Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2332502

A Neural Network Prediction Model for Used Sailboat Prices Based on Factor Analysis

Summary

The price of used sailboats is determined by multiple factors. This article applies factor analysis, correlation analysis, and BP neural network regression models to achieve price prediction of used sailboats, which can serve as an important reference for sailboat brokers.

For task 1, We began by expanding the original dataset and performing data cleaning. Then we converted the qualitative variables into dummy variables and formed a numerical matrix with the quantitative variables. We reduced the dimensions of the matrix by conducting factor analysis. We then conducted correlation analysis between the original variables and prices to determine the important influencing factors. We combined the factor scores from factor analysis to explain the contribution of each major factor to pricing.

Based on the above two steps, we used the reduced numerical matrix and price vector as inputs and built a neural network model with 10 hidden layers based on the LM algorithm. Finally, we used the original sailboat variants and parameters from the table as input and predicted the prices of the sailboat variants in the provided spreadsheet using this model. Through data analysis and comparison, the average R value obtained was 0.7123.

For task 2, we extracted sailboat variants with price records in all three regions and constructed three data subsets with the brand-variant name as the row and each of the three regions as the column. Then, we performed neural network regression on these three subsets separately and obtained three price prediction neural networks that only included the regional change factor. Finally, we used these three neural networks to predict the prices of the table data separately and used F-test, data visualization and other methods. We concluded that the regional effect was consistent in most sailboat variants, while factors such as regional preference needed further analysis and discussion.

For task 3, we first collected GDP, port number and climate data in Hong Kong as input parameters, then used neural network to obtain predicted data, and compared it with actual data collected. We calculated the R-square data and concluded that the data fitting was high. Secondly, we collected data and analyzed how GDP, port number and climate conditions affected prices. Finally, by comparing and analyzing the predicted data of monohull and catamaran, we concluded that the regional effect had a smaller impact on catamarans.

For task 4, we used exploratory data analysis methods and various data visualization software to conduct multidimensional analysis of the data, and extracted significant features such as sailboat prices increasing with age and personal customizations and brand effects promoting prices.

Keywords: Used sailboat price, BP Neural Network, Factor Analysis,

Contents

1 Introduction 4
1.1 Problem Background 4
1.2 Restatement of the Problem 4
1.3 Problem analysis 4
1.4 Our Working 6
2 Assumptions and Justifications 6
3 Notations 7
4 Key feature extraction based on factor analysis and Chi-square test 7
4.1 Data Description 7
4.2 Factor Analysis 9
4.3 Correlation Analysis and Chi-square Test 11
4.3.1 Analysis Process 11
4.3.2 Result 11
5 Price Prediction Model Based on BP Neural Network Regression. 11
5.1 The Foundation of Model 11
5.2 Algorithms used in the model 12
5.3 Training Process 13
5.4 Model Accuracy Analysis 13
5.5 Result Analysis 16
5.5.1 Task 1: The Accuracy of Price Estimates for Each Variant of Sailboat. 16
5.5.2 Task2: The Impact of Region on Prices and Its Practical Significance 16
5.5.3 Task 3: The role of Hong Kong region and its impact on monohull and catamaran boats. 19
6 Task 4: Exploratory Analysis 21
6.1 Analysis principle 21
6.2 Analysis Process 21
6.3 Inference and Conclusion 21
6.3.1 Sailing Fever. 21
6.3.2 Depreciation 22
6.3.3 Custom-made. 22
6.3.4 Brand Effect. 23
6.3.5 Country Of origin Effect(COO). 23
7 Sensitivity Analysis 23

8 Model Evaluation and Further Discussion 24
8.1 Strengths 24
8.2 Weaknesses 24
8.3 Further Discussion 25
9 Conclusion 25
References 25
Appendices 26

1 Introduction

1.1 Problem Background

Sailboats have always been popular and widely used in many fields such as leisure, sports, scientific research, and commercial activities. With the rise of the sailboat trading industry, the price of used sailboats has gradually become a hot topic for people. In fact, like many luxury goods, it will also change with many factors such as age and market demand. used sailboats are often traded through brokers as intermediaries between sellers and consumers. In-depth exploration and understanding of the price of used sailboats can better understand the used sailboat market and provide sailboat brokers with effective references and inspirations. In addition, it can also bring inspiration to shipbuilders, parts suppliers, ship sellers, and consumers, thereby further promoting the development of the used sailboat industry and the improvement of the industry chain.

