Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2331774

How Region Affects Used Sailboat Price

Summary

Used sailboats have high price volatility and heterogeneity, so it is important to develop a scientific mechanism for transparent pricing of used sailboats. Therefore, we develop a regression prediction model of sailboat prices to explore the regional influence on prices, form a set of pricing mechanisms that can explain the regional effect, and apply it to the Hong Kong sailboat market for empirical analysis.

To begin with, we develop a regression prediction model based on **BBS integrated learning** to explore the pricing mechanism. We first divide the features into two categories: regional features and sailing features, and use the **XGBoost algorithm** to select the features into 13 features with high importance. After the model regression analysis, we conclude that: 1. the top three in importance are **new product price**, **beams** and **age**. 2. the sailboat feature has **93.9**% of the weight and dominates the sailboat price. 3. the goodness-of-fit R^2 of the regression prediction model is **0.853**.

To analyze the effect of area effects on each variant, we first performe **K-means clustering** for monohull and catamaran sailboats separately. After screening out the categories with too small a sample size, we perform regression analysis for each class of sailboats separately to obtain the feature importance weights under each variant type. The **Kendall consistency test** for the feature weights for each category of sailboats shows that the overall data have a significance p-value of **0.051***, which indicates the data present inconsistency. In our analysis, the **regional influence inconsistency** represents that each region has different **consumer preferences** for sailboat variants.

Because of the inconsistent regional influence, when the model is applied to the Hong Kong market, we cluster Hong Kong and the given regions, so as to obtain a larger amount of information. After regression prediction, the regional influence on price in Hong Kong accounts for **13.3**%, and the goodness of fit between the model predicted price and the actual comparable listed price in Hong Kong is **0.87**. By comparing the effect of regional features we get that regional features are not consistent for monohull and catamaran, indicating that there is consumer preference for both types of sailboats in Hong Kong.

From the above analysis process, we also draw some interesting conclusions,such as the **effect of the characteristics of the sailboat variants** is also inconsistent for different variants of the boat.

Finally, based on the BBS integrated learning model we also prepare a **report** to sailboat brokers in Hong Kong, making recommendations for sailboat trading strategies in the Hong Kong region.

Keywords: **Regional Effects;BBS Integrated Learning;Kendall Consistency Test.**

Contents

1 Introduction

1.1 Problem Background

Unlike other capital-intensive fixed assets, used sailboats are limited collectors in nature while also being mobile and relatively liquid. As a result, the price of used boats fluctuates rapidly, and as Stopford notes: 'Typically, the price of used boats responds dramatically to changes in market conditions.' [1]

Based on this high market volatility and heterogeneity, the used sailboat market is prone to a mismatch between the information held by the buyer and the seller, which in turn leads to the destruction of the trust mechanism of the entire trading market. [1] Therefore it is important to form a scientific mechanism for transparent pricing of used sailboats.

A reasonable pricing mechanism can regulate the price chaos in the used sailboat market, make the pricing of sellers transparent, give consumers an indicator system for reference, build a good trust mechanism between buyers and sellers, and thus promote the steady upward development of the global used sailboat trading market.

1.2 Restatement of the Problem

Based on the data given in the attachment, we need to select and collect additional data on features that may affect the price of used sailboats and provide solutions to the following problems:

- Build a mathematical model to explain the listing price using the factors that affect the listing price and use the model to accurately predict price.
- Use the model to explain how regional characteristics affect the list price of used sailboats and test whether this effect is consistent across all variants.
- Perform an empirical analysis of the Hong Kong market using the model, comparing model predicted prices with actual prices in the Hong Kong market and discussing whether monohull and catamaran prices are affected by region in the same way.
- Give other statistical conclusions based on the data.
- Provide a report for sailboat brokers based on the conclusions drawn, giving a reference for market analysis and providing recommendations for buying and selling decisions.

1.3 Our Work

The work we have done in this problem is mainly shown in the following Figure 1.

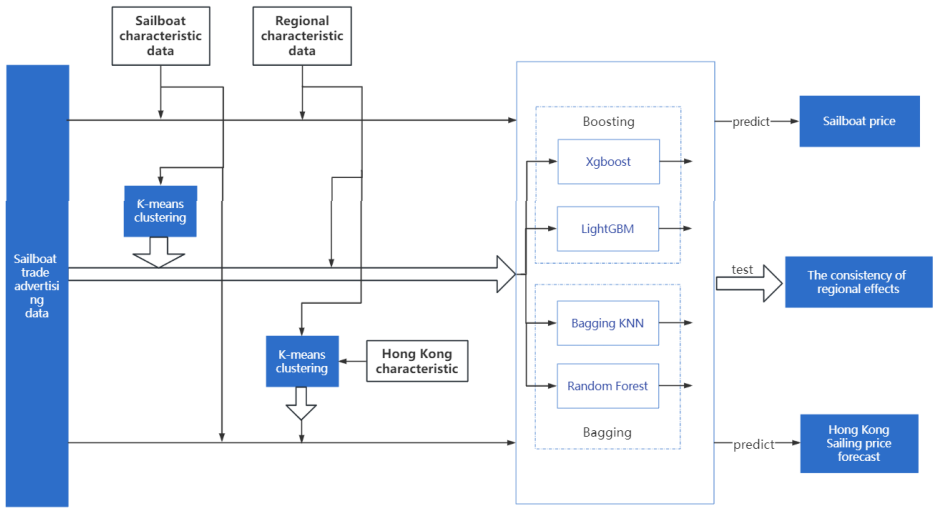

Figure 1: Overview of Our Work

2 Assumptions and Justification

We have made several assumptions in our model to optimise our model to make it more applicable in a complex realistic environment.

1. **Assumption 1: The used sailboat market is a free market.**
   ⇒***Justification: Disregarding local ship management policies and import/export constraints, the main factors determining commodity pricing are simplified to commodity supply and demand.***

2. **Assumption 2: Inflation and other economic changes in recent years are not taken into account.**
   ⇒***Justification: Major events such as Covid-19 and inflation may affect the pricing of used sailboats, and here we do not consider the impact of such external factors.***

Additional assumptions are made to simplify analysis for individual sections. These assumptions will be discussed at the appropriate locations.

