A Study on Used Sailboat Market Pricing Strategies Based on the XGBoost Model

The listing price of used sailboats has always been a complex issue due to the multiple factors. As the sailboat market continues to grow, the demand for accurate pricing of used sailboats is also increasing. Therefore, research on used sailboat pricing is becoming increasingly crucial. The aim of this report is to analyze the factors affecting the pricing of used sailing boats in Hong Kong and their importance by utilizing data visualization and machine learning algorithms, providing a scientific pricing report for Hong Kong brokers.

For Problem 1, we establish a Lasso regression model to filter various factors affecting the pricing of used sailboats, identifying key factors with significant impact on pricing. We then develop an XGBoost learning regression model based on machine learning to rank the importance of each factor. For monohull sailboats, ranking feature importance values by significance: Length, Year, Beam, GNI per capita, Sail Area, Draft and GDP. For catamarans, ranking feature importance values by significance: Year, Sail Area, Length, Beam, GDP, GNI per capita and Draft. According to the goodness-of-fit, our model demonstrates relatively high accuracy.

For Problem 2, we first establish a principal component analysis model to analyze the three variables reflecting regional factors, determine the optimal number of principal components, and incorporate the new variables into the model for XGBoost regression analysis, concluding that regional factors have a significant impact on used sailboat pricing. We then use a two-factor analysis of variance with interaction effects to explore the impact of regional effects on the pricing of different types of sailboats, finding that the influence of regional effects on different types of sailboats is significant.

For Problem 3, we first replace the values of the variables reflecting regional factors with data from the Hong Kong Special Administrative Region, select several types of sailboats with larger sales volume from different categories, and use the XGBoost model trained in the first two problems to predict the pricing of various types of used sailboats in Hong Kong, comparing the predictions with the actual values to determine the applicability of the model in the Hong Kong region, which is found to be satisfactory.

For Problem 4, we delve into deeper insights from the data based on the first three problems, obtaining various information conducive to pricing used sailboats, such as best-selling products and popular prices, and present the conclusions through a series of intuitive visualizations. This information can provide marketing suggestions for sailboat brokers.

Finally, based on our model and conclusions, we provide a report on used sailboat pricing for Hong Kong brokers.

Keywords: Machine Learning;XGBoost Regression Model; SHAP Model; Data visualization

Contents

1 Introduction

1.1 Problem Background

Sailboats are versatile vessels with growing popularity in buying and selling used boats. The pricing of used sailboats is influenced by factors such as usage duration, size, hull materials, etc. Understanding these factors and their significance is crucial for setting reasonable prices, promoting the development of the sailing boat market, providing quality options, and generating profits for brokers. A report on used sailboat pricing is needed for sailing boat brokers in Hong Kong to set appropriate prices and aid in the market's growth.

Figure 1 Sailboat

1.2 Restatement of the Problem

Given the contextual background and specified constraints outlined in the problem statement, the following issues must be resolved:

- Develop a model to explain sailboat listing prices using relevant predictors and economic data by year and region. Identify data sources and discuss the precision of each sailboat variant's price estimate.
- Develop a model to analyze regional impact on sailboat listing prices and evaluate consistency across variants. Consider practical and statistical significance of regional effects.
- Analyze the impact of the Hong Kong (SAR) region on sailboat prices using a subset of data, including both monohulls and catamarans. Model the potential impact and determine if it is consistent for both boat types.
- Identify and analyze any additional intriguing and informative insights or findings that your team has derived from the data.
- Create a report, consisting of one to two pages, for the sailboat broker in Hong Kong (SAR). Incorporate a select few graphics to aid in conveying your findings to the broker.

1.3 Our Work

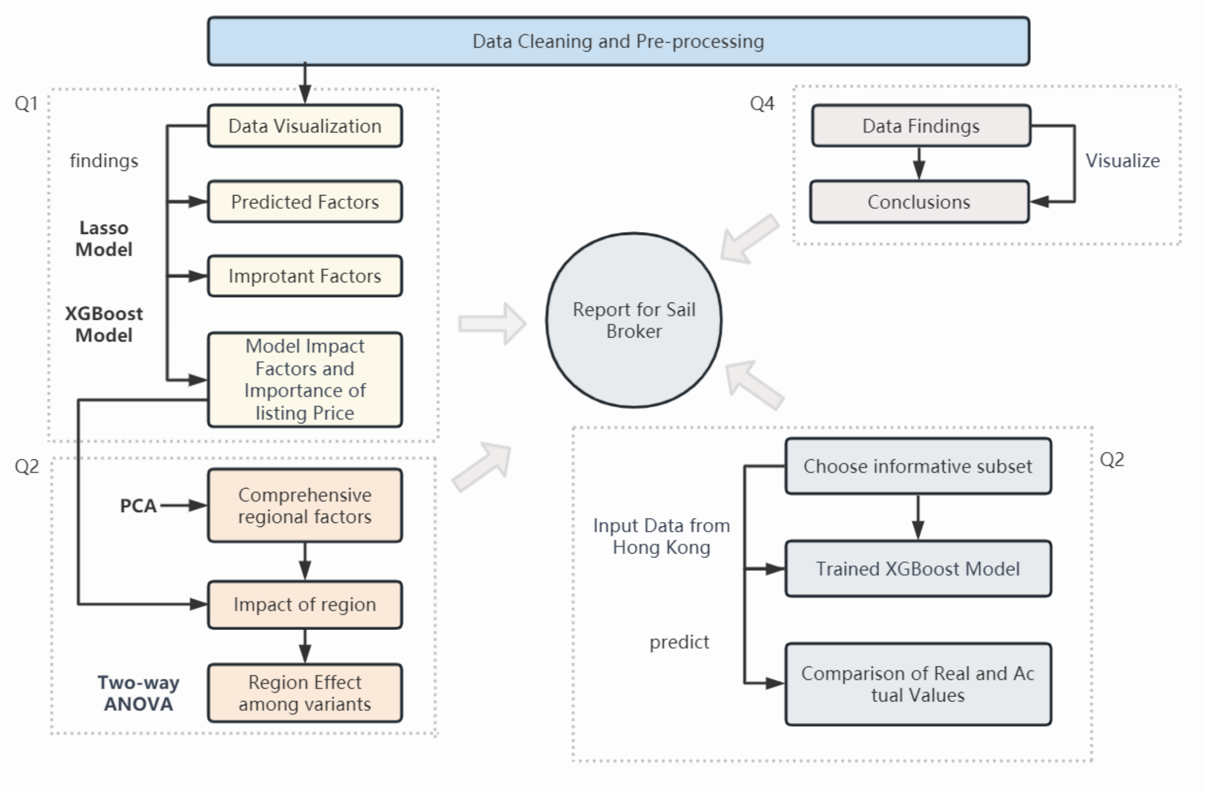

Data Cleaning and Pre-processing
Q1
Data Visualization
findings
Predicted Factors
Lasso Model
Improtant Factors
XGBoost Model
Model Impact Factors and Importance of listing Price
Q2
PCA
Comprehensive regional factors
Impact of region
Two-way ANOVA
Region Effect among variants
Report for Sail Broker
Q4
Data Findings
Visualize
Conclusions
Q2
Choose informative subset
Input Data from Hong Kong
Trained XGBoost Model
predict
Comparison of Real and Ac tual Values

Figure 1 Research ideas diagram

2 Assumptions and Justifications

Assumption 1: We hypothesize that other factors, such as sailboat features and regional economic level, may also impact the listing price of used sailboats in the spreadsheet provided.

Justification: The listing price of used sailboats is likely influenced by factors pertaining to the sailboat's features, including beam, draft and sail area. These traits have a considerable effect on the sailboat's stability, sailing speed, and control abilities, which consequently impact its market value and price. In addition, as sailboats are classified as luxury commodities, regions with greater economic development, living standards, and purchasing power are expected to have higher demand for sailboats, thereby leading to increased pricing.

Assumption 2: Assuming total GDP, per capita GNI and Average Cargo Throughput adequately measure a country's economic status.

Justification: Total GDP reflects economic status and development of a country, and can be used to compare its economic development and wealth distribution. However, a high total GDP alone may not indicate good average economic conditions for residents, so per capita GNI is also used to measure living standards and purchasing power. Average Cargo Throughput indicates the logistics and trade activity in a country or region. Active import and export trade may lead to higher demand for sailboats. Higher GDP, per capita GNI and Average Cargo Throughput indicate a better economic.

Assumption 3: We hypothesize that the age of a sailboat can serve as an indicator for measuring its depreciation status, with the understanding that a shorter usage period is associated with less value loss in the sailboat.

Justification: Due to our inability to obtain detailed data on the specific usage duration and damage conditions of each sailing vessel, we use the age of the sailboat as a proxy for measuring the loss of its value. Generally speaking, the older the sailboat, the greater the likelihood of it having been used multiple times and having sustained significant damage. Therefore, the age of the sailboat can largely represent its damage condition.

Assumption 4: Assuming that the Total GDP, GNI per capita, and Average ratio of total logistics cost to GDP are the main factors influencing the price of sailboats in a region, and that these three factors are sufficiently representative.

