Problem Chosen	2023	Team Control Number
Y	MCM/ICM Summary Sheet	2330124

Price prediction model of used sailing ship based on statistical analysis and BP neural network

Summary

Brokers act as sellers' representatives to sell their boats in Hong Kong. However, the variety of used sailboats and their geographical influence makes it difficult to rely on subjective judgment and experience in pricing. In order to maximize the interest of the seller while setting a reasonable selling price, the quantitative pricing method makes the pricing accurate and reliable by building a reasonable model.

After getting the data, we **pre-processed, cleaned the original data**, dealt with **missing and duplicate values**, and further **filtered and expanded the data**. After that, we performed a simple **statistical** analysis and visualized our results as graphs.

For problem 1, we first collected and integrated with the original data and the favorable indicators for used sailboat price prediction, including the sailboat's characteristics and the regional characteristics. Then, we screened the samples based on the previous preliminary statistical results and built a **BP neural network**-based sailboat price prediction model according to the categories of sailboats, respectively. By continuously adjusting the implied layers, we finally determined that the number of implied layers to achieve the best results for both was **11** and **10**, respectively, and the $\boldsymbol{R}^2$ of the test set was above **0.94**. We used samples not involved in training for validation, and the visualization showed good prediction results.

For Problem 2, we first conducted a **simple statistical analysis** and visualization of listing prices by region and initially found that used sailboat pricing has geographical differences. Subsequently, using quantitative data, we used **One-way ANOVA** to verify the significant effect of geographic differences. Finally, we concluded that there is a **consistent regional effect** in the sailboat variant and explained the effect accordingly.

For problem 3, for the subset of sailboats to be studied, we selected Beneteau's Oceanis series and Lagoon's 450 series to represent the characteristics of the sailboat market in the given region based on the previous **statistical analysis**. After collecting data related to Hong Kong, we first used the established model to forecast the subsets' pricing and verified the model's accuracy. On the other hand, we input the regional data of Hong Kong into the **neural network** and obtained the conclusion that the regional effect of Hong Kong is **significant.**

For problem 4, firstly, we use the **correlation analysis** method to mine the ship's characteristics and draw a correlation matrix heat map. Secondly, we conduct statistical analysis for the models and average selling prices of popular variants produced by large manufacturers and explore the price gap between monohulled boats and Catamarans accordingly. In addition, we counted the highest-priced vessels and the lowest-priced vessels sold by the major manufacturers.

Finally, we wrote a report for the broker with our models, results, and suggestions. Moreover, the advantages and disadvantages of the model are analyzed.

Keywords: BP neural network   statistical analysis   One-way ANOVA   visualization

Contents

1 Introduction 3
1.1 Problem Background 3
1.2 Restatement of the Problem 3
1.3 Our Work 3
2 Assumptions and Justifications 4
3 Notations 5
4 Data processing and analysis 5
4.1 Missing value processing 5
4.2 Duplicate value processing 6
4.3 Simple statistical analysis 7
5 Sailboat price forecasting modeling 9
5.1 Selection of additional features 9
5.1.1 Selection of sailboat features 9
5.1.2 Selection of area characteristics 10
5.2 Predict sailing prices 11
5.2.1 Model preparation 11
5.2.2 Sailboat price prediction model based on BP neural network 11
5.2.3 Training of BP Neural Network 11
5.2.4 Model solving and result analysis 14
6 Regional impact analysis based on statistical modeling 15
6.1 Fundamental statistical analysis of two types of hull prices 15
6.2 Analysis of variance 16
6.2.1 Region-based one-way ANOVA 16
6.2.2 Analysis results and interpretation 17
7 Simulation of the Hong Kong market 17
7.1 Determining a subset of sailboats 18
7.2 Data Collection 18
7.3 Examining the Hong Kong second-hand ship market 18
8 Data Mining 19
8.1 Information mining on the characteristics of the sailboat itself 19
8.2 Digging information on the sales status of major manufacturer variants 19
8.3 Other interesting information 19
9 Model Evaluation and Further Discussion 20
9.1 Strengths 20
9.2 Weaknesses 20
References 23

1 Introduction

1.1 Problem Background

Shipping is one of the few industries where significant assets can be traded in a sepa-rate and active market. Second-hand ships refer to the transaction process in which the original shipowner resells the old ships that are still within active life and still have the value of use to the new shipowner. Due to the new building's long cycle and high cost, the second-hand shipping market has an essential economic status in the shipping industry. Ac-cording to the statistics released by IMO, the volume of global second-hand ship transac-tions in recent years accounts for about two-thirds of the overall transaction volume.

Due to the unique characteristics of sailboats, the factors that affect the price of used boats are market conditions in addition to the boat's age, length, and wear and tear. As a luxury item, there is a regional effect on the selling price of sailboats. It is appropriate to measure the local shipping industry market by the region's economic development level and the re-gional demand for ships. There are some constraints in the data and information that can be collected. It is essential to sort out the factors influencing the valuation of used ves-sels for different vessel types and explore a comprehensive valuation method to understand the sailboat market.

Ship brokers play an essential role in communicating and catalyzing the lengthy pro-cedures of ship buying and selling and the low liquidity of the market. The factors in-flu-encing the price of a used vessel differ from those of a new building, as do the market con-ditions and market prices in the region, so brokers' pricing decisions differ. Brokers re-ceive a percentage of the service fee for ship transactions. A comprehensive analysis of the pric-ing decision factors for used ships is crucial to making the right decision and allows brokers to increase their earnings.

1.2 Restatement of the Problem

After fully understanding the background information and constraints of the problem, we need to address the following questions.

- Problem one, using reliable information, search for data related to other characteristics of sailboats, find regional economic data that can be used along with the existing characteristics data to explain the listed prices of sailboats in the table, and illustrate the accuracy of the model.
- Problem two is to explore whether there is a regional effect on the price of the same sailboat variant by comparing the listed prices of sailboats in different regions and to illustrate the relevance of the above discussion.
- Problem three, for monohulls and catamarans, select some information-rich variants to form a subset, respectively. Compare the table with the typical indicators of the region where the sailboat is located in Hong Kong and the other regional prices of the sailboat with the comparable listed prices in Hong Kong comprehensively. Simulate the Hong Kong market situation and explore it further.
- Problem four, look for other informative and exciting inferences and conclusions through data mining.
- Problem five, prepare a report card for Hong Kong sailboat brokers on how used boats should be priced in Hong Kong.

1.3 Our Work

This problem requires the development of a sailboat price forecasting model to forecast prices for sailboats with given characteristics and thus achieve an analysis of the overall used sailboat market situation. Our work consists mainly of the following:

(1) Cleaning the data and establishing a BP neural network-based sailboat price prediction model through simple statistical analysis, data collection, and sample selection.

(2) A series of statistical methods, such as the chi-square test, were used to explain the effect of region on the listed price, and its practical and statistical significance was illustrated based on the test results.

(3) Screening subsets from the data based on the statistical results, collecting relevant data from the Hong Kong region, and performing regional effect analysis.

(4) Data mining is used to explore the intrinsic connections in the data and visualize the presentation.

