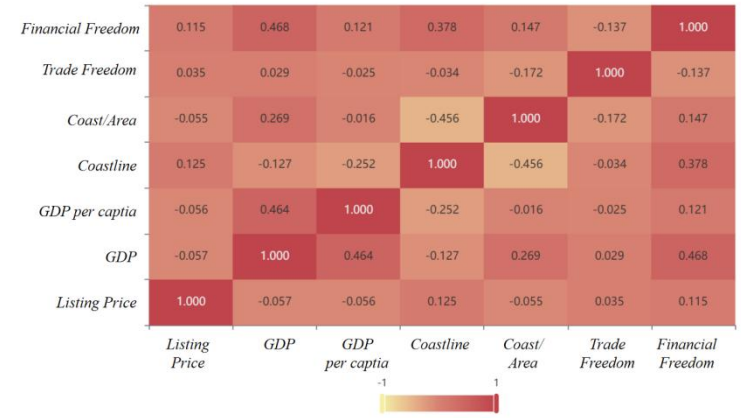


Figure 13: Pearson coefficient matrix（Catamarans）

4.3.3 Ridge regression with cross-validation

Ridge regression is a biased estimation regression method dedicated to the analysis of covariance data and is essentially a modified least.It is obtained by abandoning the unbiased nature of the least squares method at the cost of losing some information and reducing accuracy.

The regression coefficients are more realistic and reliable regression methods, and the fit to the pathological data is stronger than the least squares method.Perform a Ridge Regression analysis for each of the two cases:

- The independent variables include three important parameters：displacement、length、

age.

- The independent variables are nine parameters with the addition of geographical factors：GDP、GDP per captia、Coastline、Coast/Area、Trade Freedom、Financial Freedom

Before performing the regression, the data are standardized. Based on the mean and variance of the data, the data are changed to a standard normal distribution for easy fitting.

The cost function of the ridge regression can be expressed as

$$COST(w) = \sum_{i=1}^{N}(y_i - w^T x_i)^2 + \alpha \|w\|_2^2 \tag{4}$$

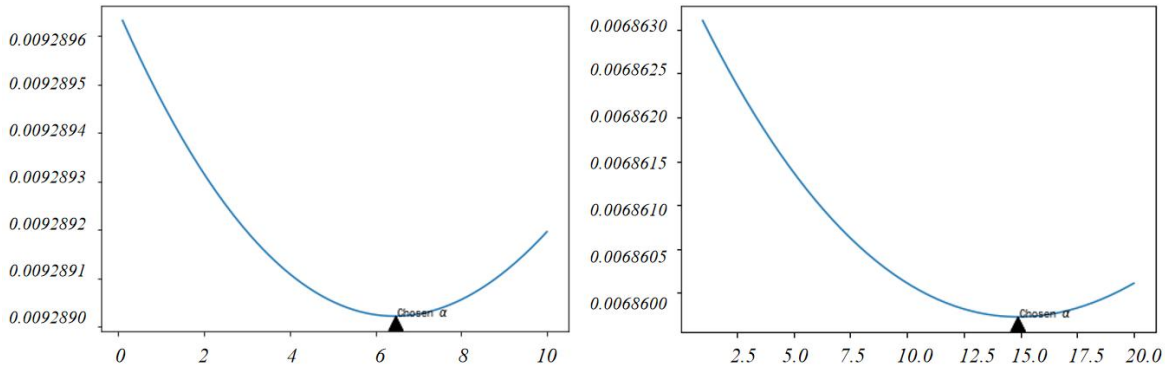
When the cost function *COST(w)* is the smallest

$$w = \arg\min_{w}(\sum_{i=1}^{N} y_i - w^T x_i)^2 + \alpha \|w\|_2^2 \tag{5}$$

Thus obtaining the analytical solution of *w*

$$w = (X^T X + \alpha I)^{-1} X^T y, \alpha \in R \tag{6}$$

In the Ridge regression process, the choice of the influence of the hyperparameter α is essential, and when we choose a suitable α , it can reduce the variance very effectively. Taking Monohulled Sailboats as an example:



(a)3 factors **(b)9 factors**

Figure 14: Ridge regression Hyperparameter selection

The variant is extracted according to a certain ratio (7:3) and is used as the training set and test set respectively.The predicted effect is shown in Table 6.