3 Notions

Table 1: Notions

Symbols	Definition	Units
A	Sailboat Age	/
B_e	Beams	ft
L	Length	ft
D_r	Draft	ft
D_{is}	Displacement	lb
SA	Sail Area	m^2
H	Headroom	ft
AT	Average Temperature	$°C$
GDP	Regional GDP	billion USD
PCI	Per Capita Income	USD
RB	Registered Boats	/
SC	Sleeping Capacity	/

4 Data Preprocessing

Before data analysis, the availability of data must be guaranteed.

- **Data Source**
  In addition to the data in the attachment, we have collected data on sailing variants and regional economic and geographic data from the following websites.

Figure 2: Sources of Data

Here is an snapshot of our data.

Make	Bali	Variant	5.4
Geographic Region	Europe	Country/Region/State	Italy
Length (ft)	53.15-55.12	Beams(ft)	28.67
Draft(ft)	4.86	Displacement(lb)	45856
Rigging	Fractional Sloop	Sail Area(sqft)	2195.84
Number of the sailing boat clubs	1174	Average Temperature	15.6
Registered Yachts	850000	GDP in billions of US dollars	1595.3
Per Capita Income 2020 (USD)	34430	Age	1
Hull Type	Catamaran	New Product Price(USD)	1525000
Headroom(ft)	4.2	Sleep Capacity	16-40
Hull Material	Fibreglass	Listing Price (USD)	811143
Electronic Equipment	Pilote auto P70S, GPS plotter AXIOM 7", MULTI I70S, VHF RAY 63 + VHF RAY MIC at steering station, AIS receiver transmitter, 12 "full touch screen at steering station *		

Figure 3: Snapshot of Data

- **Missing Value Processing**
  We use IQR outlier identification which identifies three missing values in Country/Region/State in the data and set them to null values.

- **Data Normalization**
  In the regression prediction, normalization gives equal weight to the eigenvalues.We calculate the mean ($\bar{x}$)and standard deviation(σ) of the data for each feature, and then process the data as: $x' = \frac{x-\bar{x}}{\sigma}$

5 Used Sailboat Price Forcasting

5.1 BBS Integration Model

The model uses the **Stacking** model fusion strategy to fuse the algorithm based on **Bagging** integrated learning framework and the algorithm based on **Boosting** integrated learning framework, and **LightGBM** is used as the learning algorithm.

To be specific, Random Forest and Bagging_KNN are selected as the base learners in the Bagging integrated learning framework; XGBoost and LightGBM are selected as the base learners in the Boosting integrated learning framework.

- **Random Forest**

The random forest regression model recursively divides each region into two subregions in the input space where the training data set is located and determines the output values on each subregion, constructing a binomial decision tree:[2]

1. Choose the optimal tangent variable j and the tangent point s, and solve:

$$\min_{j,s}\left[\min_{c1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right] \tag{1}$$

Iterate over the variable j, scan the cut point s for the fixed ground cut variable j, and select the pair (j,s) that minimizes Eq. 1.

2. Delimit the region with the selected pair (j,s) and determine the corresponding output value:

$$R_1(j,s) = \left\{x \mid x^{(j)} \leq\right\}, R_2(j,s) = \left\{x \mid x^{(j)} > s\right\}$$
$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m(j,s)} y_i, x \in R_m, m = 1,2 \tag{2}$$

3. Continue calling steps 1,2 for both subregions until the stop condition is met;

4. Divide the input space into M regions, R1, R2,..., RM, and generate a decision tree;

$$f(x) = \sum_{m=1}^{M} \hat{c}_m I\left(x \in R_m\right) \tag{3}$$

- **KNN**

KNN (k-Nearest Neighbor) for regression prediction uses the weighted mean of K neighbors as the prediction result.

The distance metric is as follows:[3]

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - u_i|^p\right)^{1/P} \tag{4}$$

When p is 1, the distance is the Manhattan distance, and when p is 2, the distance is the Euclidean distance (Euclidean distance).

- **XGBoost**

XGBoost keeps adding trees based on the splitting of features, and with each added tree, the residuals of the previous prediction are fitted to obtain the new function, and the model is continuously optimized by stepwise iterations.[4]

The objective function equation is:

$$L(\phi) = \sum_i l\left(\hat{Y}_i - Y_i\right) + \sum_k \Omega\left(f_k\right) \tag{5}$$

$l\left(\hat{Y}_i - Y_i\right)$ denotes the prediction error of the ith sample and $\sum_k \Omega\left(f_k\right)$ denotes a function of the complexity of the tree, the smaller the function, the lower the complexity and the stronger the generalization ability of the model, the expression is

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|g\|^2 \tag{6}$$

where T denotes the number of leaf nodes and g denotes the value of the node.

- **LightGBM**

The basic idea of LightGBM is to combine M weak regression trees linearly into strong regression trees, as shown in Eq.

$$F(x) = \sum_{m=1}^{M} f_m(x) \tag{7}$$

where F_x is the final output value; $f_m(x)$ is the output value of the mth weak regression tree.

Since the model uses Stacking model fusion strategy as well as Bagging and Boosting integrated learning framework, we name it as BBS model, and the specific implementation process of the model is shown in Figure x.