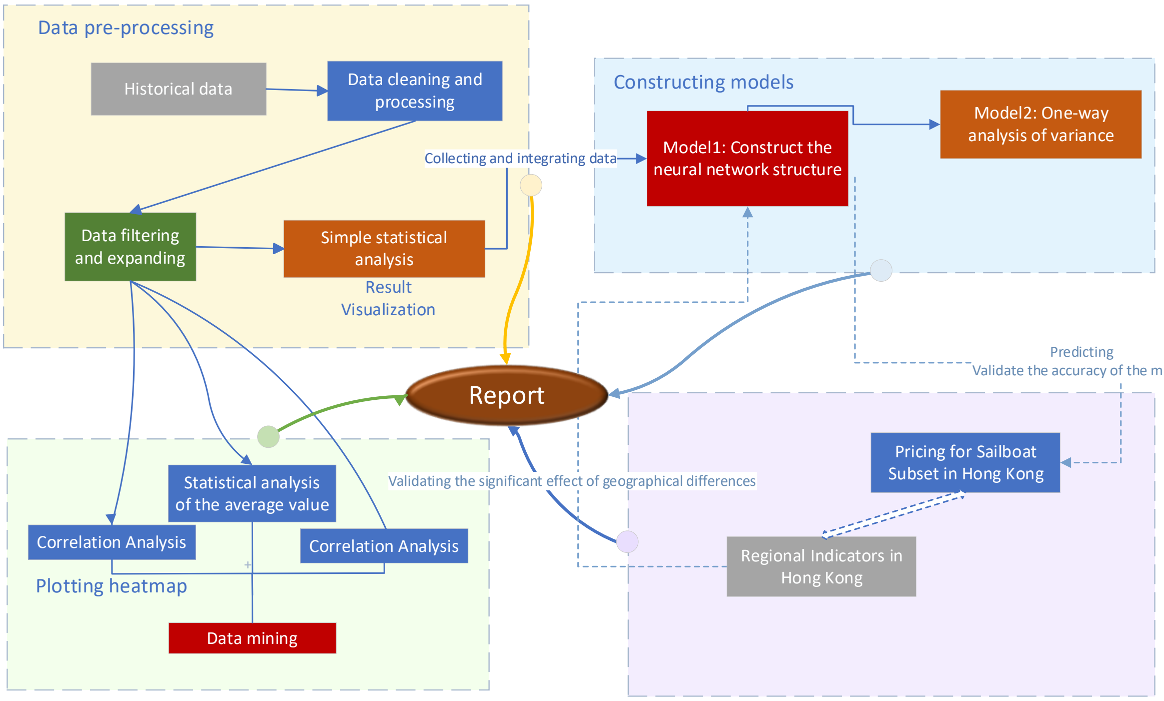

Data pre-processing
Historical data
Data cleaning and processing
Collecting and integrating data
Data filtering and expanding
Simple statistical analysis
Result Visualization
Constructing models
Model1: Construct the neural network structure
Model2: One-way analysis of variance
Predicting
Validate the accuracy of the model
Report
Pricing for Sailboat Subset in Hong Kong
Validating the significant effect of geographical differences
Regional Indicators in Hong Kong
Statistical analysis of the average value
Correlation Analysis
Correlation Analysis
Plotting heatmap
Data mining

Figure 1 Flow Chart of Our Work

2 Assumptions and Justifications

Assumption1: Because the usage time of a ship is proportional to its age, we use the age of a ship as a quantitative characteristic of usage time.

justification: The number of engine hours of ships can represent the usage time of each ship more accurately, but the data are difficult to collect and are influenced by personal factors. The number of years of ship use is easy to count and conforms to statistical laws.

Assumption2: The pricing of a vessel is determined only by the hull parameters, performance, time of use, and region of the sale in this paper. Policies and markets are stable in each region.

justification: The systematic errors caused by other factors are unbiased, and the influence decreases when the data is more significant. This small influence is ignored in this paper. Due to the uncertainty of the market, there are regular fluctuations in the prices of ships. However, the prices of ships may be strongly affected by sudden changes in external factors, such as the development of policies and the occurrence of market crises, which we ignore here.

Assumption3: The identical data are considered anomalous data and de-duplicated.

justification: Since the pricing of ships is affected by various random factors, even if the indicators are precisely the same, the price fluctuation range is extensive, and the possibility of an exact agreement is almost zero. In this paper, the identical data are considered anomalies with repeated input and are processed.

Assumption4: The area where the ship is located is where the sale is oriented.

justification: Since ships can be transported by sea, there are cases that they are sold to other regions. However, there is a large amount of missing data for these imports and exports, and we can only simulate the Hong Kong market with the characteristics of each region as an indicator for price illustration. Therefore, we consider where the ship is currently docked as the location to which it is sold.

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	Input variable
y_i	Output Variable
H_j	The hidden layer output of the neural network，Assumption
f	Intermediate layer excitation function
w_{ij}	Weight
a_j	Threshold
O_k	Predicted values for the output layer
b_k	Output layer threshold
E	the difference between the expected value and the predicted value
δ_k	Error of the same floor unit
η	A constant indicating the scale factor
n	Input layer parameter
l	Hidden layer parameters
m	Output layer parameter
μ_i	Average value
$\bar{x_j}$	The sample mean at the j-th level
x_{ij}	The i-th value of the j-th level
n_j	Sample size of the j-th level
${S_j}^2$	The sample variance of the j-th level
MSB	Variance between groups
SSB	Horizontal sum of squares
MSE	Within-group variance
SSE	Sum of squares of error terms
$\bar{\bar{x}}$	Total sample mean
n_r	Sum of the capacity of each sample

4 Data processing and analysis

Before data mining, it is necessary to ensure the integrity of the data, so we performed preliminary data cleaning and simple statistical analysis on the two forms separately to facilitate subsequent modeling.

4.1 Missing value processing

First, observing the official EXCEL sheet, we found no missing values in the 'Catamarans' form. However, there were three missing values in the 'Monohulled Sailboats' form, and their data characteristics in the original form are shown in Table 2.

Table 2 Missing values from sheet 'Catamarans'

number	Make	Variant	Length(ft)	GR	CRS	LP	Year
1588	Beneteau	Oceanis 54	54	USA	NULL	$479,805	2013
1594	Delphia	46 cc	46	Europe	NULL	$314,606	2013
1718	Bavaria	Cruiser 46	46	Europe	NULL	$201,640	2014

tips:GR means Geographic Region,CRS means Country/Region/State,LP means Listing Price (USD)

We found that the columns where the missing values are all regional variables with a missing rate of 0.12%. Since the missing data are minor and do not significantly impact the overall sailing market forecast, we chose to treat all the samples where the missing values are located as invalid data, which is the most straightforward and reasonable to exclude.

4.2 Duplicate value processing

After processing the missing data, the two forms were imported into SPSS software separately. After searching for duplicate values, we found ten duplicates in the 'Monohulled Sailboats' form and 72 in the 'Catamarans' form. We believe that these duplicates were created at the time of data collection. We eliminated the redundant data and kept only one of them for analysis. Tables 3 and 4 show some samples with duplicate values from the 'Monohulled Sailboats' and 'Catamarans' forms, respectively.

Table 3 Duplicate value from sheet 'Monohulled Sailboats '

Make	Variant	Length(ft)	GR	CRS	LP	Year
Bavaria	39 Cruiser	39	Europe	France	95961	2006
Bavaria	42 Match	41	Europe	Croatia	57091	2006
Beneteau	Oceanis 473	47	Europe	Spain	151876	2006
Beneteau	Oceanis 55	55	Europe	France	576981	2015
Beneteau	Oceanis Clipper 473	47	Europe	Spain	144586	2006
Dehler	46	46	Europe	Netherlands	460370	2018
Hanse	455	44	Europe	Croatia	245335	2017
Hanse	461	46	Europe	Italy	143371	2006
Jeanneau	Sun Odyssey 389	39	Caribbean	British Virgin Islands	159000	2018

tips:GR means Geographic Region,CRS means Country/Region/State,LP means Listing Price (USD)

Table 4 Partial duplicate value from sheet 'Catamarans'

Make	Variant	Length(ft)	GR	CRS	LP	Year
Lagoon	52	52.0	Europe	Croatia	1003140	2014
Seawind	1260	41.0	USA	Florida	595000	2018
Nautitech	40 Open	39.5	Europe	Croatia	345919	2016
Manta	42 MK IV	42.0	USA	Florida	270000	2008
Fountaine Pajot	SABA 50	49.0	Europe	Croatia	934587	2018

tips:GR means Geographic Region,CRS means Country/Region/State,LP means Listing Price (USD)

4.3 Simple statistical analysis

We created separate bar charts based on Geographic Region for the two tables. Preliminary findings show that for Monohulled Sailboats, the vast majority are located in Europe, with 75.95%, and only 24.05% are sold to the Americas and the Caribbean. For Catamarans, more than located in Europe, with 64.21%, and only 35.79% in the Americas and the Caribbean. The European market is the main target of sailboat sales, followed by the Americas and the Caribbean.

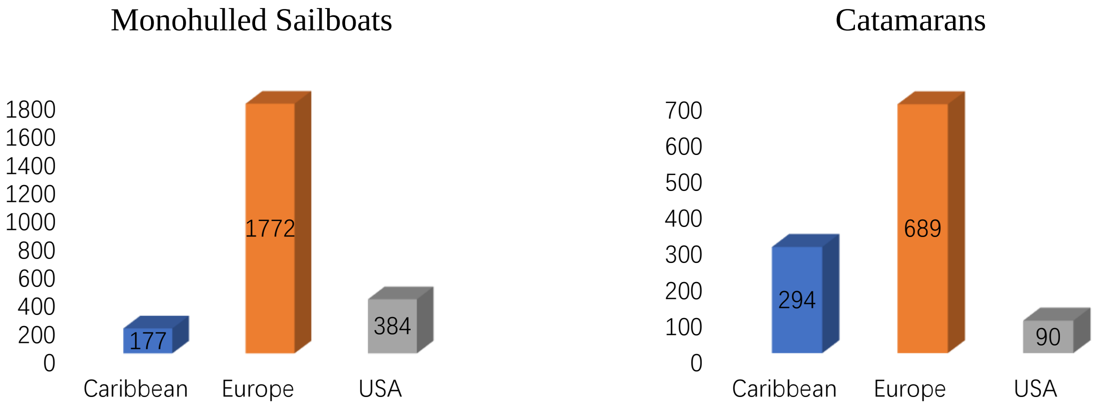


Figure 2 Sales volume histogram by geographical region

Then, the two types of sailboats were arranged alphabetically according to the manufacturers' names and numbered in order. It was counted that 61 manufacturers were producing Monohulled Sailboats and 20 manufacturers producing Catamarans, and no manufacturer in the form produced both types of boats. Their production is shown in Figure2.