Table 6: Prediction effect contrast(Monohulled Sailboats)

Number of Factors	3(only manufacturing)	9 (includle regional factors)
R^2_score (test)	0.6027	0.6756
RMSE	0.0978	0.0884

We can find the predicted effect has improved if we add regional factors From Table 6,and

the better predicted effect is shown in Figure 15.

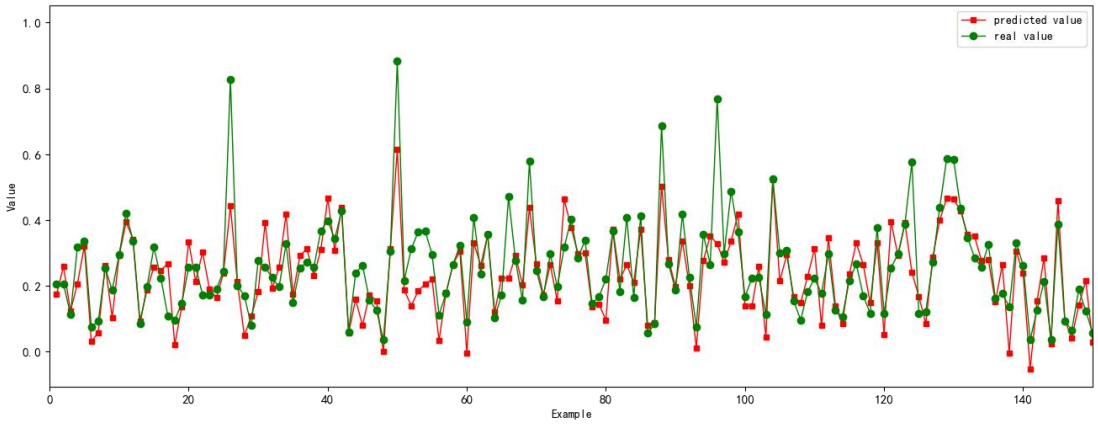


Figure 15: Prediction effect of ridge regression(Monohulled Sailboats,9 factors)

Based on the results of the ridge regression analysis, we can obtain the expression of the linear relationship for price prediction.

$$y = 0.0663x_1 + 0.0269x_2 - 0.0641x_3 + 0.0116x_4 + 0.0158x_5 + 0.0072x_6 - 0.0124x_7 - 0.0094x_8 + 0.0225x_9 + 0.2369 \quad (7)$$

The same analysis is similar for Catamarans, whose prediction results are shown below(Table 7). It also indicates that the additional geographical factor is beneficial in predicting the sailboat price, its prediction effect is shown in Figure 16.

Table 7: Prediction effect contrast(Catamarans)

Number of Factors	3(only manufacturing)	9 (includle regional factors)
R^2_score (test)	0.7365	0.7539
RMSE	0.1067	0.1032