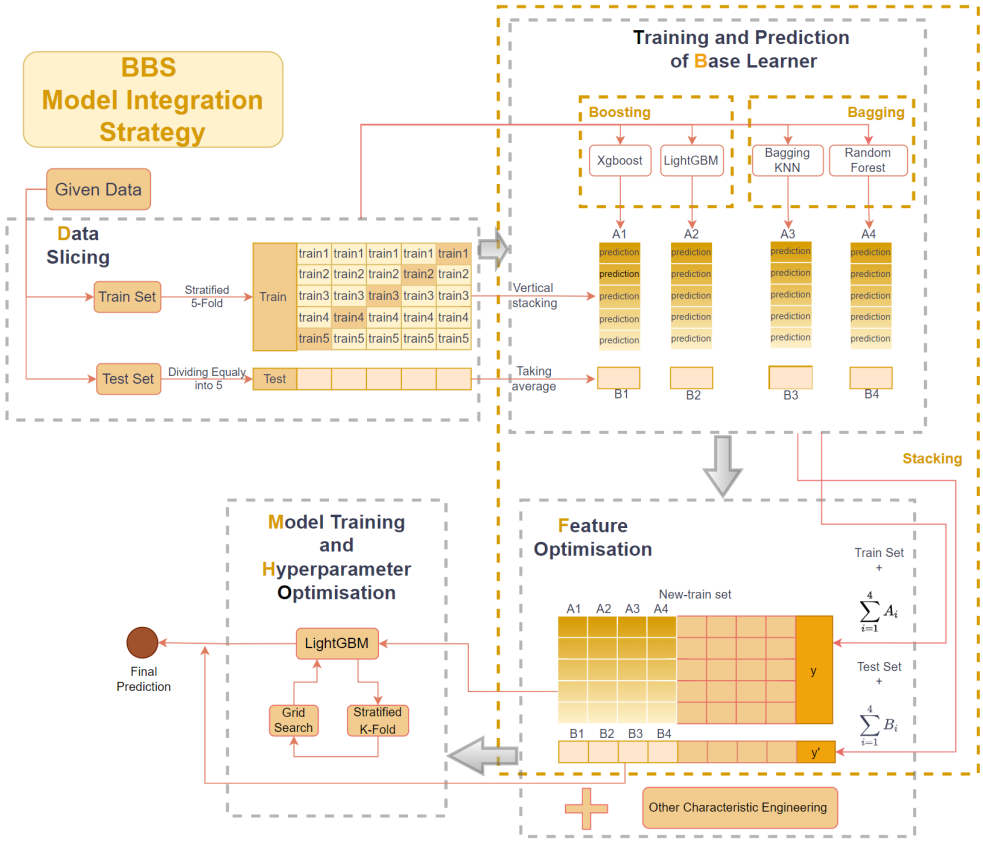

Figure 4: BBS Working Mechanism Flow Chart

5.2 Model Implementation

5.2.1 Feature Selection

When considering the factors influencing the listing price of used sailboats, we analyze both sailboat configuration parameters and regional market supply and demand.

- **Sailboat Configuration**
  In terms of sailboat configuration parameters, several drivers of sailboat prices have been identified in the literature, such as sailboat **age**, **size**, **materials**, **rigging**, **displacement**, **headroom**, **sleeping capacity**, **sail area**, and **hull type**, among others. [3] While in today's rapidly developing sailboat market, we believe that the influence on sailboat prices should also include marketing and service factors, such as **new sailboat prices**, **brand**, **electronic equipment configuration**.

- **Market Demand**
  In terms of regional market supply and demand, we believe that the supply and demand should be described in terms of the purchasing power of used sailboats. Since a sailboat is a luxury item, the level of consumption and economic development of a region will determine the amount of demand for a sailboat in that region. Therefore,we choose **Regional GDP** and **Per Capita Income** as regional characteristics. On the other hand, the stronger the sailing culture in a region, the higher the demand for sailing boats in the region is likely to be. Therefore, we select the **number of sailing registrations**, **the number of sailing clubs** and the **average temperature** of the region as regional characteristics.

In this way, we establish a price forecasting index system consisting of 19 indicators, as shown in the figure.

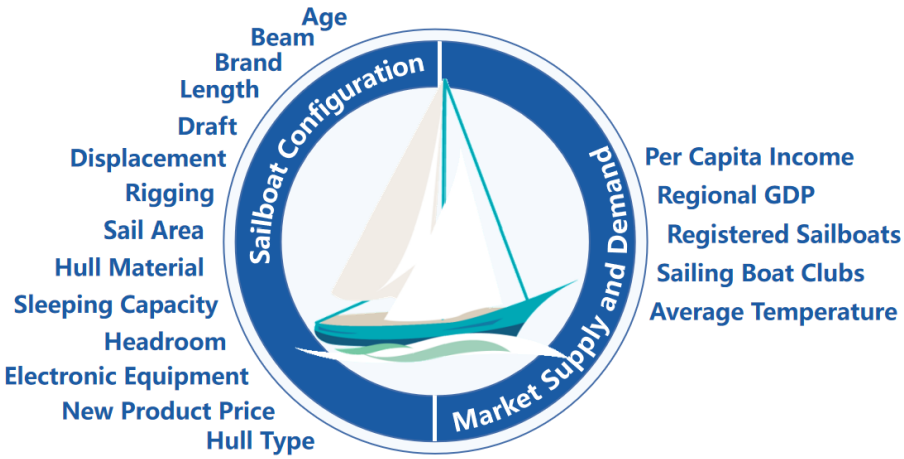

Figure 5: Feature System

Based on the feature system, we quantify the categorical variable 'brand' by replacing it it with variant new product price. Therefore, there are 17 features.

We substituted the selected 17 features into the XGBoost regression model to get the importance ranking of the features, as shown in Table

Table 2: Feature Importance Ranking

Rank	Feature	Importance	Rank	Feature	Importance
1	NP	40.60%	10	PCI	1.90%
2	B_e	23.90%	11	AT	1.00%
3	A	11.00%	12	RB	1.00%
4	SA	5.60%	13	SC	0.70%
5	D_{is}	4.70%	14	HT	0.00%
6	L	2.90%	15	HM	0.00%
7	D_r	2.60%	16	SBC	0.00%
8	GDP	2.20%	17	EE	0.00%
9	HR	1.90%	16	R	0.00%

Therefore we screen out hull type, hull material, number of regional sailing clubs, equipment equipment and rigging due to their low importance. Therefore our data is finally consisted of 13 features.

5.2.2 Listing Price Regression and Prediction

- **Data Slicing**
  In dividing the training and test sets we use ten fold sampling, with 80% of the training set and 20% of the test set. The training and validation sets are obtained using Stratified 5-fold sampling in the training set.

- **Cross-validation**
  In order to verify the performance of the classification evaluator, this paper adopts the K-fold cross-validation method, which is to divide the original data into K groups, and each subset of data is used as the validation set once, and the remaining K-1 subsets are used as the training set, so that K models can be obtained. The classification accuracy of the final validation set of the K models is averaged and used as the prediction accuracy of the regressor. The average of the classification accuracy of the final validation set of K models is used as the prediction accuracy performance index of the regressor.[5]

- **Grid Search**
  In this paper, a grid search method is used to optimize the model parameters by iterating through a loop, trying each combination of parameters, and finally returning the best combination of parameters. Assuming that there are m parameters to be tuned, and each parameter has a choice, then there are a total of combinations. Since the cycle is like traversing a grid, it is called grid search. [1] The following model is trained with the default parameters during the tuning process, and then tuned with the grid search and compared with the training error of the model before and after tuning.