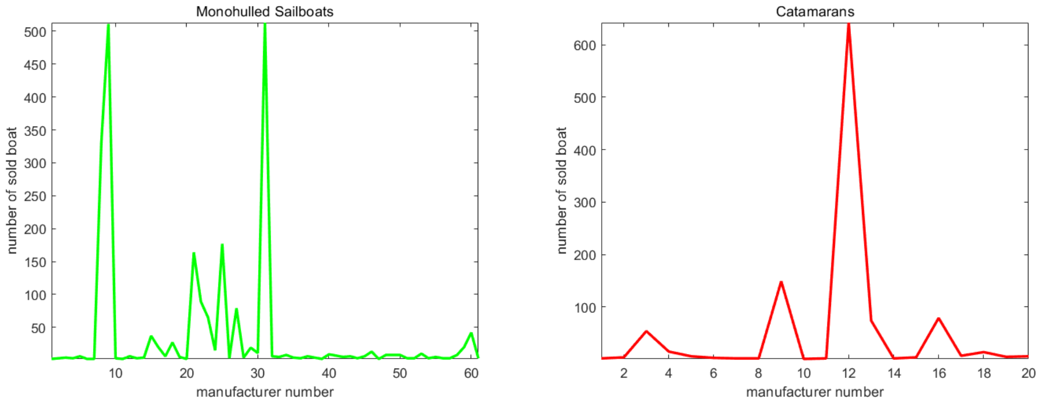


Figure 3 Manufacturer sales of both types of boats

From Figure 2, it is easy to see that manufacturers' distribution is not uniform for the same type of sailboat, and there are extreme values. To facilitate the analysis, we screened out the manufacturers that produce more than 50 of each of the two types of boats. We intend to consider these manufacturers as the primary study subjects and use the drawing of pie charts to illustrate the manufacturers' production visually.

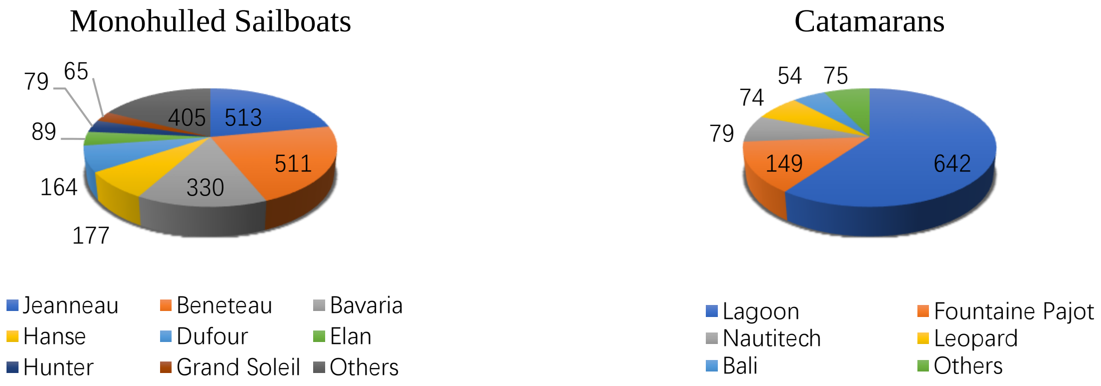


Figure 4 Sales by major manufacturers

In addition, for both types of sailboats, we also counted the significant manufacturers' boat sales in each region, and the results are shown in Table 5 and Table 6. From these two tables, we can observe each manufacturer's central sales regions and the leading boats suppliers in each region.

Table 5 Central sales area for major manufacturers of monohulls

GeographicRe-gion	Jean-neau	Bene-teau	Ba-varia	Hanse	Dufour	Elan	Hunter	Grand Soleil	Oth-ers
USA	83	110	7	8	3	3	68	0	102
Europe	392	329	307	158	141	86	7	61	291
Caribbean	38	72	16	11	20	0	4	4	12

Table 6 Central sales area for major manufacturers of Catamarans

Geographic Region	Lagoon	Fountaine Pajot	Nautitech	Leopard	Bali	Others
USA	34	18	2	10	3	23
Europe	458	83	52	19	49	28
Caribbean	150	48	25	45	2	24

For different geographic regions, we counted the sales in each country. We chose the significant selling countries in each region. The results are shown in Figures 4 and 5. The figure shows significant differences in sales in different countries or regions of the same geographical area. For example, for catamarans, Florida occupies most of the USA's market. The variability of these locations is the basis for our subsequent analysis.

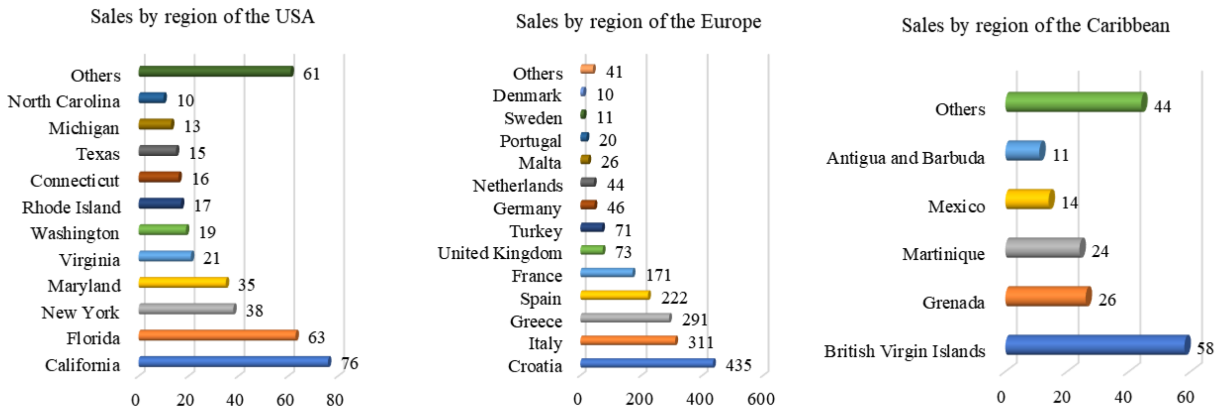


Figure 5 Individual sales of Monohulled Sailboats by country

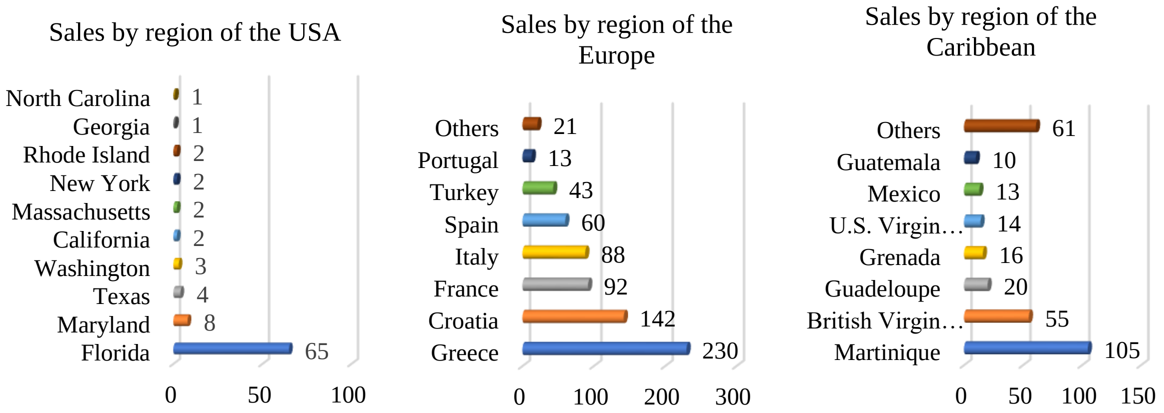


Figure 6 Individual sales of Catamarans by country

5 Sailboat price forecasting modeling

5.1 Selection of additional features

5.1.1 Selection of sailboat features

In addition to the data given in the table, by reviewing the information and the hints in the question, we have selected some characteristics that reflect the overall performance of a sailboat, and the interpretation of these characteristics is as follows:

●**Horsepower**

Horsepower is a common unit of measurement of power used in engineering, and the horsepower of a sailboat often represents the output of the sailboat's motor, an important indicator of the sailboat's engine.