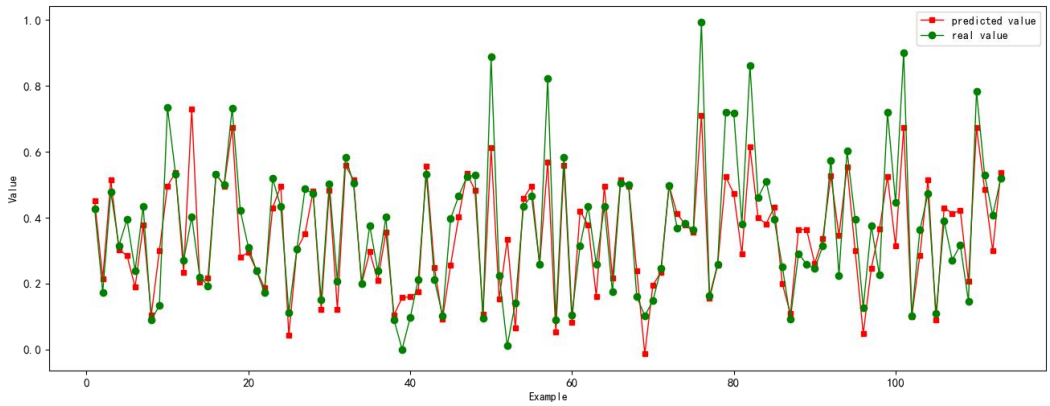


Figure 16: Prediction effect of ridge regression(Catamarans,9 factors)

And the linear relationship for price prediction:

$$y = 0.1193x_1 + 0.0185x_2 - 0.0077x_3 - 0.0020x_4 + 0.0006x_5 + 0.0189x_6 + 0.0013x_7 + 0.0044x_8 + 0.0035x_9 + 0.3442 \quad (8)$$

4.3.4 Discussion

Both from the research in Chapter 3 and the Pearson coefficient analysis in this chapter, we can see that many of the economic factors that we consider more significant do not have a direct impact on sailboat prices, but we still retain this part of the research process, while finding new conclusions from it.

When we utilize the three core factors (Displacement, age, length) analyzed in Chapter 3 as independent variables to analyze the price, due to the existence of some special samples in the original data, it does end up as a ideal prediction, and when we add these geographic factors that seem have no direct correlation, there is an improvement in the prediction results, and at the same time, this improvement is present in all variant.

It is obvious that the development of regional economies affects the volume of cargo transported on a global scale, which is mainly carried out by water. Therefore the global economic situation will affect the demand for maritime cargo transportation. The change of maritime transportation demand will directly act on the freight market, and the change of freight market will quickly reflect to the supply and demand market of sailboats, which will in turn affect the supply and demand of sailboats and eventually affect the price of sailboats[2]. However, this influence is indirect and does not play a decisive role in the price; it is still the various manufacturing factors of sailboats that play a decisive role.

5 Model applications: forecast Hong Kong (SAR) sailboat prices

To validate our model for regional effects, we searched the Hong Kong (SAR) market for information on a number of used sailboats built before 2020, including both Monohulled Sailboats and Catamarans, with data from a sailboat for sale website in Hong Kong(SAR) [8]. Then we can obtain the mean and standard deviation of sailboat prices(Table 8).

Table 8: Hong Kong Sailboat Data Characteristics

	Mean	Standard deviation
Monohulled Sailboats	341384.6	292898.0
Catamarans	820473.7	511112.5

To well model the regional effect in Hong Kong(SAR), a price interval is specified, and the data provided by sailing enthusiasts are censored to find comparable prices to the queried used sailboats in Hong Kong(SAR). The interval is *[mean-std , mean+std]*, and the censored dataset is merged with the queried Hong Kong used sailboat data, adding labels indicating that they belong to the Hong Kong(SAR) market. The purpose of the merge is to ensure that

the Hong Kong data are normalized together with other data when normalized.

After the min-max normalization on the merged data, the merged data are then split into training data and test data, with the training data being the non-Hong Kong market used sailboat data and the test data being the Hong Kong market used sailboat data.

The training data are trained by the Ridge Regression Model noted before(need to find the optimal parameter α and the following multiple regression equation is obtained :

