Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2330646

Set the Sail: Pricing of Used Sailboats

Summary

Due to the year of sailboats and the complex market conditions, the estimation for used sailboats is a difficult problem just like many luxury goods. This thesis intends to construct a reasonable model for sailing boat valuation based on the available data of used sailing boat transaction price in 2020, and fully analyze the influence of regional effects on the price, and finally give reasonable suggestions and analysis based on the market situation in Hong Kong.

For problem(a), we built **Used Sailboat Price Forecasting Model**. Based on the suggestion of the question, we extended the intrinsic and external factors affecting boat prices, found rich supplementary data and performed reasonable data pre-processing. Based on the **LightGBM**, we utilized the **SHAP**method to rank the importance of features for monohull and catamaran respectively, and retained the 11 highest scoring feature factors (see Figure 5 of the paper for details). The final sailboat price prediction models were obtained with mean square error (MSE) of 61,164.43 and 96,983.49, respectively. Based on the **multiple linear regression** approach, we performed stepwise regression analysis and obtained partial regression coefficients for monohull and catamaran, respectively, which can indicate the weights of each feature.

For question(b), using SHAP **feature importance ranking** feature importance ranking and **weight analysis** of multiple regression coefficients, we obtained that the main regional factors affecting monohull boats are **disposable income per capita** and **tourism income**; the main regional factors affecting catamaran prices are **CPI (Consumer Price Index)** and **local tourism income**. We thus performed **Kmeans clustering** for all economic regions and classified the 65 regions into 5 categories based on per capita disposable income and tourism income, and analyzed in detail how regional effects affect sailboat prices.

For question(c), we collected data on the second-hand sailboat market in Hong Kong and used our model to predict the data on monohulls and catamarans in the Hong Kong market with mean squared errors of 145,057 and 166,592, respectively.To analyze the **regional effects on Hong Kong**, we conduct a **comparison of prices** of the same variant in different regions based on data from the Hong Kong second-hand market. The prices of second-hand boats in Hong Kong are reasonably and adequately explained in terms of the main factors affecting boat prices.

Finally, on the basis of the available data analysis and our observations, we found further informative conclusions. The main factors affecting the price of used sailboats are potentially related to the application scenario of sailboats, and the volume of sailboats traded in different regions is related to the latitude of the region. In conclusion, our model gives a reasonable model for valuing used sailboats and analyzes in detail the influence of regional effects on sailboat prices. It provides a great reference value for the valuation of used boats in Hong Kong.

Key Words: Price Forecasting; LightGBM; SHAP; Multiple Linear Regression; Feature Importance Ranking; K-means Clustering

Contents

1 Introduction

1.1 Background

The value of a sailboat, being a luxury item, is subject to fluctuations due to factors such as age and market conditions. Pricing used sailboats can also be challenging, as it is influenced by regional economic and climatic conditions, proximity to the sea, and the degree of tourism development. All these factors make it difficult to determine the accurate price of each sailboat variant. Therefore, we aim to develop a pricing model for used sailboats that takes into account these influencing factors, enabling us to provide more precise estimates of their value.

1.2 Restatement of the Problem

- Based on the data file given in the question, we need to establish a functional relationship between Make, Variant, Length , Geographic Region, Listing Price, Year and used sailboat prices and construct a regression explanatory model with multiple independent variables and a single dependent variable. This model can be used to explain the effect of different regions on the price of used sailboats.
- On the basis of the attached data file, we can obtain more information about the characteristics of the sailboat (such as beam, draft, displacement, sail area, hull materials, sleeping capacity, headroom, etc.)and the economic data related to the different regions.
- Choose an informative subset of sailboats, split between monohulls and catamarans, from the provided spreadsheet.By analyzing the above economic factors, we will simulate and forecast the trading price of used sailboats in Hong Kong based on the HK's economic situation.Find comparable listing price data for that subset from the Hong Kong market.
- Analyze how the regional effect affects the price of used ship transactions and explain how the regional effect affects the price in Hong Kong. At the same time dig more relationships between data.

1.3 Our work

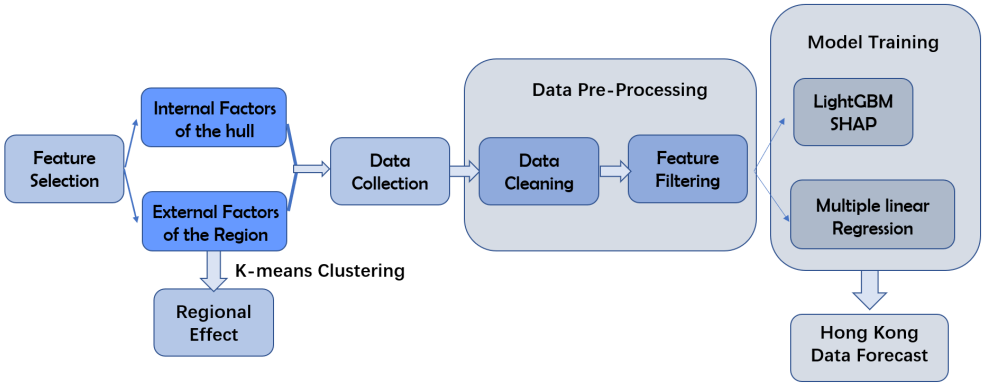


Figure 1: Processing Flow Chart

2 Assumptions and Justifications

(a)We assume that the economic factors affecting the price of used sailboats in a region are broadly divided into five components: disposable income per capita, GDP, GDP growth rate, CPI, and tourism revenue.

(b)The data given in the question is true and reliable. Since the question restricts the use of the data given in the question, our results are valid only if the data in the question are true and reliable.

3 Notations

Symbol	Description
ϕ_j	Attribution value for each feature
a_{ij}	The contribution of the jth feature of the ith sample to the result
$Price$	Dependent variable of linear regression
β	The partial regression coefficient
μ	The random error term
K	Number of clusters of Kmeans algorithm
$\rho_{X,Y}$	Pearson's correlation coefficient between X,Y

4 Used Sailboat Price Forecasting Model

4.1 Data Pre-processing

4.1.1 Data Collection

Through the preliminary analysis of the data characteristics provided by the question, we get that there are 56 Makers and 456 Variants. If used for regression analysis, the Makers and Variants number are categorical variables, and it is obviously unrealistic to use them directly for regression analysis for such a large number of categorical variables. Therefore, in order to make our model data richer and the modeling results more realistic, we collected additional characteristics data (e.g., beam, draft, displacement, sail area, hull material, sleeping capacity, and headroom) for the sailboat hull according to the requirements of the topic. These quantitative data were used to participate in the model building instead of the two definite categories of make and Variant.Part of the data is shown in the Table 1.

This question requires a model for the price of used sailboats. After our preliminary analysis, apparently, the characteristics of the data provided in the questions, the Make and Variant, the Year and length of the boat are internal factors of the boat itself, while the economic market conditions in the region where the transaction takes place can also greatly affect the selling price of a used sailboat.

Through our visualization and preliminary analysis of the data, we can clearly see that sailing vessels are generally traded in coastal or harbor areas, and that more sailing vessels are traded in areas with developed economies or thriving tourism.Details are shown in the Figure 2.

As a result of our research, we believe that the factors that can influence the price of a local

Table 1: More sailboat hull characteristics data[1]

Make	Variants	Beam(m)	Draft(m)	Displace.[2](kg)	S.A.[3](m2)	Materials	S.C. [4]	Headroom(m)
Alubat	Ovni 395	3.9	1.2	7400	89.5	Aluminum	06	15
Bavaria	38 Cruiser	3.87	1.95	7200	77.1	GRP	06	17.8
Bavaria	39 Cruiser	3.97	1.85	8300	83.4	GRP	06	18.45
Bavaria	42 Match	3.98	2.4	8400	109.7	GRP	06	20.95
Bavaria	42 Cruiser	4.29	1.95	9800	89.7	GRP	07	19.35
Bavaria	Cruiser 46	4.35	1.73	12600	112	GRP	07	20.75
Bavaria	50 Cruiser	4.49	1.99	14550	108.8	GRP	07	21.67
Beneteau	40	3.89	1.95	8260	67.8	Fiberglass	06	19.58
Beneteau	41.1	3.96	2.00	7800	74.0	Fiberglass	06	19.81
Beneteau	423	3.90	1.70	8420	80.0	Fiberglass	08	18.55
Beneteau	43	4.17	2.00	9400	94.7	Fiberglass	08	19.20
Beneteau	523	5.00	2.50	16000	136.0	Fiberglass	08	23.20
Beneteau	56	5.00	2.50	21500	178.0	Fiberglass	08	26.50
Beneteau	America 50	4.49	2.20	11500	104.0	Fiberglass	08	22.20
Beneteau	Cyclades 39.3	3.95	1.90	7500	62.5	Fiberglass	08	18.80

[1] The data is from https://sailboatdata.com/
[2] Displacement:The weitght of the volume of water displaced by a boat.
[3] Sail Area: The Total surface area of the sails of a boat when fully raised.
[4] Sleeping Capacity.

