Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2330671

Price Prediction of Used Sailboats Based on Ridge Regression with CatBoost Algorithm

Sailboats vary in value with age, changing market conditions, The aim of this report is to build a **Price Prediction Model** to help Hong Kong (SAR) sailboat brokers open up the used sailboat market. We are expected to conduct the most accurate evaluation of the transaction price of second-hand sailboats under the constraints of various objective conditions, and provide convenience for sailboat brokers and customers.

Three models are established: Model I: Ridge Regression Price Forecasting Model Based on CatBoost; Model II: Regional Impact on Prices Model; Model III: Hong Kong Shipboat Price impact model.

For Model I, the information of used sailboats from 2005 to 2019 is firstly collected, the data are classified, qualitative analysis and quantitative analysis are carried out respectively, and several indicators with greater influence are obtained. Then carry out ridge regression analysis and prediction, and use python to find out that the fitting coefficient of ridge regression is **0.76**, which proves the effectiveness of using **ridge regression model** to predict prices. Next, the **CatBoost Algorithm** is used for 10,000 machine learning trainings, and the fitting coefficient is increased to **0.9**. Finally, the historical datas of sailboat are used for fitting, and the error is found to be within **2%**, witch lays the foundation for predicting prices.

For Model II, firstly, according to the ridge regression model, we use **per capita GDP** to measure the economic level of the region, so as to distinguish the differences between regions. Then we randomly **disturb the per capita GDP of each country**, and increase the per capita GDP by 1% and 10% respectively. The result is: when the regional GDP increased by 1%, the average price of catamaran changed by **30227.8114**, and when the GDP increased by 10%, the average price of catamaran changed by **33137.7461**. The specific significance is in **6.2.3**.

For Model III, a subset of 16 metadata-rich sailboats was selected, combined with Hong Kong economic data, and ridge regression was performed on monohulls and catamarans to predict their prices in Hong Kong and compare with actual values; combined Hong Kong's economic data finally draw the conclusion that catamaran sailing ships are more sensitive to regional effects, which is quite different from the predicted results, as shown in **Table 12**.

In addition, this report also briefly analyzes the historical price of sailboats, its preservation rate, and the general situation of the sailboat market. It is found that the listing price of sailboats will **rise first and then fall with the development of time**. Realizing the maximization of the interests of buyers and sellers provides inspiration.

Eventually, a robustness and sensitivity analysis of the model is performed. It is found that the cross-validation of the model, **the data sets are roughly distributed near a line**, reflecting the robustness of the model. As for the factors that affect the model, it is found that the year and displacement, the increase of these two factors will have a greater impact on the results, laying the foundation for further in-depth research on this question.

Keywords: Ridge Regression, CatBoost, ANOVA, Spearman Correlation,per capita **GDP**

Contents

1 Introduction

1.1 Problem Background

Sailboats vary in value as they age and as market conditions change. The file contains about 3500 data36 to 56 foot sailboats advertised for sale in Europe, the Caribbean and the US December 2020. A boating enthusiast provided these data to COMAP. like most real world data set, it may have missing data or other issues that require some data cleaning before analysis.We collect data on the overall price of monohull sailboats and catamarans from 2005 to 2019, and the following **Figure 1** can be obtained after merging.

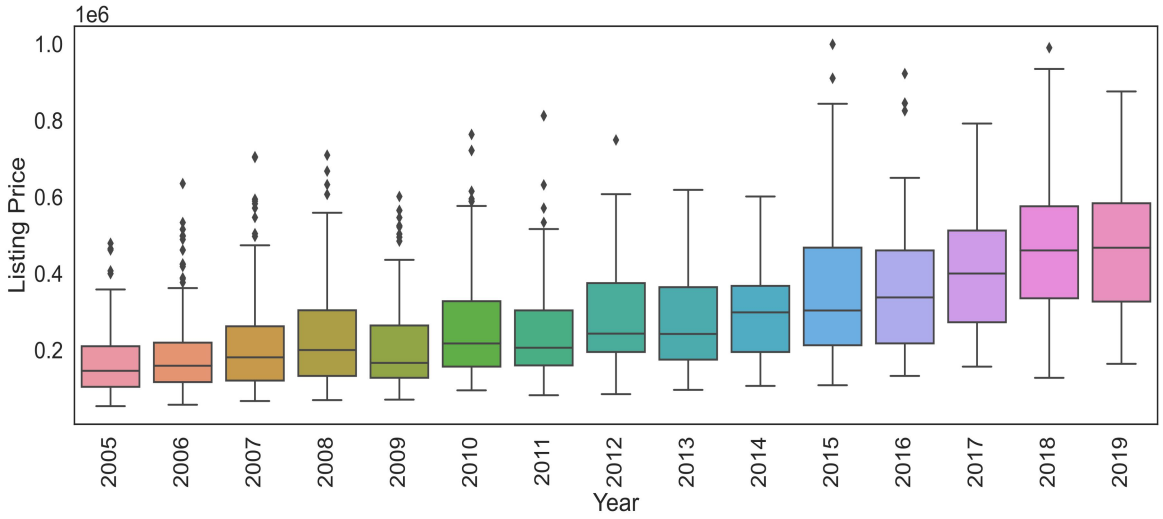


Figure 1: 2005-2019 Sailboat Price General Situation

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- Develop a mathematical model that explains the listing price of each of the sailboats in the provided spreadsheet and discuss the accuracy of the estimates for the price of each sailboat.
- Explain the effect of region on list price, illustrating the practical and statistical significance of any noted region effects.
- Discuss how modeling of a given geographic region can work in the Hong Kong (SAR) market and simulate the regional impact of Hong Kong (SAR) on the price of each sailboat in your group.
- Identify and discuss any other interesting and valuable inferences or conclusions drawn from the data.

1.3 Literature Review

This problem is mainly about developing a model to predict the price of a used sailboat

based on various parametric attributes of a used sailboat. It is necessary to combine the parameters of the sailboat itself and the economic situation of the place of sale. **Regression predictive analysis using machine learning methods** has become a hot topic in recent years. Multi-factor **analysis of variance, correlation coefficient method, feature selection method based on Lasso regression**, and **feature selection method based on decision tree** are the main solution methods of machine learning methods for establishing second-hand ship price prediction models. Solving this model has important practical significance for the sales of second-hand monohulls and catamarans in Hong Kong (SAR). This section mainly discusses the proposed model.

- First of all, since it is necessary to develop a mathematical model to predict the price of the used sailboat, in order to first determine the influencing factors required by the model, many authors divide the data into qualitative data and quantitative data, and use **analysis of variance and Spearman correlation coefficient analysis**[1] respectively .
- Secondly, In terms of prediction methods, commonly used models include linear regression model and **ridge regression model**[2], commonly used algorithms include XGBoost algorithm and **CatBoost algorithm**[3].
- Finally, explain the influence of regions on prices. The **per capita GDP** of a region can objectively reflect the local economic level[4], and the per capita GDP can be used as an important factor affecting the region to dig out the practical and statistical significance of the regional effect.
- The strengths and weaknesses of the **Ridge Regression Forecasting Model Based on CatBoost Algorithm** can be visually presented and is shown in Figure 2:

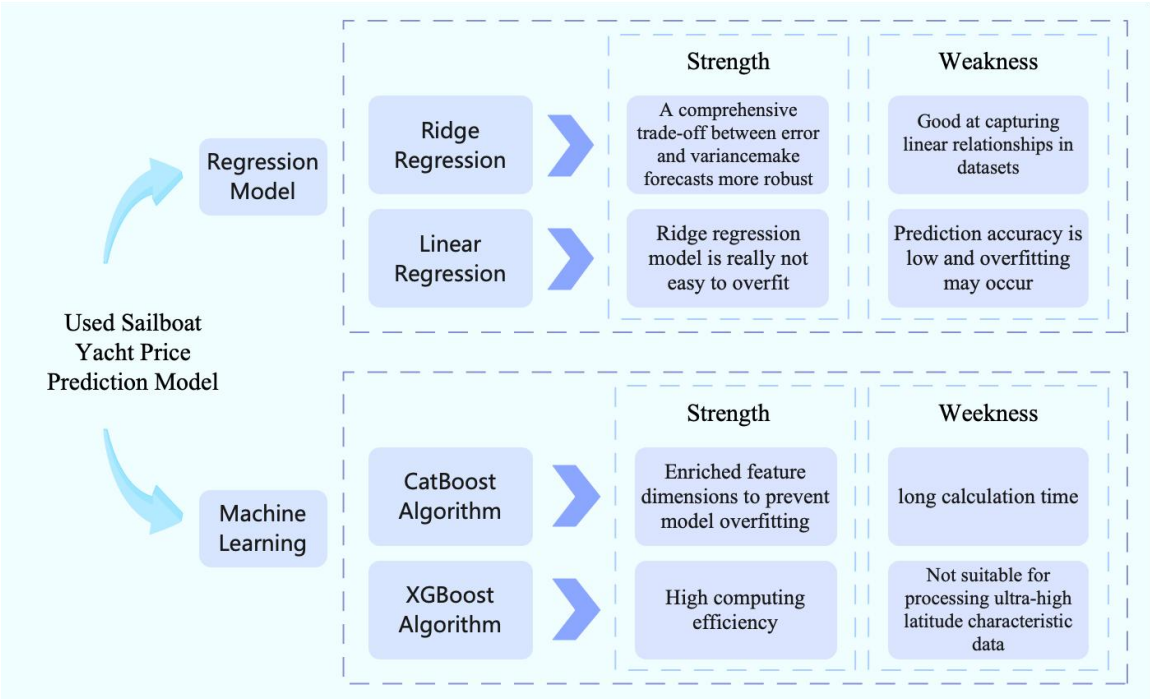


Figure 2: Literature Review

1.4 Our work

This question requires us to predict the pricing of second-hand ships based on their own attribute factors and the sales market environment, analyze regional effects, and the role of actual geographic regions in the Hong Kong (SAR) market. Our work mainly includes the following aspects:

- ✧ Firstly based on the price data given in the question and looking for other sailboat attribute data and regional economic data on relevant websites, focus on correlation analysis of length, beam, draft, and establish a prediction model for the transaction price of second-hand sailboats;
- ✧ Then use the regional per capita GDP to measure the regional effect, and find the practical and statistical significance of the regional effect by changing the size of the per capita GDP;
- ✧ Last but not least, Digging for other interesting conclusions, such as the retention rate of sailboats and so on;
- ✧ In summary, the whole modeling process can be shown as follows:

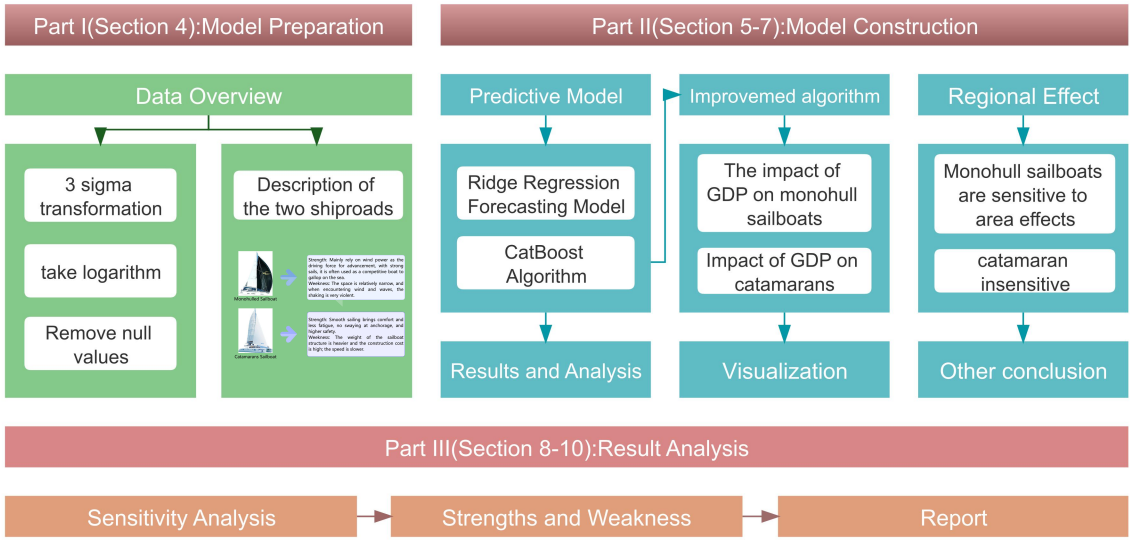


Figure 3: Model Overview

2 Assumptions and Justifications

To simplify the problem, we make the following basic assumptions, each of which is properly justified.

➢ **Assumption 1: The changes of sailboat price is predictable.**

Justification: Although the price of each sailboat may not completely follow our given CatBoost Ridge regression model, according to the analysis of a large pair of correlation indicators, the correlation coefficient of our model is extremely high, so we can predict the

used shipboat price and price changes over time .

- **Assumption 2: Consider only the factors selected in 5.1 as factors affecting price**

  Justification: The price of sailing boats is most affected by the factors selected in 5.1, while other factors have little influence. In fact, there are many factors that affect each other, but in order to simplify the model, we ignore the interaction of these factors.

- **Assumption 3: The region was free from financial crisis impacts, trade environment changes and breakthrough in sailboat technology during the study period.**

  Justification: Because the model takes into account the impact of regional changes on the price of second-hand sailboats, and then compares and analyzes the impact of economic benefits in different regions and different time periods on the price of sailboats. Technological breakthroughs of sailing boats will also have a greater impact on prices, and such comparisons are meaningful only when external conditions are consistent.

- **Assumption 4: Assume the research data is accurate.**

  Justification: We assume that the specific parameters of sailboats, fishing data and economic data of various regions collected on the website do not show obvious measurement deviations and are considered false, so a more reasonable quantitative model can be established on this basis.

3 Notations

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

Table 1: Notations used in this paper

Symbol	Description
$x_{i,j}$	The i-th sailboat j-th factor
y_i	Listing Price of the i sailboat
k_i	Regression coefficients
$\frac{\partial y_i}{\partial x_i}$	The partial derivative of the ith sailboat with respect to LIsting Price
p	statistical significance
$r_{i,j}$	The correlation coefficient of factor i to factor j
R^2	fit coefficient

4 Model Preparation

4.1 Data Overview

The data provided to us by this question is insufficient, so we need to consider other data. Collected during model building. Through the analysis of the problem, we need to collect information about other parameters of the sailing ship itself, such as draft, beam, displacement, headroom and other information, as well as economic data of each region, such as GDP per capita. Due to the large amount of data, it is not convenient to list them one by one, so it is a good way to visualize the data.

4.1.1 Data Collection

The data we use mainly includes the displacement and data of sailboats, the economic data of various countries and regions, and the distribution data of the main sales places of sailboats. The data sources are summarized in Table 2.

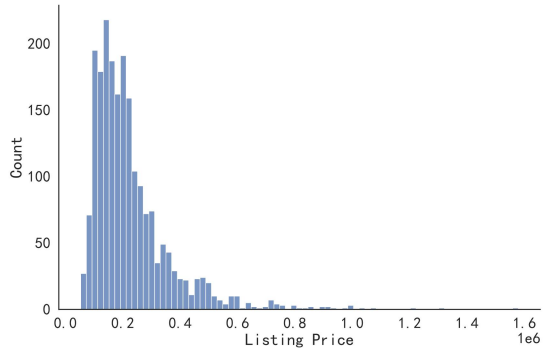
Table 2: Data source collation

Database Names	Database Websites Data	Type
Sailboat Data	http://www.sailboatdata.com/	Database
Boats	https://www.boats.com/	Trading platform
The World Bank	https://data.worldbank.org/	Goverment
Google Map	https://www.google.com/maps/	Geography
Google Scholar	https://scholar.google.com/	Academic paper

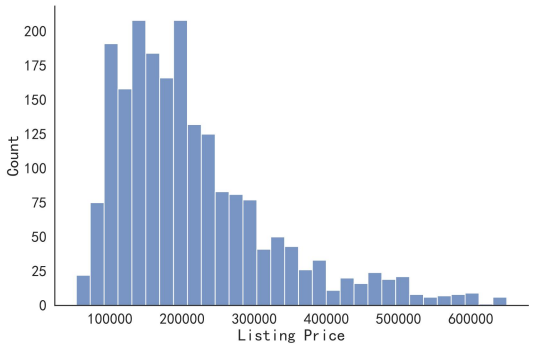
4.1.2 Data Screening

- Delete the data that has been collected but some of the influencing factors are missing.
- Perform 3sigma processing on the value of Listing Price to deal with outliers. It can be seen in Table 4 and Table 5 for the processed results.
- Label codes to derive qualitative data: geographical area, country, name (make and model).
- Scale the features.
- ln processing: listingPrice, averageCargoThroughput, gdpPerCapita, displacement.
- Divided by 10 processing: sailAria.
- Subtract year_min for each year.