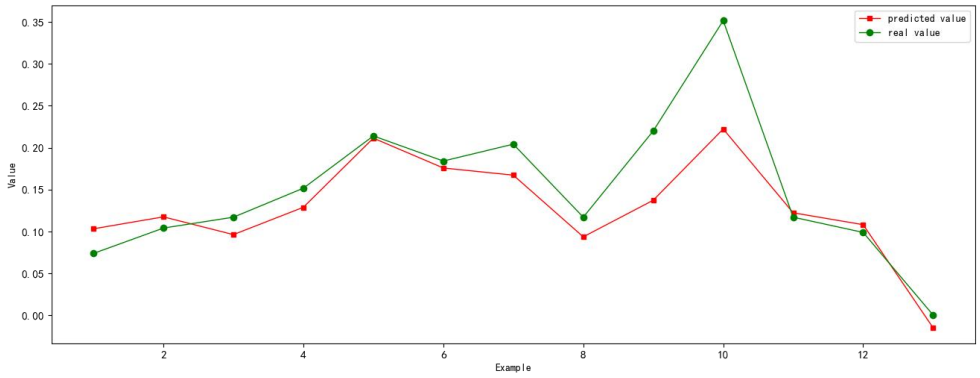
(a)Monohulled Sailboats：

$$y = 0.0407x_1 + 0.0132x_2 - 0.0358x_3 + 0.0030x_4 + 0.0078x_5 + 0.0055x_6 - 0.0031x_7 + 0.0012x_8 + 0.0175x_9 + 0.1297 \tag{9}$$

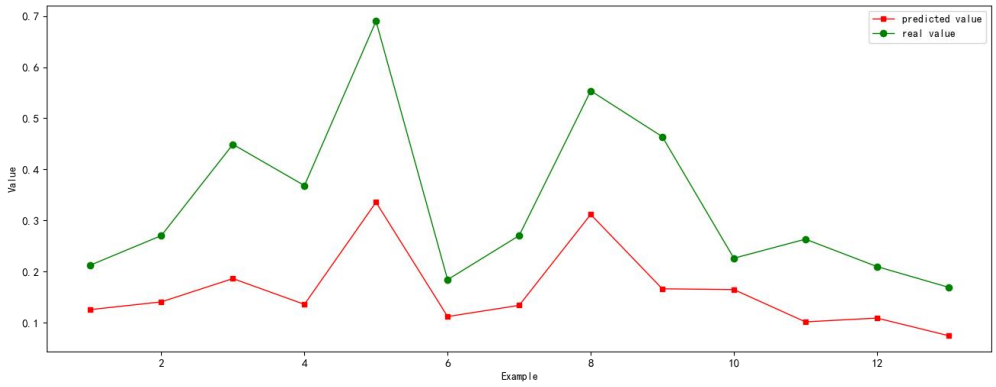
(b)Catamarans：

$$y = 0.0519x_1 + 0.0186x_2 - 0.0329x_3 + 0.0002x_4 + 0.0003x_5 + 0.0144x_6 + 0.0076x_7 + 0.0156x_8 + 0.0067x_9 + 0.1614 \tag{10}$$

Then the trained area effect model can predict the price of used boats in Hong Kong(SAR) well. We bring the test set into the model for prediction and get the following results(see Figure 17):



(a)Monohulled Sailboats



(b)Catamarans

Figure 17: Prediction effect of ridge regression(Hong Kong(SAR))

It can be seen that this model simulating the regional effect of Hong Kong, using the variables of Hong Kong's GDP, GDP per capita, coastline, Coast/area, Trade freedom, and Financial freedom to represent the regional effect, predicts the prices of both Monohulled

Sailboats and Catamarans very well with errors (R2_score , RMSE) within the acceptable range of permissibility.

6 Other conclusion in the data

6.1 The fitting effect of important factors

Based on previous analysis, we know that Displacement, Age, and Price are the most important factors. Through the analysis, we found that for Monohulled Sailboats, the linear relationship between price and two factors(Age and Displacement) is obvious. The results obtained by multiple regression fitting are shown in Figure 18.