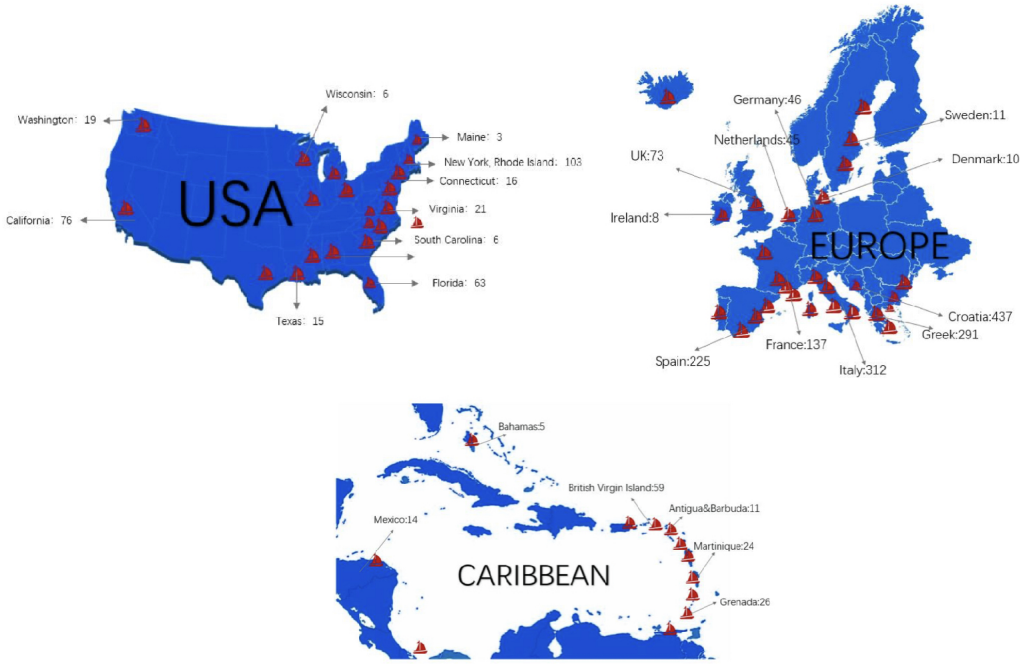


Figure 2: Number of reference prices for used sailboats by region in the given data

sailboat transaction are:Per capital disposable income, GDP, CPI(Consumer Price Index), GDP growth rate, Whether waterfront, Tourism revenue, Tourism arrivals. Some of the relevant data are shown in Table 2.

We divided the total 16 factors affecting the price of used sailboats into two main blocks: internal factors of the sailboat's intrinsic characteristics(Beam, Draft, Displacement, Sailing Area, Hull Materials, Sleeping Capacity, Headroom) and external factors of the region where the transaction

Table 2: Economic conditions at each sales location[1]

Region	PCDI(USD)[2]	GDP(billion USD)	CPI	GDP growth rate	waterfront	Tourism revenue(billion USD)	Tourism arrivals(10,000)
Alabama	45,653	224.8	266.2	2.60%	1	15.5	27.7
Antigua and Barbuda	16,200	1.6	133.2	5.30%	1	1.5	0.3
Aruba	25,067	3.1	113.8	2.90%	1	2.5	1
Bahamas	35,177	13.4	102.4	0.90%	1	4.2	1.4
Belgium	49,400	536.7	107.7	1.80%	1	10.9	11.5
Belize	4,747	1.9	103.6	-6.50%	1	0.7	0.5
British Virgin Islands	34,772	0.9	110.8	-1.70%	1	0.3	0.2
Bulgaria	9,619	59.4	100.5	3.40%	1	4.1	8.8
California	72,895	2,932.60	288.3	3.30%	1	140.6	251.4
Connecticut	78,222	289.7	264.1	2.20%	1	15.5	30.9
Cork	79568	3849.1	0.5	4.30%	1	5.5	107
Croatia	14,569	55.1	104.5	2.90%	1	10.3	19.7
Cyprus	28,876	23.8	102.4	4.00%	1	2.7	4
Denmark	61,104	306.3	110.6	2.70%	1	10.8	11.8
Dominican Republic	88.03	8,353	3.3	-6.80%	1	8.1	5.6
Estonia	33.08	23,426	2.9	7.80%	1	1.5	3.8
Florida	1,090.40	50,884	2.7	-4.10%	1	86.3	131.4
France	2,687.30	41,598	2.1	-8.20%	1	63.8	89.4
Georgia	17.83	4,467	4.7	4.50%	0	3.2	0.9
Germany	4,238.60	51,319	1.9	-4.90%	1	57.6	38.6

[1] The data is from US Bureau of Economic Analysis, US Bureau of Labor Statistics, Alabama Department of Commerce, Alabama Tourism Department World Bank, Eastern Caribbean Central Bank Central Bureau of Statistics, Aruba Tourism Authority International Monetary Fund, Bahamas Ministry of Tourism World Bank, National Bank of Belgium, Visit Flanders, Visit Brussels World Bank, Central Bank of Belize, Belize Tourism Board World Bank, British Virgin Islands Tourist Board World Bank, National Statistical Institute of Bulgaria, Ministry of Tourism US Bureau of Economic Analysis, US Bureau of Labor Statistics, California Department of Finance, Visit California US Bureau of Economic Analysis, US Bureau of Labor Statistics, Connecticut Office of Tourism World Bank, Fáilte Ireland World Bank, Croatian National Bank, Croatian National Tourist Board World Bank, Cyprus Statistical Service, Cyprus Tourism Organization World Bank, Statistics Denmark, Visit Denmark World Bank, Trading Economics, UNWTO BEA, Trading Economics, Visit Florida IMF World Bank, Trading Economics BEA, Trading Economics BEA World Bank, Trading Economics, Fáilte Ireland World Bank Jersey Statistics Unit US Travel Association US Bureau of Economic Analysis and NY State Department of Economic Development US Bureau of Economic Analysis and Visit NC Statistics Norway, Norwegian Directorate of Immigration and Visit Norway US Bureau of Economic Analysis and Ohio Development Services Agency US Bureau of Economic Analysis and Travel Oregon National Institute of Statistics and Census of Panama, Central America Tourism Agency and Panama Tourism Authority Statistics Portugal and Visit Portugal Puerto Rico Planning Board and Discover Puerto Rico [US Bureau of Economic Analysis Trading Economics World Bank, IMF, CIA World Factbook https://www.bea.gov/data/gdp/gdp-state

[2] PCDI:Per capita disposable income(USD)

takes place(PCDI, GDP, CPI, GDP Growth Rate, Waterfront, Tourism Revenue, Tourism Arrivals). The details are shown in the Figure 3.