●**Displacement(Disp)**

Displacement is the mass of water a ship discharges when loaded with cargo as designed. Displacement is usually expressed in tonnage and refers to the number of tons of water a ship discharges in the water, and the value is a measure of the size of a sailboat's scale.

●**Boat age**

The age of a boat represents the degree of aging of a hull. Generally speaking, the price of a used sailboat is negatively correlated with the age of the boat. Based on the year of manufacture in the table, we can calculate the age of the boat and include it in our metrics.

●**SA/D**

SA/D is the ratio of sailboat area to displacement. The higher this value is, the more powerful the boat is. The specific calculation formula is as follows:

$$SA/D = SA/(Disp/64)^{0.666} \tag{1}$$

Where SA denotes sail area, and Disp denotes displacement, a value below 16 would be considered underpowered, 16 to 20 indicates pretty good performance, and above 20 indicates relatively high performance.

●**Bal/Disp**

Bal/Disp is the ratio of ballast to displacement. This value represents a ship's ability to withstand the wind. A value higher than 40 means a stiffer and stronger hull. The formula is as follows:

$$Bal\,/\,Disp\;=\;Bal\,/\,Disp*100 \tag{2}$$

●Disp/Len

Disp/Len is the ratio of displacement to length. The lower the value, the less power is required to drive the boat to its rated hull speed. This value measures how light or heavy a sailboat is and is calculated as follows:

$$D\,/\,L = (Disp\,/\,2240)\,/\,\left(0.01*Len\right)^3 \tag{3}$$

If this value for the ship is less than 100, the ship is considered to be an ultralight ship. If this value is between 100 and 200, the type of this ship is a light ship. If this value is between 200 and 275, the type of this ship is a medium ship. The ship's type is heavy if this value is between 275 and 350. If this value exceeds 375, the ship's type is super heavy.

●Comfort Ratio

This is a ratio created by Ted Brewer as a measure of motion comfort. It provides a reasonable comparison between yachts of similar size and type. It is based on the fact that the faster the motion the more upsetting it is to the average person,The specific calculation is as follows:

$$Comfort\ ratio = D\,/\,(0.65*0.7LWL+0.3LOA)*Beam^{1.33} \tag{4}$$

LWL represents the hull's overall length, LOA represents the draft length of the ship, D represents the displacement, and Beam is the length of the beam.

Consider, though, that the typical summertime coastal cruiser will rarely encounter the wind and seas that an ocean going yacht will meet.Numbers below 20 indicate a lightweight racing boat;20 to 30 indicates a coastal cruiser;30 to 40 indicates a moderate bluewater cruising boat;40 to 50 indicates a heavy bluewater boat;over 50 indicates an extremely heavy bluewater boat.

5.1.2 Selection of area characteristics

For the same item, the selling price varies from region to region; therefore, it is necessary to incorporate regional characteristics into our prediction model. We have selected the following characteristics and explained them accordingly.

●GDP per capita

GDP per capita is an essential indicator of the level of economic development of a country or region. For a luxury item like a sailboat, the selling price varies significantly between countries with different GDP per capita. In addition, the market for private sailboats is only for a small group of wealthy individuals. The higher the GDP per capita, the larger this group tends to be, hence the choice of this indicator as a pricing recommendation for the sale of sailboats.For this purpose, we collected the per capita GDP of each region in 2020.

●Length of coastline

The length of the coastline is the length of the junction line between land and sea, representing to some extent the extent of the area's sea and the range in which yachts can travel. This indicator affects the demand for yachts, and the amount of demand can often determine the market price, so it is a significant indicator.

●Number of major ports

The number of main ports represents the level of maritime trade of a country or region. The number of ports is influenced by the length of the coastline and the local economic conditions, so the number of ports can reflect the comprehensive yacht purchasing power of a region.

●**Average cargo throughput**

Average cargo throughput is a basic indicator of a port's cargo throughput capacity, and this feature represents the level of maritime trade and port development of a region. The cargo throughput of different regions is limited by the number of ports, ships, trading countries, and other factors and can also synthetically indicate the purchasing power of yachts in a region.

5.2 Predict sailing prices

5.2.1 Model preparation

All indicators in the table can be divided into two types of variables - continuous and categorical. Continuous variables have numerical significance, while categorical variables indicate categories without numerical or sequential significance. Therefore, it is necessary to Perform dummy variable assignment processing to each categorical variable, such as Europe to 2, USA to 1, and Caribbean to 0, to facilitate computer processing and regression analysis.

Before building the model, we built it by region based on previous statistical analysis. The statistical analysis was used first to find the major manufacturers in each region. The central selling regions to that manufacturer's geographic area were selected for each selected manufacturer. For each model sailboat, we performed the model building by assigning values to categorical variables that could eventually be accurate to the price level of the specified sailboat in the specified region. The selection process of regions is shown in Figures 6 and 7.

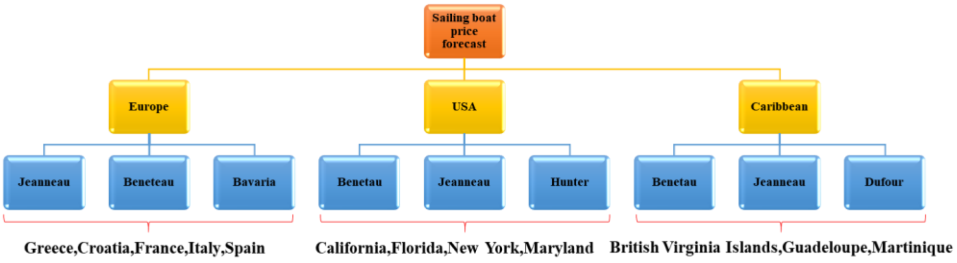

Sailing boat price forecast
Europe
USA
Caribbean
Jeanneau
Beneteau
Bavaria
Benetau
Jeanneau
Hunter
Benetau
Jeanneau
Dufour
Greece,Croatia,France,Italy,Spain
California,Florida,New York,Maryland
British Virginia Islands,Guadeloupe,Martinique

Figure 7 Schematic diagram of the zone screening process

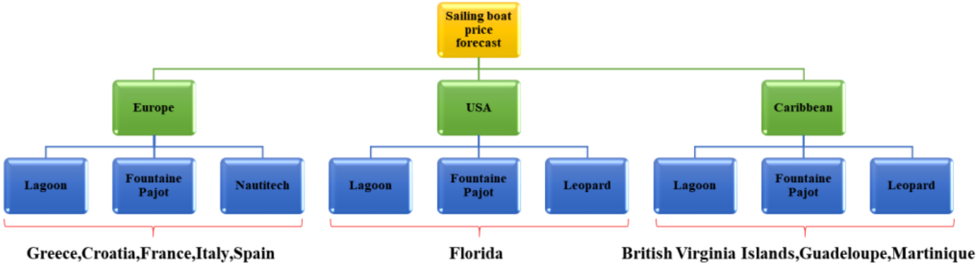


Figure 8 Catamarans of the zone screening process

5.2.2 Sailboat price prediction model based on BP neural network

The neural network is a common artificial intelligence analysis tool that can process information by imitating how the human brain thinks and processes and achieve the purpose of analysis and decision-making accordingly. With good nonlinear mapping ability, self-learning, self-adaptive ability, and strong anti-interference ability, neural networks are currently used in many fields and provide new data mining ideas. Therefore, we consider using a neural network approach to predict the price of sailboats for quantifiable indicators extracted after the screening.

5.2.3 Training of BP Neural Network

In the process of fitting analysis using the BP neural network, it is not necessary to determine the structure and parameters of the object, and it can be optimized and improved by continuously performing weight adjustment so that the best relationship between the input and output is established. The structure of this network consists of an input layer, an output layer, and an implicit layer, and the network structure is shown in Figure 8.

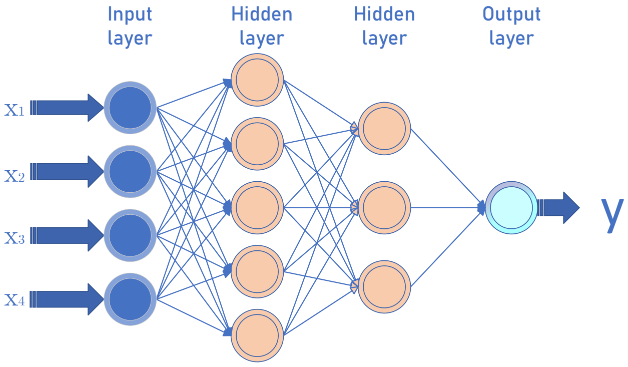


Figure 9 Neural network structure diagram

The BP neural network requires continuous forward transmission of information and backward propagation of errors during processing. The former process sends the corresponding signal to the implicit layer and continuously passes it to the output. In contrast, the latter process continuously adjusts the weights to make the theoretical and test values of the output as consistent as possible. The process of BP neural network training is as follows.