Justification: Unpredictable factors, including regional sailboat sales seasonality, drive our assumptions, including regional economic indicators Total GDP and GNI per capita. The average logistics cost to GDP ratio highlights the significance of logistics in the economy. A higher ratio may signal logistics inefficiency, leading to increased transportation and distribution costs, and affecting sailboat prices through higher production, transportation, and import costs and decreased competitiveness.

3 Notations and Definitions

3.1 Notations

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

Table 1 Notations used in this paper

Symbol	Description	Unit
x_i	The i-th quantitative factor affecting pricing, i=1,2,... ,8	
Y	Dependent variable Listing Price	

3.2 Definitions

Catamarans: A multi-hulled watercraft featuring two parallel hulls of equal size.

Monohull Sailboats: Sailboats that have only one hull usually centered around a heavy keel (the center blade).

Beam: The width of a boat at its widest point.

Draft: The minimum depth of water required to float a boat without touching the bottom.

Sail Area: The total surface area of the sails of a boat when fully raised.

Average Cargo Throughput: Average Cargo Throughput is a metric representing the volume of goods handled by ports, airports, or other transportation hubs within a specific time period. It indicates the logistics and trade activity in a country or region.

Total GDP: Total GDP, or Gross Domestic Product, is the sum of the value of all goods and services produced within a country's borders during a specific time period. It is a key indicator of a country's economic health.

GNI per capita:GNI per capita, or Gross National Income per capita, is the total income earned by a country's residents, including those living abroad, divided by the population. It is an indicator of a country's average living standards and economic well-being.

Average ratio of total logistics cost to GDP： It serves as a metric to assess the proportion of logistics costs to GDP in a country or region, from an academic standpoint.

4 XGBoost-based Pricing Model for Used Sailboats

4.1 Data Description

4.1.1 Data Collection and Pre-processing

We obtained additional data that may influence the listing price of used sailboats by web crawling. Data regarding sailboat characteristics and performance, including beam, draft, and displacement, were collected from multiple sailboat sales websites[1]-[4]. In addition, data on regional economies, like total GDP and per capita GNI, were sourced from Gampminder[5].

Due to missing values in the data obtained through web scraping, it is necessary to pre-process the data. In the case of economic data, since the sailboat prices provided in the table are from 2020, we are looking for the total GDP and per capita GNI data for the same year. To address missing values in total GDP data, we use GDP data from nearby years as a substitute. For missing per capita GNI data, we fill in the gaps with either per capita GDP or per capita GNI data from nearby years. As there is a large volume of data and only a few missing values in sailboat performance data, we directly eliminate rows with missing values using Python software.

4.1.2 Data Visualization

Due to the large amount of data, it is not possible to visually represent the relationships between the data. Therefore, we performed visualization processing on some data that may have relationships, with the aim of exploring the hidden connections between the data and discovering their value.

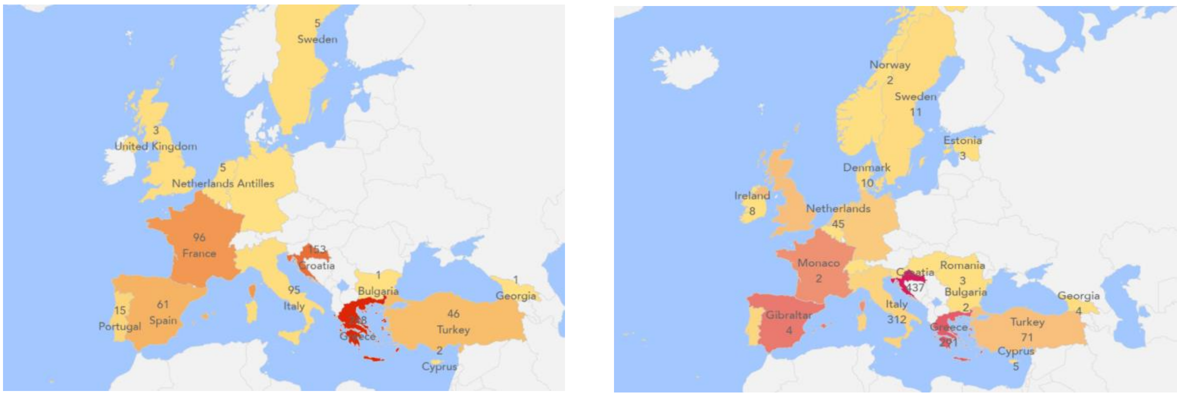
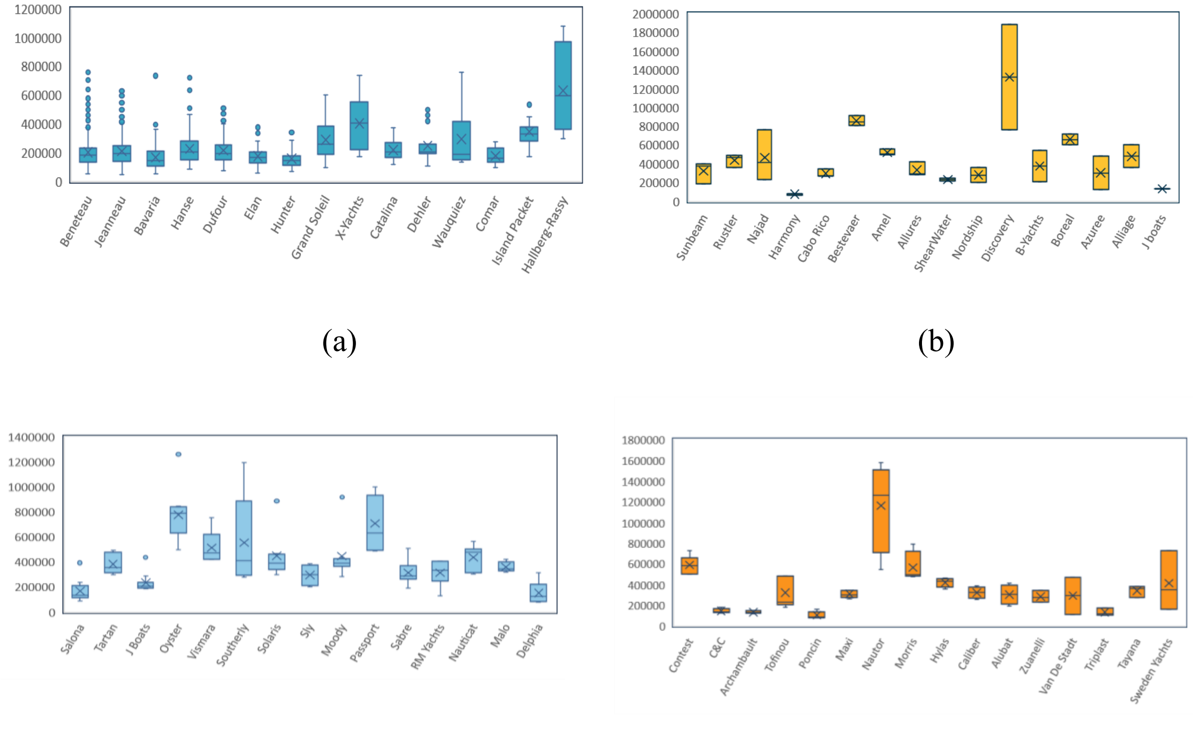


Figure 2 Distribution of quantities of sailboats: (left)Monohull (right)Catamaran

Distribution of 2020 sailboat sales across main countries/regions/states shows that Caribbean region emerges as the leader in sailboat sales among Europe, the Caribbean, and the US, highlighting its importance in the global sailboat market.

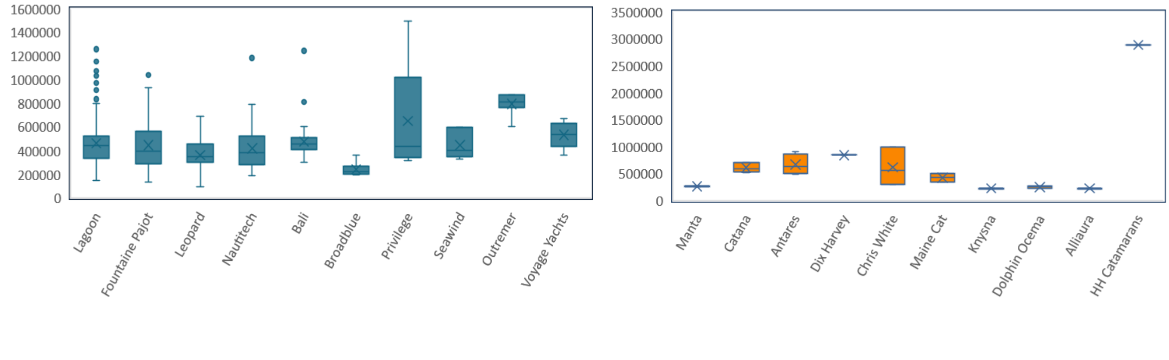


(a) (b)

(c) (d)

Figure 3 Boxplot of Make and Price for Monohull Sailboat

The four graphs display box plots of Monohull Sailboat prices for various brands. Brands are arranged by sales volume, with Beneteau having the most and J Boats having the least. The top seven brands have similar, affordable prices with limited price dispersion, while luxury brands Hallberg-Rassy, Southerly, Passport, Nautor, and Discovery have higher prices and greater price dispersion, resulting in lower sales volume.