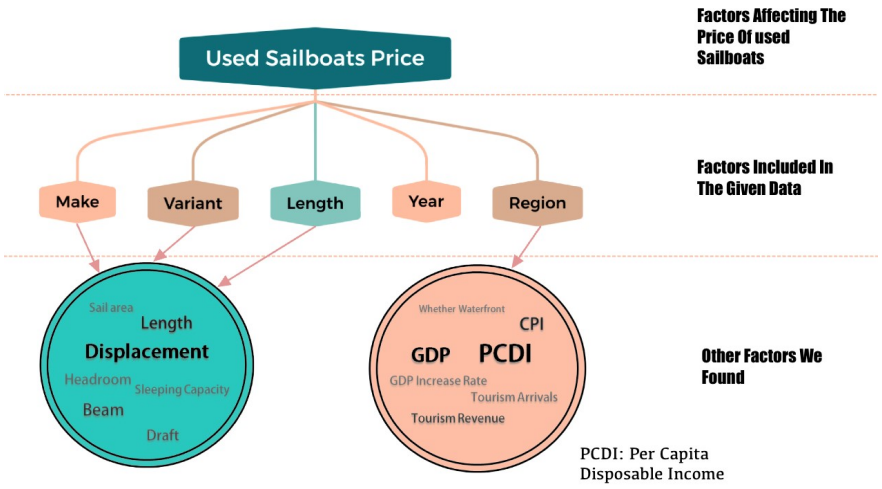


Figure 3: Factors affecting the price of a sailboat

4.1.2 Data Cleaning

(a)Missing data filtering

By filtering and extracting the raw data, we found that only three data for single-hull sailboats were missing information about the region they were located in, and we chose to simply discard them.

(b)Mean Integration

In addition, there are multiple data showing: at the same region and the same year the same boat variant were sold at different prices. For this kind of data where only the prices are different, we use the method of taking the mean value and synthesize the multiple data into one data. The original data of monohull sailboats has 2347 items, and 1837 items are left after cleaning by mean processing. The original data of Catamarans has 1146 items, and 673 items are left after cleaning by mean processing.

(c)Clear items with missing economic data

We have added some additional relevant characteristics. When collecting data, we found that many regions could not obtain economic data directly because of untimely statistics or political reasons, so we also considered eliminating these data items. After this step, a total of 1707 clean monohull sailboat data and 573 clean catamarans data were finally obtained for subsequent modeling.

4.2 Feature Filtering Based on SHAP Method

4.2.1 LightGBM-based model training

LightGBM is an improved version of **XGBoost**. The idea is to discretize the continuous floating-point features into k discrete values and construct a histogram of width k. Then traverse the training data and calculate the cumulative statistics of each discrete value in the histogram. For feature selection, it is only needed to iterate through the discrete values of the histogram to find the optimal segmentation point.

LightGBM contains only one decision tree, and the process of generating the tree discards XGBoost's **level-wise** and uses **leaf-wise** with a depth limit, as shown in the figure. leaf-wise is more efficient than level-wise, finding the leaf with the greatest splitting gain from the current leaf each iteration. Then the splitting is done again and so on. Compared with level-wise, the error can be reduced and better accuracy can be obtained with the same number of splits. However, it may grow a deeper decision tree and produce overfitting, so a depth limit should be imposed. When a set of sample values is input, LightGBM can use level recursion to get the prediction result.

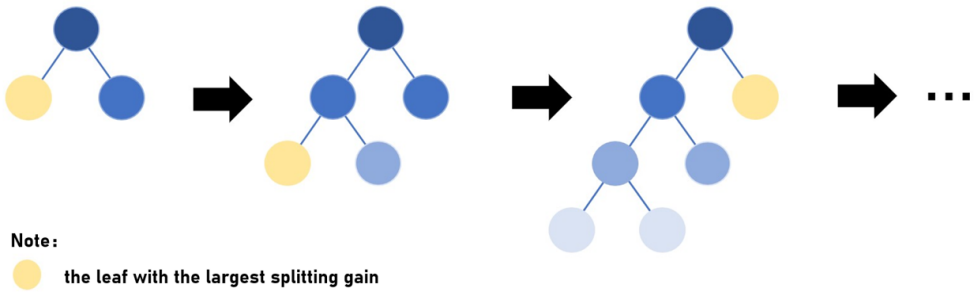


Figure 4: The leaf-wise method of LightGBM generating tree

The traditional feature importance algorithm based on XGboost is based on three importance

indicators: **weight**, **cover**, and **gain**. When the value of the indicator is larger, the importance of the variable is higher, and its specific meaning is explained as follows:

- **weight**:Sum of the number of nodes used to segment features in all trees
- **cover**:Average gain of features for segmentation
- **gain**:The number of samples covered by the leaf nodes below the feature divided by the number of times the feature is used to split

In general, weight will give higher values for numerical features. This is because the more variations in consecutive values, the more space can be cut when the tree splits, and that is used more often. If the tree is split by a certain feature, and the increment of entropy is larger, then the stronger the importance of that feature is when using gain calculation. And when cover is used, the closer the split is to the root of the tree, the larger its value is. Therefore the choice of different importance metrics can have a large impact on the results, and the feature importance ranking may be different.

To avoid the instability of the results obtained by traditional methods, we finally chose to use the **SHAP** method to give the feature importance.

4.2.2 SHAP method based on imputation of addable features

Since we independently searched a lot of extended sailboat data, for which the influence on the final used sailboat price is doubtful, we need to perform feature filtering before conducting regression analysis.

For traditional multiple linear regression methods, multicollinearity is a problem that needs to be concerned and solved, otherwise it will greatly affect the accuracy of our regression results. We can use Person correlation coefficient analysis to eliminate the characteristic variables with high correlation. And numerous machine learning algorithms are generally black box models, which generally have the problem of poor interpretability, so we use the SHAP method, hoping to give the importance ranking of numerous features.

SHAP is an additivity interpretation model inspired by Shapley value. SHAP interprets the predictive value of the model as the sum of the imputed values of each input feature. For each prediction sample, the model produces a prediction value, and the SHAP value is the value assigned to each feature in that sample. Assuming that the i-th sample is x_i, the j-th feature of the i-th sample is $x_{i,j}$, the predicted value of the model for the i-th sample is y_i, and the baseline (i.e., the mean of the target variables for all samples) for the entire model is y_{base}, then the SHAP values obey the following equation:

$$y_i = y_{\text{base}} + f\left(x_{i,1}\right) + f\left(x_{i,2}\right) + \cdots + f\left(x_{i,k}\right)$$

where $f(x_i, 1)$ is the SHAP value of $x_{i,j}$. Intuitively, $f(xi, 1)$ is the contribution value of the 1st feature in the ith sample to the final prediction value yi. When $f(xi, 1) > 0$, it means that the feature enhances the prediction value and also has a positive effect;conversely, it means that the feature makes the predicted value lower and has the opposite effect.The biggest advantage of the

SHAP value is that SHAP is able to reflect the influence of the features in each sample, and also shows the positive and negative nature of the influence.

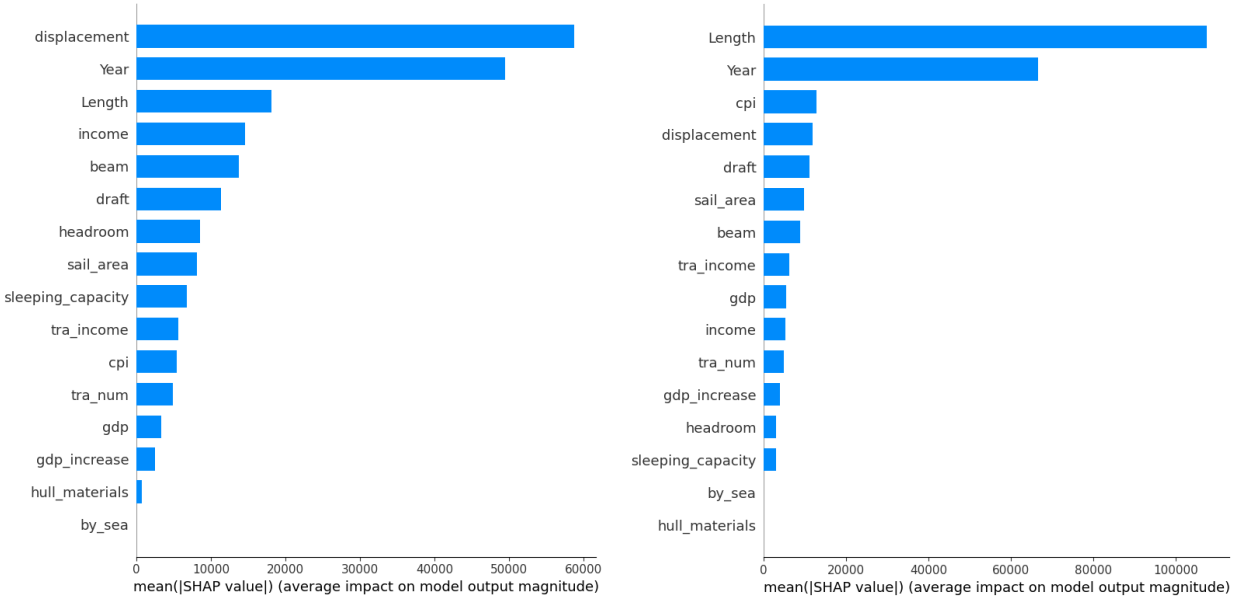
Specifically, when the model is nonlinear or the input features are not independent, the SHAP values should compute weighted averages for all possible feature rankings. SHAP combines these conditional expectations with the classical Shapley values from game theory into an attributed value ϕ_j for each feature, according to the following equation[1][2]:

$$\phi_j = \sum_{S \subseteq \{x_1, \cdots, x_p\} \setminus \{x_j\}} \frac{|S|!(p - |S| - 1)!}{p!} \left(f_x \left(S \cup \{x_j\} \right) - f_x(S) \right)$$

where $\{x_1, \cdots, x_p\}$ is the set of all input features, p is the number of all input features, $\{x_1, \cdots, x_p\} \setminus \{x_j\}$ is the set of all possible input features excluding $\{x_j\}$, and $f_x(S)$ as the prediction of the feature subset S.

By performing SHAP analysis on all of the factors we listed, we obtained the results of monohulled sailboats and Catamarans are shown in the figure 5. Since the SHAP scores of the last few elements are low, they have little impact on the results. To save computational resources, we discard the last five features and take only the first eleven features for the subsequent regression analysis, which are **displacement, year, length, beam, per capital disposable income, draft, sleeping capacity, headroom, sailing area, tourism revenue and CPI** for monohulled sailboats and **length, displacement, year, beam, tourism revenue, per capital disposable income, draft, sleeping capacity, tourism arrivals, sailing area, GDP and CPI** for Catamarans.

Based on the features that have been selected, we proceed to fit the model regression using



(a) Monohulled sailboat (b) Catamarans

Figure 5: SHAP feature importance ranking

lightGBM. Taking the catamaran as an example, the Figure 6 shows the histogram of the predicted and true values fitted by lightGBM before and after excluding the features. We take the mean square error(MSE) as the indicator, and according to the calculation, the mean square error of the monohull is reduced from 72,569.96 to 61,164.43; the mean square error of the catamaran is reduced from 98,877.00 to 96,983.49.

4.3 Multiple linear regression model

Obviously this question is a typical regression problem, and we can first try to use the traditional multiple linear regression method. Assume that the explanatory variable $Price$ has a linear relationship with multiple explanatory variables$(x_1, x_2, x_3, ..., x_{16})$ and is a multiple linear function of the explanatory variables.

$$Price = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

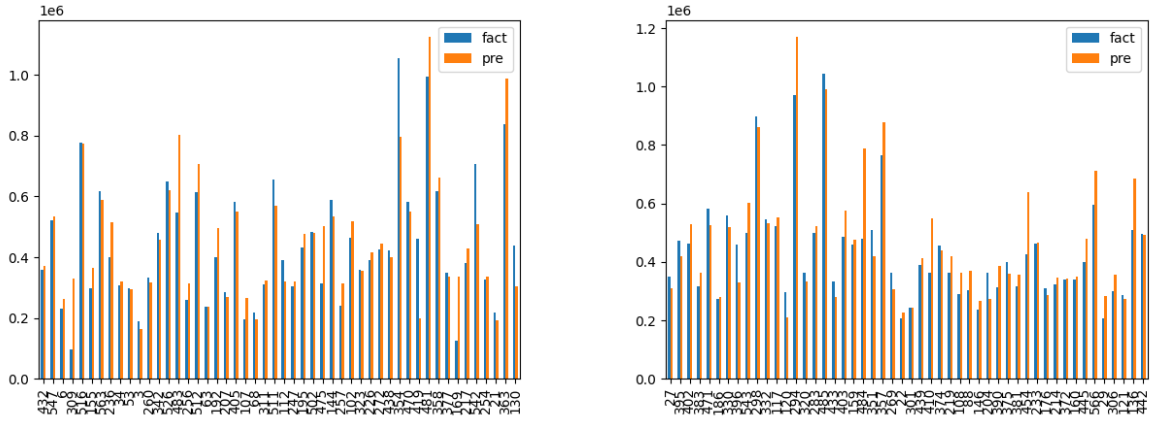
Notes: For convenience, we set the above 16 factors as $x_1, x_2, x_3, ..., x_{16}$, $price$ is the explanatory variable, the partial regression coefficient β is the 16 unknown parameters, β_0 is the constant term, and μ is the random error term.

For n observations, the system of equations takes the form:

$$Price_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mu_i$$

The partial regression coefficients are all unknown and can be estimated using the sample observations$(Price_i, x_{1i}, x_{2i}, ..., x_{ki})$. If the calculated parameter estimates are $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$, and the position parameters$\beta_0, \beta_1, \beta_2, ..., \beta_k$ in the overall regression equation are replaced by the parameter estimates,, then the multiple linear regression equation is:

$$\widehat{Price_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \cdots \widehat{\beta}_k x_{ki}$$



(a) Training results without feature removal

(b) Training results of feature removal

Figure 6: Comparison of true and predicted values of lightGBM training results

When establishing a multiple linear regression model, in order to ensure that the regression model has excellent explanatory power and predictive effect, attention should first be paid to the selection of independent variables, and the guidelines are:

- The independent variables must have a significant effect on the dependent variable and be closely linearly correlated
- The independent variables should have a certain degree of mutual exclusivity, that is, the correlation between the independent variables should not be higher than that between the independent variables and the dependent variable
- The independent variables should have complete statistical data

Here we use OLS (ordinary least squares) for parameter estimation, we first try to perform regression analysis on all 16 independent variables. The results are shown in Fig.We can find that when OLS regression is first applied, $R^2 = 0.937$ indicates that 93.70% of the dependent variable $Price$ (composite score) can be determined by the model. The p-value of the hypothesis test $(prob(F - statistics) = 5.77e - 260)$ tends to 0. This indicates that our model results are significant, that is, price is significantly linearly related to $x_1, x_2, ..., x_{16}$. In addition to that, $P > |t|$ reflects the linear significant relationship between each independent variable and $Price$. It can be found that the hull material, headroom, income, gdp, etc. are larger in this item, which means that these variables have insignificant linear relationship with y. We used **step-wise regression** to remove it.

The principle of step-wise regression is that only one item can be eliminated at an iteration, generally the one with the largest p-value is eliminated, and then the above modeling process is repeated with the remaining independent variables.Until all p-values are less than or equal to 0.05, these remaining independent variables are the ones we need.The final step-wise regression results are showed in the Table 4. For monohulled boat, $R^2 = 0.876$, the p-value also tends to 0. Moreover, the $P > |t|$ items have small values, which means our model has significant relationship,

Table 3: First Linear Regression Model Evaluation

Factors	Coef	Std Err	t	P>|t|
Beam	0.1291	0.048	2.673	0.008
Draft	0.0903	0.018	4.973	0.000
Displacement	0.1510	0.034	4.467	0.000
Sail Area	-0.1693	0.036	-4.646	0.000
Hull Materials	-0.0254	0.026	-0.993	**0.321**
Sleeping Capacity	-0.0538	0.016	-3.351	0.001
Headroom	-0.0075	0.026	-0.284	**0.777**
Length	0.4640	0.042	11.112	0.000
Income	0.0129	0.027	0.487	**0.627**
GDP	0.0059	0.019	0.306	**0.760**
CPI	-0.0261	0.026	-0.994	**0.321**
GDP Increase	-0.0203	0.019	-1.053	**0.293**
Whether Waterfront	-0.0476	0.026	-1.830	0.068
Travel Income	0.0973	0.032	3.012	0.003
Travel Num	-0.0181	0.039	-0.469	**0.639**
Year	0.2180	0.014	15.514	0.000

and makes sense.The final fitting function is:

$$\begin{aligned} price = (&-5.323x_1 + 0.777x_2 + 7.463x_3 - 1.589x_6 - 0.6214x_7 + 2.441x_8 - 1.735x_9 \\ &+1.156x_{10} + 0.6808x_{11} + 0.6632x_{12} - 0.2698x_{15} + 2.069x_{16}) \times 10^5 \end{aligned} \tag{1}$$

For catamarans, $R^2 = 0.959$, all the hypothesis tests have passed. The final function is:

$$\begin{aligned} price = (&1.34x_1 + 1.345x_2 + 1.955x_3 - 2.38x_4 - 0.7931x_6 + 6.589x_8 + 1.092x_{11} \\ &-0.4179x_{12} + 0.6892x_{14} + 3.043x_{16}) \times 10^5 \end{aligned} \tag{2}$$

From the above data, we find that the multiple linear regression fits well for catamarans, but not for monohulls. Comparing the results of lightGBM and multiple linear regression methods, for the correctness of the subsequent modeling, we recommend using lightGBM to ensure the accuracy of the predicted data.