After 3sigma transformation, the data of monohull and catamaran are shown in the Figure 4 and Figure5 below:

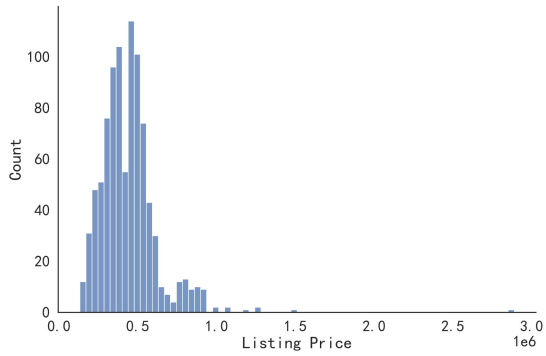


(a) The most likely location of fishes

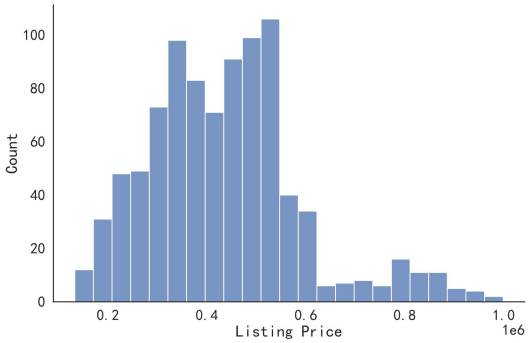


(b) The location of fishes (std)

Figure 4: The data of monohull sailboat after 3sigma transformation



(a) Catamaran Original Listing Price



(b) Catamaran Price after 3sigma

Figure 5: The data of the catamaran after 3sigma transformation

4.2 Description of two kinds of sailboat

The two types of sailing ships have different types, but they have similar attributes, but they have different equipment functions, and they have different advantages. A brief schematic diagram is shown in Figure 5:

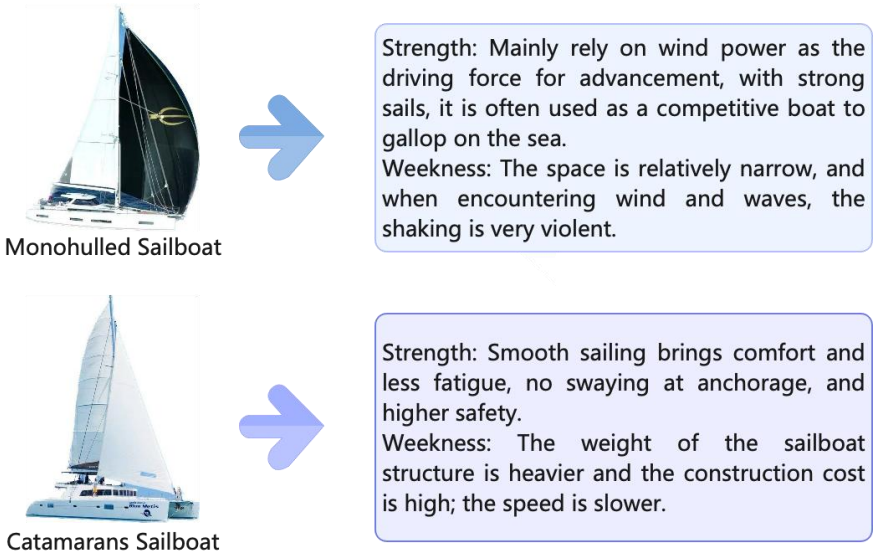


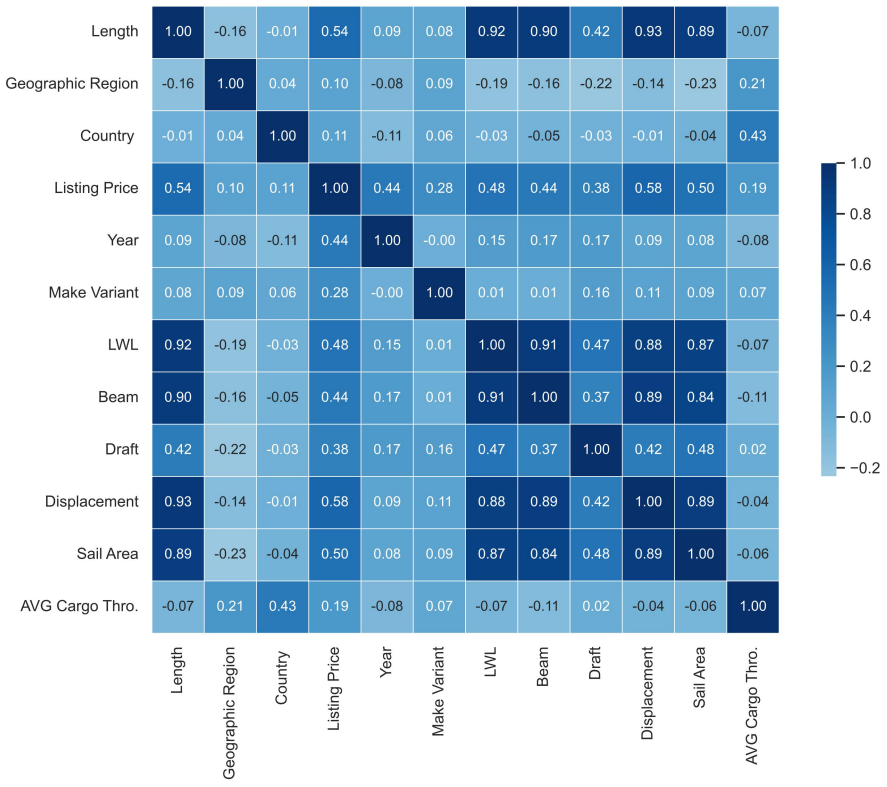
Figure 5: Two Types of sailing ships

5 Model I: Sailboat Price Prediction Model

The price of a sailboat is determined by a variety of factors. By dividing the attributes of sailboats into two types: discrete and continuous, and performing **variance analysis** and **Spearman correlation coefficient analysis** respectively, it is found that the main influencing factors are brand, length, country, year, beam, draft, displacement, and sailing area. Both local per capita GDP and average freight volumes have a significant impact on sailboats. Shipping prices. Therefore, for such a complex model, a regression method is used to solve it. Formally, the CatBoost algorithm can simultaneously consider the spatiotemporal correlation of each variable, and maximize data mining without introducing exogenous factors. Therefore, the forecasts of the least-squares-based ridge regression model are good approximations for price simulations.

5.1 Normalization of Influencing Factors

Divide all influencing factors into two categories: discrete categorical data and continuous quantitative data, and perform variance analysis and Spearman correlation coefficient analysis respectively. In order to visualize the results, we use python to get the following Figure7:



Divide Make and Country/Region/State into qualitative data, and divide Length, Year, cargo throughput, regional per capita GDP, beam, displacement, rigging, sail area, hull materials, and engine hours into quantitative data.

Step1. Perform one-way analysis of variance for each discrete type of data, and select

data types with significant differences.

Step2. Calculate the Pearson correlation coefficient for all continuous quantitative data and perform correlation analysis to select the data type with significant correlation.

Step3. Integrate the above two data types, and remove collinear VIF > 10.

Final selection: Make, Country, Year, cargo throughput, regional per capita GDP, beam, displacement, sail area as the influencing factors of the research.

5.2 Ridge Regression Price Forecasting Model

5.2.1 Ridge Regression Model Establishment

Using the ridge parameter, we can construct the following ridge regression model.

$$\ln(y_i) = \sum_{i=1}^{2} k_i \ln(x_i) + \sum_{i=3}^{11} k_i x_i \tag{2}$$

Where m is the number of samples; n is the number of features. The cost function of ridge regression is a convex function, so the global optimal solution can be obtained by using the gradient equal to 0.

Commonly used regression models to solve multiple regression problems mainly include ridge regression, principal component regression and partial least squares regression. The Generation Function is as follows:

$$J(k_i) = \frac{1}{2m}\left[\sum_{j=1}^{m}(k_i x_j - y_j)^2 + \lambda \sum_{i=1}^{n} k_i^2\right] \tag{1}$$

Where Set step size $step = 0.01$, Ridge regression coefficient $\lambda = 10$, $x_i = 10$ and the number of iterations t=10000;

The specific process of ridge regression algorithm is as follows:

Algorithm 1: Ridge Regression Forecasting Algorithm

Input: $step = 0.01, \lambda = 10, x_i = 10, t = 10000$

Output: k_i

for i = 1 to t **do**

Constructing the Standard Equation for Linear Regression: $\theta = (X^T X)^{-1} X^T y$

Solve using gradient descent method

Then go down according to the gradient corresponding to the initial point $x_i - step\dfrac{\partial y(x_i)}{\partial x_i}$

Then judge whether to meet the cost $J(k_i)$

Update the initial point k_i size.

end

5.2.2 Calaculation Results

Through python, we analyze the ridge regression and find that r2 changes with the change of the ridge parameter, and finally we can get the following Figure 6.