5.3 Results and Evaluation

- **Regression Results**

After the model regression, we obtain the ranking of the coefficients of the influence of different features on the price of sailing boats, as shown below.

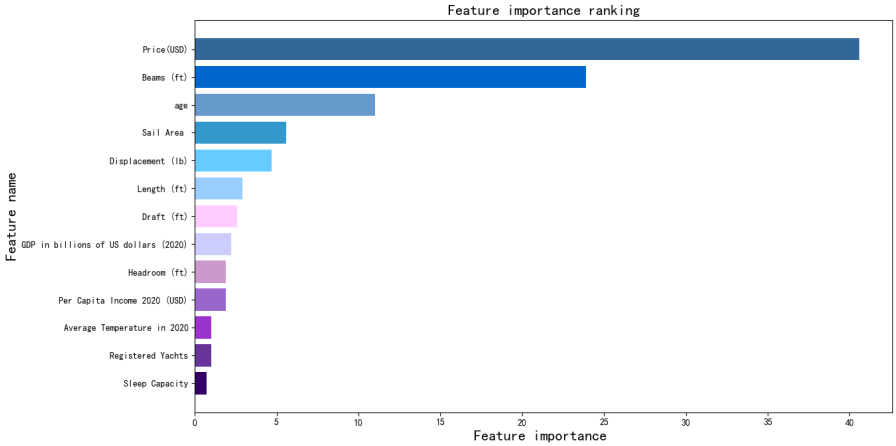

Figure 6: Feature Coefficients

According to the top graph of the model output, we can see that the first seven influences with large characteristics all come from sailing variant differences. Overall, price differences from sailing variant differences, etc. account for **93.9**% and price differences from regional differences account for **6.1**%.

Among the sailing variant effects, new product prices have the largest effect on used sailing boats, with an influence factor of 0.4. This result is also corroborated in the literature,xx where it is noted that the price correlation between used and new sailing boats is extremely high and close to substitution due to the commodity nature of sailing boats.

- **Prediction Results**

After regression we obtained the factors influencing the price for each feature, which allowed us to predict the price of used sailboats through the model, the results of which are shown below:

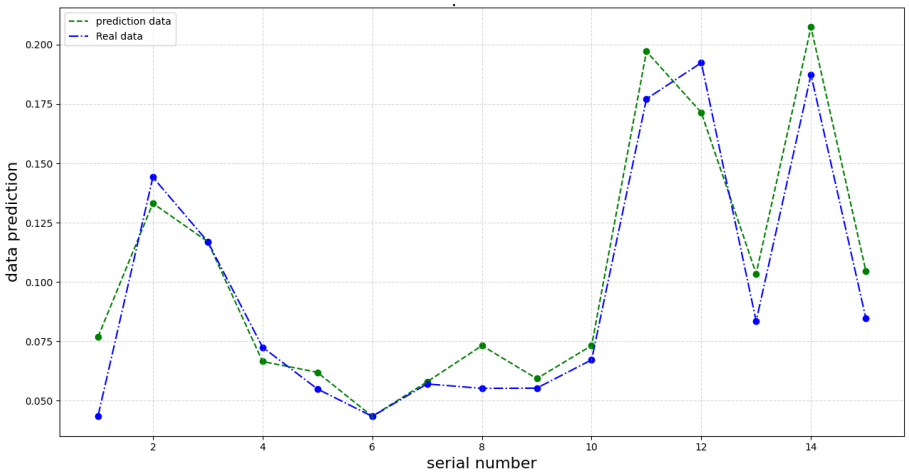

Figure 7: Model Backtest Prediction

In the above figure, the green line represents the prediction result and the blue line represents the original data, we can judge from the similarity of the two lines that the prediction accuracy of the model is relatively good.

- **Model Evaluation**

We chose five indicators, MSE, RMSE, MAE, MAPE, and R^2, to evaluate the model. As shown in the table below, the model has lower MSE, RMSE, MAE, and higher R2 in the training set, validation set, and test set, so we believe that the model regression prediction performance is relatively good and the above model results are relatively reliable.

Table 3: Model Performance Evaluation

Sets	MSE	RMSE	MAE	MAPE	R^2
Train Set	0	0.01	0.007	11.086	0.979
Validation Set	0.001	0.029	0.015	19.421	0.82
Test Set	0.001	0.026	0.015	18.758	0.853

6 Regional Impact Discovering

6.1 Sailboat Variants Clustering

In order to determine whether region has an effect on the price of a sailboat, the idea is to weaken the variation in the price of a sailboat due to the factors of the boat itself. To do this, we chose to cluster the different variants of boats such that the price fluctuations of each sailboat originate mainly from regional influences.

We first perform a frequency analysis of the data and conclude that 80.946% of the sailboat variants have a sample size of less than 30 and are not suitable for regression analysis alone, so we cluster the sailboat variants using K-means Clustering.

- **K-means Clustering**

Step 1: Randomly select K centers, denoted as $\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_k^{(0)}$;

Step 2: Define the loss function: $J(c, \mu) = \min \sum_{i=1}^{M} \|x_i - \mu_{c_i}\|^2$;

Step 3: Let t = 0, 1, 2, ... be the number of iterative steps and repeat the following process until J converges:

1. For each sample xi, assign it to the nearest center:

$$c_i^t < - \operatorname{argmin}_k \left\|x_i - \mu_k^t\right\|^2 \tag{8}$$

2. For each type of center k, recompute that:

$$\mu_k^{(t+1)} < - \operatorname{argmin}_\mu \sum_{i:c_i^t=k}^{b} \|x_i - \mu\|^2 \tag{9}$$

In order to make the value of K reasonable, the Gap Statistic method is used here, and the expression is:

$$Gap(k) = E\left(\log D_k\right) - \log D_k \tag{10}$$

The physical meaning of Gap Statistic is the difference between the loss of a random sample and the loss of an actual sample, the larger the Gap the better the clustering. [1]

6.2 Cluster Regression

We divided the clustered sailing variants into monohull 1-4 as $M_1 - M_4$ and catamaran 1-4 as $C_1 - C_4$, and regressed each category as a whole by substituting it into the model developed in the previous section.

6.2.1 Cluster Charateristics

As mentioned above, after clustering we divided the monohull ships into four categories as shown in the figure.

Meanwhile, the catamarans are divided into four categories, which are shown in the following diagram.

