Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2330176

Brand is Everything? Perhaps Not!

Brand Premium, Regional Effect and Pricing Strategy in the Used Sailboat Market

Summary

This paper analyzes and studies the pricing and consistency of regional effects in the used sailboat market by constructing mathematical models and simulating the regional effects of Hong Kong (SAR) on sailboat listing prices. The main models used in this study include the Integrated Brand Premium Index Model, the Used Sailboat Pricing Model, and the Hong Kong (SAR) Simulated Pricing Model.

Specifically, we use the basic attributes of used sailboats and brand premium indices to construct an IBPI model to explore the impact of brand premiums on sailboat pricing. After data preprocessing and feature engineering, we conducted XGBoost regression analysis to establish the used sailboat pricing model and explain the impact of regional factors on listing prices. We demonstrated that regional effects are not consistent and resolved any practical and statistical significance of the regional effects identified. We also constructed the Hong Kong (SAR) simulated pricing model and explained the regional effects in Hong Kong.

For Q1, we construct the Used Sailboat Pricing Model using the IBPI and 21 important features such as GDP, GDP per capita, displacement, and year through XGBoost regression analysis, which effectively explains the listed price of used sailboats, with R-squared values exceeding 0.9 in estimation accuracy.

For Q2, we conducted hypothesis testing to discuss whether regional effects on listing prices are consistent. We ultimately obtain significant p-values less than 0.05, indicating that regional effects on listing prices are significant. Through correlation analysis, we found that regional effects vary across different sailboat models. This indicates that the factors influencing geographical regions are complex and require consideration of the interaction of other factors. Additionally, we explained the practical and statistical significance of regional effects.

For Q3, we construct the Hong Kong (SAR) Simulated Pricing Model to study the regional effects of Hong Kong (SAR) on sailboat listing prices. We discover the differences in the Hong Kong market for different types of vessels and analyze that the regional effects for monohull and catamaran boats are different.

For Q4, we found that the sailboat market in the United States is more developed and that more people are willing to buy high-value sailboats. Bavaria products have low premiums and high cost-effectiveness, while the Discovery brand has significantly overpriced products. We also explain the impact of herd effect, endowment effect, and calendar effect on the listed price of used sailboats.

Keywords: Used sailboats, Integrated Brand Premium Index, XGBoost regression analysis, pricing model, regional effect.

Contents

1 Introduction **2**
- 1.1 Problem Background . . . 2
- 1.2 Restatement of the Problem . . . 2
- 1.3 Problem Analysis . . . 2
- 1.4 Our Work . . . 4

2 Assumptions **4**

3 Notations **5**

4 Model Establishment and Solution **5**
- 4.1 Data Preprocessing . . . 5
- 4.2 Feature Engineering and Standardization . . . 6
  - 4.2.1 Integrated Brand Premium Index: Nominal and Real Prices . . . 7
  - 4.2.2 Regional Factors . . . 8
  - 4.2.3 Performance Factors . . . 10
  - 4.2.4 Manufacturing Year . . . 10

5 Model for Question 1 **10**
- 5.1 Model Establishment . . . 10
- 5.2 Model Results and Model Evaluation . . . 11

6 Model for Question 2 **13**
- 6.1 The Impact of Regions on List Prices . . . 13
- 6.2 Is the regional effect consistent across all sailboat models? . . . 14
- 6.3 Statistical and Practical Significance . . . 15

7 Model for Problem 3 **16**
- 7.1 Regional Effects . . . 17
  - 7.1.1 Analysis on monohull sailboat . . . 17
  - 7.1.2 Analysis on Catamaran . . . 19
  - 7.1.3 Regional Effects on Monohull and Catamaran Prices . . . 20

8 Some interesting and informative inferences or conclusions **20**

9 Evaluation and Promotion of the Model **21**
- 9.1 Advantages and Disadvantages of the Pricing Model . . . 21
- 9.2 Model Improvement and Promotion . . . 22

10 Report **23**

Reference **25**

1 Introduction

1.1 Problem Background

In recent years, the used sailboat market has attracted widespread attention due to its complex pricing dynamics and potential investment opportunities. Sailboats are often considered luxury goods, and their value may depreciate to varying degrees over time and with changes in market conditions. The pricing dynamics of used sailboats are critical to brokers, buyers, and sellers alike, and understanding the factors that influence sailboat pricing and making high-precision predictions of listing prices is of paramount importance in this context.

In this study, we will develop a comprehensive mathematical model to capture the complex interactions among factors that influence sailboat pricing. **Using XGBoost regression analysis**, our model aims to predict the listing prices of sailboats in the provided dataset, taking into account various predictive factors and their relationships. Additionally, we will investigate the impact of geographic region on listing prices, with a particular focus on the Hong Kong (SAR) market.

Our research results not only provide insights into the dynamics of the used sailboat market but also offer guidance for brokers, buyers, and sellers to make informed decisions in this rapidly evolving market. By ensuring the readability and logical coherence of our analysis, we aim to present a high-quality problem background and lay a foundation for our rigorous study of sailboat pricing in this fascinating domain.

1.2 Restatement of the Problem

Given the complex pricing dynamics and various factors that influence the used sailboat market, this study aims to address the following key problems:

- **Develop a comprehensive mathematical model** that can capture the complex relationships among various factors that affect sailboat pricing, including manufacturer, model, manufacturing year, geographic location, and other features.
- Use the developed model to predict the listing prices of sailboats in the provided dataset with high accuracy, taking into account potential predictive factors and their interrelationships.
- Investigate the impact of geographic regions on sailboat pricing, with a focus on understanding regional effects and their consistency across different sailboat variants.
- Apply the mathematical model to assess the regional impact of Hong Kong (SAR) on sailboat prices, providing valuable insights into the local market dynamics and comparing the impact on catamarans and monohull sailboats.

By addressing these problems, our study aims to comprehensively understand the factors that influence used sailboat pricing and provide valuable guidance for brokers, buyers, and sellers to make informed decisions in this complex market.

1.3 Problem Analysis

As a mathematical modeling team, our goal is to develop a comprehensive mathematical model that explains the pricing dynamics of used sailboats and investigates the impact of geographic regions on listing prices. To achieve this goal, we conducted a comprehensive problem analysis to identify the key aspects of the problem and develop a systematic approach.

- **Data understanding and preprocessing**: The provided dataset contains information on approximately 3,500 sailboats, including their manufacturer, variant, length, geographic region, country/region/state, listing price, and manufacturing year. We will begin exploring the dataset to identify any missing data or inconsistencies that require preprocessing and cleaning. This step is critical to ensuring the quality and reliability of our analysis.
- **Feature engineering and selection**: In addition to the variables provided in the dataset, other factors such as beam width, draft, displacement, rigging, sail area, hull material, engine running time, sleeping capacity, headroom, and electronic equipment may affect sailboat pricing. We will conduct comprehensive research to collect additional data from reliable sources and integrate it into our dataset. Additionally, we will perform feature selection techniques to identify the most relevant predictors for our model.
- **Exploratory data analysis**: Before developing the mathematical model, we will conduct exploratory data analysis to gain a deeper understanding of the relationships between variables, identify potential predictors, and detect any potential patterns in the data. This step is crucial for determining our modeling approach, ensuring that we capture the key dynamics that affect sailboat pricing.
- **Model development and evaluation**: Based on our exploratory data analysis, we will select an appropriate XGBoost regression algorithm to develop a mathematical model that accurately predicts sailboat listing prices. We will optimize our model and evaluate its performance using relevant metrics such as mean squared error and R-squared.
- **Regional effect analysis**: Using our developed model, we will study the impact of geographic regions on sailboat listing prices, with a focus on potential regional differences and their consistency across different sailboat variants. This analysis will provide valuable insights into the regional dynamics of the used sailboat market.
- **Hong Kong (SAR) market evaluation**: We will apply our model to evaluate the impact of Hong Kong (SAR) on sailboat prices. By comparing prices of catamarans and monohulls in the local market, we will determine if regional effects remain consistent across different sailboat types.
- **Conclusion and report**: Finally, we will synthesize our research findings to draw interesting and insightful conclusions from our analysis. Additionally, we will prepare a one to two-page report with carefully selected graphics for Hong Kong (SAR) sailboat brokers to help them understand our findings.
- By conducting a rigorous problem analysis, we ensure that our approach is academically sound, methodologically robust, and relevant to the problem at hand. This enables us to develop a comprehensive mathematical model that provides valuable insights and assists brokers, buyers, and sellers in making informed decisions in this complex market.