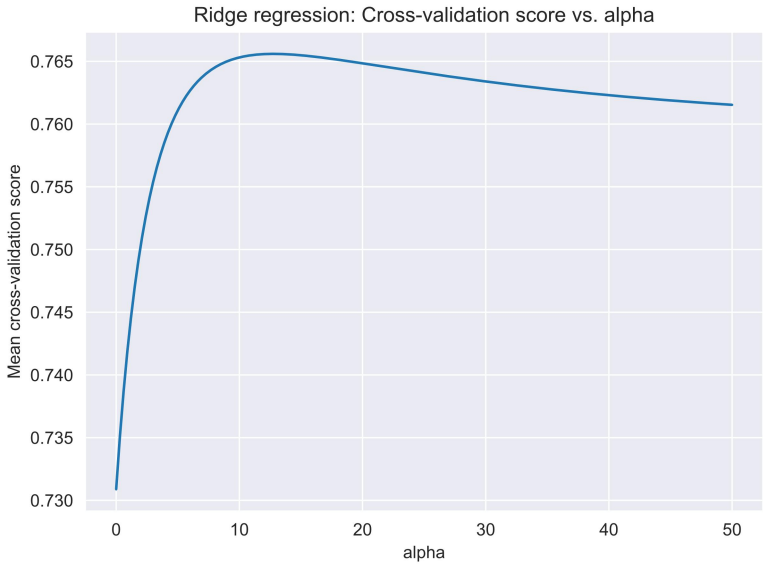


Figure 7: Ridge parameter optimization

Ridge regression coefficients are shown in the Table 3.

Table 3: Ridge Regression Coefficient

Parameter name	Parameter Size	Parameter name	Parameter Size
Length	0.05303	width	0.03009
Country	0.00455	Draft	0.00523
Year	0.5123	Displacement	0.08109
Name	0.00089	per capita GDP	0.00008
Waterline	-0.02758	Sail Area	0.00208
Type	0.53009		

5.3 Price Prediction Model based on CatBoost Algorithm

5.3.1 Parameter Estimation

In this paper, Python software is used to predict and analyze the various data of the sailing ship using the **CatBoost Algorithm**. Firstly, the cross-validation and grid search method are used to calculate the optimal parameters of the model (**seen in 5.3.2**), so that the model can obtain the optimal classification accuracy parameters, and then the model is trained based on the CatBoost Algorithm.The specific algorithm flow is as follows:

Algorithm 2: Catbosst Machine Learning Algorithms

Input: $M_{i,j}, \theta, \frac{\partial L}{\partial s}$

Output: $M_{i',j}$

for i = 1 to 10^5 **do**

Step1. We trained a model $M_{i,j}$, where $M_{i,j}(\theta)$ represents the model learned from the first j samples in the sequence σ_i for the θth sample Prediction.

Step2. In each iteration t, the algorithm samples a sequence σ_i from $\sigma_i \cdots, \sigma_s$, and builds the learning tree T of step t based on it.

Step3. Calculate the corresponding gradient $grad_{i,\sigma(r)-1}(r) = \frac{\partial L(y_r, s)}{\partial s}\Big|_{s=M_{ij}(\theta)}$ based on $M_{i,j}(\theta)$.

Step4. Using cosine similarity to approximate the gradient G, for each sample r, take the gradient $grad_{i,\sigma(r)-1}$.

Step5. In the process of evaluating candidate split nodes, the leaf node value $\delta(r)$ of the rth sample is obtained by averaging the gradient values of the first p samples of all samples of $leaf_i(r)$ that belong to the same leaf as r.

Step6. When the tree structure of the t-th iteration is determined, it can be used to upgrade all models $M_{i,j}$.

end

5.3.2 Calaculation Results

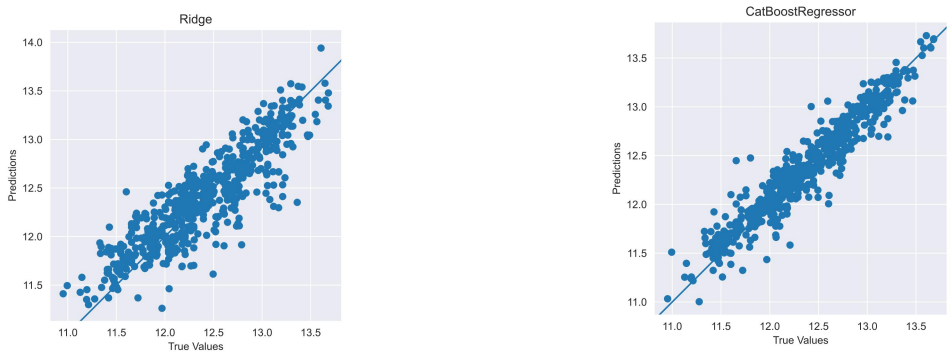
We can get CatBoost Parameters as shown in the Table 4 below:

Table 4: CatBoost Parameters

learning_rate	iterations	depth
0.0469	1000	4
l2_leaf_reg	border_count	early_stopping_rounds
1	32	20
random_strength	bagging_temperature	subsample
0.1	0.1	0.5

5.4 The Results

In order to intuitively display the accuracy of Ridge regression and CatBoost algorithm, we use python to draw the following Figure 8.



(a) CarBoost algorithm (b) Ridge regression

Figure 8: Fit Comparison between Ridge regression and CarBoost algorithm

The specific parameters are as following Table 6:

Table 6: CatBoost regression model accuracy results

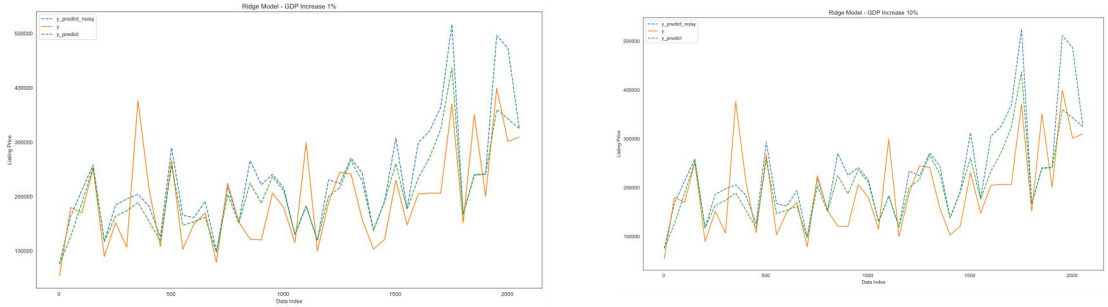
[CatBoostRegressor]CatBoost Regression model r2_score:	0.916828164
[CatBoostRegressor]CatBoost Regression model mse:	0.026835207
[CatBoostRegressor]CatBoost Regression model mae:	0.120981616
[CatBoostRegressor]CatBoost Regression model mape:	0.009822888

It can be seen from the table that the fitting coefficient (r2) is very large, which reflects the good fitting effect, and the low standard deviation reflects the good stability of the model.