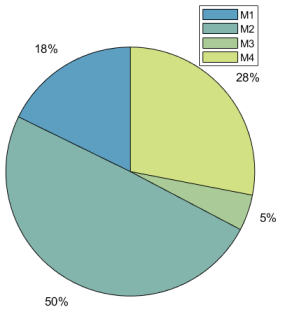

Figure 8: Monohull Categories

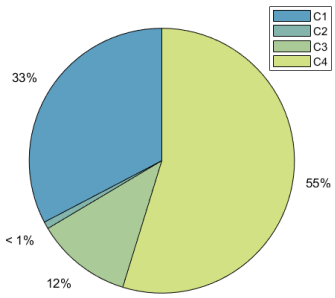

Figure 9: Catamarans Categories

It can be seen that the sailboats in each category fluctuate less in price after clustering, so we have a better clustering effect, and to a certain extent, the influence of the boat's own characteristics on the price is mitigated.

However, since the frequency of categories M2 and C3 are 4 and 10, respectively, the data are too small to achieve regression, so we remove these two categories. The remaining six categories C1, C2, C4, M1, M3, and M4 are substituted into the model for regression.

6.2.2 Regression Result

After regressing the six categories of sailboats, the coefficients of regional characteristics explaining the prices of the six sailboats were obtained as shown in the following table.

Table 4: Regional Feature Coefficients on Cluster

Cluster	PCI	AT	GDP	RB
C1	7.70%	6.60%	3.90%	3.60%
M1	9.21%	3.79%	2.66%	2.15%
C2	12.80%	8.80%	6.80%	4.20%
M3	2.38%	11.88%	10.02%	3.31%
C4	4.20%	3.60%	17.30%	6.90%
M4	1.16%	1.36%	1.45%	1.65%

6.3 Regional Effect Consistency Analysis

6.3.1 Kendall Consistency Test

The Kendall's W test, known as Kendall's concordance analysis, is a coefficient of concordance or coefficient of agreement, which is used to measure the degree of concordance. Kendall's W is the result of normalization of Friedman's statistic, which takes values between 0 and 1, and is used to measure the consistency of different measures, the closer the coefficient is to 1, the higher the consistency. The formula is as follows.[6]

$$\mathrm{W} = \frac{12\ \mathrm{S}}{\mathrm{K}^2\left(\mathrm{N}^3 - \mathrm{N}\right)} \tag{11}$$

where S is the sum of squares of the differences between the rank sum and its mean, K is the number of groups for the rank assessment, and N is the number of subjects whose rank was assessed. The chi-square test was used to construct the statistic $\chi^2 = K(N-1)rw$, where the degrees of freedom were N-1.

Judgment criterion:

If $\chi^2 < \chi^2(\mathrm{df})\alpha$then there is $100(1-\alpha)\%$ certainty that the K variables are not correlated; if $\chi^2 \geqslant \chi^2(\mathrm{df})\alpha$ then there is $100(1-\alpha)\%$ certainty that the K variables are correlated.[7]

6.3.2 Consistency Result and Analysis

After performing the kendall consistency test on the prices in Table 4, we obtain analytical results shown in Table 5.

The results of the Kendall coefficient consistency test showed that the overall data had a significance p-value of 0.051*, which did not show significance at the level of 0.051*, and the original hypothesis could not be rejected, so the data could not show consistency.

6.3.3 Practical Implications of Regional Effect Inconsistency

Since regional influence primarily describes the supply and demand of regional consumers, our understanding of the inconsistent regional influence across all variants is that consumers in a given region have **preferences** for different variants of sailboats, and thus the same regional characteristics do not have the same degree of influence on sailboat prices across variants.

This means that the coefficients we obtain for the regional influence on different hulls can explain the propensity of consumers in a given region to purchase used sailboats. For example, if the effect of region A on the price of luxury sailboats differs significantly more than the effect on the price of racing sailboats, then in a way it could indicate that consumers in region A have a greater preference between racing and luxury sailboats.

7 Empirical Analysis on Hongkong

7.1 Selection of Informative Subset

In selecting informative subsets, due to the inconsistency of regional influences, a subset that is informative enough for Hong Kong should be similar to the regional characteristics of Hong Kong. Therefore, we choose to put Hong Kong into the regional dataset of our previous model for clustering and find the regional category that is most similar to Hong Kong.

Table 5: Kendall's Consistency Test Result

Cluster	Rank Mean	Median	Kendall's W factor	χ^2	P
C1	3.75	0.053			
M1	3	0.032			
C2	5	0.078	0.55	11	0.051*
M3	4	0.067			
C4	4.25	0.056			
M4	1	0.014			

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively.

7.1.1 Region Clustering

We added Hong Kong to the regional dataset for clustering and obtained four categories of regions, of which Hong Kong is located in the Area 4.

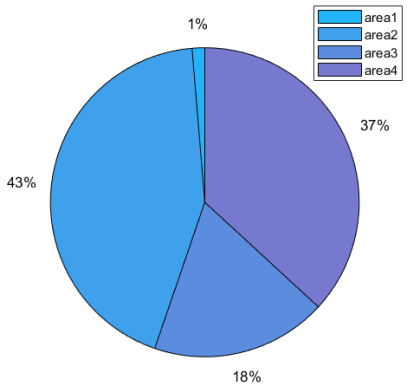

Figure 10: Region Cluster Categories

7.2 Regression and Prediction of Hong Kong Sailboat Price

The fourth class region is substituted into the model as a subset of information for learning, and regression is performed on this subset to obtain the regional impact coefficient of the Hong Kong class region. Meanwhile the corresponding sailing variants are predicted in the secondary prices in Hong Kong region, and the prediction accuracy is obtained by comparing the real data of Hong Kong.

7.2.1 Regression Result of Hong Kong

- **Monohull Price Regression**

After the model regression, we obtain the ranking of the coefficients of the influence of different features on the price of monohull in Hongkong, as Table 6 shown below.

Table 6: Feature Importance Ranking in Monohull

Rank	Feature	Importance	Rank	Feature	Importance
1	A	16.30%	8	SA	4.60%
2	NP	15.90%	9	H	4.50%
3	L	14.70%	10	GDP	4.00%
4	D_{is}	13.30%	11	RB	2.60%
5	B	10.50%	12	PCI	2.00%
6	D_r	5.20%	13	SC	1.70%
7	AT	4.70%			

From the above chart, it can be concluded that the main factors affecting the price of used monohull in Hong Kong market are age, new product price, length, displacement and beams. Meanwhile, the influence of regional characteristics of Hong Kong on prices are 4.7% for average temperature; 4% for GDP; 2.6% for number of registered sailboats; and 2% for per capita income, respectively.