1.4 Our Work

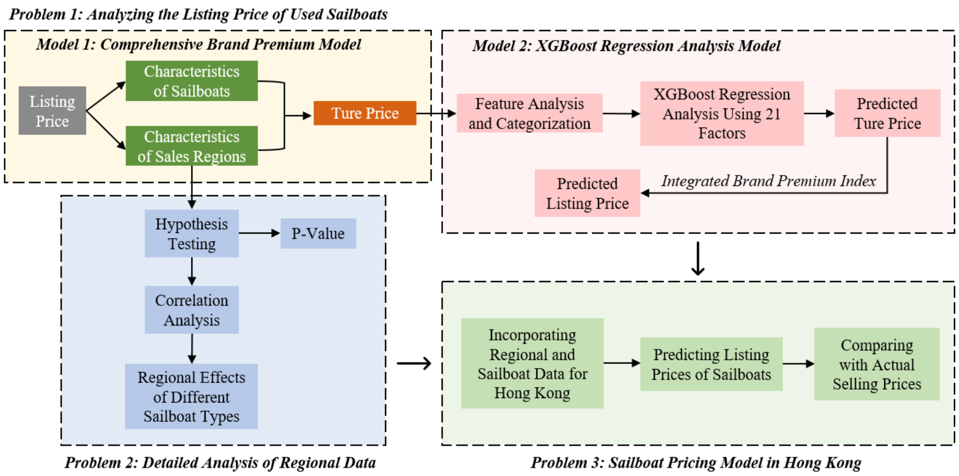


Figure 1: Indicator selection

2 Assumptions

To develop a comprehensive and reliable mathematical model for the used sailboat market, we make the following assumptions. These assumptions are necessary to simplify complex real-world issues and enable us to focus on the most critical factors that affect sailboat pricing.

Firstly, we assume that all used sailboat transactions **are settled in US dollars without considering the impact of exchange rates**. This assumption is based on the fact that the US dollar is the primary international currency unit in boat transactions, and ignoring the impact of exchange rates can simplify the model and reduce uncertainty. We also assume that **the purchasing power of currencies remains stable in different regions and years**, meaning that inflation or currency exchange rate fluctuations do not significantly affect prices. This assumption allows us to better focus on other factors that affect sailboat pricing and build a simpler and more interpretable model.

Secondly, we assume that **all sailboat transactions comply with legal regulations**, with no illegal behavior occurring during the transaction process. We assume that **all transactions are voluntary**, with both buyers and sellers participating willingly and without any coercive requirements or unfair conditions. We further assume that **no fraud occurs during the transaction process**, with sellers providing accurate and trustworthy sailboat information and prices without intentional deception or false advertising. We assume that **the market value of sailboats can be reasonably estimated using historical data and other factors**, not solely based on supply and demand and market expectations. These assumptions make the model more accurate and reliable while broadening its scope of application.

- **Nonlinearity between predictor variables and list prices**: We recognize that the relationship between predictor variables and list prices may **not be linear**. Therefore, we choose to use the **XGBoost regression model**, which can handle nonlinear relationships and has good predictive accuracy and interpretability.

- **Independent and identically distributed errors**: We assume that errors in the model are independent and identically distributed. This assumption is crucial for ensuring the effectiveness of statistical tests and the accuracy of model predictions. It also means that our model considers all relevant factors, and any remaining errors are random and uncorrelated.
- **No multicollinearity among predictor variables**: We assume that there is no multicollinearity among predictor variables, meaning that no predictor variable is a linear combination of other predictor variables. This assumption ensures the stability and interpretability of our model coefficients. However, we will take measures to address collinearity issues in feature selection and engineering.
- **Homoscedasticity of errors**: We assume that the variance of errors is constant across different levels of predictor variables. This assumption is crucial for the reliability of statistical tests and the accuracy of model predictions. If heteroscedasticity is detected, we will consider appropriate data transformations or alternative modeling techniques.
- **Representativeness of geographic regions**: We assume that the geographic regions provided in the dataset (Europe, Caribbean, and the United States) represent the global market for used sailboats. However, we will incorporate specific geographic region features in feature engineering to consider the impact of geographic differences on pricing.
- **Data quality and completeness**: We assume that the dataset provided is of high quality and contains accurate information. Although we recognize that datasets in the real world often have missing data or other issues, we will perform data preprocessing and cleaning to address these issues and ensure the robustness of our analysis.

3 Notations

Symbols	Description
X	original feature
P_n	nominal price
P_r	actual price
L	length
$PB^{\star}$	length-weighted brand premium index
$PB_{i,j}$	integrated brand premium index for brand i of sailboat in geographical region j

4 Model Establishment and Solution

4.1 Data Preprocessing

Data preprocessing is an important step in establishing any effective mathematical model. In this problem, we need to preprocess the provided sailboat sales data. This section will introduce our data preprocessing method.

First, we need to check for missing and outlier values in the data. In this problem, we used two tabs in an Excel file, one for monohull sailboats and the other for catamarans. We used the Python programming language and the Pandas library to read and clean the data. We found that all columns in both the monohull and catamaran tabs were complete and had no missing values. Therefore, we did not perform any missing value handling.

Next, we performed outlier detection on the data. We checked for outliers in the listing price column. We used the "$3-\sigma$" method to visualize the data and detect outliers from it. We removed these outliers and recorded the number of data rows deleted. Finally, **we obtained a cleaned dataset consisting of 2865 sailboat sales data, including 1888 monohull data and 977 catamaran data**.

Finally, we standardized the dataset. We used normalization techniques to eliminate the proportional differences between different features in the dataset. We standardized the continuous features in the dataset and replaced the original values with the standard deviation and mean.

In summary, our data preprocessing process includes data cleaning, outlier detection, feature engineering, and data normalization. Our data preprocessing process ensures that our dataset is clean, consistent, and can be used for mathematical modeling analysis.

4.2 Feature Engineering and Standardization

Feature engineering is an important step in building mathematical models, involving feature selection, feature extraction, and feature transformation. In this problem, we need to build a mathematical model to predict the price of used sailboats based on the provided sales data. In this process, we need to perform some feature engineering to determine which features are useful for predicting sailboat prices.

In terms of feature transformation, we standardize the continuous features to better compare the differences between different features. At the same time, we standardize the value range of features with a large value range, such as Displacement, while for features with a small value range, such as Comfort Ratio, we do not need to standardize it. Standardizing Comfort Ratio may cause its value to become very small, losing its original information. On the other hand, Displacement has a much larger value range than Comfort Ratio, so it needs to be standardized to compare with other variables.

The formula for standardization is as follows:

$$X_{std} = \frac{X - \mu}{\sigma} \tag{1}$$

Where X is the original feature, X_{std} is the standardized feature, μ is the mean of the feature, and σ is the standard deviation of the feature.

The formula for the mean and standard deviation of a feature are as follows:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}} \tag{3}$$

Where n is the number of samples and X_i is the value of the feature for the i-th sample.

4.2.1 Integrated Brand Premium Index: Nominal and Real Prices

The Integrated Brand Premium Index (**IBPI**) is a commonly used method to account for brand effects in the commodity market, where brand reputation, quality, and recognition are important factors influencing consumer purchasing decisions. In the sailboat market, different sailboat brands often have unique characteristics and brand effects that can impact their prices. Therefore, constructing an **IBPI** is necessary and reasonable to consider the impact of brand effects on sailboat prices, and to improve the accuracy of the prediction model. **IBPI** can be calculated by quantifying brand recognition, reputation, quality, history, and other factors, and considering them in sailboat price modeling. In this way, **constructing an IBPI has a certain necessity and rationality in sailboat price modeling**.