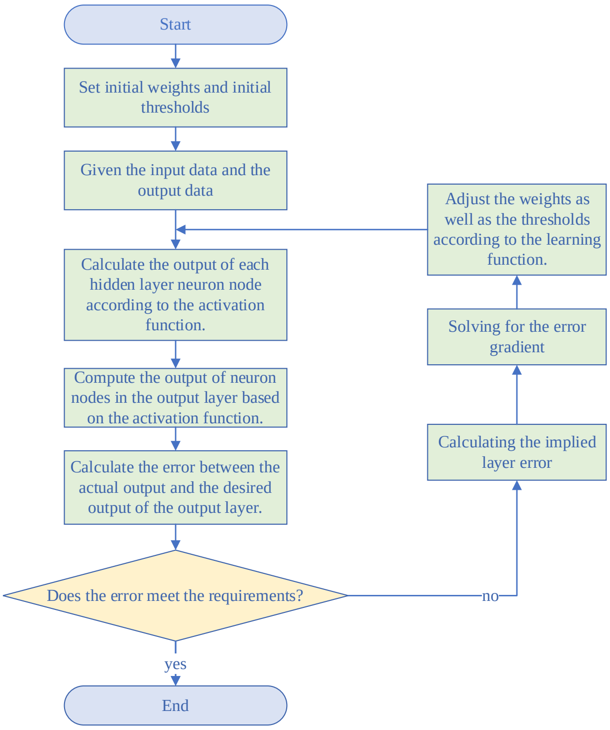

Start
Set initial weights and initial thresholds
Given the input data and the output data
Calculate the output of each hidden layer neuron node according to the activation function.
Compute the output of neuron nodes in the output layer based on the activation function.
Calculate the error between the actual output and the desired output of the output layer.
Does the error meet the requirements?
yes
End
no
Calculating the implied layer error
Solving for the error gradient
Adjust the weights as well as the thresholds according to the learning function.

Figure 10 General flow chart of neural network training

(1)BP initialization, $X(x_1, x_2, \ldots, x_n)$, $Y(y_1, y_2, \ldots, y_m)$ denote the input and output variables, which are closely related to the structure and complexity of the network. The input, implicit, and output layer parameters n, l, and m; w_{ij} are the connection weights, and a represents the intermediate layer threshold. The appropriate learning rate and excitation function parameters should also be set before the problem is processed through the neural network.

(2)Calculate the output of the intermediate layer, where the solution process is based on the input sequence information X, connecting the weights w_{ij} and the threshold a, and the intermediate layer output H is calculated; the corresponding expressions are as follows.

$$H_J = f(\sum_{i=1}^{n} x_i w_{ij} - a_j) \tag{5}$$

where f denotes the intermediate layer excitation function, take $f(x) = 1/(1 + exp(x))$

(3)The output of the output layer is then computed. In this process, the predicted value O_k is obtained based on the output H-connection weights w_{ik} of the output layer.

$$O_k = f(\sum_{i=1}^{l} H_j w_{jk} - b_k) \tag{6}$$

Comparing the theoretical and actual results after the neural network processing and calculating the error, the corresponding expressions are as follows.

$$E = \frac{1}{m}\sum_{k=1}^{m} (y_k - o_k)^2 \tag{7}$$

We have the following equation expanding the above error definition equation to the hidden layer.

$$E = \frac{1}{m}[f(\sum_{j=1}^{l} H_j w_{jk} - b_j) - o_k]^2 \tag{8}$$

Further expansion to the input layer yields.

$$E = \frac{1}{m}\{f[\sum_{j=1}^{l}(\sum_{j=1}^{l} x_j w_{jk} - a_j)\, w_{jk} - b_k - o_k]\}^2 \tag{9}$$

A specific analysis shows that the input error in this network is a function of the weights w_{ij}; w_{jk} of each layer, so as to adjust the weights effectively by constantly adjusting E.

$$\begin{aligned} \Delta w_{jk} &= -\eta \frac{\partial E}{\partial W_{jk}} \\ \Delta w_{ij} &= -\eta \frac{\partial E}{\partial W_{ij}} \end{aligned} \tag{10}$$

The negative signs in the two equations above specifically indicate gradient descent, and the constant $\eta \in (0,1)$ indicates the scale factor.

According to the above analysis, the description of the learning algorithm can be obtained as follows.

(1) Initialize the corresponding learning factors and iterations, as well as the network layer parameters n, m ; and obtain the corresponding hidden layer parameters based on the parameters by trial-and-error analysis.

(2) Input forward propagation process: compare the results, and if the corresponding error is found to exceed a certain threshold value, proceed to the next step.

(3)Error back propagation: Rooted in minimizing the error continuously modify the weights. Firstly, calculate the error δ_k for the same layer of cells; secondly, adjust the weights and thresholds according to the following formula. For the connection weights, the correction formula is

$$w_{jk}(t+1) = w_{jk}(t) + \eta \delta_k o_j \tag{11}$$

5.2.4 Model solving and result analysis

In constructing the model, we used the BP neural network toolbox in MATLAB to achieve the prediction of the price of a specific sailboat. By continuously adjusting the parameters, we finally obtained more satisfactory results, and the specific steps are as follows.

●Parameter adjustment

The number of hidden layers often affects the accuracy of BP neural networks. To meet the prediction requirements, we train separately for two types of sailboats. The input data for both are the raw data from the tables and the indicator data queried via the Internet, and the output data are the prices of sailboats. Multiple training iterations are performed by continuously modifying the number of hidden layers. The following figure shows the topology of the Monohulled Sailboats network training.

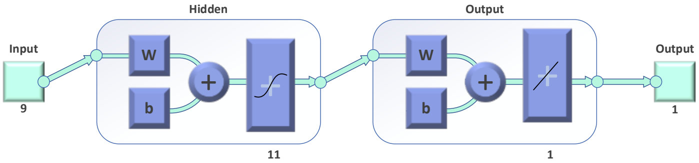


Figure 11 Monohulled Sailboats BP neural network training topology

To avoid overfitting or underfitting, 70% of the sample was used for training, 15% for training, and the remaining 15% for testing, using the mean square error to measure the network performance. For Monohulled Sailboats, the number of implied layers is set to 11, while for Catamarans, the number of implied layers is set to 10 for best results.

●Model test

The fitting accuracy of predicting Monohulled Sailboats using the BP neural network is shown in the figure below. The figure shows that the R^2 of the training set, test set, and validation set are all greater than 0.94, and the overall R^2 also reaches 0.96, which indicates that the regression effect is good and can be used to predict the price of sailboats.

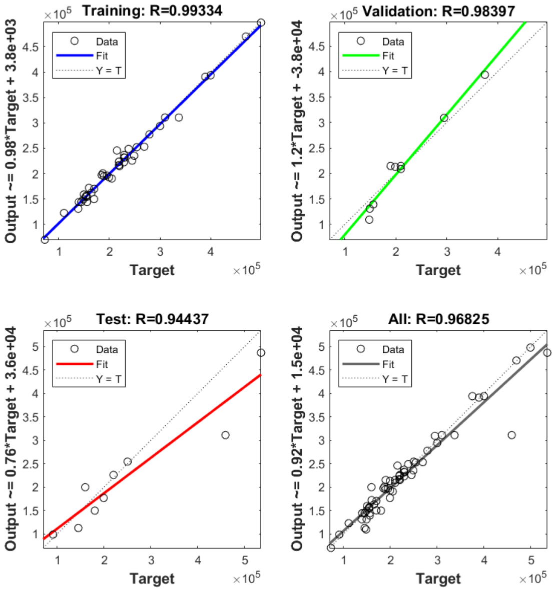
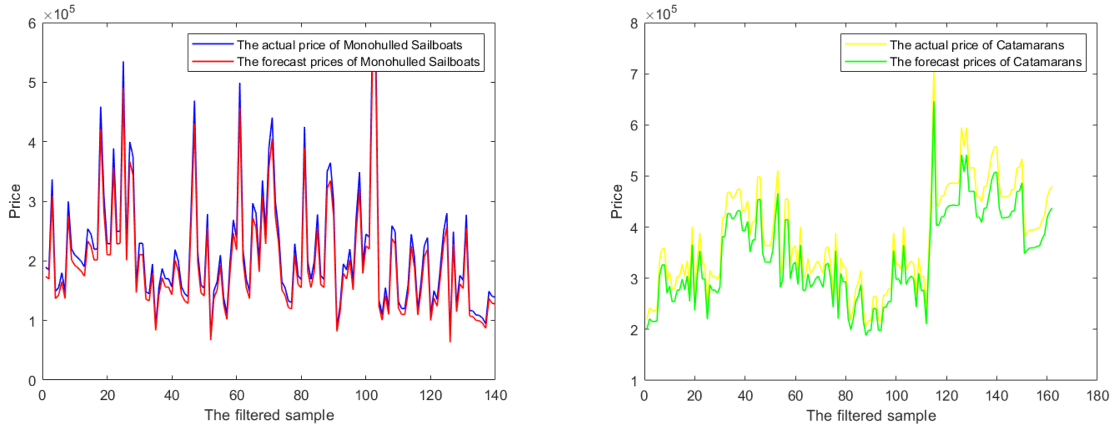


Figure 12 Fit Accuracy Plot for Monohulled Sailboats

●**Accuracy analysis**

To better demonstrate the BP neural network training results, we compared the actual and predicted prices of the screened monohull and catamaran sailboats, respectively, and drew the line graphs below. It can be seen from the graph that the predicted prices are almost identical to the actual prices, which again proves that the model has very high accuracy and is a guide for the prediction of sailboat market prices.