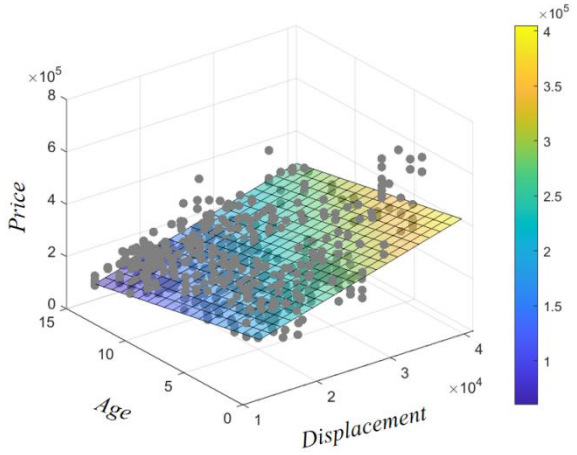


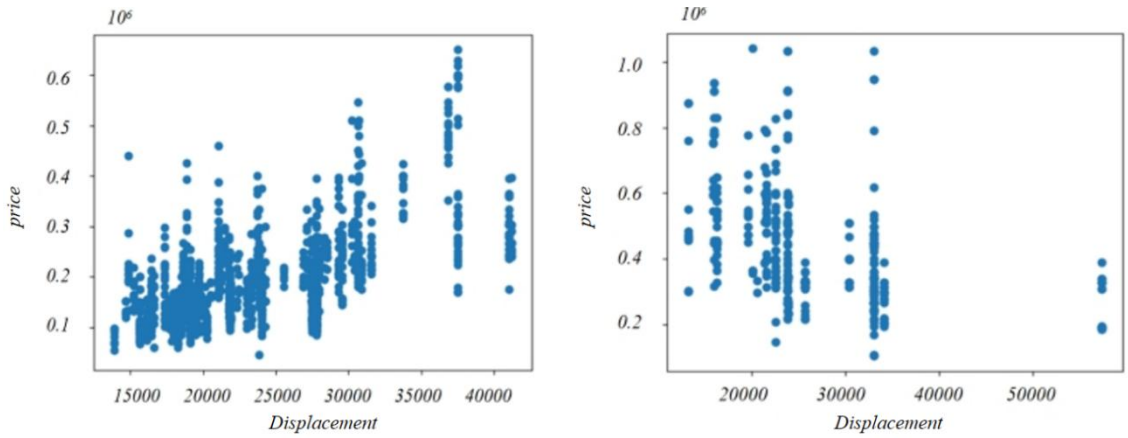
Figure 17: Three-factor fitting effect(Monohulled Sailboats)

The fitting plane is

$$y = 7.87x_2 - 9667.14x_3 + 92571.55 \tag{11}$$

6.2 Extreme value point of displacement

Displacement, as the most important factor affecting the price, also represents the cargo carrying capacity of the sailboat. Generally speaking, the relationship between sailboat displacement and sailboat deadweight tonnage is positively proportional. However, if we further look at the relationship between displacement and price, a scatter plot is drawn as shown in Figure 18.



(a) Monohulled Sailboats **(b)Catamarans**

Figure 18:Scatter plot of price and displacement

We can see that for Monohulled Sailboats, the overall price shows an increasing trend as the displacement increases, but when the displacement reaches around 37,000, the price starts to have a decreasing trend. For Catamarans, the positive correlation is basically invisible, and the negative correlation is basically observed when the displacement is larger than 15000. The reason for this is that the number of such sailboats is small, which is not enough to fit the complete trend and only reflects a specific negative correlation. It should be noted that we look for displacement data that are fully correlated with variants, so the data will be concentrated in specific displacements.

This conclusion can also prove that the sailboat price does not change proportionally to the sailboat's cargo capacity [9] For a particular sailboat type, according to the summary of chartering practice, each sailboat type has an optimal cargo capacity, and if the displacement is larger than that, the sailboat price will be lower than that of the optimal cargo capacity. For the two kinds of sailboats discussed in this paper, the optimal cargo capacity of Monohulled Sailboats is higher.

6.3 Effect of sailboat age

We analyze a large number of sailboat variants and make the average price curve with age, which is shown in Figure 19.