6 Model II: Regional Impact on Prices Model

6.1 Impact of per capita GDP on Price

We use the established ridge regression model to analyze regional differences, which are generally reflected by per capita GDP. Therefore, we continuously change the size of per capita GDP while other factors remain unchanged, and check the changes in Listing Price to measure the impact of regions on pricing. We will increase the per capita GDP of each country by 1% and 10%, respectively. The corresponding prices of various single and double sailboats are shown in the following Figure 9 and Figure 10:



(a) GDP change 1 percent (b) GDP change 10 percent

Figure 9: The impact of per capita GDP changes on monohull sailboats

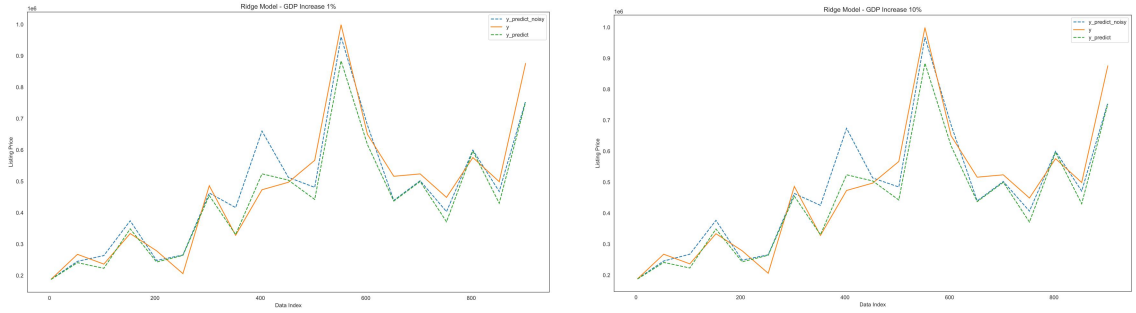
We made the data obtained from the test into the following Table 7:

Table 7: The impact of per capita GDP changes on monohull sailboats

monohull sailboat Model	Average price change	Average price change rate
1%	2261.6387	3.15%
10%	25238.0053	12.26%

From the table, it can be seen that when the regional per capita GDP increases by 1%, the average price of monohull ships has changed by 22961.6387. When the per capita GDP increases by 10%, the average price of monohull ships has changed by 25238.0053, indicating that regional effects have a significant impact on monohull ships.

The results for the catamaran are shown in the Figure 10 below:



(a) per capita GDP change 1 percent (b) per capita GDP change 10 percent

Figure 10: The impact of per capita GDP changes on catamaran

We made the data obtained from the catamaran test into the following Table 8:

Table 8: The impact of per capita GDP changes on catamaran

monohull sailboat Model	Average price change	Average price change rate
1%	3227.8114	2.8753%
10%	33137.7461	11.5371%

Overall, it can be seen that the regional effects on single and twin sailboats are not consistent. When per capita GDP increases to a certain extent, the average price growth rate of twin sailboats on per capita GDP (3227.8114, 33137.7461) is higher than that of single sailboats (2261.6387, 25238.0053), but the average price change rate (2.8753%, 11.5371%) is lower than that of single sailboats (3.1523%, 12.2580%).

As shown in the figure, it can be seen from the figure that when per capita GDP and national regional random disturbances occur, the price also fluctuates differently, but the region has an impact on the price.

6.2 The Impact of Regional Effects

6.2.1 Normality Test Results

To discuss whether regional effects are consistent across different sailboats, one-way ANOVA is used to determine whether regional effects have a significant impact on prices. The normality test results is in Table 9.

Table 9: Normality Test Results

Variable name	sample size	median	average value	standard deviation	skewness	kurtosis	S-W inspection	K-S test

Predicted price	2059	0.251	0.251	0.029	0.076	-0.331	0.998(0.003***)	0.02(0.356)

Analysis item	Inter group difference	Total dispersion	Eta(Partial η^2)	Cohen's f
Predicted price	0.349	1.751	0.199	0.499

It can be seen from the table that Eta is greater than 1.8 and Cohen's value is greater than 0.4, and the significance P value is 0.000***, which is significant horizontally and has a significant impact on the Listing Price (USD). There is a Main effect. Therefore, there is a significant difference between regional per capita GDP and the Listing Price, indicating that there is a significant difference between the country and the predicted price.

6.2.2 Practical meaning

Its practical significance is expressed as that in high economic regions, regional effects have a significant positive impact on single and double sailing ships, while in low economic development regions, regional effects have a significant negative impact on single and double sailing ships. That is, when the regional economic per capita GDP increases, the price of single and double sailing ships in the region also increases. When the regional economic per capita GDP decreases, the price of single and double sailing ships in the region also decreases.

6.2.3 Statistical Significance

Subsequently, analysis of variance was conducted on regional per capita GDP and Listing Price for single and double sailboats, and the results are as following Table 10 :

Table 10: Analysis of variance statistical chart

Variable name	sample size	average value	standard deviation	skewness	kurtosis	S-W test
GDP	2970	1002.394	1040.103	0.881	-0.376	-------
Predicted price	2970	0.269	0.039	0.17	-0.616	-------
GDP Paired Forecast Price	2970	1002.125	1040.097	0.881	-0.376	0.835(0.000***)

Paired Variables	Mean ± Standard Deviation			t	df	P	Cohen's d
	2970	1002.394	1040.103	0.881	-0.376	-------	
GDP Paired Forecast Price	1002.394±1040.103	0.269±0.039	1002.125±1040.064	52.508	2969	0.000***	0.963

It can be seen from the table that P is less than 0.05 and Cohen's value is greater than 0.8, with a significant P value of 0.000***, which is significant horizontally and has a significant impact on the Listing Price (USD). There is a Main effect. Therefore, there is a significant difference between regional GDP and the Listing Price.

Therefore, its statistical significance is expressed as a significant difference between Country/Region/State and Listing Price for different regions, and the closer the region is to the sea, the greater its significance.

7 Model III: Hong Kong Shipboat Price impact model

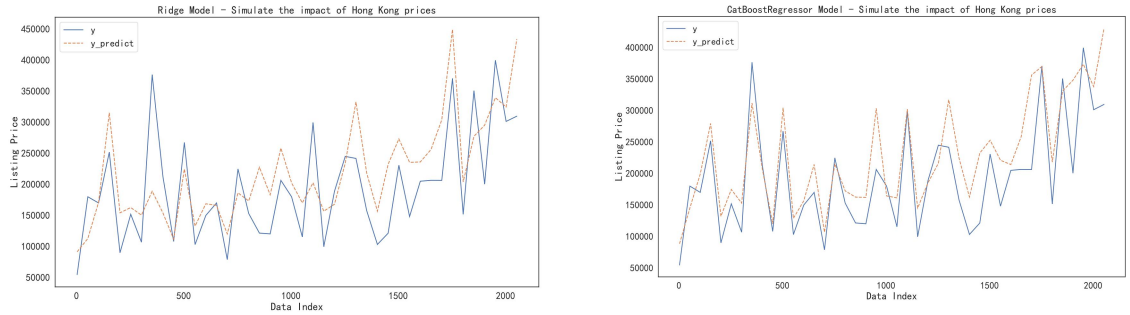
7.1 The difference between Monohulls and Catamarans

In the original data, a subset of 16 types of sailboats with abundant quantities is selected, as shown in the following Table:

Table 11: 16 Types of sailboats

Make	region	Variant	Make	region	Variant
Jeanneau	Asian	Prestige 36	Numarine	Asian	Fly 55
Jeanneau	Asian	42	Beneteau	Asian	Oceanis 38
X-Yachts	Asian	X-55	Beneteau	Asian	MC5
Beneteau	Asian	Monte Carlo 47 Fly	Nautor	Asian	Swan Swan 60
Beneteau	Asian	Sense 43	Bavaria	Asian	R40
Jboats	Asian	111	Nautique	Asian	G23
Bavaria	Asian	Sport 39	Beneteau	Asian	Oceanis 51.1
Hershine	Asian	Hershine	Nautor	Asian	Swan 54

Useing the ridge regression model to predict the price of monohull sailboats and catamarans in Hong Kong, and compare them with their real prices, as shown in the Table 11 below:



(a) Ridge Regression (b) CatBoost Algorithm

Figure 11: Accuracy of Monohulls and Catamarans Price Forecast

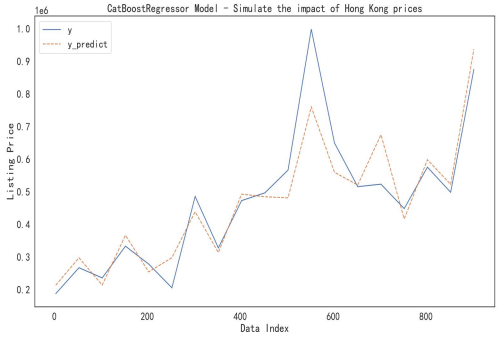
Table 12: Error Analysis Table

	Mean square deviation	Sum of squared residuals
monohull sailboat	1361.25	2562.36
Catamaran sailboat	13712.95	22648.34

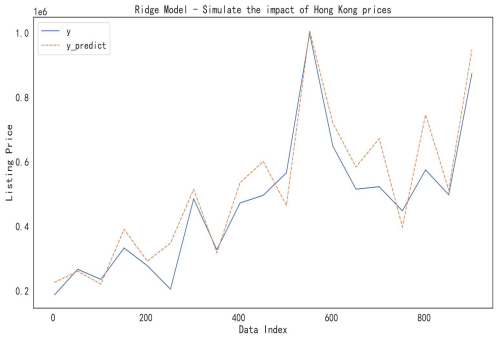
It can be seen that the ridge regression model is relatively accurate in predicting prices

in the Hong Kong region, with little error (1361.25) for single hull sailboats.