4.4 Explanation of regional effects on the price of used sailboats

4.4.1 Explanation of "regional influence on prices"

From the SHAP feature importance ranking chart already in the previous section, we can visualize that: for monohulls, the regional factors affecting the trading price of sailing boats are **local disposable income per capital** ; for catamarans, the regional factors are mainly **CPI, local tourism income**. Details can be seen in Figure5.

From the values of the final partial regression coefficients of the multiple linear regression, the weights of the different factors influencing the price of sailboats and whether they have a positive or negative effect can be obtained.Obviously, for monohull boats, **local disposable income per capita and local tourism income** all have an effect on prices and are positively correlated; for catamarans, **CPI and local tourism income** has a greater effect and is also positively correlated. Detailed data can be found in Table 4.

Combining the results of these two analyses, we can conclude that regional effects also affect the price of used sailboats to some extent, and the most important factors are the **local disposable income per capita for monohulls** and **CPI and the local tourism income for catamarans**. The higher the income from tourism and the higher the level of economic development of a region, the higher the price of sailboats in that region.

This is understandable from a practical use perspective. Monohull sailing is generally used for individual sports training and racing, which is closely related to local personal disposable income, while catamaran sailing is generally used for tourism and leisure vacations, which is closely related to local tourism income and CPI. This can also reflect the correctness of the results of our data analysis.

4.4.2 K-means clustering based on regional economic level

To further analyze the impact of regional effects, we clustered the locations of sailboat transactions by economic level. We used the Kmeans method for clustering. Since the main regional

Table 4: Evaluation of linear regression models for monohulled sailboats(left) and Catamarans(right)

Factors	Coef(10^5)	Std Err(10^4)	t	P>|t|
Beam	-5.323	4.74	-11.309	0.000
Draft	0.777	2.47	3.154	0.002
Displacement	7.463	4.18	17.977	0.000
Sleeping Capacity	-1.589	2.28	-6.483	0.000
Headroom	0.6214	0.8694	7.025	0.000
Length	2.441	2.47	9.773	0.000
Income	1.735	1.6	12.436	0.000
GDP	1.156	1.78	5.183	0.000
CPI	0.6808	1.29	-5.411	0.000
GDP Increase	0.6623	2.41	3.372	0.001
Travel Num	-0.2698	1.53	2.013	0.044
Year	2.069	0.8653	24.444	0.000

Factors	Coef(10^5)	Std Err(10^4)	t	P>|t|
Beam	1.34	5.28	4.033	0.000
Draft	1.345	2.3	6.145	0.000
Displacement	1.955	3.89	4.707	0.000
Sail Area	-2.38	4.57	-5.364	0.000
Sleeping Capacity	-0.7931	2	-4.301	0.000
Length	6.589	4.7	14.448	0.000
CPI	1.092	4.7	14.448	0.000
GDP Increase	-0.4179	4.7	14.448	0.000
Travel Income	0.6892	2.07	6.862	0.000
Year	3.043	1.74	17.293	0.000

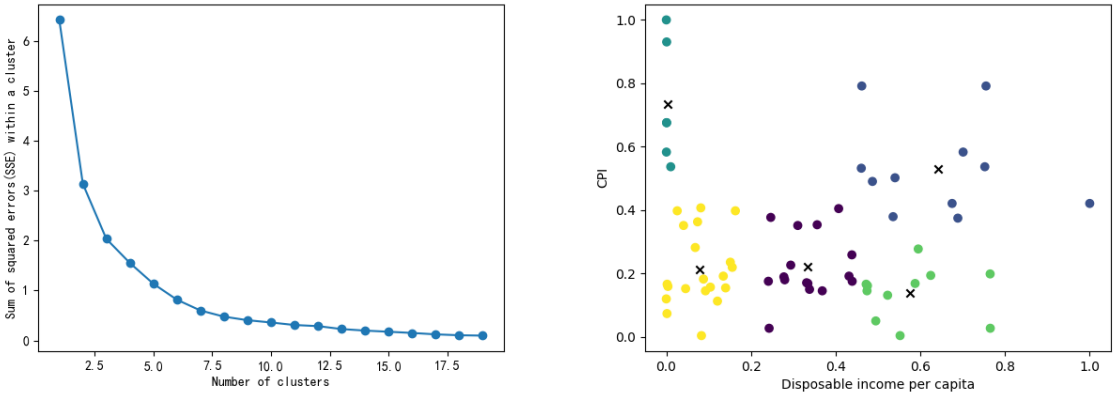
factors affecting the price of sailing boats are tourism income and local disposable income per capita, we clustered them in these two dimensions.

In order to find the optimal parameter k in the Kmeans algorithm, we draw the elbow figure with SSE as an indicator. We chose the elbow inflection point (k=5) as the number of clusters we divided, in order to avoid errors in the results due to overfitting of the model. The results of the final clustering analysis are shown in Fig 7.

4.4.3 Regional effects on sailboat prices

Using this method, we divided all 65 regions into five categories. (Note that here the classification labels have no real meaning and their value does not reflect the economic level.) To visualize the effect of regional economies on sailboat prices, we selected the same manufacturer and variant of the boat and analyzed its prices in different economic regions in cross-section.

As we can see from the Fig8, the Hunter 38 model sells for a differentiated price in five regions and, of course, there are certain fluctuations within the same economic category, but within



(a) Elbow figure based on SSE

(b) Kmeans clustering results

Figure 7: Kmeans clustering method

(a) (b)

Figure 8: Comparison of sailboat prices in different regions

our acceptance range. The Bavaria 39 Cruiser also has distinguishable price fluctuations in two different economic regions. In addition to this, California in the United States is the province with the highest GDP in the world and it tops the list with a high level of economic development. As you can see from the various data, the prices of sailboats in California are significantly higher than in other regions. This reflects the fact that regional effects do have an impact on prices, and is consistent with the conclusions drawn from our model.

5 Analysis of Used Sailboat Prices in Hong Kong Market

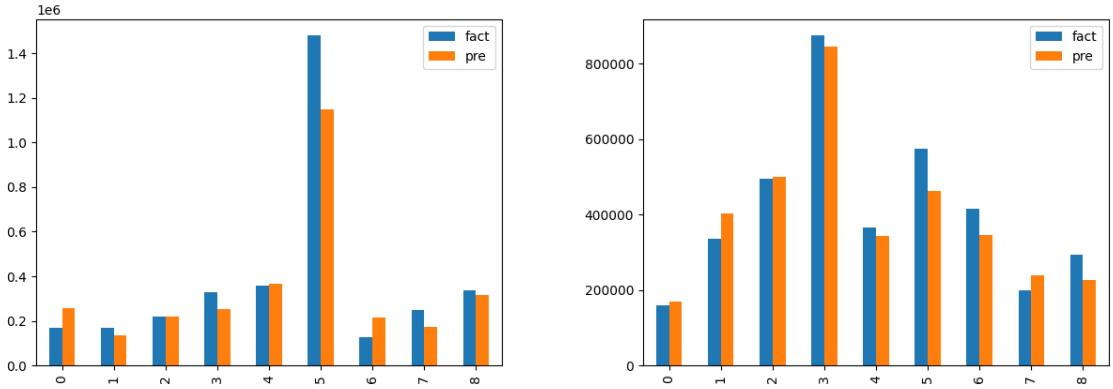
5.1 Application of our model to Hong Kong data

We collected the data of the used boat market in Hong Kong, and selected the part of the provided data that overlap with it in terms of manufacture and variants as a representative subset. These selected data are not too much and shown in the Table 5.We apply the previously trained model to the Hong Kong data to give predicted values, as shown in the Figure 9.It can be seen that our prediction results are closer to the actual values, and the prediction performance is excellent.The MSE are 145057 and 166592 respectively.