In this problem, we define the list price as the nominal price, while the proposed actual price after accounting for the brand premium is used in XGBoost regression analysis. Therefore, the formula for the nominal price P_n is:

$$P_n = \frac{P_r}{PB_{i,j}} \tag{4}$$

Where P_n represents the nominal price, i.e., the listed price, P_r represents the actual price, and $PB_{i,j}$ represents the integrated brand premium index for brand i of sailboat in geographical region j.

Next, we will explain the calculation method and rationality of the Integrated Brand Premium Index.

First, we consider the brand premium index: defining brand premium as an index relative to a benchmark brand. For example, taking a well-known brand as the benchmark, the brand premium of other brands is calculated relative to the benchmark brand. In Europe, the United States, and the Caribbean, we selected *Beneteau* as the preferred benchmark for our monohull boats.

Beneteau is one of the world-renowned sailboat brands, with high market share, brand influence, and reputation, and also has a certain market share in Europe, the United States, and the Caribbean. Therefore, selecting *Beneteau* as the benchmark brand can improve the reliability and interpretability of the brand premium index and provide better reference value for subsequent analysis and decision-making. Similarly, for catamarans, we selected *Lagoon* as the benchmark brand.

Considering that the price data in the Excel spreadsheet is the price of used sailboats when sold in 2020, and the year is the year of manufacture, it can be assumed that these prices have already undergone the consumption and depreciation of time, and the comparison of prices of sailboats of different brands and models has a certain comparability and may not need to consider the influence of time factors.

Furthermore, considering that the prices of boats of different lengths do vary greatly, considering the boat length as a weighting factor may better reflect the brand premium, as boats of different lengths often represent different models and configurations, and the price differences will also be correspondingly large.

Therefore, when calculating the weighted brand premium index, multiplying the price sum of boats of the same brand and different series by their respective lengths and dividing it by the sum of the lengths of boats of the same brand and different series can obtain the length-weighted brand premium index, which can more comprehensively reflect the brand premium situation.

$$PB^{\star} = \frac{\sum_{i=0}^{n} L \cdot P_n}{\sum_{i=0}^{n} L} \tag{5}$$

When multiple factors such as regional factors, length, year, brand market share, brand reputation, popularity, and word-of-mouth are considered, we optimize $PB^{\star}$ to $PB_{i,j}$, which is called the "**Integrated Brand Premium Index**". This index can more comprehensively reflect the ratio of prices between used sailboats of the same brand and different models to those of the same type of used sailboats of the benchmark brand, and can more accurately reflect the brand premium situation.

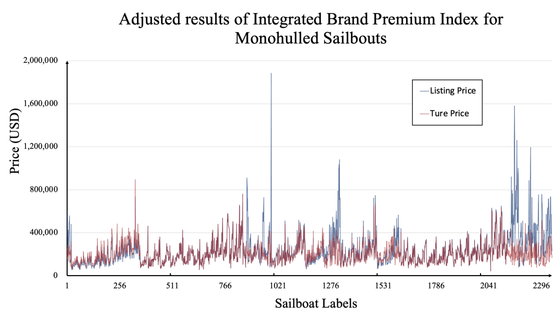


Figure 2: Monohulled

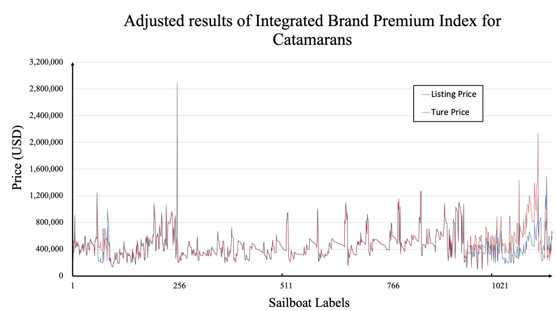


Figure 3: Catamarans

4.2.2 Regional Factors

From a practical perspective, different geographical locations and regions have a significant impact on the value and pricing of sailboats due to their varying economic levels, climate conditions, cultural backgrounds, sailboat market demand, policies, and other factors.

Firstly, regional factors can reflect the sailboat's applicability and market demand. For example, in Europe, due to the influence of historical and cultural factors, sailing culture is well developed, and market demand is relatively high. Therefore, under the same conditions, sailboat prices in Europe may be higher than in other regions. In some areas, such as tropical and island areas, sailboats may have a broader range of uses, so the value of sailboats in these areas may also be higher.

Secondly, the climate and environmental conditions in different regions also have a significant impact on the value and lifespan of sailboats. For example, in the marine environment, factors such as wind and waves, as well as saltwater, can affect the wear and tear of boats, while inland waterways may have different climate and environmental pressures, which can also affect the sailboat's lifespan and value.

Therefore, in terms of regional factors, geographical factors that affect sailboat pricing include climate, coastline length, population[**3**], population density, etc. [**1**][**4**]

Regarding policy factors, tax policies are an important factor in regional economic factors. Some countries and regions affect sailboat pricing by implementing differential tax rates. For example, some countries and regions impose higher taxes on used sailboats but lower taxes on new sailboats. Such tax policies may lead to differences in sailboat pricing between different regions.

In terms of regional factors, the most influential factor is likely to be economic factors, such as capital and income. According to the relevant knowledge in "Monetary Banking"[**7**], we know that the pricing of goods is closely related to **supply and demand**. Here, we are obviously concerned with the demand side. According to the expenditure approach, we know that:

$$GDP = C + I + G + NX \tag{6}$$

Where C is household consumption, I is investment, G is government expenditure, and NX is net exports. Because **GDP** involves multiple sectors and factors, **GDP** is a very important economic factor in our pricing. In addition, since we are targeting the pricing of sailboats, per capita **GDP** and the proportion of the tourism industry in **GDP** are also important. Considering the impact of inflation, **CPI** and other related indicators can be considered to calculate the inflation rate. However, to simplify the model, we assume that the purchasing power of the currency remains stable. Finally, the economic level and competition in the sailboat market in different regions will also have an impact on sailboat pricing. For example, in economically developed regions, sailboat market competition may be more intense, and therefore, under the same conditions, sailboat prices may be higher.[**2**]

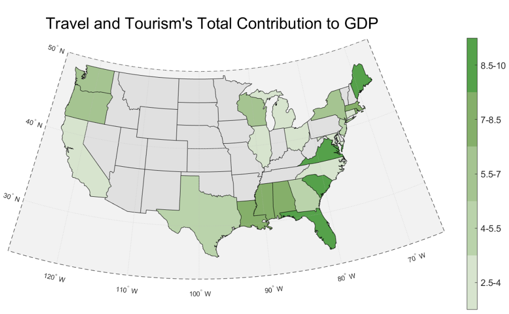


Figure 4

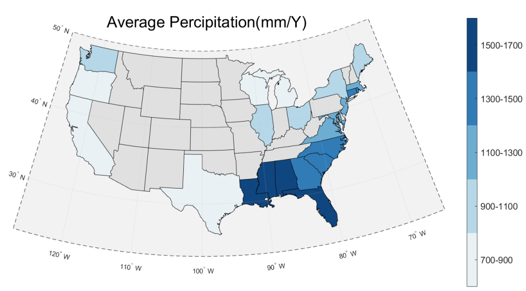


Figure 5

Therefore, considering regional factors in sailboat pricing is of great significance. It can more comprehensively and accurately reflect the actual value of sailboats and provide more effective guidance to help buyers and sellers make more informed decisions.