(a)Monohulled Sailboats **(b)Catamarans**

Figure 13 Comparison chart of true and predicted values

6 Regional impact analysis based on statistical modeling

Statistical modeling refers to building statistical models and exploring bulk data processing using various statistical analysis methods with computer statistical analysis software as a tool. Statistical modeling can reveal the factors behind the data, interpret socio-economic phenomena, or make predictions or judgments about economic and social development. For this question, we first conducted a fundamental statistical volume analysis of the prices of two types of hulls by geographic region. One-way and double-way ANOVA was used to explore the effect of region on prices and derive the actual and statistical significance of regional effects.

6.1 Fundamental statistical analysis of two types of hull prices

First, we conducted a fundamental statistical analysis of the two sailing prices in different geographic regions, including basic statistics such as mean, median, and standard deviation. The results are shown in the figure below.

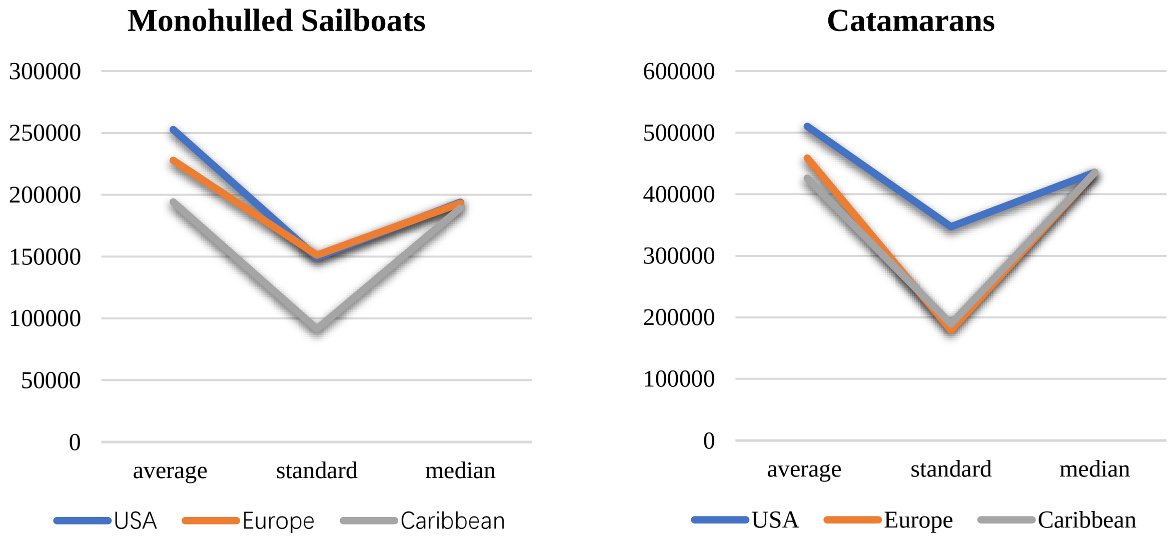


Figure 14 Description of the basic statistics of the two sailboat prices

From the figure, we can see that monohull and catamaran sailboats have a significant difference in the mean prices depending on the region. At the same time, there is no significant difference in the median. For monohull sailboats, Europe and the United States have similar standard deviations, indicating a similar degree of price volatility. In contrast, the Caribbean has a more stable degree of price volatility. For catamaran sailboats, Europe and the Caribbean have a similar degree of price volatility and are lower than the U.S. Based on this. We initially explored the existence of the influence of different regions on prices.

6.2 Analysis of variance

ANOVA is mainly used to verify whether there is a significant difference between the means of two groups of samples or more than two groups of samples. In the test, the index under investigation is called the test index, and the conditions affecting the test index are called factors. Factors can be divided into two categories. One is the number of artificially controllable measurements and the uncontrollable random factors. If only one-factor changes during the test, it is called a single-factor test; if more than one-factor changes, it is called a multi-factor test. In this question, a single-factor ANOVA is conducted for the re-gion and price of the two hulls, respectively, and the specific steps are as follows.

6.2.1 Region-based one-way ANOVA

In this problem, one-way ANOVA was conducted for the regions and prices of the two hulls separately, and the steps are as follows.

(1)Establishing assumptions

The hypothesis to be tested is called the original hypothesis, denoted as H_0, and the hypothesis opposed to H_0 is called the opposing hypothesis, denoted as H_1. To compare whether the means of each aggregate is consistent is to test whether the means of each aggregate is equal, and let the mean of the ith aggregate be μ_i, then:

The hypothesis test is H_0: $\mu_1 = \mu_2 = \ldots = \mu_c$.

The alternative hypothesis is that H_1: μ_1, μ_2, ..., μ_c are not all equal.

If H_0 holds, there is no significant difference between the c totals. This indicates that factor B has no significant effect on the indicator, all X_i can be considered as coming from the same aggregate, and random factors only cause the differences among individual X_i;

If H_0 does not hold, then in the total deviation, in addition to the differences caused by random factors, it also includes the differences generated by the effect of different levels of factor B. If the differences arising from the action of different levels of factors are much more significant than those caused by random factors, factor B is considered to have a significant effect on the indicator; otherwise, it is considered to have no significant effect.

(2)Calculate the sample mean and sample variance.

$$\overline{x_j} = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$
$$S_j{}^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \overline{x_j})^2}{n_j - 1} \tag{12}$$

Where $\overline{x_j}$ is the sample mean at the jth level, x_{ij} is the i-th value at the jth level, n_j is the sample volume at the jth level, and $S_j{}^2$ is the sample variance at the jth level.

(3)Calculate the variance between groups

The between-group variance, denoted as MSB, is the mean square of the B factor with:

$$MSB = \frac{\sum_{j=1}^{c} n_j(\overline{x_j} - \bar{\bar{x}})^2}{c-1}$$
$$\bar{\bar{x}} = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n_j} x_{ij}}{n_T} \tag{13}$$

where $\sum_{j=1}^{c} n_j(\overline{x_j} - \bar{\bar{x}})^2$ is the horizontal sum of squares, denoted as SSB, $c-1$ is the degree of freedom of SSB, $\bar{\bar{x}}$ denotes the total sample mean, and n_r denotes the sum of each sample capacity.

(4)Estimation of within-group variance

The within-group variance is denoted as MSE, and the MSE is calculated as

$$MSE = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})^2}{n_r - c} \tag{14}$$

Where $\sum_{j=1}^{c}\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})^2$ is called the sum of squared error terms and is denoted as SSE. $n_r - c$ is the degree of freedom of SSE.

(5)Construct the F-statistic for testing

$$F = MSB/MSE \sim F(c-1, n_T - 1) \tag{15}$$

Suppose the c overall means are not equal. In that case, the between-group variance MSB will be greater than the within-group variance MSE, and H_0 can be rejected when the value of F is large enough to reach a particular critical value. The size of the critical value is determined by the given α and degrees of freedom. So, when the significance level is given as α, the rejection domain of F is $F > F\alpha(c-1, n_T - c)$.

6.2.2 Analysis results and interpretation

For monohull and Catamarans in different geographic regions, the results of the analysis of variance of the region on sale price are shown in the table below.

Table 7 One-way ANOVA results for Monohulled Sailboats

Variable Name	Variable Value	Sample size	Average value	Standard deviation	F	P
Listing Price(USD)	Europe	1772	227788.274	151157.162	10.016	0.0002
	Caribbean	177	194134.695	91512.105		
	USA	384	252737.594	149198.949		

Table 8 One-way ANOVA results for Catamarans

Variable Name	Variable Value	Sample size	Average value	Standard deviation	F	P
Listing Price(USD)	Europe	689	458689.298	179024.322	6.551	0.001
	USA	90	510222.378	347214.161		
	Caribbean	294	426012.296	189295.679		

From the table, a one-way ANOVA with p-values less than 0.05 for both hulls, classified by geographic region, indicates a significant effect of region on the listing price and that this effect is consistent across all sailboat variants.