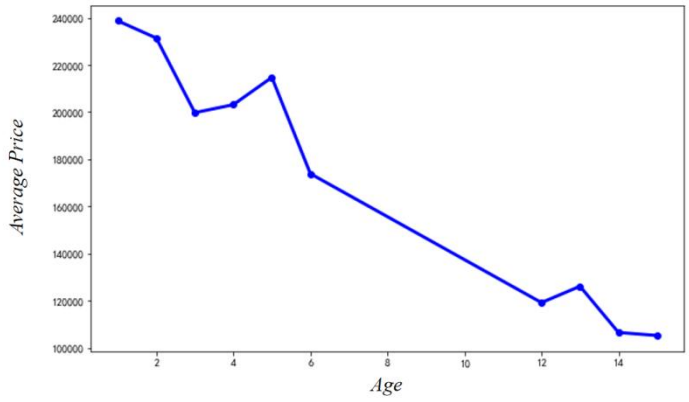


Figure 19: Line chart of the relationship between price and age

The variation of the depreciation rate of vessels with age shows a clear downward trend of the selling price of used vessels as a percentage of the original price as the age of the vessel changes. Interestingly, we find that the depreciation rate of sailboats from year 5 to 12 is the largest at 42.9%.

Therefore, for sailboat company brokers, trading a vessel before the fifth year retains the maximum economic value of the sailboat. And for the buyer, there is the option to buy after the sixth year to acquire the vessel at the lowest possible price.

7 Model evaluation and further discussion

7.1 Strengths

- The price model includes various additional attributes of the used sailboats to screen the decisive factors, and have a more comprehensive understanding of the sailboat's condition.
- In addition to the analysis of the explicit effects of the factors on prices, the relationships between the factors and the intrinsic patterns of the factors are also considered.
- Combine the characteristics of quantitative and definite class variables, the correlation of physical significance and the realistic context in the process of data processing, making the conclusions of greater value for brokers.

7.2 Weaknesses

- Geographical factors are complex and diverse, and many are difficult to measure in data, which are in the process of dynamic change over time, so the model may not fully describe the ancillary effects of region on sailboat price.
- The additional attributes of the sailboat are fully correlated with both variant and year, so it is difficult to learn the intrinsic differences of the same variant.

7.3 Further work

To find a more comprehensive data set and the economic factors that are more relevant to the maritime economy, while understanding the differences in the usage scenarios of various sailboat variants, thereby having a more comprehensive understanding of the used sailboat market in the context of reality.

8 Conclusion

Based on the data provided by sailboat enthusiasts and abundant reliable data from the Internet, we conducted a price analysis of the used sailboat market and established a price prediction model. On this basis, we further analyzed the effect of geographical factors on price and established a more comprehensive price model, which can make more accurate price predictions using nine factors. This model performs well in predicting sailboat prices in Hong Kong (SAR). In addition, we also analyzed the data to play other intrinsic patterns.

The price model we have developed is a good reference to assist brokers in their decision-making. It also helps governments and companies understand the dynamics of the ocean economy.

References

[1] https://sailboatdata.com
[2] Xiao, Q.j.(2013).Research on ship price evaluation method based on BP neural network, Shanghai : Shanghai Jiaotong University.
[3] https://unctad.org/
[4] www.heritageofthomasville.com
[5] Kalpana, G., Durga, A. K., Reddy, T. A., & Karuna, G.(2022). Predicting the Price of Pre-Owned Cars Using Machine Learning and Data Science,International Journal For Research.
[6] Nicolls, William, et al.(2020).Defining and measuring the US ocean economy, Was hington: Bureau of Economic Analysis.
[7] Central Intelligence Agency, ed. (2011).The World Factbook 2011. Central Intelligence Agency.
[8] https://www.boats.com/boats-for-sale/?boat-type=sail&country=hong-kong
[9] Alex Adamou.Behind the Screens:the Mathematics of Vessel Valuation,VesselValue.com.

Used Sailboat Price Market Analysis Report

COMAP conducted a research on the selling prices of used sailboats produced from 2005-2019 and analyzed the determinants affecting the used sailboat market, including indirect role of geographical characteristics on the prices of used boats.