(a) (b)

Figure 4 Boxplot of Make and Price for Catamarans

The two figures display box plots of catamaran prices for various brands, arranged by sales volume. Lagoon leads with the most sales while HH Catamarans has the least. The top three brands have similar prices and limited price dispersion, while luxury brands Privilege and

Outremer have higher prices, with Privilege having greater price dispersion. These two brands have lower sales volume due to their luxury status in the double-sail sailboat market.

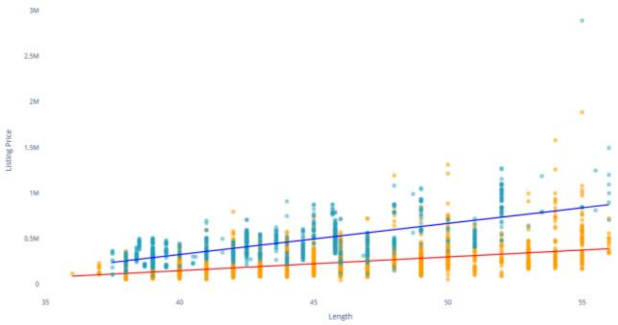


Figure 5 Scatter plot and trend line of price and length

The scatter plot shows length and price relationship, with Monohull Sailboats in orange and Catamarans in blue. The trend lines indicate a stronger correlation between length and price for sailboats, with a higher average price for Catamarans at the same price point.

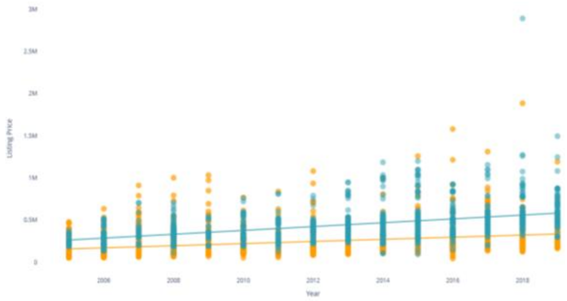


Figure 6 Scatter plot and trend line of price and year

The figure displays the relationship between sailboat manufacturing year ('Year') and price, with Monohull Sailboats in orange and Catamarans in blue. Younger sailboats have higher market value and prices. The graph shows prices increase with increasing 'Year', with a stronger trend for Monohull Sailboats.

4.2 The Establishment of Model 1

To address the high correlation among features and large number of features with limited impact on sailboat prices, we use the Lasso model for feature selection and the XGBoost model for regression analysis and computation optimization. Additionally, the SHAP model is utilized for interpretable analysis and presentation of results.

4.2.1 Lasso Regression Model

Lasso regression is a shrinkage method that uses *L1* regularization and simplifies the model by compressing some coefficients and setting others to zero, avoiding overfitting. In this study, the dependent variable *Y* is the Listing Price, and categorical data (Make, Variant, Country/Region/State) are transformed into quantitative data (Beam, Draft, Sail Area, 2020 GDP, per capita GNI) to be included in the pricing factors. Lasso regression is used to select important factors affecting the Listing Price.

4.2.2 XGBoost Regression Model

XGBoost is a machine learning algorithm based on gradient boosted decision trees, suitable for both classification and regression problems. In this paper, the important factors screened by Lasso regression are used as input independent variables x_i, and Listing Price is set as the dependent variable Y. The input data is divided into training and test sets in a 7:3 ratio. The feature importance of each variable is obtained through XGBoost machine learning regression, and the accuracy of the model is evaluated using the Mean Absolute Percentage Error (MAPE).

1. Establish the objective function

$$\mathrm{Obj}(x) = \rho(x) + L(x) \tag{1}$$

In this case, *(x)* is the loss function, is the regularization term, and x represents the model parameters. The regularization term includes both *L1* regularization and regularization. *L1* regularization can lead the model to produce sparse solutions, that is, settin *L2*g the weights of some irrelevant features to 0, there by achieving the purpose of feature selection.

2. Construct the prediction formula

$$\hat{y}_i = \sum_{k=1}^{K} f_k\,(x_i) \tag{2}$$

In this case, is the predicted value, is the k-th tree, and is the feature of the i-th sample. During the training process of each tree, XGBoost uses the gradient boosting algorithm to train the residuals as new labels, thereby gradually improving the fitting ability of the model.

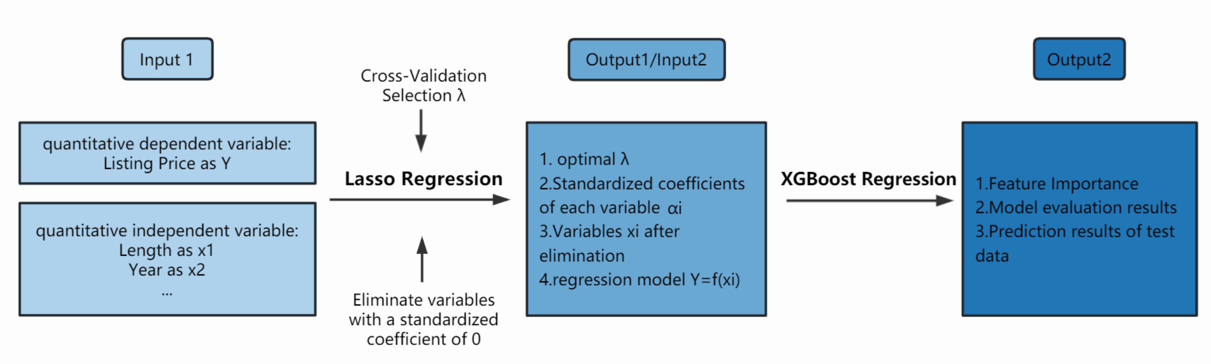

Input 1
quantitative dependent variable: Listing Price as Y
quantitative independent variable: Length as x1 Year as x2 ...
Cross-Validation Selection λ
Lasso Regression
Eliminate variables with a standardized coefficient of 0
Output1/Input2
1. optimal λ
2.Standardized coefficients of each variable αi
3.Variables xi after elimination
4.regression model Y=f(xi)
XGBoost Regression
Output2
1.Feature Importance
2.Model evaluation results
3.Prediction results of test data

Figure 7 Workflow diagram of Lasso and XGBoost Model

4.2.3 SHAP Modle

The SHAP (SHapley Additive exPlanations) model is an innovative method for interpreting machine learning model predictions. It is grounded in game theory and utilizes Shapley values to quantify the contribution of each feature to a specific prediction. The core concept of the SHAP model is to decompose the explanation of a prediction into the contributions of each feature. By calculating the Shapley values of each feature, we can quantify their respective contributions to a single prediction. The calculation of Shapley values adheres to principles such as efficiency, symmetry, null contribution, and linearity.

4.3 The Solution of Model 1

Sailboats can be categorized into monohull sailboats and catamarans. As evident from the data visualization section, there are significant differences in the attributes and prices of these two types of sailboats, necessitating separate analyses for each type.

For both monohull sailboats and catamarans, we first employ Lasso regression models to perform regression analysis on all the collected quantitative data. Utilizing SPSS and Python for data processing, we obtain the optimal values for λ and the standardized coefficient α. Based on these standardized coefficients, we identify the key factors that influence the pricing of used sailboats.