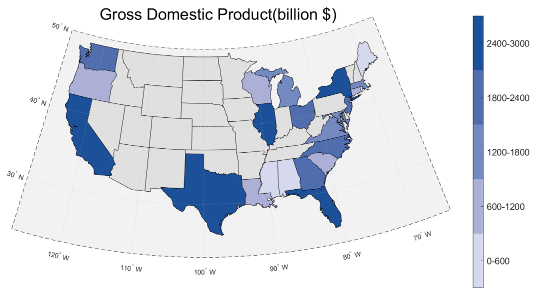


Figure 6

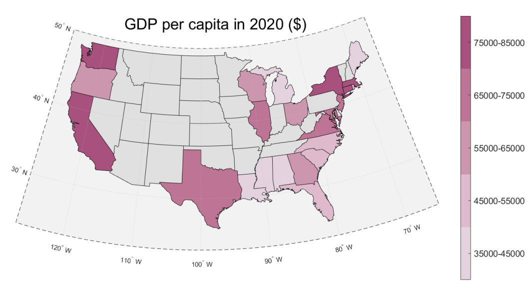


Figure 7

4.2.3 Performance Factors

In pricing sailboats, performance factors play a significant role. Sailboat performance factors include but are not limited to speed, stability, seaworthiness, comfort, and various parameters. These factors are crucial for the safety and comfort of navigation and also represent various characteristics required by the vessel.[**5**]

Additionally, performance factors can also serve as an essential basis for comparing different sailboats. When purchasing a sailboat, buyers typically compare performance differences among different models and brands. Therefore, considering performance factors in the pricing model can help more accurately reflect the actual value of sailboats and provide effective guidance for buyers and sellers to make more informed decisions.

4.2.4 Manufacturing Year

The manufacturing year is another important factor that affects sailboat prices. Generally, newer sailboats have higher prices because they typically have more modern features and technology, as well as longer lifetimes. In addition, the manufacturing year also reflects market demand and trends, such as the recent focus on sustainability and environmental protection, which may affect the prices of older boats. Therefore, considering the manufacturing year factor in modeling is crucial for accurately predicting sailboat prices. By modeling the manufacturing year factor, we can determine the price trends of sailboats of different years and the relative prices of sailboats of different years during a specific period. This helps sailboat brokers and buyers understand market trends and reasonable price ranges, and also helps manufacturers develop appropriate pricing strategies and improve product design and features to enhance the product's market competitiveness.

5 Model for Question 1

5.1 Model Establishment

When we need to predict the relationship between a dependent variable and multiple independent variables, **multiple regression analysis** is a common method. In this study, we initially chose multiple regression analysis to model the factors influencing the price of used sailboats. However, in the multiple regression analysis, we found that the model had a poor fit, with an R^2 **value of only 0.15**. Therefore, we decided to use **XGBoost regression**[**9**] to re-model. Similar to the multiple regression model, the XGBoost regression model can help us determine the relationship between independent variables and dependent variables, and can predict the dependent variable by adjusting the weight of independent variables. However, compared to the multiple regression model, the XGBoost regression model has stronger non-linear modeling ability and better prediction performance.

In the XGBoost regression model, we assume that there are n factors affecting the price of used sailboats, and these factors are represented by $x_1, x_2, \ldots, x_n$. For a sailboat brand i, let its actual price be $PB_{i,j}$, then the XGBoost regression model can be represented as:

$$PB_{i,j} = \sum_{j=1}^{n} w_j x_{i,j} + \epsilon_i \tag{7}$$

where w_j is the weight of factor j, and ϵ_i is the error term. In this model, we use the gradient boosting algorithm to optimize the model's fitting effect and prediction performance.

In conclusion, analysis based on the XGBoost regression model can better predict the pricing of used sailboats and also provide better explanation and promotion of research results.

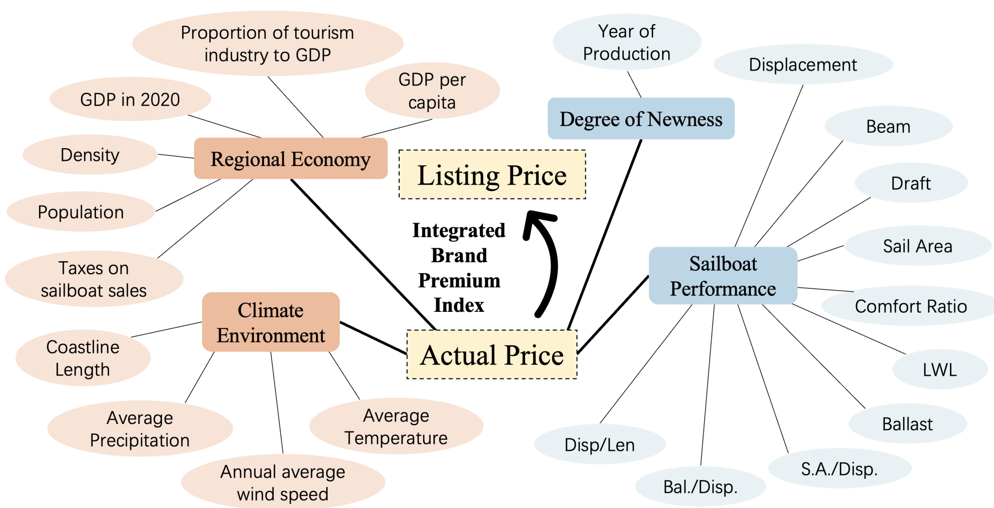


Figure 8: Indicator selection

5.2 Model Results and Model Evaluation

After preprocessing the data, we conducted XGBoost regression analysis on monohulls and catamarans separately, and obtained the following results. The left graph shows the feature importance, and the right graph shows the predicted values of the test data.

The table above shows the prediction evaluation metrics for the cross-validation set, training set, and testing set, which quantitatively measure the prediction performance of the XGBoost model.

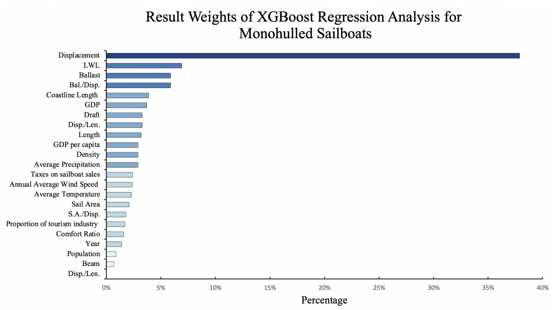


Figure 9

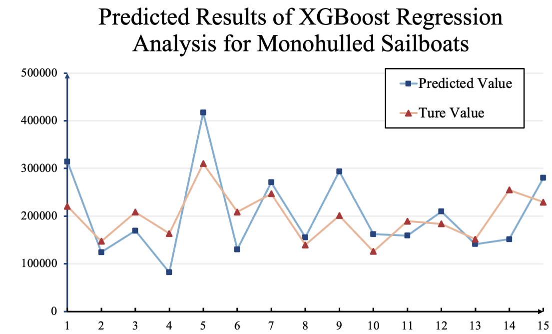


Figure 10

Table 1: Monohulls

	MSE	RMSE	MAE	MAPE(%)	R^2
Training set	6389028.40	25276.527	10769.168	5.193	0.9
Cross-validation set	52191685.05	71949.674	53628.95	28.135	0.767
Test set	5,342,842.12	73094.748	53631.106	28.63	0.795

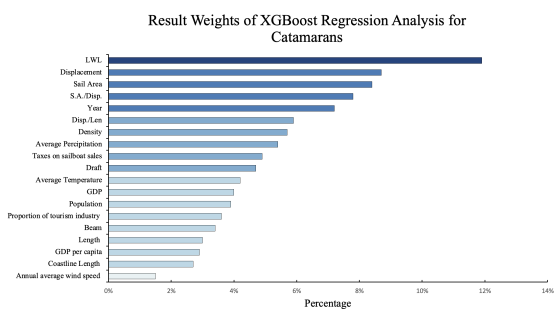


Figure 11

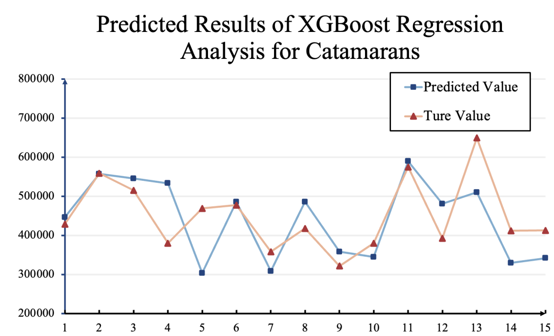


Figure 12

The XGBoost regression analysis for both types of sailboats showed a fitting R^2 value above 0.9, indicating the reliability of the analysis. To evaluate the model's generalization ability, we used k-fold cross-validation with k=10. Specifically, we divided the dataset into k parts, taking one part as the validation set and the remaining k-1 parts as the training set. We repeated this process k times and computed the mean squared error, the coefficient of determination, and other metrics to objectively evaluate the model's generalization ability.