Then replace the area in the original data with Hong Kong, and make a price difference map. By comparing with Hong Kong's high-economic development areas, you can get the difference in the regional impact of each sailboat (including monohull sailboats and catamaran sailboats).



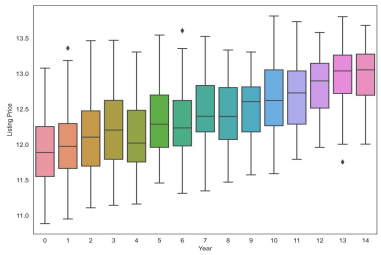
(a) The monohulls in the original data are valid in the Hong Kong area.



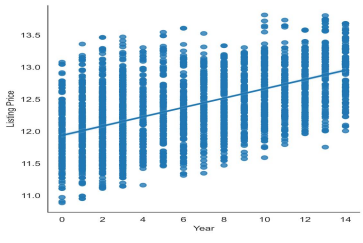
(b)Regional effects of catamarans in Hong Kong in the original data.

Figure 12 :Visualization of area effects

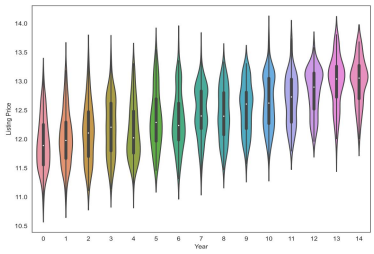
7.2 Other interesting and valuable conclusions



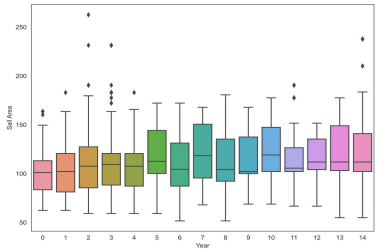
(a)Correlation of Year to Listing Price 1



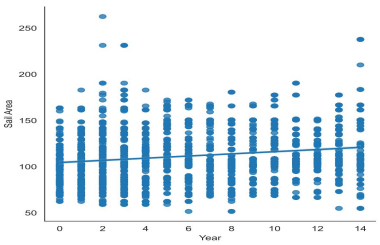
(b)Correlation of Year to Listing Price 2



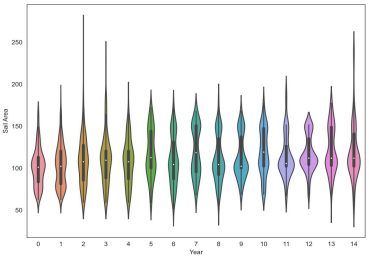
(c)Correlation of Year to Listing Price 3



(d) Correlation of year to sailing area 1



(e) Correlation of year to sailing area 2



(f) Correlation of year to sailing area 3

Figure 13 Other interesting and valuable conclusions

- ✧ The listing price has a synchronous increase trend with the increase of the year, and the fourth and fifth graphs show that the sailing area has a slow growth trend with the year.
- ✧ With the increase of the year, the increase in the value preservation rate of the sailboat is the main reason for the increase in its listed price. Therefore, in the process of trading in the second-hand sailboat market, the value preservation rate is a point that needs attention, and it determines the market price. The value of sailing boats of the same model.
- ✧ Considering the progress of the manufacturing process and the increase in people's

demand for large sailboats, it is very important to pay attention to market demand, which will affect the future listing price of sailboats.

8 Test the Model

8.1 Sensitivity Analysis

Studies have shown that changing any selected variable will have a profound impact on the results, which shows that the variables selected by the model have a great impact on the model, reflecting the credibility of the model.

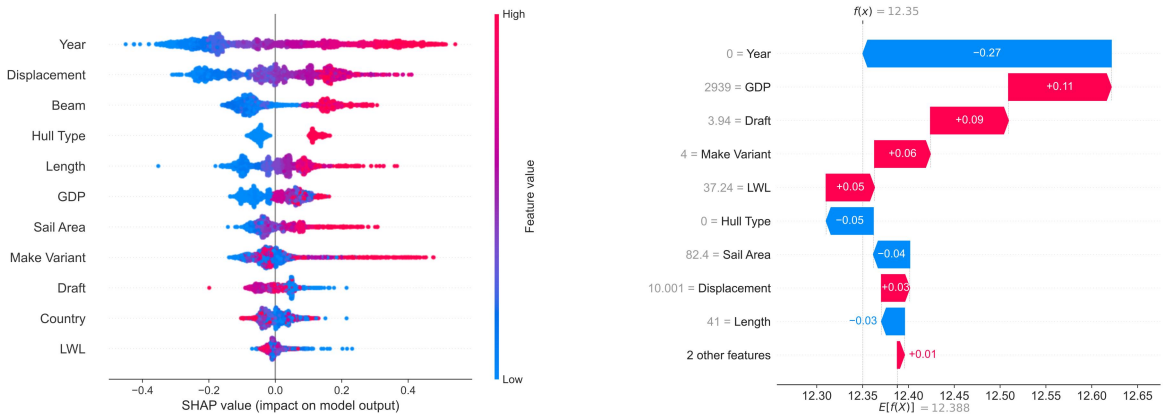


Figure 14: Sensitivity analysis for some factors

8.2 Robustness Analysis

For model 1, Cross-validation is a model evaluation method whose main purpose is to evaluate the generalization performance of the model, that is, the performance on new unseen data. Cross-validation methods are commonly used in problems such as model selection, hyperparameter tuning, and feature selection.

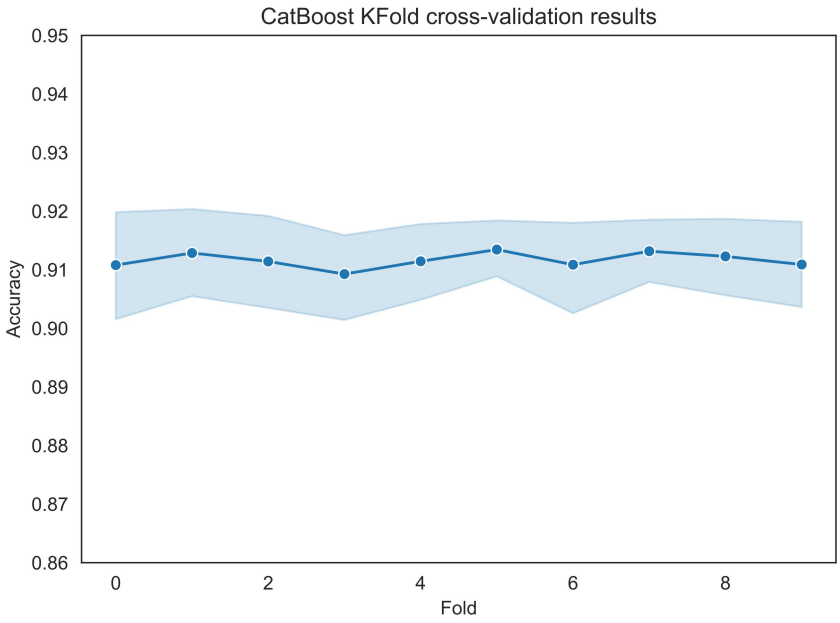


Figure 15: Robustness Analysis of the Model

It can be seen that the range of variation is roughly distributed on a straight line,

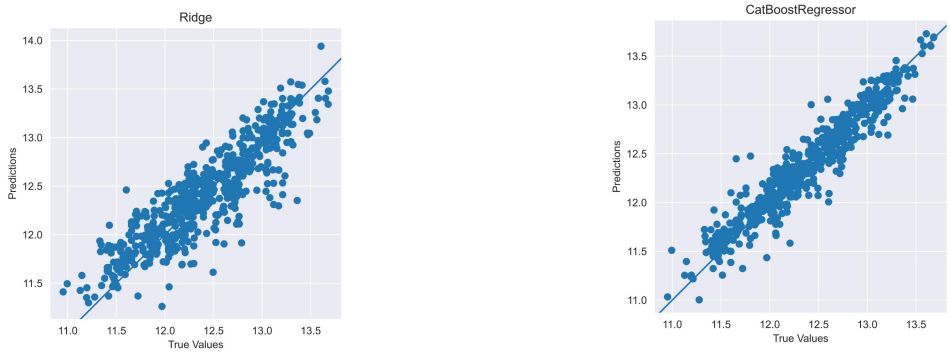
indicating that the overall fitting situation is stable. It shows that using our CatBoost-based ridge regression prediction model, similar prices can be obtained, which verifies the decisive role of factors such as year and displacement on prices, and reflects the stability of the model.