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:

- Use the official dataset and the data we searched for to design a model to explain the listed prices of the sailboats provided in the table, and discuss the accuracy of the price estimation.
- Build a model and explain the impact of the sales region on the listed price.
- Apply our established model in the Hong Kong market. Select a suitable dataset and collect comparable listed price data in the Hong Kong market. Evaluate the regional impact of Hong Kong on sailboat prices in the selected dataset and analyze the relationship between catamarans and monohulls.
- Share other interesting and meaningful conclusions.
- Write a report for the sailboat brokers in Hong Kong Special Administrative Region to help them better understand the sailboat market.

1.3 Problem analysis

For task 1:

Task 1 requires us to establish a model to explain the listing prices of each sailboat in the spreadsheet and discuss the accuracy of our price estimates for each type of sailboat. Due to issues such as missing data and a small data set in the original data, we first need to obtain data from reliable websites, expand the data set, and clean the data, converting qualitative variables such as region and variant into dummy variables to construct an analyzable data matrix. Due to the large number of relevant factors and high matrix dimensions,

dimensionality reduction is necessary. After the reduced factors are named and identified, we also need to conduct a correlation analysis between each variable and the price.

To discuss the accuracy of our price estimates for each type of sailboat, we need to establish a model to predict the price of the sailboat. First, we need to analyze the collinearity relationship between variables and select a regression model with minimal influence from multicollinearity. To improve the accuracy of the model prediction, a BP neural network regression model is considered. By comparing the calibrated price with the predicted price, we can obtain relevant conclusions.

For task 2:

Task 2 requires us to analyze the impact of regions on the listing price. We can control other variables to observe the data distribution of price indicators by changing the regional factors. The sailboat variant data set recorded in all three regions is divided and three price-prediction BP neural networks are trained and established respectively. Then, the data set is predicted using three neural networks respectively, and the three prices for each sailboat variant using different regional data training neural network models can be obtained and compared to derive conclusions.

For task 3:

Task 3 requires us to analyze how regional factors are applied in the Hong Kong region in the established model, and based on the selected data for each individual sailboat, analyze the factors influencing the single-hull and catamaran sailboats and determine whether the impact is consistent. First, we find the data related to the region in the model established for the Hong Kong region, use the neural network to obtain the predicted data, and compare it with the actual data collected to draw conclusions. Next, we collect information about regional factors and analyze how these factors influence the price. Finally, we compare and analyze the predicted data of single-hull and catamaran sailboats to draw conclusions.

For task 4:

Task 4 requires us to derive interesting and insightful conclusions from the data. We use exploratory data analysis methods and various data visualization software to conduct multidimensional analysis of the data and extract significant features to derive conclusions.

For task 5:

Task 5 requires us to prepare a report with carefully selected graphics for Hong Kong brokers. For this task, we need to present the above conclusions in a straightforward and understandable manner, accompanied by illustrations.

1.4 Our Working

Our main workflow is shown in the following diagram.