Table 5: Hong Kong Used Boat Market Data[1]

Monohulled Sailboats				Catamarans			
Make	Variant	Year	Price	Make	Variant	Year	Price
Bavaria	Cruiser 46	2015	169,000	Fountaine Pajot	Bahia 46	2020	299000
Beneteau	Oceanis 43	2009	168,500	Fountaine Pajot	Athena 38	2004	159000
Beneteau	Oceanis 46	2011	220,000	Fountaine Pajot	Belize 43	2019	399000
Hanse	445	2013	330,000	Fountaine Pajot	Orana 44	2014	335000
Hanse	455	2017	359,000	Fountaine Pajot	Helia 44	2014	495000
Hanse	588	2019	1,480,000	Fountaine Pajot	SABA 50	2016	875000
Jeanneau	Sun Odyssey 42	2012	125000	Lagoon	440	2008	365000
Jeanneau	Sun Odyssey 45	2010	249000	Lagoon	500	2012	575000
Jeanneau	Sun Odyssey 509	2013	339000	Lagoon	420	2008	415000
				Lagoon	380	2012	200000
				Lagoon	380 S2	2011	295000

[1] The data is from https://www.yachtworld.com/boats-for-sale/asia/hong-kong



(a) Forecast of Hong Kong monohull data

(b) Forecast of Hong Kong catamaran data

Figure 9: Comparison of Forecast and True Values of Hong Kong Used Boat Data

5.2 Analysis of regional effects in Hong Kong

Based on the above analysis, we can obtain that for monohull sailing boats, the main factors affecting their prices are disposable income per capita and CPI, while for catamaran sailing boats, we find that the main influencing factors are local tourism income and CPI. from the collected data for Hong Kong, we specify the boat manufacturer and variant and study only its regional economic effects (CPI and disposable income per capita), trying to specifically analyze the regional effects in Hong Kong.

For the monohulls, we show the prices of sailboats with manufacturer Jeanneau and variant Sun Odyssey sold in different regions at the same time (2010). As can be seen from the Table 6, the same boat sells for the same price in Hong Kong as in Washington and is higher than in the remaining other regions. It is not difficult to explain from the analysis of regional economic factors we provide: Hong Kong has a higher CIP than Washington, but its per capita disposable income is lower than that of Washington, which, combined with the size of the influence of CPI and per capita disposable income, leads to similar selling prices for the same type of sailboat in Hong Kong and Washington. However, when comparing Hong Kong to other cities, such as the UK, Hong Kong has a higher CPI and per capita disposable income, and naturally has a higher selling price.

Table 6: Prices of $Jeanneau$ Sun Odyssey45 in different regions

Make	Variant	Country/Region/State	Year	CPI	Income	Price (USD)
Jeanneau	Sun Odyssey 45	France	2010	2.1	2687.3	181036
Jeanneau	Sun Odyssey 45	Spain	2010	2.68	30562	194384.3333
Jeanneau	Sun Odyssey 45	United Kingdom	2010	1.91	42330	193199
Jeanneau	Sun Odyssey 45	Washington	2010	1.58	79533	249000
Jeanneau	Sun Odyssey 45	Hong Kong	2010	2.1	48004	249000

For the catamarans, we selected the sailboat with manufacturer Fountaine Pajot and variant SABA 50. The results of the analysis are similar to those for monohulls, except that the disposable income per capita is replaced by the income from local tourism. Since for catamarans, the CPI is heavily weighted, the selling prices in the two regions end up being essentially equal due to the higher CPI in Hong Kong, despite the fact that tourism revenues in Greece are about five times higher than in Hong Kong.

Table 7: Prices of $Fountaine\ Pajot$ SABA50 in different regions

Make	Variant	Country/Region/State	Year	Tour. Income	CPI	Listing Price
Fountaine Pajot	SABA 50	British Virgin Islands	2016	0.3 bUSD	1.11	595000
Fountaine Pajot	SABA 50	Spain	2016	83.4 bUSD	2.68	922450
Fountaine Pajot	SABA 50	Greece	2016	19.9 bUSD	1.1	825350
Fountaine Pajot	SABA 50	Hong Kong	2016	4.8 bUSD	2.1	875000

6 Further Discussion——Exploratory Data Analysis

6.1 Correlation Analysis

To explore more informative conclusions, we tried data mining with correlation analysis.The Pearson correlation coefficient between two variables is defined as the covariance between the two variables:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y}$$

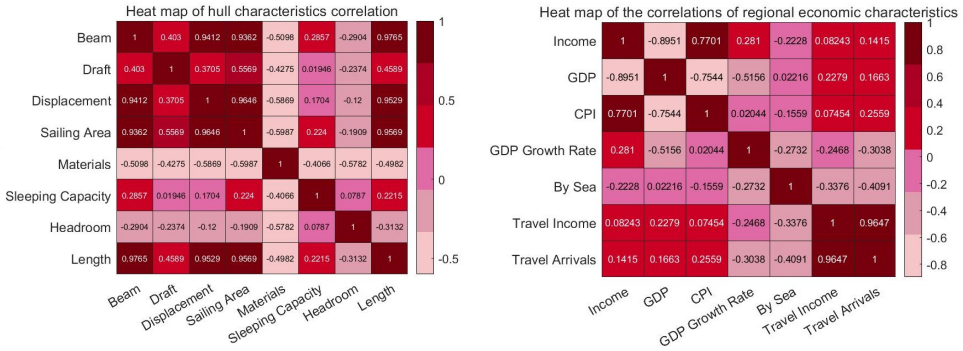
where the correlation coefficients are broadly classified as:

$$|\rho_{X,Y}| = \begin{cases} 0.9 \sim 1.0 & Extremely\ strong\ related \\ 0.6 \sim 0.9 & Strongly\ related \\ 0.4 \sim 0.6 & Moderately\ related \\ 0.2 \sim 0.4 & Weakly\ related \\ 0 \sim 0.2 & Extremely\ Weak\ related \end{cases}$$

We attempted a correlation analysis of the original 16 influencing factors, trying to explore any potential link between them.The final correlation heat map is shown in figure 10.

As can be seen in Figure 10(a), there is a strong correlation between the beam and length, sail area, and displacement of a sailboat. In other words: the beam, length and sail area of a sailboat increase according to a certain proportional range, and their increase together makes the sailboat larger. However, Sleeping capacity and headroom are relatively independent.

As seen in Figure 10(b), there is a strong correlation between PCDI (per capita disposable income) and the GDP and CPI (Consumer Price Index) of the region. It is also clear that there is a strong correlation between the number of tourists and tourism income. The factor of proximity to the sea, on the other hand, has a weaker association with other factors.



(a) (b)

Figure 10: Heat map of correlation of all factors

6.2 Other connections hidden in the data

1. The regional factors affecting the price of monohull sailing boats and catamaran boats are not entirely consistent.

In the SHAP analysis of the various factors affecting monohull and catamaran prices, we found that among the regional factors, monohull prices are most influenced by PCDI (per capita disposable income), while catamaran prices are most influenced by tourism income and CPI(As shown in the figure 5).

To explain this phenomenon, we looked up the application scenarios of monohull sailing boats and catamarans separately. We found that catamarans are preferred for recreational vacations because of their larger space and more comfortable living environment, and thus they are more closely related to tourism income and also to the local price level (CPI). Monohulls, on the other hand, are commonly used for personal training or racing with their smaller berthing space and the ability to sail better in any wind direction. Therefore, it is not difficult to imagine that the price of monohull sailing boats is more closely related to the PCDI.