1. Determining the optimal value of λ.

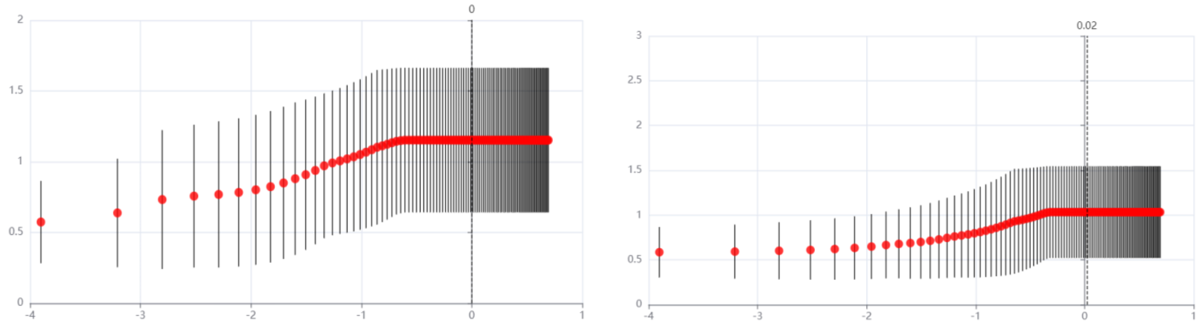


Figure 8 Regression Cross-Validation Plots: (L) Monohull , (R) Catamaran

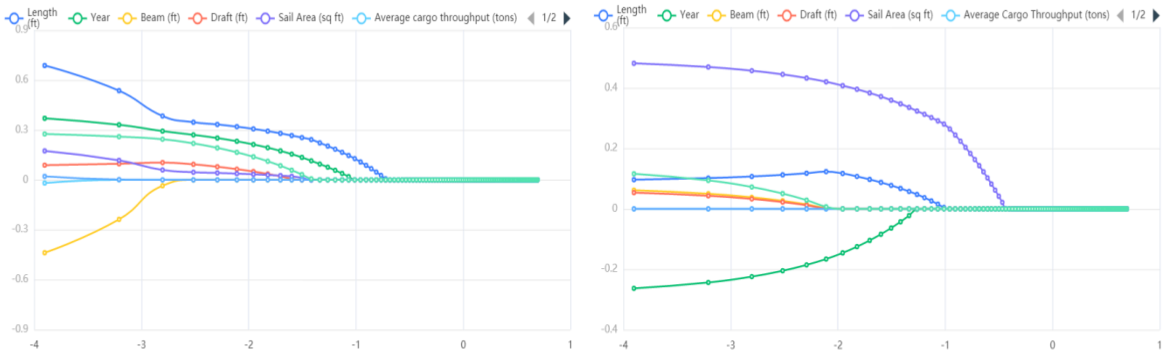


Figure 9 Lambda and Regression Coefficient Plots: (L) Monohull, (R) Catamaran

The graph shows that as λ approaches 0, the mean squared error stabilizes and coefficients approach 0, making $\lambda=0$ the optimal value. This indicates that Lasso regression is similar to linear regression. The analysis reveals that the non-standardized average cargo throughput coefficients are 0, implying minimal impact on used sailboat prices and necessitating exclusion before XGBoost regression.

2. Standardized Coefficients of Various Quantitative Data.

Through the Lasso regression analysis, the standardized coefficients of the various variables can be obtained as shown in the following table.

Table 2 Lasso regression standardized coefficient table

Monohulled Sailboats			Catamarans		
Variable Name	**Standardized Coefficient**	**R²**	**Variable Name**	**Standardized Coefficient**	**R²**
Intercept	-28497949.398	0.642	Intercept	-512703.173	0.619
Length	24721.688		Length	4627.239	
Year	14163.589		Year	-11676.937	
Beam	-82397.628		Beam	8848.364	
Draft	12143.906		Draft	31657.173	
Sail Area	120.343		Sail Area	349.558	
Average cargo throughput	0		Average Cargo Throughput	0	
GDP	13.175		GDP	-1.186	
GNI per capita	26.225		GNI per capita	1.724	

As can be seen from the table, all factors, with the exception of average cargo throughput, have an impact on the pricing of used sailboats.

3. Result of Regression Model

Upon analyzing the data, the standardized regression model for the listing price of used monohull sailboats can be expressed as $Y = -28497949.398 + 24721.688x_1 + 14163.589x_2 - 82397.628x_3 + 12143.906x_4 + 120.343x_5 - 0.0x_6 + 13.175x_7 + 2.225x_8$ while the standardized regression model for the listing price of used catamaran sailboats is $Y = -512703.173 + 4627.239x_1 - 11676.937x_2 + 8848.364x_3 + 31657.173x_4 + 349.558x_5 + 0.0x_6 - 1.186x_7 + 1.724x_8$. In these models, $x1$- $x8$ represent Length, Year, Beam, Draft, Sail Area, Average cargo throughput, GDP, and GNI per capita, respectively.

The regression models are then used to fit the data, and the specific fitting effect graphs are presented below.

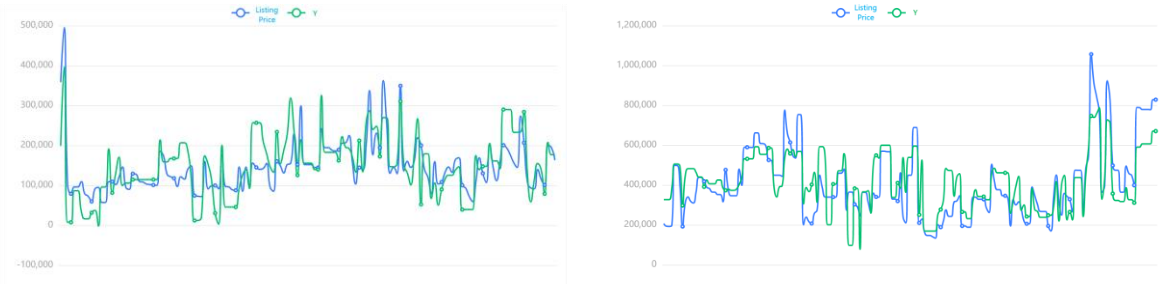


Figure 10 Lasso Regression Fitting Results (L) Monohull Sailboats (R) Catamaran Sailboats

The seven variables impacting used sailboat prices were determined and analyzed using the XGBoost machine learning regression method. The method fits the data with a 30% test set and calculates the goodness-of-fit values to determine the model's accuracy for both sailboat prices. The constructed model is analyzed with the SHAP method to understand the impact of each factor on prices.

(1) Monohull Sailboats

Firstly, an XGBoost regression analysis was conducted on Monohulled Sailboats, resulting in the feature importance of each variable on the Listing Price, as illustrated in the figure below.

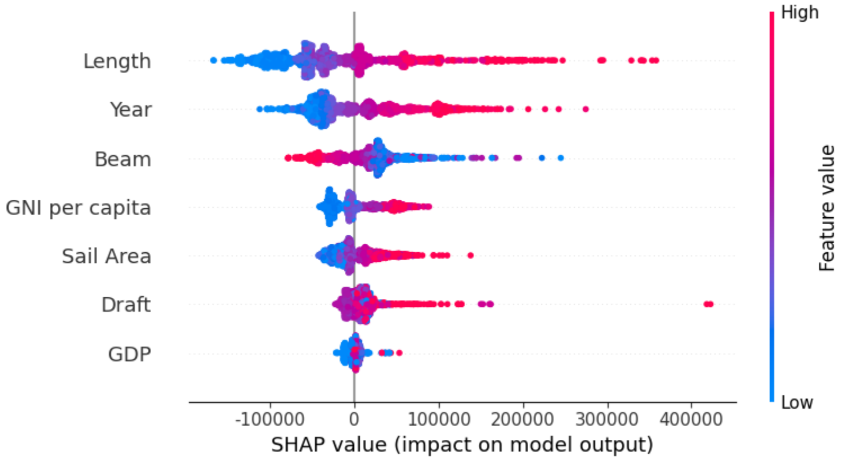


Figure 11 SHAP Values Plot for Each Factor

The figure shows feature importance ranking is Length > Year > Beam > GNI per capita > Sail Area > Draft > GDP. The positive relationships between sailboat prices and factors except Beam and GDP. Length, Year, GNI per capita, Sail Area, and Draft are positively related to the Listing Price. A longer sailboat length and a newer Year indicate higher difficulty in manufacturing and higher costs, resulting in higher used sailboat prices. The accuracy of prediction is indicated by MAPE and R^2, with good results from the 70:30 train-test split.

Table 3 Model Evaluation Results Table

	MAPE	**R^2**
Training Set	0.167	0.8393956
Test Set	0.186	0.7633030

As can be seen from the data in table , the MAPE values of the training set and the test set are very close and both are less than 1, while the R2 values are both greater than 0.6. This indicates that the XGBoost training performs well and is suitable as a good model for fitting the Listing Price.

The fitting results for some of the data are shown in the charts below.

Table 4 Partial Fitted Data Values.

Predicted Values	Actual Values	Predicted Values	Actual Values
304863.95	349900	261103.75	236866
186704.25	209588	499536.4	464162
156579.39	121348	432325.91	449377
133452.72	121257	117131.85	168000
133452.72	121227	126209.625	133617

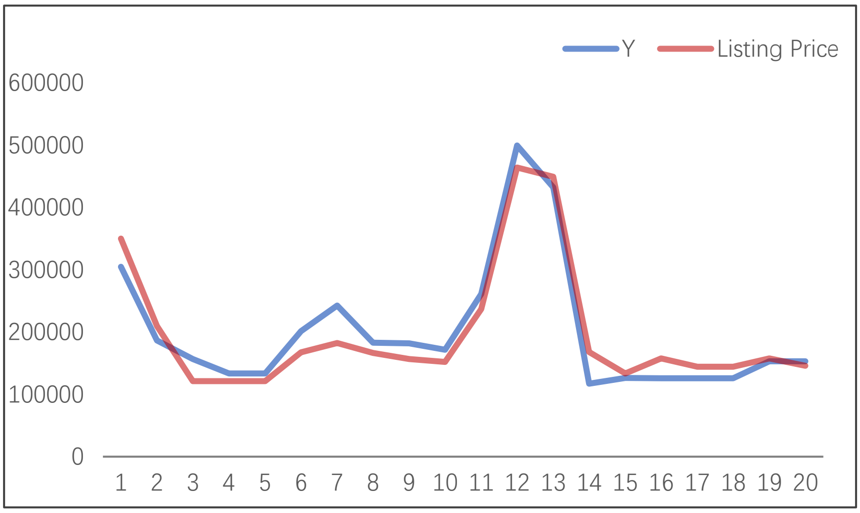


Figure 12 Partial Test Data Prediction Results Plot

(2) Catamarans

Firstly, we continue to perform the XGBoost regression analysis on catamaran sailboats to determine the impact of each variable on the Listing Price. The specific data is illustrated in the figure below.