Overall, the main factor affecting the price is still the characteristics of the monohull itself, accounting for 86.7%.

- **Catamaran Price Regression**

After the model regression, we obtain the ranking of the coefficients of the influence of different features on the price of catamaran in Hongkong, as Table 7 shown below.

Table 7: Feature Importance Ranking in Catamaran

Rank	Feature	Importance	Rank	Feature	Importance
1	A	23.80%	8	L	5.80%
2	NP	9.90%	9	PCI	5.70%
3	SA	8.70%	10	H	5.50%
4	B	8.40%	11	SC	4.20%
5	D_r	7.80%	12	RB	3.70%
6	D_{is}	7.80%	13	AT	2.50%
7	GDP	6.00%			

From the above chart, it can be concluded that the main factors affecting the price of used catamaran in Hong Kong market are age, new product price, sail area, beams and draft. Meanwhile, the influence of regional characteristics of Hong Kong on prices are 2.5% for average temperature; 6% for GDP; 3.7% for number of registered sailboats; and 5.7% for per capita income, respectively.

Overall, the main factor affecting the price is still the characteristics of the monohull itself, accounting for 82.1%.

7.2.2 Consistency Test

The regional influence of the above monohull and catamaran is summarized in the following table, which shows that the regional influence has inconsistency on monohull and catamaran prices.

Table 8: Regional Feature Coefficients on Monohull and Catamaran

Cluster	PCI	AT	GDP	RB
Monohull	2.00%	4.70%	4.00%	2.60%
Catamaran	5.70%	2.50%	6.00%	3.70%

7.2.3 Prediction Result of Hongkong

We forecast separately for the monohull and catamaran datasets and integrate the forecast prices in both sets to obtain the forecast curves as follows.

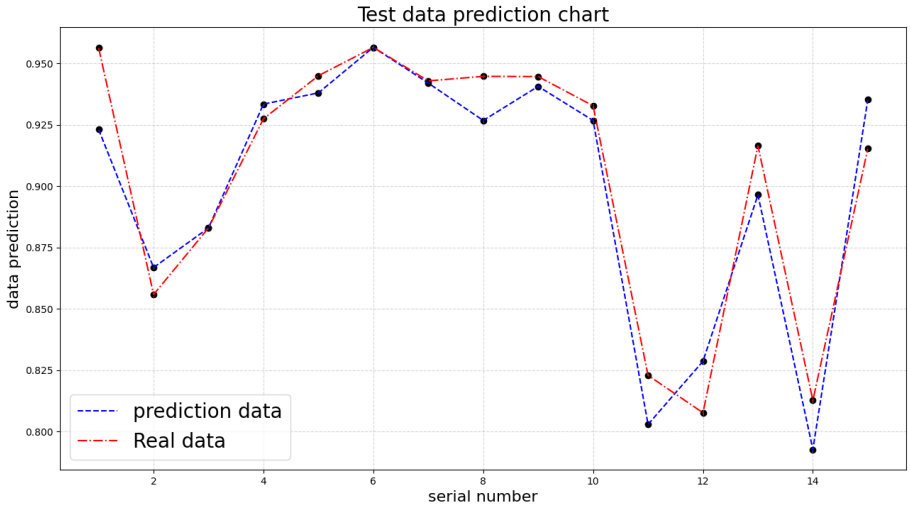

Figure 11: Comparison between Predicting Price and Real Price

8 Model Statistical Inference and Practical Implications

In the process of processing the data and regression predictions, we also found some interesting phenomena.

8.1 Regional Clustering Hierarchy

Analyzing the clustering of countries in the three regions, we find that all U.S. states are classified in Areas 1 and 4, and most of the richer states are classified in Area 4. European countries are more diverse, with a significant number of European countries in Areas 1, 2, and 4, while almost all countries in the Caribbean are classified in Area 2.

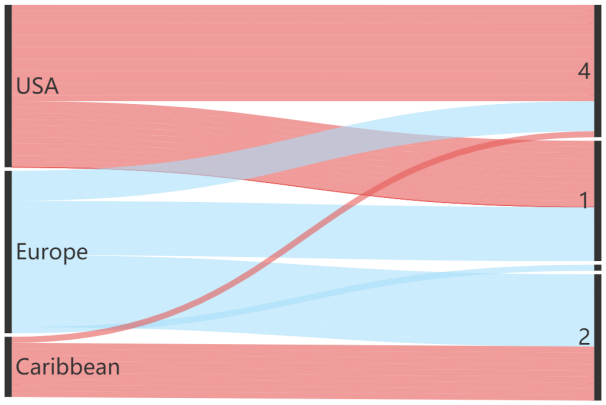

Figure 12: Regional Clustering Sankey Map

8.2 Outliers Monaco

When clustering regions, Monaco occupies a separate category when the clusters are 3,4,5. We have carefully analyzed this phenomenon and found that Monaco is the country with the highest per capita wealth in the world, with a per capita wealth that exceeds that of Liechtenstein, which is in second place, by about 170%, thus causing it to occupy a separate category.

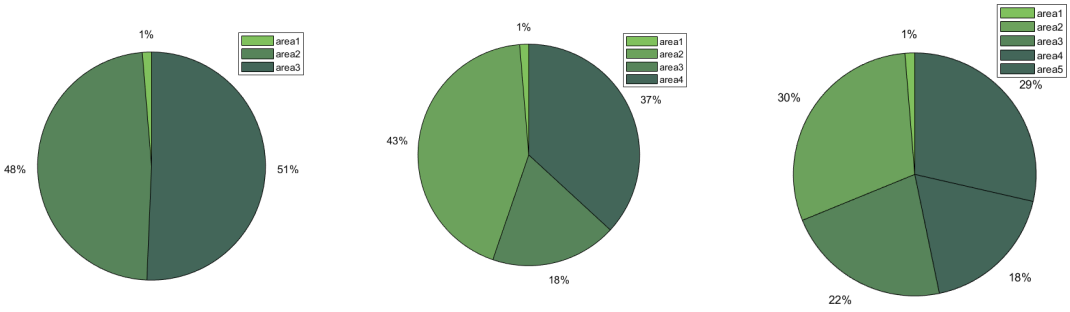

Figure 13: Cluster Categories when n=3,4,5

8.3 Inconsistent Influence of Sailingboat Feature

In the analysis of the clustering of sailing variants, we found that the top five impact characteristics were not consistent across sailing variants, as shown in the figure below.