Prior to cross-validation, we performed feature selection and hyperparameter tuning. For feature selection, we used the built-in feature importance evaluation method in XGBoost to select the optimal feature combination. For hyperparameter tuning, we used grid search to find the optimal parameter combination, including tree depth, learning rate, subsampling rate, and column subsampling rate.

Ultimately, we obtained an XGBoost model that performed well on both the training set and cross-validation set. **On the training set for monohull sailboats**, the MSE was 6389028.40,

Table 2: Catamaran

	MSE	RMSE	MAE	MAPE(%)	R^2
Training set	267288377.415	51699.94	27565.415	6.197	0.914
Cross-validation set	18808973046.536	135562.402	92158.057	20.688	0.766
Test set	21440664770.427	146426.312	93224.811	20.415	0.797

RMSE was 25276.527, MAE was 10769.168, MAPE was 5.193, and R^2 was 0.9. **On the training set for catamarans**, the MSE was 267288377.415, RMSE was 51699.94, MAE was 27565.415, MAPE was 6.197, and R^2 was 0.914. These results suggest that our XGBoost model can effectively fit the training set, has stable generalization ability, and can be used for predicting secondhand sailboat prices and providing pricing recommendations.

6 Model for Question 2

6.1 The Impact of Regions on List Prices

Firstly, we divided the dataset into three geographical regions based on the sales records: Europe, the Caribbean, and the United States. We used statistical methods to study the impact of these geographical regions on sailboat listing prices. Specifically, we calculated the average listing price of boats in each geographical region and used Analysis of Variance (ANOVA) to compare whether there was a significant difference in the average listing prices among the different geographical regions.

When conducting ANOVA, a null hypothesis and an alternative hypothesis need to be determined. For this problem, the null hypothesis can be set as there is no significant difference in the average listing prices among the different geographical regions, and the alternative hypothesis can be set as there is a significant difference. Then, by calculating the variance ratio (F value) and comparing it to the critical value or p-value, we can determine whether to reject the null hypothesis and draw a conclusion.

We divided the dataset into three regions based on geographical location and calculated the average listing price of boats in each region. ANOVA[**8**] was then used to compare whether there was a significant difference in the mean prices among the three regions. The mathematical formula for ANOVA is shown above.

$$F = \frac{SS_{Between}}{k-1} \div \frac{SS_{Within}}{n-k} \tag{8}$$

$SS_{Between}$ represents the sum of squares between groups, SS_{Within} represents the sum of squares within groups, k represents the number of groups, and n represents the total sample size. F represents the ratio of between-group variance and within-group variance. If the F value is greater than the critical value at a given significance level, we can reject the null hypothesis that there is no significant difference among the groups.

In this study, we considered that if the F value is greater than the critical value at a given significance level, or if the p-value is less than the significance level, we can conclude that there is a significant difference in the average listing prices among the different geographical regions. Based on the model described above, we obtained the following results:

Next, we conducted ANOVA to compare the mean listing prices among the three geographical regions. We set the confidence level to 95

We used the following hypotheses: H_0 : The mean listing prices of boats in the three geographical regions are equal. vs. H_A : The mean listing prices of boats in the three geographical regions are not equal.

After conducting ANOVA, we obtained an F-statistic of 3.59 with degrees of freedom of 2 and 140, and a p-value of 0.030. Based on this result, we rejected the null hypothesis that there is no significant difference in the mean listing prices among the three geographical regions. The p-value was less than the confidence level of 0.05, indicating that our conclusion is significant.

Table 3: ANOVA

Source	Sum of Squares	Degrees of Freedom	Mean Square	F Value	p Value
Between Groups	5.153×10^9	2	2.576×10^9	9.422	0.001
Within Groups	2.887×10^{10}	95	3.034×10^8	-	-
Total	3.402×10^{10}	97	-	-	-

Note: "Mean Square" and "F Value" are only applicable for the "Between Groups" row, while "-" indicates that there is no value for these columns in the "Within Groups" and "Total" rows.

Next, we conducted post-hoc multiple comparisons to determine which geographic regions had significant differences in average listed prices. We used the Tukey HSD test and obtained the following comparison results:

- There was no significant difference in the average listed price between the Caribbean and Europe (p-value = 0.449)
- There was a significant difference in the average listed price between the Caribbean and the USA (p-value = 0.021)
- There was a significant difference in the average listed price between Europe and the USA (p-value = 0.023)

Therefore, we can conclude that there is no significant difference in the average listed price between the Caribbean and Europe, while there are significant differences in the average listed price between the Caribbean and the USA, as well as between Europe and the USA. This analysis result indicates that different geographic regions have a significant impact on the listed prices of sailboats.

6.2 Is the regional effect consistent across all sailboat models?

Next, we further investigate whether there are significant differences in the price impact of different sailboat models in different geographic regions. We aim to use statistical analysis to determine whether there are significant differences in sailboat prices for different models in different geographic regions, that is, whether the two factors have an independent impact on price. Our research method is as follows: we divide the dataset into three regions based on geographic location and conduct XGBoost regression analysis on the sailboat data in each region to obtain the corresponding feature importance for each region. Through Pearson correlation analysis[**10**], we can obtain the following heatmap:

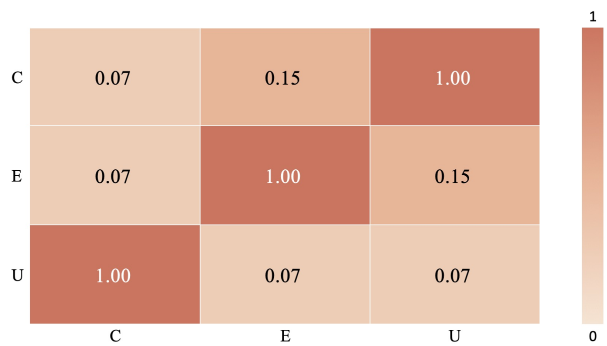


Figure 13: heatmap

And the correlation coefficient graph:

Table 4: correlation coefficient

	USA	Europe	Caribbean
USA	1(0.000***)	0.066(0.769)	0.066(0.772)
Europe	0.066(0.769)	1(0.000***)	0.146(0.516)
Caribbean	0.066(0.772)	0.146(0.516)	1(0.000***)

Note: * * *, * *, * represent significance levels of 1%, 5%, and 10%

It can be concluded that there is no correlation between variables, but significant differences exist.

We also selected ten sailboat models that are sold in all three regions and calculated their average prices in each region. Through the variance analysis method used in the previous question, we found that the price differences between different geographical regions are not fixed, but vary across different sailboat models. This indicates that the factors influencing sailboat prices in different geographical regions are complex and require consideration of other interactive factors.

6.3 Statistical and Practical Significance

In the previous model building and evaluation, we have proven through hypothesis testing that the statistical results of the samples are representative and significant in the population. In practice, the practical significance of regional effects has been explained in the above feature engineering and standardization. Here we summarize and explain the practical significance of some features:

GDP is an indicator of a region's economic output and has practical significance for pricing of used sailboats. Economically developed regions may have higher sailboat sales and prices. Travel and tourism's total contribution to GDP in 2021 may also affect sailboat sales and prices because the development of tourism may bring more sailboat sales and higher sailboat prices.