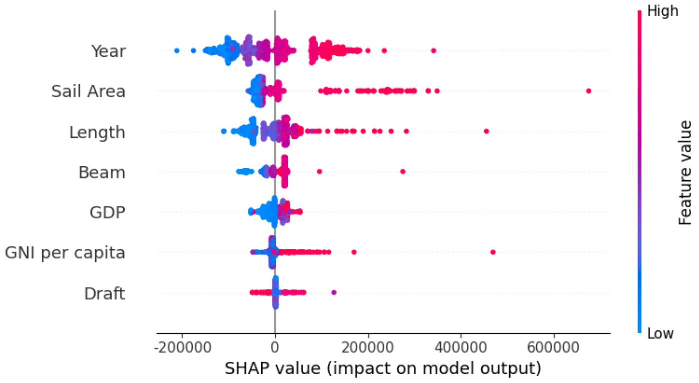


Figure 13 SHAP Value Plot of the Factors

The graph shows that The ranking of factors affecting sailboat prices is Year > Sail Area > Length > Beam > GDP > GNI per capita > Draft. For Catamarans, Length, Year, GNI per capita, Sail Area, Draft, Beam, and GDP have a positive effect on price. Year is the most important factor, with higher Year values corresponding to higher prices. The accuracy of the prediction is indicated by MAPE and R^2 values, obtained after a 7:3 train-test split.

Table 5 Model Evaluation Result Table

	MAPE	R^2
Training Set	0.092	0.9264301
Test Set	0.121	0.8157493

As can be seen from the data in the table, the MAPE values of the training set and test set are very close and both less than 1, while the R^2 values are both greater than 0.8, indicating that the XGBoost training has a good performance and is suitable as an excellent model for fitting the Listing Price. The fitting performance of some of the results is shown in the following graph.

Table 6 Partial Fitting Result Table

Predicted Values	Actual Values	Predicted Values	Actual Values
405736.1	399324	536704.28	510302
536704.28	557773	516704.28	498530
536704.28	557773	506704.28	497146
533423.9	546539	503423.9	484598
533423.9	545648	493423.9	483808

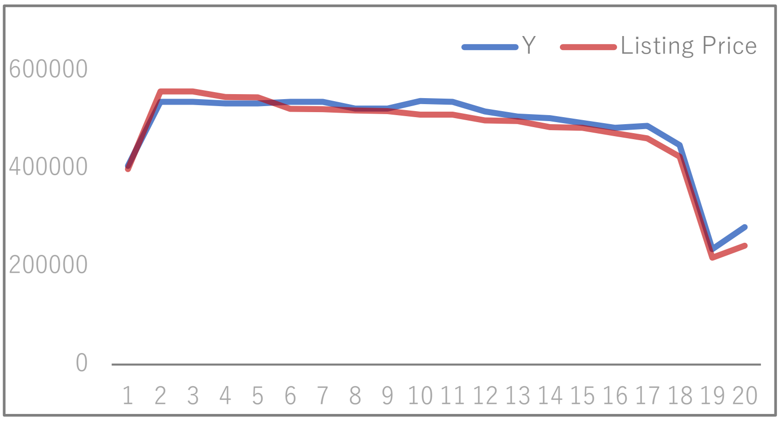


Figure 14 Partial Test Data Prediction Results Plot

5 Exploring Regional Effects on Used Sailboat Pricing through PCA and Two-Factor ANOVA

5.1 The Establishment of Model

To assess the impact of regions on listing prices, we consider three factors – Total GDP, GNI per capita, and Average ratio of total logistics cost to GDP. We employ PCA to reduce dimensionality, aiming to represent regions with one or two composite indicators. Subsequently, we use the XGBoost model to analyze the importance of regional factors on pricing alongside other indicators. Utilizing the SHAP model for visual representation of results. Lastly, we apply Two-way ANOVA to examine interactions between variant and regional factors affecting pricing.

5.1.1 Principal Component Analysis (PCA) model

Principal Component Analysis (PCA) is a statistical method designed to reduce data dimensions, extract significant features, and retain the majority of the information within a dataset. This technique is widely used in areas such as data compression, data visualization, and pattern recognition.

5.1.2 Two-way Analysis of Variance

Two-way ANOVA with interaction is a statistical method for studying the effects of two categorical independent variables (factors) on a numerical dependent variable, as well as the interaction between the two independent variables. It includes main effect tests and interaction effect tests. Main effect tests examine the individual influence of each independent variable on the dependent variable, while interaction effect tests investigate the impact of the interaction between the two independent variables on the dependent variable. Significant interaction effects indicate that the combined impact of the independent variables on the dependent variable cannot be simply explained by their main effects alone.

5.2 The Solution of model 2

Before conducting a principal component analysis on the collected region-related factors, the KMO test and Bartlett test were performed to analyze whether principal component analysis can be used. The analysis results for monohull sailboats and catamarans are shown in the following table.

Table 7 Translate: KMO Test and Bartlett Test

Monohulled Sailboats			**Catamarans**		
KMO values		0.812	**KMO values**		0.810
Bartlett Sphericity Test	**Aproxmate Chi-Square**	1748.163	**Bartlett Sphericity Test**	**Approximate Chi-Square**	874.175
	df	3		**df**	3
	P	0.000***		**P**	0.000***

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

As shown in the table, the KMO values are all greater than 0.6 and the P values are all less than 0.05, indicating that these factors have a correlation and can pass the significance test, and principal component analysis can be performed. The scree plot obtained from the principal component analysis is shown in the following graph.

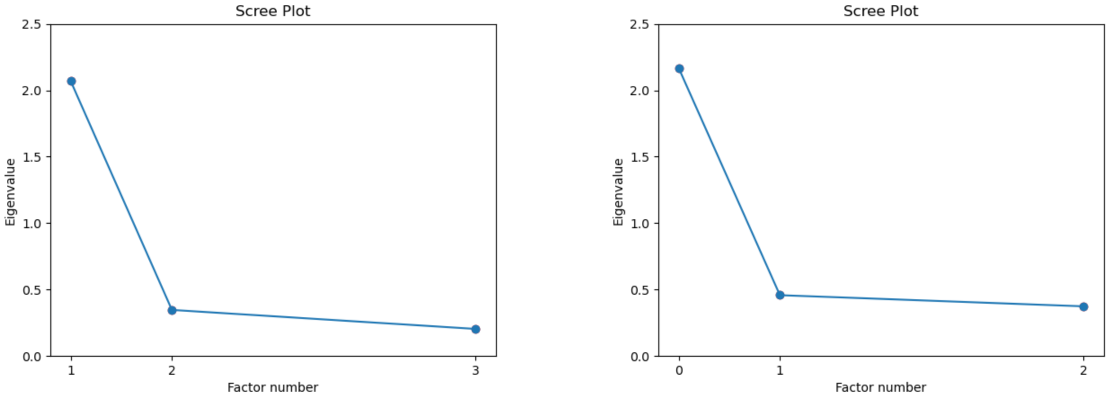


Figure 15 Monohulled Sailboats(Left)、Catamarans(Right)Principal Component Analysis Scree Plot

Combine regional factors into one principal component, named Comprehensive regional factors, which meets the requirement of PCA with over 70% variance explained. Use XGBoost regression analysis to determine the importance and impact of this new variable on the listing price of Monohulled Sailboats and Catamarans.

(1) Monohull Sailboats

Through XGBoost machine learning regression analysis, we obtained the feature importance of each factor and reflected the positive or negative impact of each factor on the Listing Price through the SHAP values. The results are shown in the following graph.

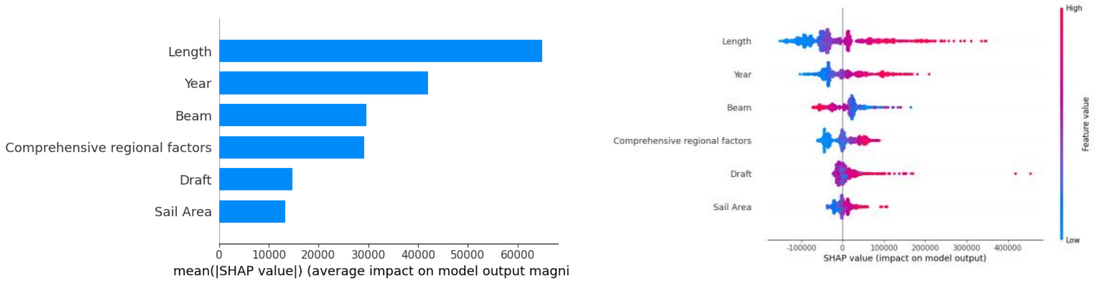


Figure 16 Feature Importance Plot of Each Factor

The Length factor has the greatest impact on the Listing Price, with regional factors also having a significant effect. The SHAP value data shows that the regional factors have a positive impact.