9 Conclusion

9.1 Summary of Results

9.1.1 Result of Problem 1

In order to intuitively display the accuracy of Ridge regression and CatBoost algorithm, we use python to draw the following Figure 16.



(a) CarBoost algorithm (b) Ridge regression

Figure 16: Fit Comparison between Ridge regression and CarBoost algorithm

The specific parameters are as following Table 13:

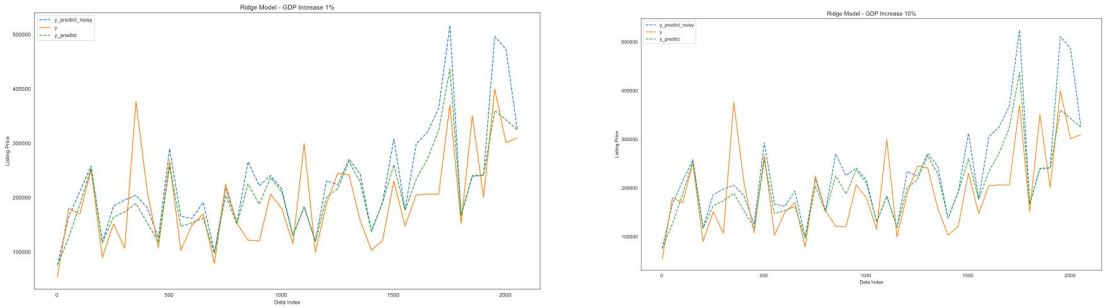
Table 13: CatBoost regression model accuracy results

[CatBoostRegressor]CatBoost Regression model r2_score:	0.916828164
[CatBoostRegressor]CatBoost Regression model mse:	0.026835207
[CatBoostRegressor]CatBoost Regression model mae:	0.120981616
[CatBoostRegressor]CatBoost Regression model mape:	0.009822888

It can be seen from the table that the fitting coefficient (r2) is very large, which reflects the good fitting effect, and the low standard deviation reflects the good stability of the model.

9.1.2 Result of Problem 2

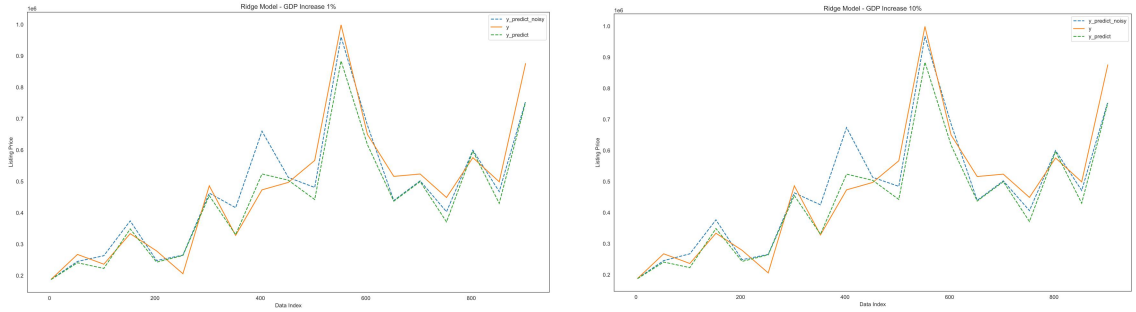
The results for the monohull sailboats are shown in the Figure 17 below:



(a) GDP change 1 percent (b) GDP change 10 percent

Figure 17: The impact of per capita GDP changes on monohull sailboats

The results for the catamaran are shown in the Figure 18 below:



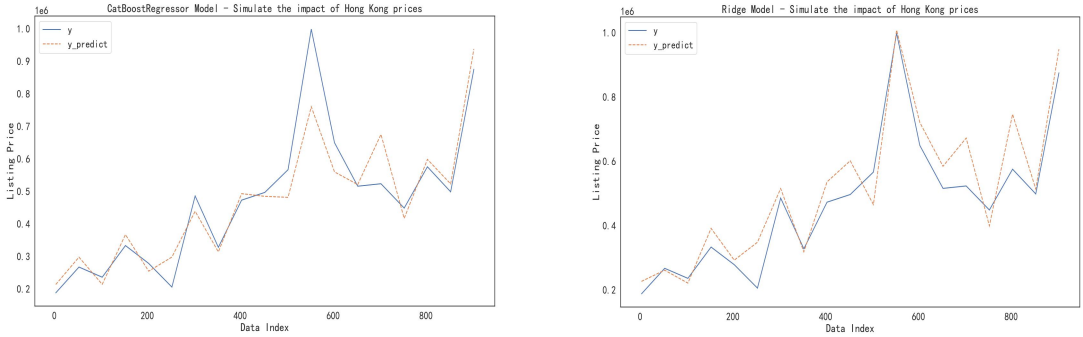
(a) GDP change 1 percent (b) GDP change 10 percent

Figure 18: The impact of per capita GDP changes on catamaran

- Its practical significance is expressed as that in high economic regions, regional effects have a significant positive impact on single and double sailing ships, while in low economic development regions, regional effects have a significant negative impact on single and double sailing ships. That is, when the regional economic per capita GDP increases, the price of single and double sailing ships in the region also increases. When the regional economicper capita GDP decreases, the price of single and double sailing ships in the region also decreases.
- Its statistical significance is expressed as a significant difference between Country/Region/State and Listing Price for different regions, and the closer the region is to the sea, the greater its significance.

9.1.3 Result of Problem 3

the sailing data of the data is changed from the area to the area to the area to the area to the price from the difference map through the comparison with the development of Hong Kong's high economic development and development of the development area of the area Sailboat sailboat sailboat sailboat sailboat change sailboat change sailboat.



(a) The monohulls in the original data are valid in the Hong Kong area.

(b)Regional effects of catamarans in Hong Kong in the original data.

Figure 19 :Visualization of area effects

Obviously, some of the monohull sailing boats are sensitive to regional effects, and the difference between the predicted price and the actual price is too large, while the catamaran sailing boats are not sensitive to the regional effect, and the difference between the predicted

price and the actual price is not too large.

9.1.4 Result of Problem 4

Through the statistical arrangement of the price information, the following figure can be obtained. It is not difficult to see that after the production of a new sailboat, the listing price will first increase, and then the overall price trend will decrease.

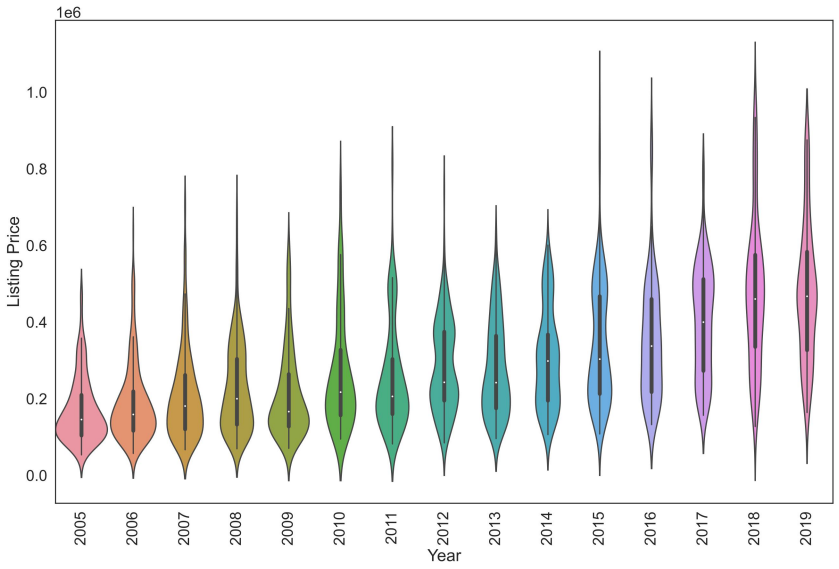


Figure 20: 2005-2019 Sailboat Price General Situation

9.2 Strengths

- The price prediction model based on **Ridge Regression** and **CatBoost Machine Learning** is scientific and reasonable, and can pass various statistical tests. The resulting predictions have a reliable statistical description;
- In the price forecasting model of sailboats, we classified and quantitatively analyzed the factors that need to be considered on the price, such as sailboat beams, displacement, and regional factors, to make our model more **sufficient and specific**;
- Based on **large sample data**, the scientific prediction and evaluation of Hong Kong Special Administrative Region pricing is convenient for sailing brokers to make rational decision-making pricing according to the actual situation in Hong Kong;
- The **sensitivity analysis** of the model shows the effectiveness of the model under different parameter combinations and proves the robustness of the model and the article **visualization** works excellent.