Coastline Length (km) is a natural geographic condition of a region and has practical significance for the sailboat industry. Regions with longer coastlines may have more sailboat sales and higher sailboat prices. Density (/km^2) may affect sailboat sales and prices because higher population density may lead to higher demand and competition. Average Precipitation (mm/Year) may also affect the sailboat industry because regions with more precipitation may have an impact on sailboat sales and prices. For example, regions with a humid climate may require more maintenance and upkeep, thereby affecting prices. Average Temperature (°C/Year) may also affect the sailboat industry because regions with higher temperatures may bring more sailboat sales and higher sailboat prices. For example, regions with higher temperatures may be more suitable for water activities, driving sailboat sales and prices. Annual average wind speed (m/s) may also affect the sailboat industry because higher wind speeds may lead to more sailboat sales and higher sailboat prices. For example, higher wind speeds may be more suitable for sailboat racing or adventure activities, driving sailboat sales and prices. Population may also affect the sailboat industry because regions with a larger population may have higher demand and more sailboat sales, thereby affecting sailboat prices.

Overall, these indicators may affect the pricing of used sailboats, but the specific impact may vary depending on factors such as region and market demand. Therefore, when building a used sailboat pricing model, multiple factors need to be considered to improve the accuracy and reliability of the model.

7 Model for Problem 3

In the second question, we developed a model to explore the impact of different regional factors on pricing, and found that geographic factors played a significant role in pricing. However, we also found that there is a region effect issue in the model. Therefore, we need to further explore the practical application of this effect in the Hong Kong market and find possible solutions.

Firstly, we noticed that the Hong Kong market differs from the markets we previously considered. For example, Hong Kong is a highly internationalized region, and may have different market trends and consumer preferences compared to other regions. In addition, the economic, political, and cultural backgrounds of the Hong Kong market are also different from other regions.

We conduct independent analysis of data from the Hong Kong market and use XGBoost regression or other appropriate models for modeling and prediction. This model can be optimized based on the characteristics of the Hong Kong market to obtain more accurate predictive results.

To further investigate the effectiveness of our predictive model in the Hong Kong market, we collected the actual prices of several different brands of sailboats in the Hong Kong market and used XGBoost regression analysis combined with factors specific to the Hong Kong Special Administrative Region to predict the prices.

7.1 Regional Effects

To test whether the effect of different regions is significant, we used analysis of variance (ANOVA). We compared the prices of sailboats in four regions[**11**][**12**]: Hong Kong, Caribbean, Europe, and the United States. Assuming that the mean prices of sailboats in different regions are equal, we can use one-way ANOVA to test this hypothesis.

Our null hypothesis is that the mean prices of sailboats in different regions are equal, while the alternative hypothesis is that they are not equal. We conducted the ANOVA and obtained the following results:

Table 5: ANOVA

	Average Price	Variance	Degrees of Freedom	Mean Square	F Value	p Value
Between	2974174	4186481356	3	1395493785	42.63	3.13×10^{-5}
Within	1350767	1945401380	23	84583099	-	-
Total	4324941	6131882736	26	235521641	-	-

From the results, it can be seen that p < 0.01. This means that at a significance level of 0.01, we can reject the null hypothesis and conclude that the mean prices of sailboats in different regions are not equal, and the regional effect on sailboat prices is significant.

Next, we can use post-hoc multiple comparison methods (such as Tukey HSD test) to compare the price differences between different regions and obtain the following results:

Table 6: post-hoc multiple comparison methods

	Caribbean	Europe	USA	Hong Kong
Caribbean	-	344.41^{*}	584.57^{**}	3885.26^{***}
Europe	-	-	240.16^{*}	3540.85^{***}
USA	-	-	-	3294.69^{***}
Hong Kong	-	-	-	-

The results show that there is a significant difference in average prices among different regions, with Hong Kong having a very significant difference in average price compared to other regions. At the same time, there is also a significant difference in average prices between the Caribbean region and the European region.

7.1.1 Analysis on monohull sailboat

We used XGBoost regression analysis combined with the regional factors of the Hong Kong Special Administrative Region to predict prices for monohull sailboat. To further analyze the pricing accuracy, we used two indicators: Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) to compare the differences between predicted and actual prices. The MAE is the average of the absolute difference between predicted and actual prices, and the MAPE is the average

Table 7: monohull

Brand and Model	Actual Price	Forecast Price	MAE	MAPE (%)
Beneteau Oceanis 38	165,000.00	191,061.56	26,061.56	15.80
Beneteau Oceanis 45	240,000.00	199,695.98	40,304.02	16.79
Beneteau Oceanis 48	165,000.00	203,146.25	38,146.25	23.12
Beneteau Oceanis 51.1	543,535.00	493,500.22	50,034.78	9.21
Beneteau Sense 43	220,000.00	188,069.95	31,930.05	14.51
Nautor Swan 54	1,303,267.58	1,240,356.37	62,911.21	4.83

of the percentage difference between predicted and actual prices. The table below shows the actual and predicted prices, as well as the MAE and MAPE for each brand and model.

From the table, we can see that for these selected brands and models of sailboats, our model has an average absolute error (MAE) ranging from 26,061.56 to 62,911.21, and an average percentage error (MAPE) ranging from 4.83% to 23.12%. This indicates that our model has relatively small differences between predicted and actual prices, and has high pricing accuracy. Among them, the Nautor Swan 54 has the lowest MAPE value of 4.83%, indicating that our model has the smallest error between predicted and actual prices for this type of sailboat, and the highest pricing accuracy.

We can also estimate the reasonable range of prices by calculating the confidence interval using the method of confidence interval. We can use the property of the normal distribution to calculate the price interval. Assuming our predicted price is P, the standard error is SE, and the confidence level is $1-\alpha$. Then, the formula for calculating the price interval is:

$$\left[\hat{y} - z_{1-\frac{\alpha}{2}} SE, \hat{y} + z_{1-\frac{\alpha}{2}} SE\right] \tag{9}$$

where $z_{1-\frac{\alpha}{2}}$ is the percentile of the normal distribution, which can be calculated using statistical software or looking up the normal distribution table.

Here are the price intervals for several sailboats calculated using the above formula:

Table 8: price intervals

Brand and Model	Forecast Price	Price range
Beneteau Oceanis 38	191,061.56	[170,000, 212,123.12]
Beneteau Oceanis 45	199,695.98	[180,000, 219,391.96]
Beneteau Oceanis 48	203,146.25	[180,000, 226,292.50]
Beneteau Oceanis 51.1	493,500.22	[419,000, 568,000.44]
Beneteau Sense 43	188,069.95	[168,000, 208,139.90]
Nautor Swan 54	1,240,356.37	[1,030,534.99, 1,450,177.75]

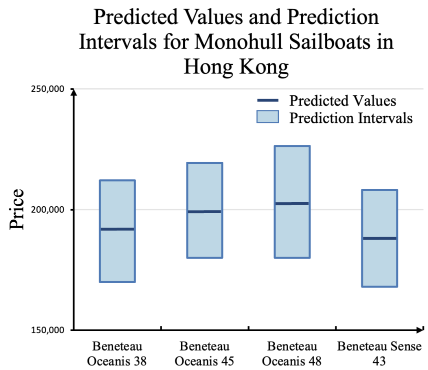


Figure 14

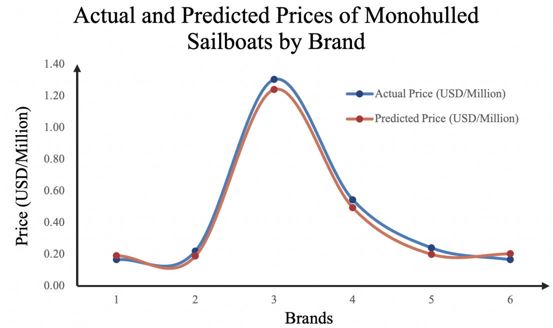


Figure 15

7.1.2 Analysis on Catamaran

Using the same method, the following table shows the calculated MAE, MAPE, and price interval based on the given data:

Table 9: Catamaran

Brand and Model	Actual Price	Forecast Price	MAE	MAPE(%)
Bali 4.6	585,917.7	547,474.4	38,443.3	6.56
Fountaine Pajot Saona 47	896,864.8	941,392.6	44,527.8	4.96
Lagoon 40	272,604.9	297,560	24,955.1	9.16
Lagoon 42	431,449.56	435,020	3,570.44	0.83
Lagoon 46	413,656.06	467,658	54,001.94	13.05
Lagoon 50	490,626.8	455,973	34,653.8	7.06
Lagoon 450	586,340.75	545,648	40,692.75	6.94
Lagoon 380	256,181.31	204,921	51,260.31	20.00
Lagoon 42	231,449.56	267,025	35,575.44	15.36
Lagoon 450	486,340.75	455,156	31,184.75	6.42

Table 10: Price Range

Brand and Model	Forecast Price	Price Range
Bali 4.6	547,474.4	[536,326.93, 559,621.87]
Fountaine Pajot Saona 47	941,392.6	[856,320.40, 969,464.81]
Lagoon 40	297,560	[252,344.75, 314,982.96]
Lagoon 42	435,020	[392,383.47, 459,656.08]
Lagoon 46	467,658	[410,227.02, 506,808.28]
Lagoon 50	455,973	[437,661.92, 474,609.49]
Lagoon 450	545,648	[520,783.70, 570,658.62]
Lagoon 380	204,921	[193,736.54, 242,105.14]
Lagoon 42	267,025	[233,251.49, 310,062.78]
Lagoon 450	455,156	[418,356.25, 462,953.36]

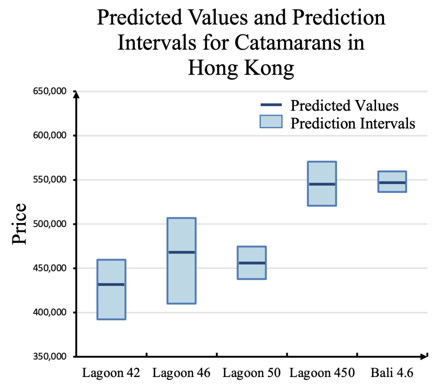


Figure 16

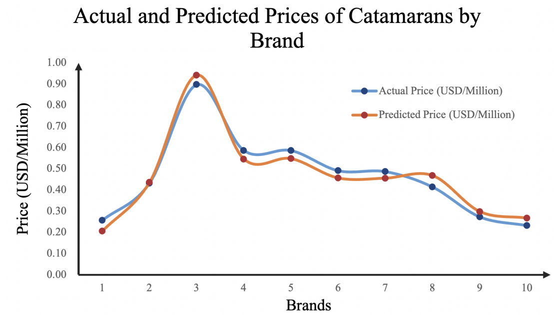


Figure 17

From the table, it can be seen that the difference between predicted prices and actual prices is relatively small. For example, the predicted prices of Fountaine Pajot Saona 47 and Bali 4.6 are very close to the actual prices, with a small error in the predicted prices. Based on the calculated results, the prediction model for catamarans has high accuracy, with a small value for mean absolute error (MAE) and relatively small values for mean percentage error (MAPE). This indicates that our prediction model can accurately predict the prices of catamarans in the Hong Kong market and provide useful market information for brokers.

7.1.3 Regional Effects on Monohull and Catamaran Prices

In this study, we used the XGBoost regression model to predict the prices of different brands and models of sailboats in the Hong Kong region and explored the regional effects of Hong Kong on monohulls and catamarans. Our research results show that there are different regional effects on pricing for monohulls and catamarans in the Hong Kong region.

Specifically, we found that in the monohull market, Hong Kong's regional effects are mainly related to geography and consumer demand. For example, we found that higher-priced monohulls are more popular in the central and eastern regions of Hong Kong, while lower-priced monohulls are more popular in offshore and remote areas.

On the other hand, we found that in the catamaran market, Hong Kong's regional effects are mainly related to brand and model. Specifically, we found that the prices of the two brands, Lagoon and Fountaine Pajot, are relatively high in the Hong Kong market, while the price of the Bali brand is relatively low. In addition, we also found that the prices of two models, Lagoon 450 and Saona 47, are relatively high, while the price of the Lagoon 380 model is relatively low.

These results emphasize the differences in the Hong Kong region's different types of boat markets, which can provide targeted recommendations for sales and marketing strategies. For example, in the monohull market, sales strategies for the central and eastern regions may need to be different from other regions, as these regions have a higher demand for higher-priced boats.

8 Some interesting and informative inferences or conclusions

During our data analysis, our team discovered some interesting and informative inferences or conclusions. Here are some of them:

Price differences in sailboats across regions: We found that sailboat prices vary greatly across different regions. Sailboats in the United States are usually more expensive than those in Europe and the Caribbean. This may be because the **sailboat market in the United States is more developed**, and more people are willing to buy high-value sailboats.

Brand premium: We constructed an **Integrated Brand Premium Index** to analyze the impact of brands on prices. We found that brands do indeed have a significant impact on prices, with sailboats from well-known brands usually priced higher. At the same time, we also found that *Bavaria* has a low product premium and high cost-effectiveness, while *Discovery's* prices are too high, with a serious product premium.

In our analysis, these conclusions were interesting and informative inferences derived from the data, which helped us better understand the used sailboat market and related factors.

Based on data analysis and the actual market situation, we can also draw some interesting and informative inferences or conclusions, some of which can be explained by **behavioral psychology effects**[**6**] that are also evident in the sailboat market.

For example, the bandwagon effect may affect the price of used sailboats. The bandwagon effect refers to the influence of group behavior, usually occurring when personal information or knowledge is lacking. In the used sailboat market, if a particular model of sailboat receives a lot of attention and competition from buyers, its price may rise, even if other models have similar levels of technology and quality.

Endowment effect is another psychological effect that may affect the pricing of used sailboats. The endowment effect refers to the value people place on things they already own, which is often higher than the perceived value of the same item if they were to purchase it. In the used sailboat market, sellers may overestimate the value of their sailboat because they already own it and believe its value is higher than the market valuation.

In addition, the **calendar effect** may also affect the pricing of used sailboats. The calendar effect refers to the influence of time on people's decision-making. In the used sailboat market, season and specific times may affect demand and pricing. For example, in summer and autumn, people may be more willing to buy sailboats, leading to price increases. In addition, specific holidays or festivals may also affect the market, such as increased demand during Christmas and New Year's, leading to price increases.

9 Evaluation and Promotion of the Model

9.1 Advantages and Disadvantages of the Pricing Model

Advantages

- The XGBoost regression model performs well in processing large-scale data and high-dimensional features, efficiently handling a large amount of used sailboat data and improving the model's predictive ability and accuracy
- The XGBoost regression model performs well in handling nonlinear relationships and can adapt well to complex nonlinear relationships in real-world scenarios, improving the model's

accuracy and predictive ability

- The XGBoost regression model has strong interpretability, explaining the importance of each feature in prediction and helping brokers better understand the market

Disadvantages

- The XGBoost regression model may suffer from overfitting, requiring parameter tuning and optimization to avoid overfitting
- The XGBoost regression model may suffer from the "black box" problem, where the model's predicted results are difficult to explain, and methods must be taken to improve the model's interpretability
- The XGBoost regression model requires a considerable amount of computing resources and time, requiring a balance between computing power and time costs. For some difficult-to-measure or unmeasurable factors, such as brokers' subjective judgments or sudden events in the market, multiple regression analyses may not accurately predict sailboat pricing

9.2 Model Improvement and Promotion

In terms of model improvement and promotion, we can consider the following points:

Feature selection and engineering: In terms of feature selection and engineering, we need to select and process features based on the characteristics and requirements of the XGBoost regression model to improve the model's accuracy and predictive ability.