Next, we used MAPE and R2to reflect the accuracy of the data prediction. After training and testing the training set and test set at 7:3, the following fitting efficiency values were obtained.

Table 8 Model Evaluation Results Table

	MAPE	R^2
Training Set	0.170	0.8243968
Test Set	0.192	0.7499291

As shown in the table, the MAPE values of the training set and the test set are very close and both less than 1, and the R2values are both greater than 0.7, indicating that the XGBoost training results are good.

(2) Catamarans

The same process as above, first we obtained the feature importances of each factor and reflected the positive and negative impacts of each factor on the Listing Price through the SHAP values, as shown in the following graph.

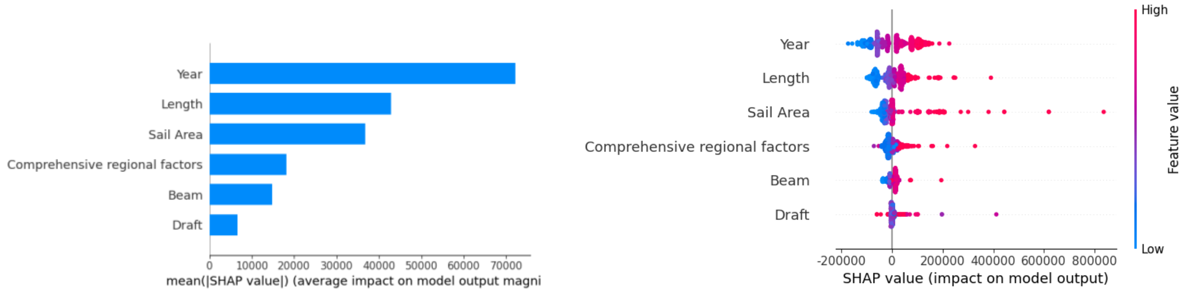


Figure 17 The graph of feature importance of each factor

The graph shows Length has the biggest impact on Year, while Comprehensive regional factors have a significant impact on Listing Price of Monohulled Sailboats, with a positive effect as shown by SHAP values. Our prediction accuracy was measured using MAPE and R^2 and we got fitting efficiency values after splitting the data 7:3 for training and testing.

Table 9 Model evaluation results table

	MAPE	R^2
Training Set	0.089	0.9225824
Test Set	0.111	0.7486544

The data in the table reveals that the MAPE values of the training set and test set are very close and both are less than 1, and R^2values are all greater than 0.7, indicating a good performance of the XGBoost training.

To understand the sailboat market and provide evidence for used sailboat pricing, we present you this report. We then explain the practical implications of our findings. As our results show, the sailboat's attributes such as Year, Length, and Sail Area have greater significance in affecting pricing compared to the Comprehensive regional factors. This aligns with the principle that the value of goods determines their price, suggesting the sailboat's attributes play a decisive role. Our conclusion is thus practical and in line with reality.

To provide a statistical explanation for the above results, we conduct a two-factor analysis of variance to explore the impact of different regions and different types of sailboats on pricing. The results of the one-factor and Catamaran variance analysis are shown in the following table.

Table 10 (L)Monohulled Sailboats(R)CatamaransVariance Analysis Table

Monohulled Sailboats						**Catamarans**				
dependent variable :ListingPrice						**dependent variable** :ListingPrice				
source	**III SS**	**df**	**MS**	**F**	**sig.**	**III SS**	**df**	**MS**	**F**	**sig.**

Modified model	4.05E+13[a]	1192	3.40E+10	12.688	.000	3.20E+13	354	9.05E+10	26.977	.000
intercept	2.29E+13	1	2.29E+13	8562.372	.000	3.27E+13	1	3.27E+13	9746.372	.000
Variant	2.80E+13	361	7.76E+10	29.004	.000	1.93E+13	88	2.20E+11	65.411	.000
Country/Region/State	9.69E+11	63	1.54E+10	5.748	.000	4.94E+11	41	1.20E+10	3.588	.000
Variant*Country/Region/State	3.88E+12	768	5.05E+09	1.887	.000	3.10E+12	224	1.38E+10	4.121	.000
error	2.66E+12	996	2.68E+09			2.20E+12	656	3.36E+09		
sum	1.56E+14	2189				2.26E+14	1011			
corrected total sum	4.31E+13	2188				3.42E+13	1010			
R squared = .938 (Adjusted R squared = .864)						a. R-squared = .936 (Adjusted R-squared = .901)				

The data in the table shows that the F values of the factors Variant, Country/Region/State, and Variant * Country/Region/State are relatively high and the significance P values are all less than 0.05, indicating that the model has statistical significance and regional factors have a significant impact on the pricing of used sailboats.

6 Used Sailboat Price Prediction Based on the Trained XGBoost Model

6.1 The Establishment of Model 3

In order to obtain a sailboat subset with the largest amount of information, we use the initial screening criterion of sample size greater than 25 in the variant and obtain subsets of monohull sailboats and Camatarans sailboats. Afterwards, we use Python crawling techniques to gather data from multiple sailboat websites in Hong Kong (such as http://www.luxboating.com/pre-owned-boat.html) in order to obtain as many prices for the sailboats in the selected subset as possible. Finally, we combine the price data we have obtained to determine the selected sailboat subset.

After collecting data, we obtained the Hong Kong region's 2020 per capita GNI, GDP, and average ratio of total logistics costs to GDP, which were 48630 US dollars, 344.9 billion US dollars, and 3.2%, respectively. Then, we replaced all regional-related influencing factors with Hong Kong data and kept other influencing factor data unchanged. We then input all data on factors affecting the sale price of sailboats into the previously constructed and trained XGBoost regression prediction model and use the two training models built in Python to predict the sale prices of single-hulled and double-hulled sailboats. Finally, we export the predicted prices of sailboats in the subset in Hong Kong and compare them with the actual sailboat prices collected.

6.2 The Solution of Model 3

Initially, we selected the sailboat types in the variant variable with a sample size greater than 25, and conducted regression prediction using XGBoost, and compared the results with the collected real price data to check the fitting goodness of fit. The fitting result is shown in the following chart.

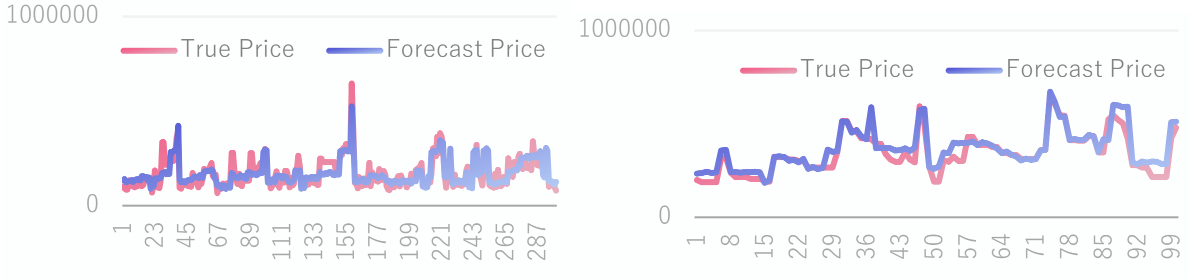


Figure 18 Model Fitting Performance Plot

Table 11 Model Evaluation Results Table

	MAPE	R^2
单体帆船	0.095	0.8025596
双体帆船	0.138	0.7762451

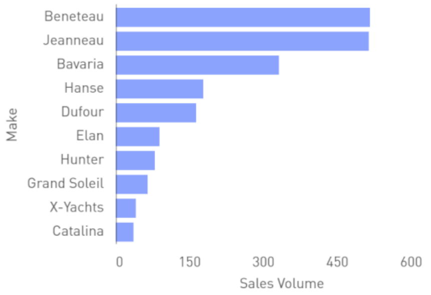
The data in the chart shows that the model has a good fit for both single-hulled and double-hulled sailboats, indicating that the previously established XGBoost model still has high applicability in the Hong Kong Special Administrative Region. Additionally, the data in the table shows that the model has better fitting results for monohull sailboats compared to catamrans sailboats.

7 Interesting and informative Conclusions

After our team processed a substantial amount of data, built models, and derived conclusions, we gained a deeper understanding of the sailboat market and sailboat pricing, as well as discovered some intriguing findings. We have summarized several conclusions about best-selling sailboats with the aim of providing brokers with more valuable information.

Conclusion 1: Monohull sailboats are more popular than catamaran sailboats.

Based on the data provided in the study, there are a total of 62 brands of monohull sailboats, while there are only 20 brands of catamaran sailboats, with the sales volume of monohull sailboats far exceeding that of catamarans.