9.3 Possible Improvements

- The prediction of shipboat can be more accurate if we have more complete data;
- Some approximate analytical methods are used to model the prediction of used sailboat prices, which in extreme cases may turn out to be the opposite of reality.

References

[1] Li Fuqiang, Peng Haili, Yang Xi, Zhang Wenjing. Used car price prediction model and impact analysis based on deep learning [J]. Journal of Automotive Engineering, 2021,11(05):379-385.

[2] Tong Jia. Risk research and game analysis of second-hand car transactions [D]. Beijing Jiaotong University, 2010.

[3] Xiao Dongling. Research on Appraisal and Evaluation Method of Used Cars [D]. Chang'an University, 2007.Jansen, T., Campbell, A., Kelly, C., Hatun, H., & Payne, M. R. (2012). Migration and fisheries of North East Atlantic mackerel (Scomber scombrus) in autumn and winter. *PLoSOne*, 7(12).

[4] Economic Growth and Macro Stability Research Group, Chen Changbing, Zhang Ping, Liu Xiahui, Zhang Ziran. Urbanization, Industrial Efficiency and Economic Growth [J]. Economic Research, 2009, 44(10): 4-21.

[5] Ren Jiayi, Wang Aiyin. Research on Stock Price Prediction Using Transformer Model with Causal Attention [J/OL]. Computer Engineering and Application: 1-11[2023-04-03].http://kns.cnki.net/kcms/detail/ 11.2127.TP.20230313.1626.032.html

[6] Wang Runzhou, Zhang Xinsheng, Wang Minghu. Price prediction of agricultural products based on signal decomposition and deep learning [J]. Journal of Agricultural Engineering, 2022,38(24):256-267.

[7] Lin Yu, Yu Yuanyuan, Zhang Xi, Yue Yuying, Liu Xun. Research on Crude Oil Futures Price Prediction Based on Error Correction and Deep Reinforcement Learning [J]. System Engineering Theory and Practice, 2023, 43(01): 206-221.

[8] Ma Xiaojun, Song Yanqi, Chang Baishu, Yuan Mingyi, Su Heng. Application Research on P2P Default Prediction Model Based on CatBoost Algorithm [J]. Statistics and Information Forum, 2020,35(07):9-17.

[9] Chen Xuanxuan. Research on multi-factor quantitative stock selection strategy based on CatBoost algorithm [D]. Shanghai Normal University, 2020. DOI: 10.27312/d.cnki.gshsu.2020.001427.

Most Accurate Price Prediction

— Exploring the sailboat market in Hong Kong

With the development of the world economy, the second-hand sailboat market has also become hot. Hong Kong is an important shipping hub and one of the most competitive cities internationally and in the Asia-Pacific region. Known as the "Pearl of the Orient", it is also one of the richest and most economically developed regions in the world. Therefore, it is imperative to create a second-hand sailboat market in Hong Kong, and it is particularly important to study the pricing of sailboats in the second-hand market.

Ridge Regression

In the research, our team established a ridge regression used sailboat price prediction model based on the CatBoost algorithm. The price of the sailboat in this model is related to factors of the sailboat. The final regression coefficient of the model fitted by ridge regression is 0.76. After CatBoost machine learning, the fitting effect is more than 0.8 and close to 0.9, the effect is very good. Then we made predictions for the monohull sailing boats and catamaran sailing boats in Hong Kong.

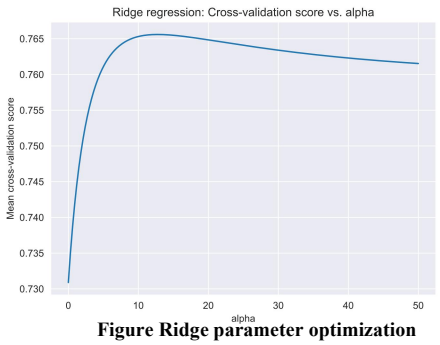


Figure Ridge parameter optimization

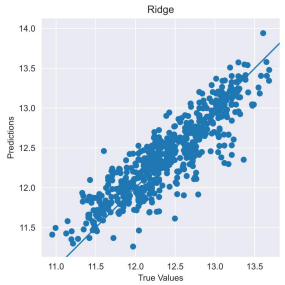


Figure (a) CarBoost algorithm

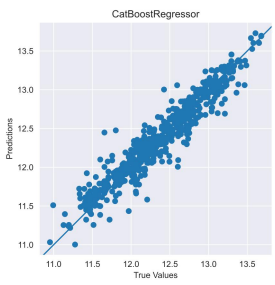


Figure (b) Ridge regression

The prediction results are shown in the figure below. The accuracy of the model is high, and the results Accurate, and we found that when the country's per capita GDP increases by 1%, the price of monohull sailboats increases by 3.15%, and the price of catamaran sailboats increases by 2.87%. And in the course of the research, it is found that regardless of the function of monohull sailboats, the price of catamaran sailboats increases over time. Growth, the first two years fell, and then rose, and this trend meets the consistency test, which has extraordinary value, so our team sincerely invites you to cooperate with us to make a big deal in the second-hand sailboat market.

(Editor: Team #2330671)

(03th April 2023)

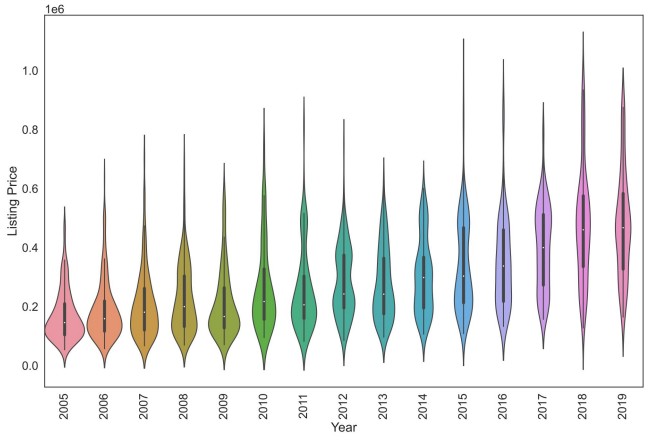
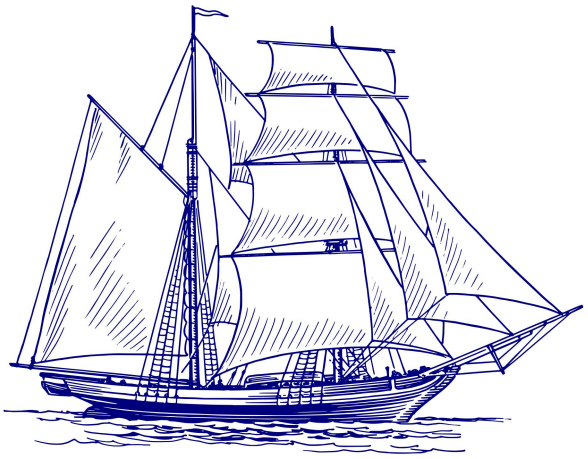


Figure: 2005-2019 Sailboat Price General Situation



Appendices

Appendix 1

Introduce: Tools and software

Paper written and generated via Office 2023.
Graph generated and calculation using Visual Studio Code.

Appendix 2

Introduce: CatBoost Algorithm by Python code (Excerpts)

```
# Train and test each model and output R-squared values
for model in models:
modelName = type(model).__name__
title = f'[{modelName}]{modelName}'
print(f'-----------{title}-----------')
# if type(models).__name__ == 'LGBMRegressor':
# model.fit(X_train, y_train, feature_name=feature_names)
# else:
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
# r2 = r2_score(y_test, y_pred)

# Predict and visualize results
plt.scatter(y_test, y_pred)
plt.title(type(model).__name__)
plt.xlabel('True Values')
plt.ylabel('Predictions')
plt.axis('equal')
plt.axis('square')
# plt.xlim([9e4, 1e5])
# plt.ylim([9e4, 1e5])
_ = plt.plot([-1e10, 1e10], [-1e10, 1e10])
plt.tight_layout()
plt.savefig(f'{modelName}.svg')
plt.show()
```