7 Simulation of the Hong Kong market

The question tests our model using data from Hong Kong and discusses the impact of the

Hong Kong region on the price of a subset of sailboats. We will validate the precision as well as the accuracy of our model by looking at the forecasts for the Hong Kong market.

7.1 Determining a subset of sailboats

For the monohulled boats, we have examined the Oceanis range of sailboats sold to Europe by Beneteau as a subset of our range. For the catamarans, we include as our subset the 450 series of sailing boats sold to Europe by Lagoon. This is because the previous statistical analysis showed that the variants of these two manufacturers have a significant share of the European market and are, to some extent, representative of the characteristics of the sailing market in the given region.

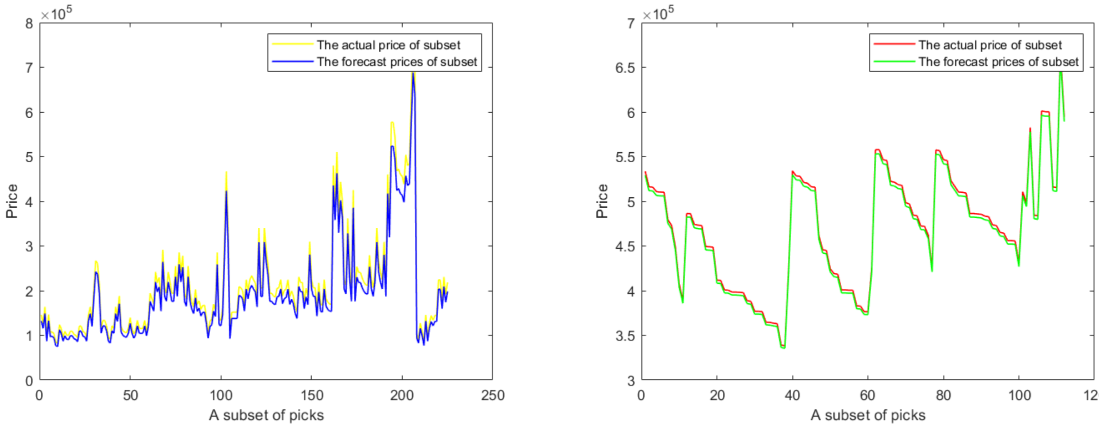
7.2 Data Collection

Based on the analysis of the first two questions, we collected data about Hong Kong's GDP per capita, coastline length, number of essential ports, and average cargo throughput, as well as the Oceanis series of sailboats and 450 series of sailboats located in Hong Kong that we found on the Internet as the objects we wanted to simulate.

7.3 Examining the Hong Kong second-hand ship market

● step1 Input the filtered subsets to the BP neural network

Problem 1 has already found the optimal parameters of the BP neural network. We input a subset into the network to get the predicted values of the two subsets, and the results of their comparison with the actual values are shown in the figure below.



(a)subset of Monohulled Sailboats **(b)subset of Catamarans**

Figure 15 Subset prediction results

● step2 Using Hong Kong data for forecasting

We input all the data indicators of Hong Kong into the neural network and explore the impact of Hong Kong on a subset of price regions by comparing the predicted results with the actual results. The prediction results are shown in the table below.

Table 9 Hong Kong Region Simulation Results

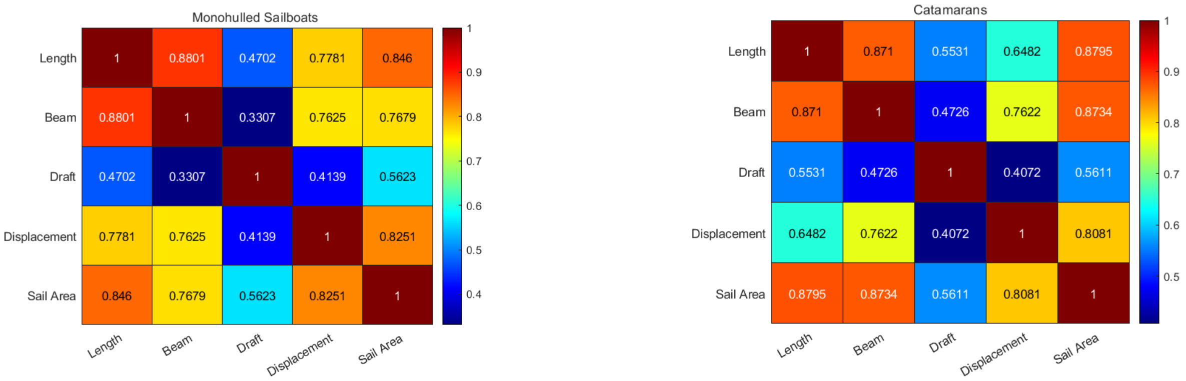
	Real price	Forecast prices	relative error
Oceanis 38	$133215	$114598	13.9%
450	$685000	$569981	12.8%

By comparison, our BP neural network model predicts prices more accurately for a given subset. However, when we input the data from Hong Kong into the BP neutral network, our prediction results have a deviation of about 10%. This suggests that Hong Kong has a regional effect influence on the pricing of both sailboats. Ultimately, we conclude that for both monohull and catamaran sailboats, prices are generally higher in Hong Kong, with a more robust regional effect.

8 Data Mining

8.1 Information mining on the characteristics of the sailboat itself

In order to explore the characteristics of the sailboat itself, we collected data about the length of the sailboat, the length of the beam, the displacement, the draught depth, and the sail area, and divided into Catamarans and Monohulled Sailboats for correlation analysis, and the correlation matrix heat map is shown below, from which it can be seen that for both sailboats, the length of the boat and the length of the beam, the sail area, the draught depth of the boat and the sail area show an apparent positive correlation, which shows that Monohulled Sailboats and Catamarans follow similar technical guidelines in their design, which is in line with the technical needs of sailboats in specific environments.



Monohulled Sailboats

	Length	Beam	Draft	Displacement	Sail Area
Length	1	0.8801	0.4702	0.7781	0.846
Beam	0.8801	1	0.3307	0.7625	0.7679
Draft	0.4702	0.3307	1	0.4139	0.5623
Displacement	0.7781	0.7625	0.4139	1	0.8251
Sail Area	0.846	0.7679	0.5623	0.8251	1

Catamarans

	Length	Beam	Draft	Displacement	Sail Area
Length	1	0.871	0.5531	0.6482	0.8795
Beam	0.871	1	0.4726	0.7622	0.8734
Draft	0.5531	0.4726	1	0.4072	0.5611
Displacement	0.6482	0.7622	0.4072	1	0.8081
Sail Area	0.8795	0.8734	0.5611	0.8081	1

Figure 16 Heat map of the correlation matrix between the characteristics of the two sailboats

8.2 Digging information on the sales status of major manufacturer variants

As with other products, large manufacturers tend to be market-oriented, so looking at the pricing of popular sailboat models from large companies makes sense. For Monohulled boats, we looked for popular sailboat models from the Jeanneau, Beneteau, and Bvaria manufacturers, whose models and average selling prices are shown in the table below.

Table 10 Average pricing of popular models by major manufacturers

Monohulled Sailboats			Catamarans		
Jeanneau	Sun Odyssey	$204648	Lagoon	450	$495647
Beneteau	Oceanis	$212950			
Bavaria	Cruiser	$164352			

For monohulled sailboats, the pricing range for the major manufacturers is $160,000 to $200,000. For Catamarans, it is $500,000, indicating that catamarans are more expensive to build than Monohulled Sailboats.

8.3 Other interesting information

As a luxury item, there is a relationship between the price of a boat and the region where it is sold. When two boats in identical conditions are sold to different regions, the prices still differ after converting the exchange rate. We selected two monohull manufacturers, Beneteau and Jeanneau, and three catamaran manufacturers, Lagoon, Fountaine, and Leopard, respectively. The average selling price of each manufacturer in each region was calculated.