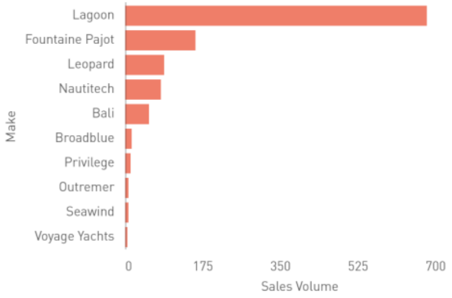


Figure 19 Best-selling Monohull sailboats

Figure 20 Best-selling Catamarans

Conclusion 2: Beneteau, Jeanneau, Bavaria, and Lagoon are the most popular sailboat brands.We present the top 10 best-selling brands for both monohull and catamaran sailboats. Among monohull sailboats, the top three most popular brands are Beneteau, Jeanneau, and

Bavaria. For catamaran sailboats, Lagoon is the most popular brand, with the second most popular brand's sales volume significantly lagging behind.

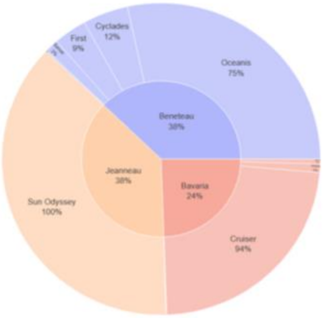


Figure 21 Sunrise Graph of Best-Selling Monohull Sailboat Series

Conclusion 3: The best-selling sailboat series of Beneteau, Jeanneau, and Bavaria are Oceanis, Sun Odyssey, and Cruiser, respectively. The Oceanis series focuses on cruising and comfort, while Sun Odyssey prioritizes performance for long-distance sailing and comfort. The Cruiser series is designed for ease of use, comfort, and high performance for families and sailing enthusiasts. The popularity of these sailboats highlights a demand for high-performance, comfortable, and easy-to-use vessels for cruising.

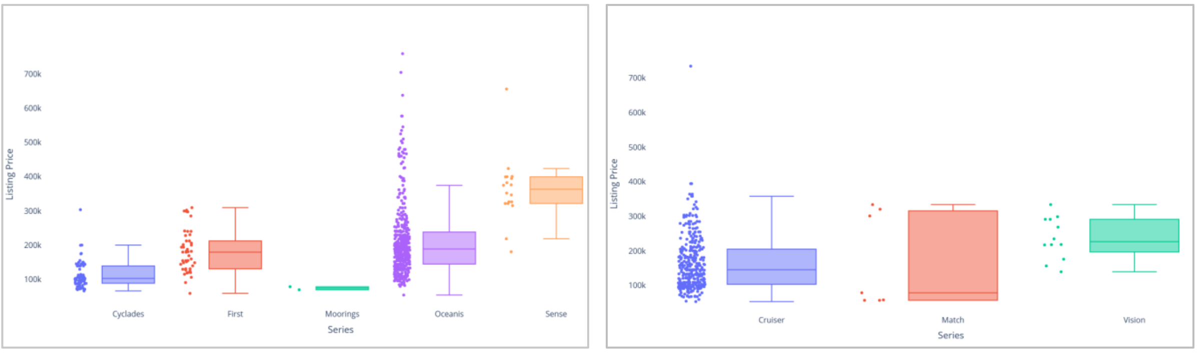


Figure 22 Box Plot of Prices for Series of Best-Selling Make: (L)Beneteau (R)Bavaria

Conclusion 4: Sailboat series prices vary due to design and performance differences. Price box plots for different series within Beneteau and Bavaria brands show variation. Beneteau's Oceanis, First, Cyclades, and Moorings are priced in descending order. Bavaria's Vision is the highest, while Cruiser and Match have similar prices. Moorings target yacht charters, while Sense offers luxury at a premium. Oceanis provides performance, comfort, and durability. First focuses on racing, while Cyclades are more affordable. Vision prioritizes luxury, while Match focuses on racing and Cruiser emphasizes practicality and comfort at a lower cost.

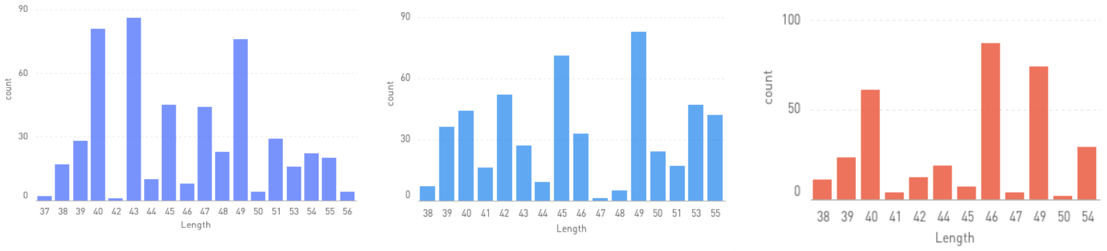


Figure 23 Hot-selling Sailboat Size Count：(A)Beneteau (B)Jeanneau (C)Bavaria

Conclusion 5: For Beneteau, 43, 40, and 46 ft sailboats are best-sellers. For Jeanneau, it's 49, 45, and 42 ft. And for Bavaria, 46, 49, and 40 ft are the most popular. Brokers can promote these models by advertising their benefits, using social media, and exhibiting at events. They can also build strong ties with manufacturers to secure priority access to these top-selling models and increase competitiveness.

8 Model Evaluation and Further Discussion

8.1 Strengths

1. Interpretability: SHAP values provide a method for quantifying the importance of model features and also allow for the creation of positive and negative effect plots for each indicator, making the predictions of the XGBoost model more interpretable and understandable.

2. Flexibility: XGBoost supports various objective functions and evaluation metrics, which can be adjusted based on the actual problem. Combined with SHAP, it can better understand the interaction between features.

3. Robustness: XGBoost has robustness to outliers and missing data, and SHAP values can also help identify outliers that have a significant impact on the prediction result.

4. Excellent visualization: We deeply study the data, extract effective data features, visualize the features, and explore the patterns in the data.

8.2 Weaknesses

The XGBoost model has multiple hyperparameters that need to be adjusted, increasing the difficulty and time for model tuning. Not using more machine learning models to further explore the factors affecting the sale price of sailboats.

References

[1]Portofino Yachting, April 1,2023, https://portofinoyachting.com/.

[2]Irmak Yachting. (April 1.2023.). Bavaria 42 Cruiser. Retrieved Month Day, Year, from https://www.irmakyachting.com/

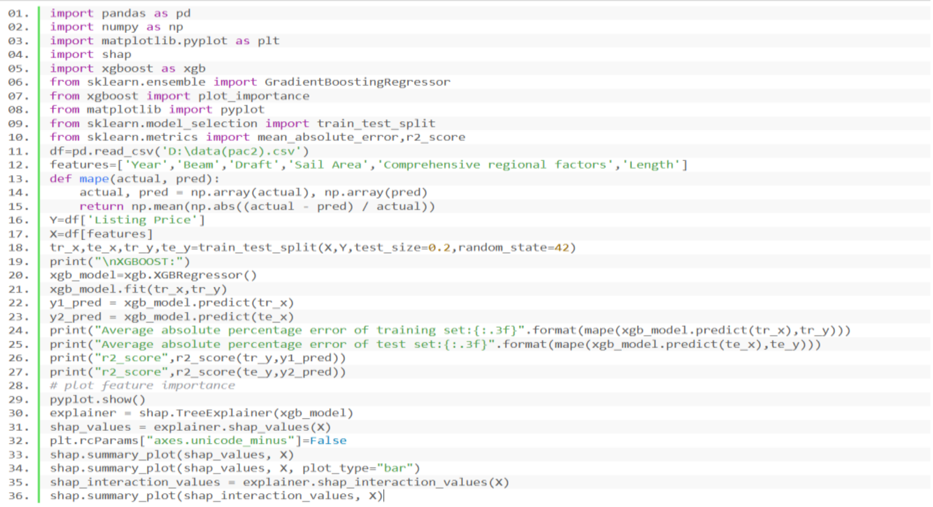
[3]Inautia. (April 1.2023.). Homepage. Retrieved Month Day, Year, from https://www.inautia.com/

[4]Clipper Marine. (April 1.2023.). Homepage. Retrieved Month Day, Year, from https://www.clippermarine.co.uk/

Appendices

Appendix 1

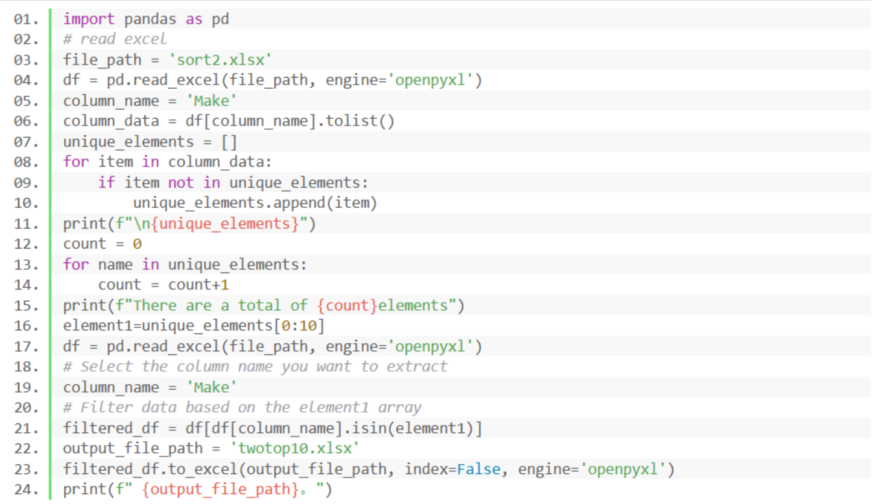
Introduce: Python code for XGBoost Regression Model