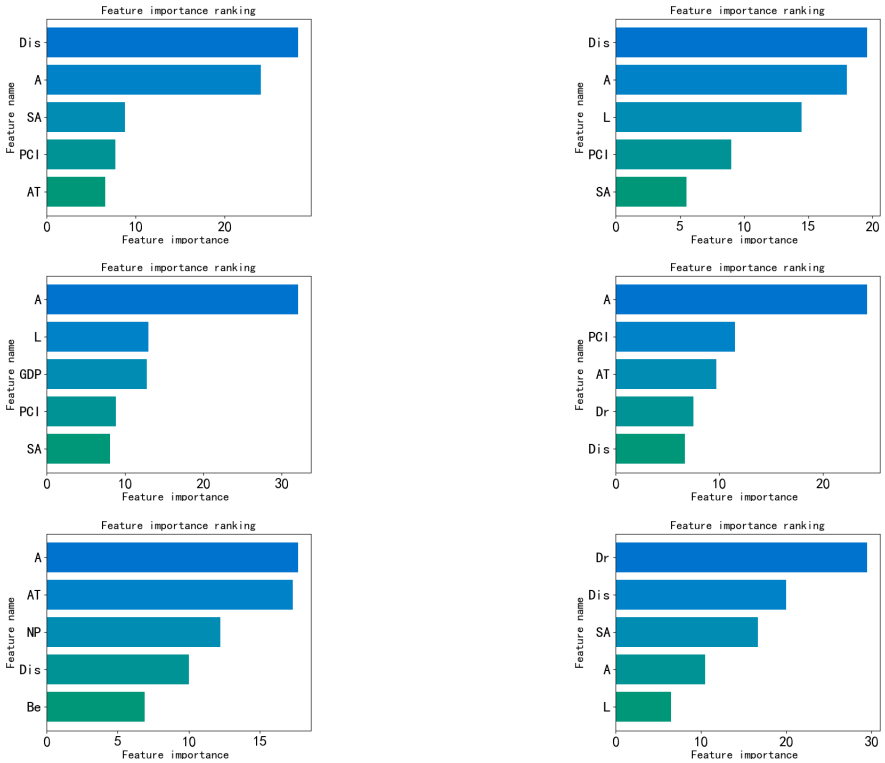

Figure 14: Top 5 Important Feature in Sailboat Cluster

As seen in the figure, the influence of sailboat characteristics on the different sailboat variants is also inconsistent, with displacement accounting for the first major influence in C1,M1 and age in C2,M3.

9 Sensitivity Analysis on Hyperparameter

In this paper, in order to achieve a large harmony of the overall model bias and variance when training the integrated learning model,the grid search method is used to adjust the parameters of the model in order to find the parameters with the best performance, so that the models have high prediction accuracy and generalization ability.

The following figure shows the validation curve of some model tuning parameters, the horizontal axis of which is a series of values of a hyperparameter and the vertical axis is a series of values of a hyperparameter. The horizontal axis of the curve is a series of values of a hyperparameter, and the vertical axis is the accuracy rate, so that the accuracy rate of the model can be compared under different parameter settings.

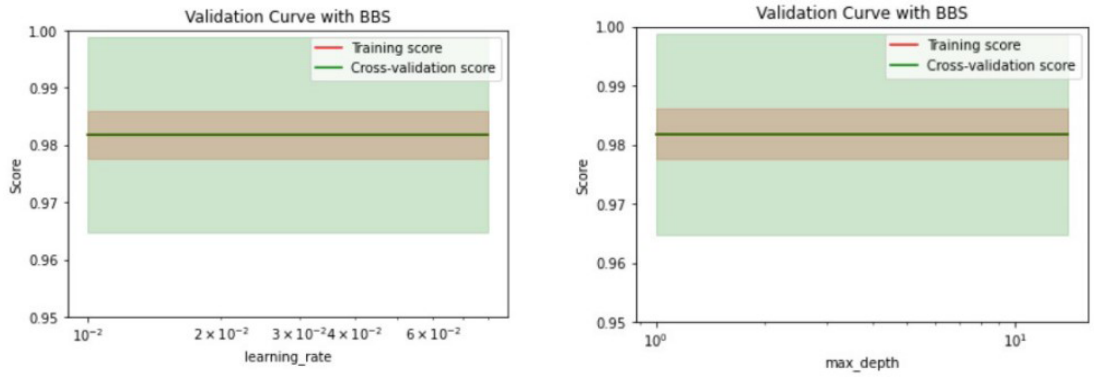

Figure 15: Validation Curve for Hyperparameter Optimization

10 Strength and Weakness

10.1 Strengths

Our model offers the following strengths:

- **Substantial data supplementation of the given data set to make the data more reliable and the results more accurate.**

- **Instead of coding regional variables on feature transformation, they are quantified as economic factors.**

10.2 Weaknesses

Our model has the following limitations and related improvements:

- **In the regression of the sailboat variant, the regression effect was overfitted due to the small number of samples in the clustered species, so some clusters had to be deleted.**

References

[1] Roar Adland, Haiying Jia, Hans Christian Olsen Harvei, and etc. Second-hand vessel valuation: an extreme gradient boosting approach. *Maritime Policy & Management*, 50(1):1–18, 2023.

[2] Abishek L. Gavin, P.K. Venkatesh Prasanna, C. Veena Vedha, and A. Sinduja. Data analysis and price prediction of stock market using machine learning regression algorithms. 124:409–417, 3 2023.

[3] Sarah Rodgers, Alexander Bowler, Fanran Meng, Stephen Poulston, Jon McKechnie, and Alex Conradie. Probabilistic commodity price projections for unbiased techno-economic analyses. *Engineering Applications of Artificial Intelligence*, 122:106065, 2023.

[4] Xiaowei Gu. Self-adaptive fuzzy learning ensemble systems with dimensionality compression from data streams. *Information Sciences*, 634:382–399, 2023.