2. Catamarans are not suitable for high latitudes. If the sailboat is to be driven to higher latitudes, then a monohull will be more suitable than a catamaran.

We counted the number of ships in each region in the data given in the question, including monohulls and catamarans. As shown in the Figure 11 below, monohull sailboats sell well in higher latitudes, such as Wisconsin, Maine and Sweden. In particular, there is a large market in New York and Rhode Island. Catamarans, on the other hand, are largely not for sale in these higher latitudes, and in New York there are only 4 catamarans compared to 112 monohulls.When the latitude is below 30 and into the Caribbean, there are significantly more catamarans than monohulls, specifically, 378 catamarans almost twice as many as monohulls in this latitude region.

We thus went to the relevant information and found that the insulation effect of monohull hull is better than that of catamaran. This explains why catamarans are more often sold in lower latitudes.

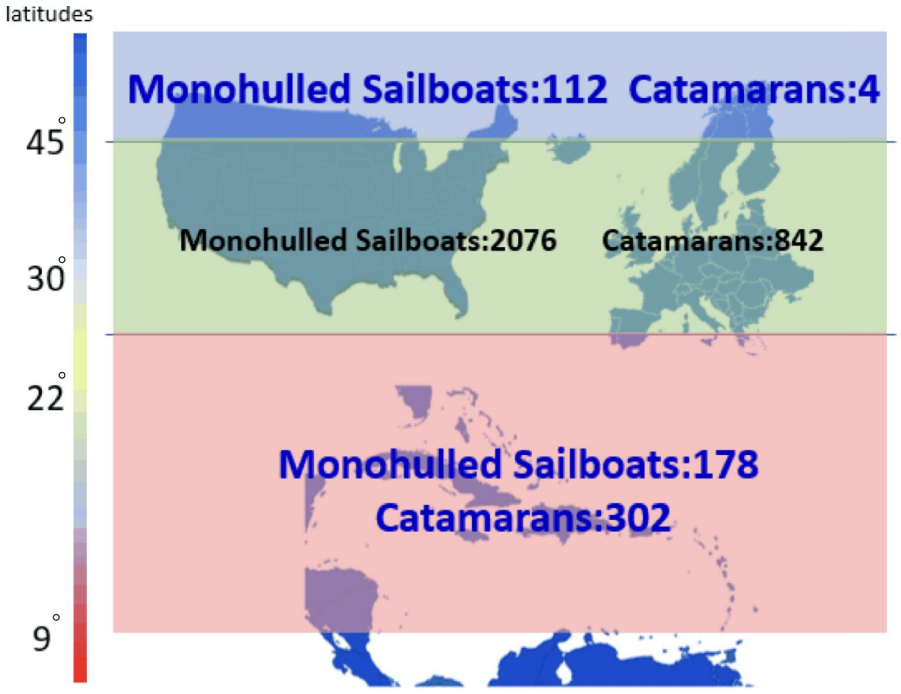


Figure 11: Number of reference prices for used sailboats by region in the given data

7 Strengths and weaknesses

Strengths:

- We use the SHAP model, which gives interpretable importance ranking for black-box models of machine learning and is more appropriate to the requirements of this question;
- LightGBM, an improved algorithm of XGBoost, consumes less computational resources and also yields good prediction results;
- In explaining the effect of regional effects on prices, we cleverly use Kmeans clustering to visualize the economic factors between regions in a more intuitive form, which is more explanatory for the model

Weakness:

- Our multiple linear regression model was poorly trained to fit a monohulled sailboat, and the final prediction accuracy for the Hong Kong data was not satisfactory. We assume that the linear regression model is too simple to fully fit the existing training data and appears to be under-fitted, and that a more complex regression model could be considered for fitting.
- We reasonably explain the effect of regional effects on the price of used sailboats in Hong Kong, but do not give a quantitative level of impact or a specific mathematical model.

8 Report on the pricing of used sailboats

8.1 Purpose of the report

The price of used sailboats in Hong Kong is influenced by a number of factors, which makes it difficult to price them reasonably. In order to solve the pricing problem in the Hong Kong used sailboat market, we developed a used sailboat pricing model (including monohull and catamaran sailboats). By collecting data on the second-hand sailboat market and refining the factors influencing the price of second-hand sailboats to 16 quantitative indicators (including beam, draft, displacement, sail area, hull material, sleeping capacity, headroom, disposable income per capita, GDP, GDP growth rate, CPI, proximity to the sea, tourism revenue, annual tourism arrivals), our model explains well the prices of approximately 3,500 boats advertised for sale in Europe, the Caribbean, and the United States in December 2020.

By collecting data on the used sailboat market in Hong Kong and comparing it to our model's predicted results for used sailboat prices in Hong Kong, we found that the two largely matched, proving that our model is also very useful in the Hong Kong market. Here is our detailed analysis.

8.2 Findings

- We found through our model that the top five most important factors for pricing monohull sailboats are Displacement, Year of Production, Length, PCDI, Beam, while the top five most important factors for catamaran sailboats are Length, Year, CPI, Displacement. Draft, as shown in the figure below.

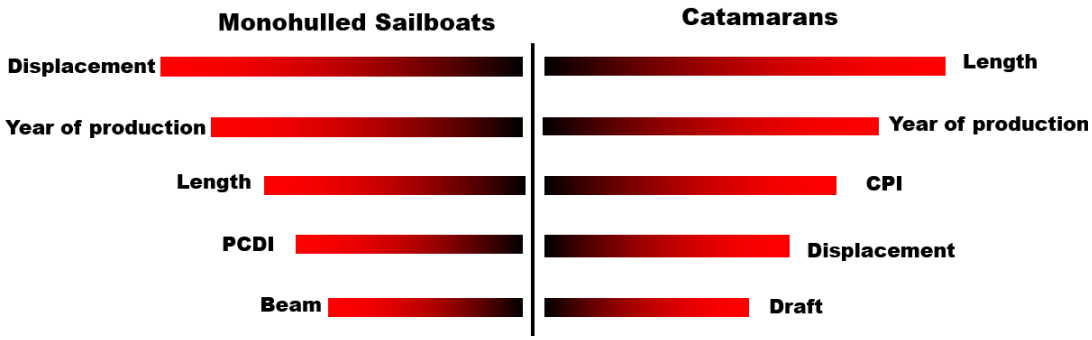


Figure 12: Top five influencing factors for monohulls and catamarans

From this figure we find that four of the top five most important factors for both monohulls and catamarans are factors of the boat itself, including the displacement of the boat, year of production, length, etc. If we consider only regional factors, the price of a monohull is most affected by the PCDI (per capita disposable income), while the price of a catamaran is most affected by the CPI (price level). For Hong Kong, considering the high CPI and PCDI, the prices of both catamarans and monohulls in Hong Kong should be higher than most regions. This is also taken into account in our pricing model.

- Monohull sailing boats are mainly used for personal training and sports competition, their audience is mainly individual users, and the main economic factor affecting their prices is PCDI; while catamaran sailing boats are mainly used for recreation and tourism vacation,

their audience is mainly recreational organizations, and the main economic factors affecting their prices are CPI and local tourism income.

- Taking the latitude factor in different regions into account, we also get the following interesting conclusions: In regions with higher latitudes, the market for catamarans tends to be smaller than the market for monohulls, as shown in the figure 13.Considering the low latitude of the Hong Kong region, which is within the Tropic of Cancer, there is likely to be a greater market for catamarans in the Hong Kong region.This is also taken into account in our pricing model.

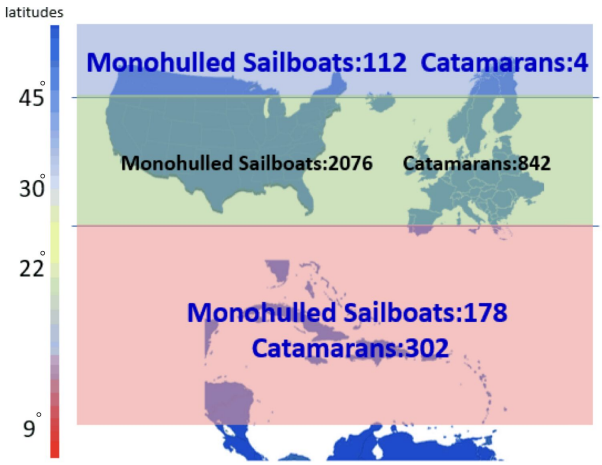


Figure 13: Comparison of sales volume of monohull and catamaran in different latitudes in 2020

8.3 Used Sailboat Pricing App for Hong Kong Region

We have developed a used sailboat pricing app based on our model in order to make it easier for you to quickly price a specific boat. All you need to do is enter the manufacturer, variant and year of manufacture and the app will automatically give you a suggested price for that model in Hong Kong. The APP resource package is available in the appendix.