✧ Core factors

Overall, the price influencing factors for each sailboat types are similar but different. The core factors affecting the price of used sailboats are Displacement, Age and Length, although there are differences in the ranking of the three factors for both types of sailboats.Meanwhile, the price is proportional to Displacement in a certain range and inversely proportional after the turning point. GDP, Country/Region/State and other factors also play a supporting role. However, in the used sailboat market, the boat's attributes are still essential to determining the price.The analysis of the influencing factors of the two types of sailboats is shown in Figure 1.

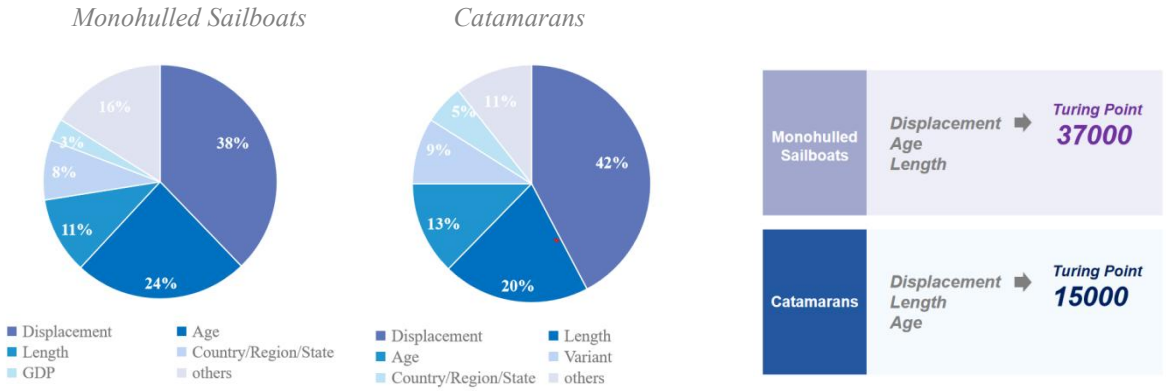


Figure 1: Price Influencing Factors

✧ Regional effect

In our study, we found that the geographic factors such as coastline length, ratio of coastline length to area, and climate all affect the price of used sailboats. Meanwhile, sailboat is an essential carrier of marine economy, are influenced by economic factors such as GDP, GDP per captia, Trade Freedom, and Financial Freedom.

The degree of influence of geographical factors for Monohulled Sailboats is shown in Figure 2, which shows that the Financal Freedom should be especially considered in the investment or purchase process.For Catamarans, the geographical influence is relatively small, and the primary consideration is the manufacturing characteristics of the boat.

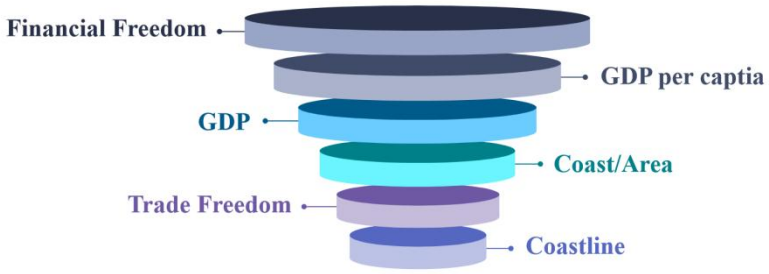


Figure 2: Geographical Factors of Monohulled Sailboats

However, using only geographic factors that include geography and economics is not enough to accurately forecast prices.The best way to forecast price is to combine the six geographic factors with three core internal factors (Displacement, Length, and Age).We are able to accurately predict prices using Ridge Regression.

✧ Conclusion about Hong Kong(SAR)

Hong Kong(SAR) is rich in islands, has a high ratio of coastline to regional area, and also has a very high degree of economic freedom and trade freedom. Therefore, if the analysis is performed with data sets from other geographic regions, there is a large inaccuracy, which is manifested by the overall high prediction results, but the trend is still consistent. In particular, the ratio of actual to predicted prices for Monohulled Sailboats is about 1.23, while Catamarans is about 2.11.