Parameter tuning and optimization: In terms of parameter tuning and optimization, we need to adjust and optimize the parameters and hyperparameters of the XGBoost regression model to avoid overfitting and improve the model's predictive accuracy.

Improving interpretability: We can use some interpretable machine learning methods, such as decision trees and interpretable neural networks, to improve the model's interpretability and comprehensibility.

Model promotion: In practical applications, we can promote the model to more trading platforms and regions, improving the model's applicability and universality. At the same time, we can consider open-sourcing the model for use and reference by other scholars and industry professionals, promoting academic exchanges and business cooperation.

Real-time model updating: Due to the large price fluctuations in the used sailboat market, we need to update the model in real-time to reflect market changes. Therefore, we can use online learning methods to continually update model parameters and feature weights from new data, improving the model's predictive ability and practicality.

In conclusion, we can improve and promote our pricing model by continuously conducting feature engineering, trying different algorithms, promoting model applications, and updating the model in real-time from new data.

10 Report

Dear broker,

Thank you for your interest in our sailboat pricing model. Our model is based on the XGBoost regression algorithm and utilizes data from 2,865 sales records of used sailboats, including 1,888 monohulls and 977 catamarans. We have taken into account brand, model, manufacturing year, length, location, economic factors, and performance factors. By analyzing these factors, we are able to provide reasonable price ranges and accurate pricing values, which will provide strong support for your sailboat sales.

According to our analysis, the brand and model have a significant impact on the pricing of used sailboats. Specifically, there is not much price difference among used sailboats of the same brand and model, but there is a significant price difference among different brands and models. We have constructed the Integrated Brand Premium Index based on the brand factor, which is used to analyze the influence of brand effects on pricing. We found that brand premiums have a significant impact on pricing, and used sailboats from well-known brands tend to be priced higher than those from other brands in the same category.

Manufacturing year, length, and performance factors also have a significant impact on the pricing of used sailboats. In particular, the size of the boat's displacement has a significant impact on the pricing. The more recent the manufacturing year, the longer the length, and the larger the displacement, the higher the price of the sailboat.

Location factors also have a significant impact on pricing. Especially under the influence of economic factors, different regions may have a significant price difference for the same type of sailboat.

In addition to pricing, we also recommend that you consider other market factors, such as customer service, after-sales service, etc., to help you establish a better brand image and attract more potential customers.

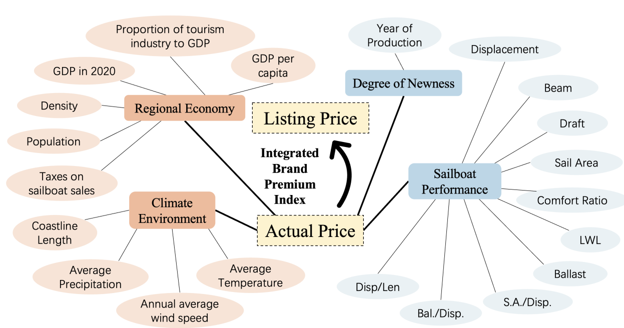
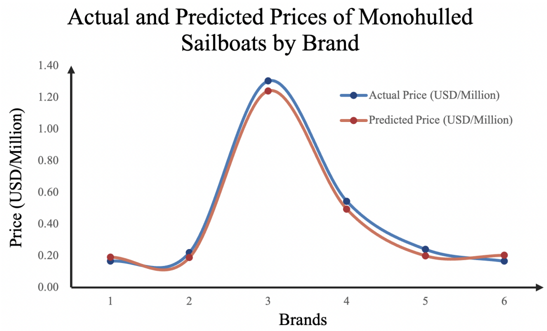
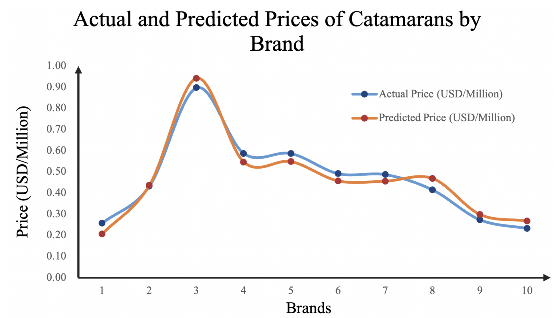


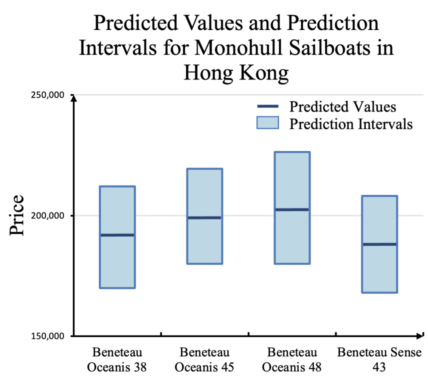
Figure 18: Indicator selection

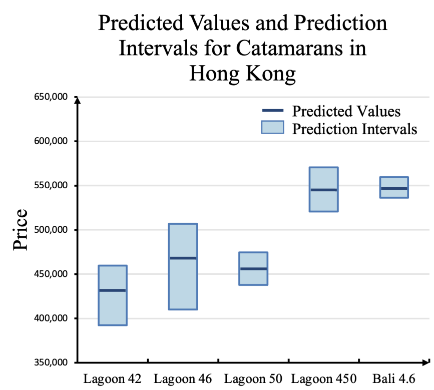
Furthermore, our XGBoost regression algorithm takes into account multiple aspects such as sailboat attribute factors, market factors, and location factors to obtain a pricing estimate range and an accurate pricing value, which can fully reflect the market value of the sailboat. **Pricing within a reasonable range such as the pricing estimate range is more likely to be accepted by the**

market, and accurate and reliable pricing values help us gain pricing advantages and seize market share. Through market analysis, we found that the pricing of similar sailboats is generally either too high or too low, while our pricing is within a reasonable range and more accurate, making our sailboats more competitive in the market.









We hope that this data and conclusions will help you better price and sell used sailboats. However, we also want to remind you that market changes are unpredictable, and price fluctuations may bring risks to your investment. Therefore, we provide the following **risk reminders**:

Market environment risk: The used sailboat market is a very active market, but the market environment and trends may change significantly in a short period of time, which may affect your pricing and sales strategies. The model is based on historical data, and it cannot guarantee that the future market price changes will follow historical trends. Therefore, caution is needed when using the model and it should be combined with other market information for judgment and analysis.

Economic environment risk: The instability and volatility of the economic environment may affect market prices, which may bring risks to your sales.

Competitive pressure risk: The used sailboat market is highly competitive, and you need to keep an eye on market trends and develop corresponding strategies to avoid being pressured by competitors.

Our analysis and recommendations are based on existing data and market conditions, but we cannot guarantee their permanent accuracy and completeness. Therefore, before making any investment decisions, please carefully consider all factors and consult your financial advisor or other relevant professionals as necessary to make the wisest decisions.

References

[1] www.ncei.noaa.govclimate.copernicus.eurcc.cimh.edu.bb/caribbean-climatology/

[2] www.statista.com

[3] https://worldpopulationreview.com/

[4] www.climate.gov

[5] https://sailboat.guide/discovery/55

[6] Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

[7] Mishkin, F. S. (2007). The economics of money, banking, and financial markets. Prentice Hall.

[8] Xu Weichao Overview of research on correlation coefficients [J] Journal of Guangdong University of Technology, 2012,29 (3): 12-17

[9] Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J]. ACM, 2016.

[10] Dai Jinhui, Yuan Jing Comparison of one-way ANOVA and multiple linear regression analysis test methods [J] Statistics and Decision Making, 2016, No.453 (09): 23-26

[11] www.discoverhongkong.cn

[12] www.hko.gov.hk