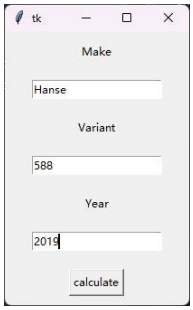


Figure 14: Number of reference prices for used sailboats by region in the given data

References

[1] Lundberg S , Lee S I . A Unified Approach to Interpreting Model Predictions[J]. 2017.

[2] Boehmke B , Greenwell B . Interpretable Machine Learning[M]. 2019.

A appendix

Question(a)

```
import pandas as pd
import numpy as np
import pandas.core.frame
import lightgbm as lgb
import shap
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
import pickle

df = pd.read_excel('data.xlsx').astype(float)
X: pd.core.frame.DataFrame = df.drop(labels='Price', axis=1)
Y = df['Price']

X_normal: pd.core.frame.DataFrame = (X - X.min()) / (X.max() - X.min())
X_corr = X_normal.corr()
X_corr.to_excel('Correlation between x.xlsx')

thresh = 0.9
ind_X = []
X_param = X_corr.columns
for i in range(len(X_param)):
    if np.sum(np.abs(X_corr.values[i + 1:, i]) > thresh) == 0:
        ind_X.append(X_param[i])
X_ind = pd.DataFrame(X_normal[ind_X])
param_list = []
param_num = []

X_train, X_test, Y_train, Y_test = train_test_split(X_ind, Y, test_size=50)

params = {
    'force_col_wise': 'true',
    'objective': 'regression',
    'min_data_in_leaf': 10,
    'seed': 0
}

dtrain = lgb.Dataset(X_train, Y_train)
dtest = lgb.Dataset(X_test, Y_test)

num_rounds = 2000
model = lgb.train(params, dtrain, num_rounds, valid_sets=[dtrain, dtest])
```

```
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X_test)
fig: plt.Figure = plt.figure(1)
shap.summary_plot(shap_values, X_train, plot_type="bar", show=True)
# ------------------------------lightGBM----------------------------------------

remove_list = ['sail_area', 'hull_materials', 'by_sea', 'tra_income', 'Price']
# remove_list = ['Price']
X: pd.core.frame.DataFrame = df.drop(remove_list, axis=1)
Y = df['Price']

X_normal: pd.core.frame.DataFrame = (X - X.min()) / (X.max() - X.min())
X_corr = X_normal.corr()
thresh = 0.9
ind_X = []
X_param = X_corr.columns
for i in range(len(X_param)):
    if np.sum(np.abs(X_corr.values[i + 1:, i]) > thresh) == 0:
        ind_X.append(X_param[i])
X_ind = pd.DataFrame(X_normal[ind_X])
X_train, X_test, Y_train, Y_test = train_test_split(X_ind, Y, test_size=50)
params = {
    'force_col_wise': 'true',
    'objective': 'regression',
    'min_data_in_leaf': 10,
    'seed': 0
}

dtrain = lgb.Dataset(X_train, Y_train)
dtest = lgb.Dataset(X_test, Y_test)
num_rounds = 2000
model = lgb.train(params, dtrain, num_rounds, valid_sets=[dtrain, dtest])
Y_pre = model.predict(X_test)
pre = pd.DataFrame()
pre['fact'] = Y_test
pre['pre'] = Y_pre
pre.plot.bar()

# ----------------------------------
df_hongkong = pd.read_excel('data_hongkong.xlsx')
X_hongkong = df_hongkong.drop(remove_list, axis=1)
Y_hongkong = df_hongkong['Price']
X_normal_hongkong: pd.core.frame.DataFrame = (X_hongkong - X.min()) / (X.max() - X.min())
Y_pre_hongkong = model.predict(X_normal_hongkong)
pre = pd.DataFrame()
pre['fact'] = Y_hongkong
pre['pre'] = Y_pre_hongkong
pre.plot.bar()
print(np.sqrt(np.sum((Y_hongkong - Y_pre_hongkong) * (Y_hongkong - Y_pre_hongkong))))
quit()

# --------
```

```
model = LinearRegression()
model.fit(X_train, Y_train)
Y_pre = model.predict(X_test)
pre = pd.DataFrame()
pre['fact'] = Y_test
pre['pre'] = Y_pre
pre.plot.bar()
print('Parameter weights')
print(model.coef_)
print('Model intercept')
print(model.intercept_)

plt.show()
model = sm.OLS(Y_train, X_train)
result = model.fit()
print(result.summary())
```

Question(b)

```
import openpyxl
import math
import copy
import matplotlib.pyplot as plt
import numpy as np
from sklearn.cluster import KMeans
excel = openpyxl.load_workbook('C:\\Users\\Zoeric\\Desktop\\cluster0.0.xlsx')
table = excel['Sheet1']
income=[];  gdp=[]; cpi=[]; gdp_increase=[];by_sea=[];
tra_income=[];tra_num=[];ratio=[]
for row in range(2, 66):
    income.append(table.cell(row=row, column=3).value)
    # gdp.append(table.cell(row=row, column=4).value)
    cpi.append(table.cell(row=row, column=5).value)
X=[]
for i in range(len(income)):
    X.append([income[i],cpi[i]])
    # X.append([income[i],gdp[i],cpi[i],gdp_increase[i],by_sea[i],tra_income[i],tra_num[i]]
X = np.array(X)
# distortions = []
# for i in range(1,20):
#     kmModel = KMeans(n_clusters=i)
#     kmModel.fit(X)
#     distortions.append(kmModel.inertia_)
# plt.plot(range(1, 20), distortions, marker="o")
# plt.rcParams['font.sans-serif'] = ['SimHei']
# plt.rcParams['axes.unicode_minus'] = False
# plt.xlabel("Number of clusters")
# plt.ylabel("Sum of squared errors(SSE) within a cluster")
# plt.show()
module = KMeans(n_clusters=5)
y_pred =module.fit_predict(X)
cen = module.cluster_centers_
print(module.cluster_centers_)
plt.scatter(X[:, 0], X[:, 1], c=y_pred)
```

```
plt.scatter(cen[:, 0], cen[:, 1], c='black',marker='x')
plt.xlabel("Disposable income per capita")
plt.ylabel("CPI")
plt.show()
print(y_pred)
```

APP For Broker

```
import tkinter as tk
import pickle
import lightgbm

# print(pickle.load(open('dict_boat', 'rb')))
# quit()
root = tk.Tk()
label1 = tk.Label(root, text='Make')
label1.pack(padx=10, pady=10)
entry1 = tk.Entry(root)
entry1.pack(padx=10, pady=10)
label2 = tk.Label(root, text='Variant')
label2.pack(padx=20, pady=10)
entry2 = tk.Entry(root)
entry2.pack(padx=20, pady=10)
label3 = tk.Label(root, text='Year')
label3.pack(padx=30, pady=10)
entry3 = tk.Entry(root)
entry3.pack(padx=30, pady=10)
def get_name():
    make = entry1.get()
    variant = entry2.get()
    year = entry3.get()
    with open('dict_boat', 'rb') as f:
        dict_boat: dict = pickle.load(f)
    value = dict_boat[make + variant]
    with open('model', 'rb') as f:
        model: lightgbm.basic.Booster = pickle.load(f)
    print('The estimated price is:,
          model.predict([[value[0], value[1], value[2], value[3], value[4], value[5],
          value[6], value[7], 48004, 366.1, 1.4, -0.061, 1, 4.8, 4157, year]]))
button = tk.Button(root, text='yes', command=get_name)
button.pack(padx=10, pady=10)
root.mainloop()
```