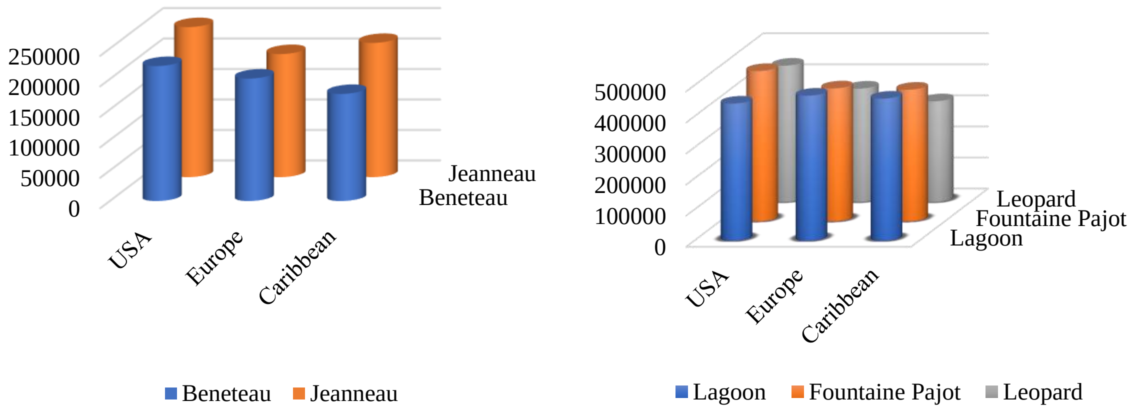


Table 11 Statistical analysis chart of the sales of the two major large manufacturers of sailboats in each region

It can be seen that except for Lagoon, which is the most expensive catamaran sold in Europe, the other four significant manufacturers sell their boats at the highest price in the United States. Except for Jeanneau, which sells the cheapest monohull in Europe, the other four significant manufacturers sell the cheapest boats in the Caribbean. Based on this, we can see that the pricing strategies of the famous brands are different for the regions

9 Model Evaluation and Further Discussion

9.1 Strengths

1. We use the BP Neural Network Model. Compared with other models, such as the Arima and the gray prediction models, the model can adequately approximate arbitrarily complex nonlinear relationships. At the same time, the model has a solid ability to synthesize information, which can handle both quantitative and qualitative information, and coordinate multiple input information well, which is suitable for the problematic situation of judging the relationship between indicators in this problem.

2. Using the parameters of the sailboat itself as the pricing indicator is not realistic. We consider the more common influencing factors in life and use performance as the guide to constructing a pricing model, which makes the model more realistic and convincing.

3. When constructing a decision model, by using both historical data and additional data for decision doing work, the model is more reliable than using When constructing a decision model, by using both historical data and additional data for decision-making work, the model is more reliable than using one type of data alone.

4. The results are intuitive and concise when interpreting raw data using data visualization techniques.

5. Our pricing model is widely applicable and can help brokers grasp the pricing standards of the Hong Kong second-hand sailboat market.

9.2 Weaknesses

1. Neural networks are "black-box" in nature, meaning we cannot know how and why they produce a specific output.

2. Due to limited search data, the indicators used cannot fully describe the used boat prices, which may reduce the accuracy of our model.

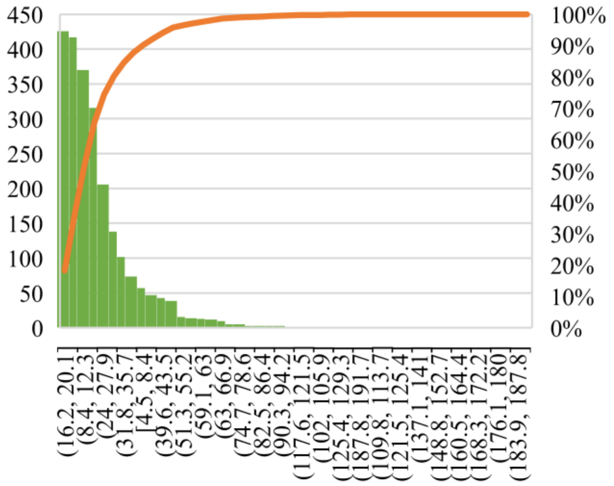
Dear Hong Kong Sailboat Broker:

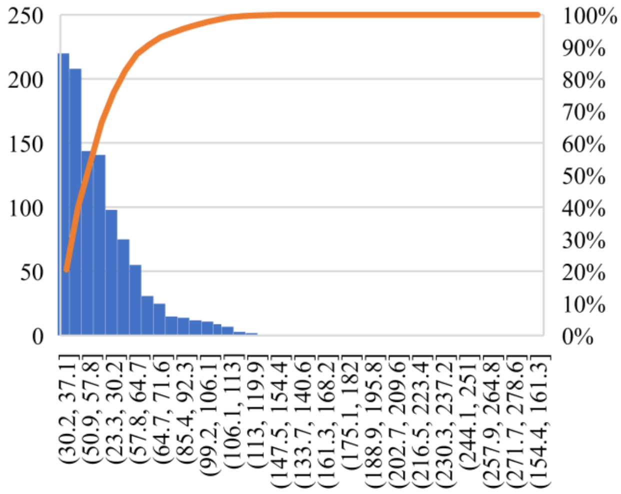
In today's increasingly competitive market, a better understanding of the Hong Kong market and the development of appropriate pricing solutions are critical to increasing your sailboat turnover and profits. In response to the request, our team here is pleased to have the opportunity to present our research findings and related recommendations, which will give you some insights on pricing. We believe several points need to be considered when setting prices for sailboats in Hong Kong:

1. Focus on the properties and performance of the sailboat itself.

In addition to the characteristics of sailing boats, such as length and age in the table, we additionally selected the metrics of sailing boat properties, including beam length, draft depth, sail area Horsepower, and displacement, and based on these metrics, we obtained the metrics that reflect the performance of the boat by calculation, including SA/D, Bal/Disp, Disp/Len, and Comfort Ratio. Using these indicators, we developed a mathematical model using a bp neural network to predict the price of a monohull and catamaran. The result is that the predicted price is almost the same as the actual price, indicating that all these indicators strongly correlate with the price, so your pricing must be based on the boat's characteristics.

We also found that boat designs tend to follow similar technical guidelines. You can use our metrics to consider the use of the boat and then perform a pricing analysis.





2. Focus on the characteristics of the Hong Kong region itself.

Using statistical analysis, we found that the average price of monohull and catamaran sailing boats varies by region, with each region experiencing different price fluctuations. A one-way ANOVA indicates a significant effect of region on the listed price. We examined the relationship between specific characteristics of regions and listing prices separately. Under our generalization, the characteristics of regions that affect ship prices include Per capita GDP, coastline length, number of major ports, and average cargo throughput. by geographic modeling, the relationship between these characteristics and prices corresponds to the case of Hong Kong, we got a more reasonable price for the ship in the Hong Kong market. Taken together, we find that the prices of ships in Hong Kong are slightly higher than in the three known regions. By comparison, we find that the estimated prices are very close to the searchable sailboat listing prices in Hong Kong, justifying our pricing recommendation to you.

3. Focus on the manufacturers who make sailboats.

Large manufacturers and popular models tend to guide pricing in the marketplace. When you sell a boat with little or no previous history of sale, you will struggle with reasonable pricing. We have found that monohulls and catamarans from more prominent manufacturers have a range of prices and are somewhat market-oriented. You can compare the situation of popular boats with a lot of available data with the situation of niche boats you want to sell and decide on a reasonable price.

4. Focus on the regional effects in the areas where sailboats are made.

As a luxury item, the price of a boat is often related to the characteristics of the region in which it is located. After research, we found that the pricing strategies of the major manufacturers are different for the region. Suppose a customer from outside Hong Kong wants to entrust you with a boat to sell in Hong Kong. In that case, we recommend that you adjust your selling price in Hong Kong according to the manufacturer's local pricing strategy to ensure that you can maximize your profit.

Besides, after our research, we found that the average boat price difference between the Caribbean and Hong Kong is the largest. After calculating the tariffs and freight costs, you can sell your boat from the Caribbean at a lower price to Hong Kong at a higher price and profit from the more significant difference.

These are the suggestions and strategies that our team offers to you. Thank you again for taking the time to read our suggestions.

We hope our model and these suggestions will be helpful to you!

Sincerely,
MCM Team Members

References

[1].https://sailboatdata.com/

[2].https://www.yachtworld.co.uk/

[3].https://www.hongkongyachting.com/

[4].Ge Zhang, Tianxiang Luo, Witold Pedrycz, Mohammed A El-Meligy, et al. Outlier processing in multimodal emotion recognition[J], IEEE Access 8, 55688-55701, 2020.

[5].Ali Azhar, Method for Estimating Price of Second hand Ship with Multi Method[J], IOP Conference Series: Materials Science and Engineering 1052 (1), 012011, 2021.

[6].Kai Cui, Xiang Jing, Research on prediction model of geotechnical parameters based on BP neural network[J], Neural Computing and Applications 31, 8205-8215, 2019.

[7].Charles M Judd, Gary H McClelland, Carey S Ryan, Data analysis: A model comparison approach to regression, ANOVA, and beyond[M],Routledge, 2017.

[8].Won Chang, L Alan Winters, How regional blocs affect excluded countries: The price effects of MERCOSUR[J], American Economic Review 92 (4), 889-904, 2002.

[9].William Seabrooke, Eddie CM Hui, William HK Lam, et al. Forecasting cargo growth and regional role of the port of Hong Kong[J], Cities 20 (1), 51-64, 2003.