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import shap
import xgboost as xgb
from sklearn.ensemble import GradientBoostingRegressor
from xgboost import plot_importance
from matplotlib import pyplot
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error,r2_score
df=pd.read_csv('D:\data(pac2).csv')
features=['Year','Beam','Draft','Sail Area','Comprehensive regional factors','Length']
def mape(actual, pred):
    actual, pred = np.array(actual), np.array(pred)
    return np.mean(np.abs((actual - pred) / actual))
Y=df['Listing Price']
X=df[features]
tr_x,te_x,tr_y,te_y=train_test_split(X,Y,test_size=0.2,random_state=42)
print("\nXGBOOST:")
xgb_model=xgb.XGBRegressor()
xgb_model.fit(tr_x,tr_y)
y1_pred = xgb_model.predict(tr_x)
y2_pred = xgb_model.predict(te_x)
print("Average absolute percentage error of training set:{:.3f}".format(mape(xgb_model.predict(tr_x),tr_y)))
print("Average absolute percentage error of test set:{:.3f}".format(mape(xgb_model.predict(te_x),te_y)))
print("r2_score",r2_score(tr_y,y1_pred))
print("r2_score",r2_score(te_y,y2_pred))
# plot feature importance
pyplot.show()
explainer = shap.TreeExplainer(xgb_model)
shap_values = explainer.shap_values(X)
plt.rcParams["axes.unicode_minus"]=False
shap.summary_plot(shap_values, X)
shap.summary_plot(shap_values, X, plot_type="bar")
shap_interaction_values = explainer.shap_interaction_values(X)
shap.summary_plot(shap_interaction_values, X)
```

Appendix 2

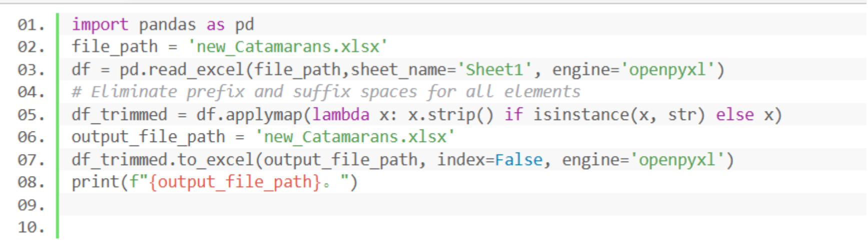
Introduce: Python code to match the 'Make' to filter the raw data



```
import pandas as pd
# read excel
file_path = 'sort2.xlsx'
df = pd.read_excel(file_path, engine='openpyxl')
column_name = 'Make'
column_data = df[column_name].tolist()
unique_elements = []
for item in column_data:
    if item not in unique_elements:
        unique_elements.append(item)
print(f"\n{unique_elements}")
count = 0
for name in unique_elements:
    count = count+1
print(f"There are a total of {count}elements")
element1=unique_elements[0:10]
df = pd.read_excel(file_path, engine='openpyxl')
# Select the column name you want to extract
column_name = 'Make'
# Filter data based on the element1 array
filtered_df = df[df[column_name].isin(element1)]
output_file_path = 'twotop10.xlsx'
filtered_df.to_excel(output_file_path, index=False, engine='openpyxl')
print(f" {output_file_path}。")
```

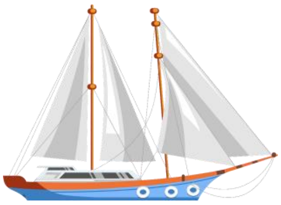
Appendix 3

Introduce: Python code for removing the prefix and suffix spaces from the raw data



```
import pandas as pd
file_path = 'new_Catamarans.xlsx'
df = pd.read_excel(file_path,sheet_name='Sheet1', engine='openpyxl')
# Eliminate prefix and suffix spaces for all elements
df_trimmed = df.applymap(lambda x: x.strip() if isinstance(x, str) else x)
output_file_path = 'new_Catamarans.xlsx'
df_trimmed.to_excel(output_file_path, index=False, engine='openpyxl')
print(f"{output_file_path}。")


```

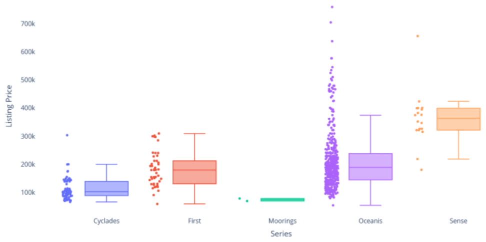


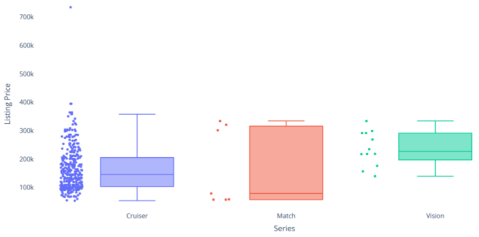
Report for sailboat Broker

Conclusion

About popular sailboat

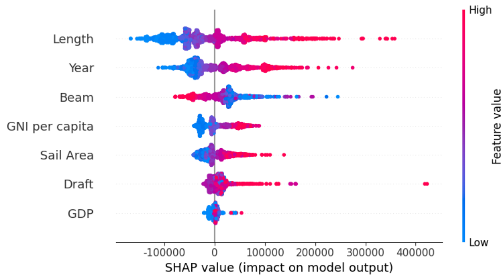
- ✧ Monohull sailboats are more popular than catamaran sailboats.
- ✧ Among monohull sailboats, the top three most popular brands are Beneteau, Jeanneau, and Bavaria.
- ✧ For catamaran sailboats, Lagoon is the most popular make.
- ✧ The best-selling sailboat series of Beneteau, Jeanneau, and Bavaria are Oceanis, Sun Odyssey, and Cruiser, respectively.
- ✧ Sailboat series prices vary due to design and performance differences.
- ✧ For Beneteau, 43, 40, and 46 ft sailboats are best-sellers.
- ✧ For Jeanneau, it's 49, 45, and 42 ft. And for Bavaria, 46, 49, and 40 ft are the most popular.

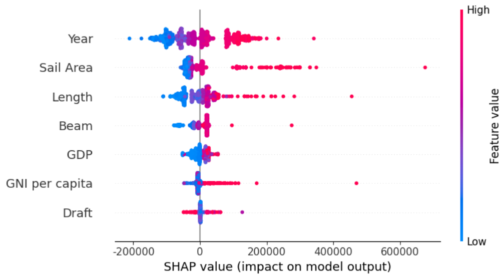




Important Features Impacting Lising Price

- ✧ For monofull sailboat,feature importance ranking is Length > Year > Beam > GNI per capita > Sail Area > Draft > GDP.
- ✧ For Catamarans, feature importance ranking is Year > Sail Area > Length > Beam > GDP > GNI per capita > Draft.

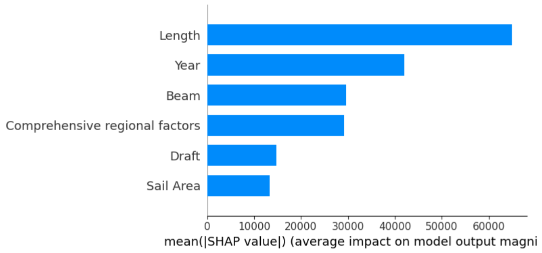


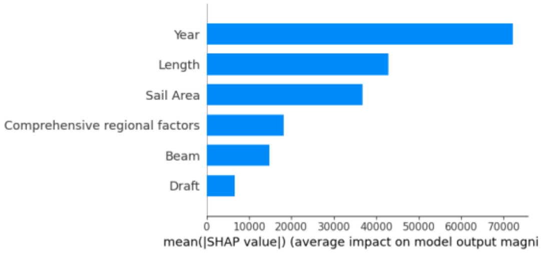


Impact of Regional Factors on Sailboat Pricing

✧ Comprehensive regional factors have some effects on the listing price the same for both catamarans and monohull sailboats.

✧ Year, Length, and Sail Area have greater significance in affecting pricing compared to the Comprehensive regional factors.





Applicability of the Conclusion to HK market

✧ The model and conclusion we established have a relatively high degree of applicability to the Hong Kong sailboat market

Suggestion

➢ Prioritize recommending popular sailboats, popular series, and popular sizes to consumers, as they are more likely to be attracted.

➢ Based on the factors affecting the pricing of sailboats and the importance of these factors that we have obtained, we provide a reasonable and scientific pricing for the used sailboat.