[5] Thomas D. Gregory, Michael L. Perry, and Paul Albertus. Cost and price projections of synthetic active materials for redox flow batteries. *Journal of Power Sources*, 499:229965, 2021.

[6] Marina, Milenkovi, Draenko, Glavi, Milica, and Marii. Determining factors affecting congestion pricing acceptability. *Transport Policy*, 82:58–74, 2023.

[7] Yao Yingle, Li Jian, and Sun Bing. Simulation of the interpolation algorithm for fitting the processing of incomplete data with missing sequences. *Computer Simulation*, 40(1):523–527, 1 2023.

REPORT

From: Team 2331774

Date: April 4, 2023

By analyzing 2020 used sailboat advertising information from the US, Europe, and the Caribbean, combined with sailboat characteristics data from the sailboat database SailboatData.com and regional economic characteristics data from the IMF, World Bank, and BEA, we provide you with a used sailboat pricing model based on BBS integrated learning regression predictions to guide you in your Hong Kong used sailboat market.

1 Factors influencing the price of used sailboats in Hong Kong

Our model regresses the price data of monohull and catamaran separately to get the top five factors influencing the price of monohull in Hong Kong are age, new product price, length, displacement and beams, and all the influencing factors are shown in Figure 1, and the larger the font in the figure represents the larger the influencing factors. Similarly, the top five factors influencing the price of catamaran in Hong Kong are age, new product price, sail area, beams and draft, and all the influencing factors are shown in Figure 2.

Figure 1: Influence Factors of Monohull

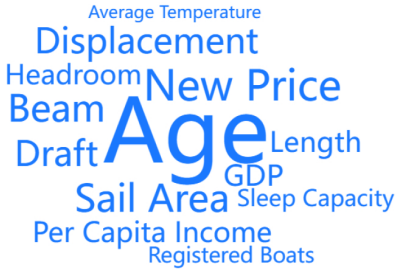

Figure 2: Influence Factors of Catamaran

2 Regional Effect Inconsistency in Monohull and Catamaran

Our model regressions show that the regional characteristics of Hong Kong affect 4.7% of average temperature; 4% of GDP; and 2.6% of the number of registered sailboats among monohulls; Similarly, the regional characteristics of Hong Kong affect 2.5% average temperature; 6% GDP; 3.7% number of registered sailboats; and 5.7% per capita income in catamarans.

This regional impact inconsistency shows a significant preference for monohulls and catamarans among consumers in the Hong Kong region.

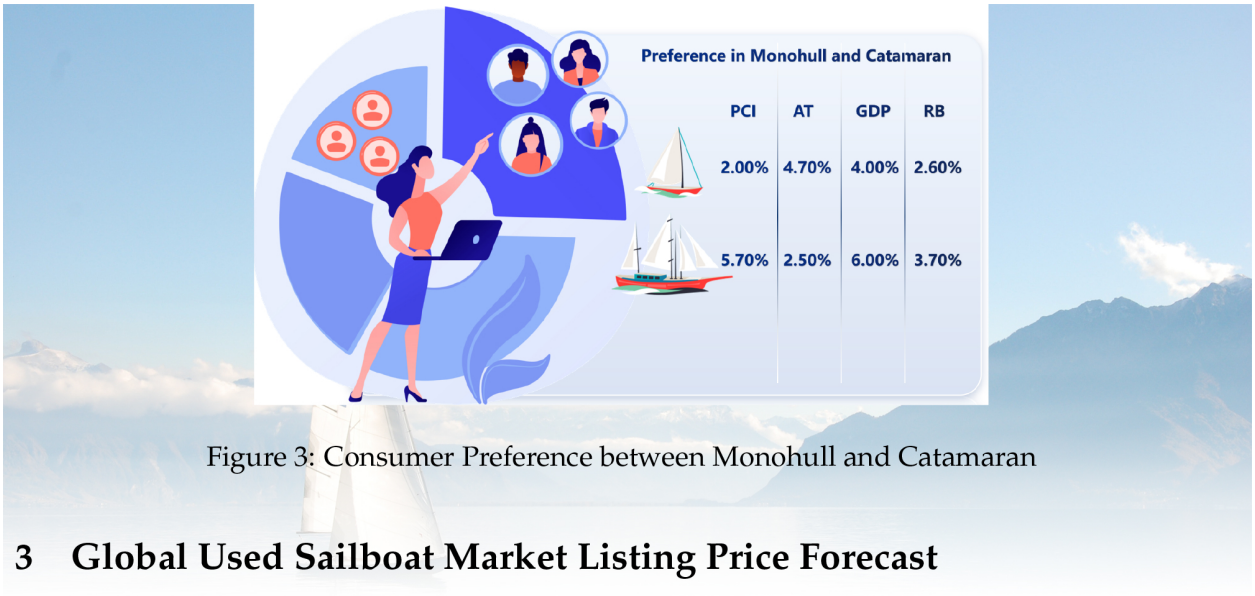

Figure 3: Consumer Preference between Monohull and Catamaran

3 Global Used Sailboat Market Listing Price Forecast

Our model performs a learning regression and prediction of global prices with a prediction accuracy of 85% and can be used to guide import and export trading decisions in Hong Kong.

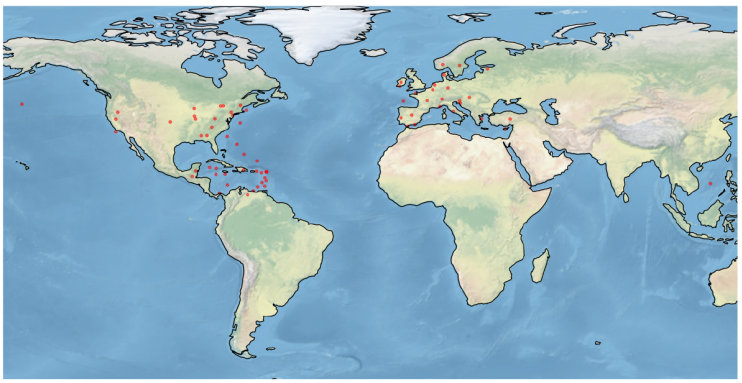

Figure 4: Map of model predictable regions

Due to the inconsistency of global regional influences, used sailboat prices are influenced by different factors in different regions. Our model does this by providing transparent pricing for used sailboat prices in different regions, advancing regional development, promoting inter-regional cooperation, and contributing to the realization of regional economic integration.