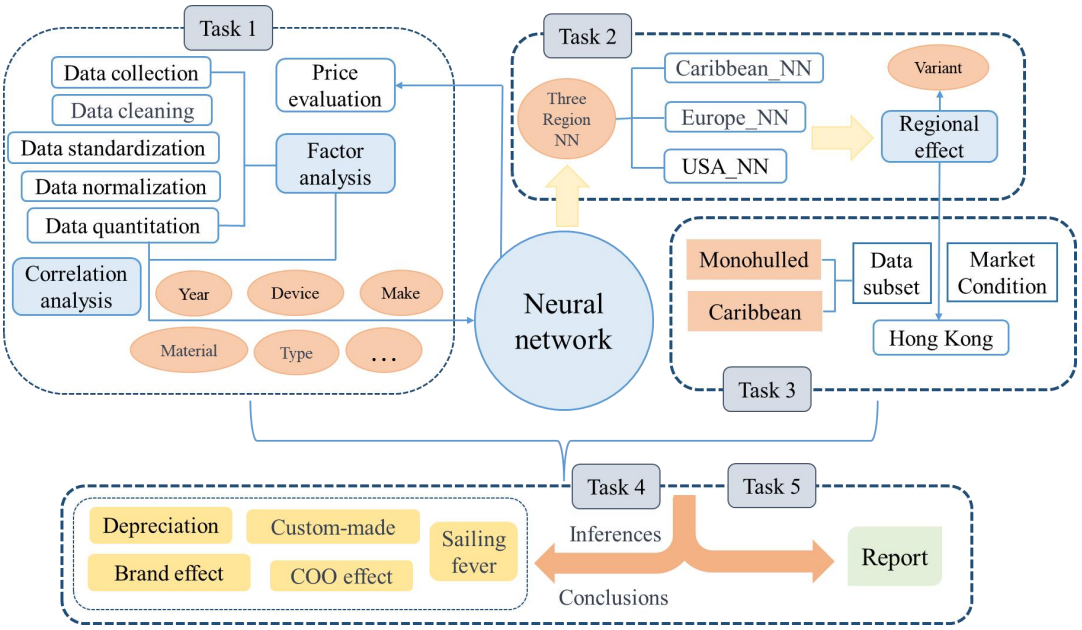


Figure 1: Our Work

2 Assumptions and Justifications

Based on the adequate analysis of problems, we make the following properly justified assumptions to simplify our models.

1. We assume that the listing price of sailboats is not influenced by the valuations of the pricing party.

   --This is because the listing price is primarily regulated by the market and is minimally affected by subjective factors.

2. We assume that sailboats of the same variant can be replaced by their average price.

   --This is because sailboats of the same variant are almost identical in objective conditions, and price changes are not too significant, thus it is reasonable to estimate using the average price.

3. We assume that prices are not influenced by cultural background factors.

   --This is because cultural background factors are difficult to quantify in the model.

3 Notations

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

Table 1: Notations used in this paper

Symbol	Description
Y	Manufacturing year
M_a	Make
M_t	Material
R_g	Region
D_v	Device
T_p	Type
N	The number of samples
X_i	Characteristic factor of predicted price
f	Neural network hidden layer
w_{ij}	The weight from node i to node j
O_i	Outputted predicted price number

4 Key feature extraction based on factor analysis and Chi-square test

4.1 Data Description

For the purpose of obtaining more accurate results, we enriched our dataset and performed necessary data processing and operations based on the following three steps:

1. **Data collection:** Based on the provided dataset, we searched for additional data using websites such as https://sailboatdata.com/ (please refer to Appendix 1 for details), according to the proportions of the regions included in the provided dataset.

We collected over 20,000 extra data points to expand our dataset, and added other features such as beam, draft, displacement, S.A., material, and GDP. Finally, we obtained a dataset with 32,984 data points and rich features.

2. **Data cleaning:** In the dataset, some feature values are missing or abnormal. For example, in the official dataset, there are 3 data missing their Country/Region/State information, and the data volume is small. We filtered them and deleted them from our worksheet. For those data with unusually high prices, we consulted relevant materials to determine whether their prices were abnormal. For example, in the official dataset, the price of HH 55_Catamarans is 2,890,000$, which is tens of times higher than many other catamarans. We doubted its correctness, but after consulting the materials, we found that it was indeed worth this price, so we kept it. For data with the same features except for the price, we replaced the price with the average value.

3. **Data quantitation:** Since some features in the dataset are qualitative, such as Make, Variant, Region, and State, but they still have a significant impact on our model building, we used dummy variables to quantify them.

4. **Data standardization:** We use the Z-Score normalization method to standardize the features.

   Standard deviation formula ：

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x-\mu)^2} \tag{1}$$

   z-score standardized conversion formula ：

$$Z = \frac{x-\mu}{\sigma} \tag{2}$$

   Through the above steps, we have transformed the numerical scales of each feature to be the same, eliminated the dimensional differences between the features, so as to better compare their weights and contribution rates.

5. **Data normalization:** We perform data regularization to encourage the model to produce sparse weight vectors, which helps with feature selection.

   $L1$ regularization formula:

$$L = E_{in} + \lambda \sum_{j} |w_j| \tag{3}$$

4.2 Factor Analysis

Factor analysis is a statistical method for analyzing multidimensional variables based on the correlation between variables and reducing them to a small number of common factors. [1] After data processing such as data quantification, we obtained 539 factors.

We first conducted a KMO and Bartlett's Test and found that the KMO value was 0.813 (close to 1), and the significance level of Bartlett's Test was 0 (less than 0.05), indicating that the data factors had strong correlation and were suitable for factor analysis.

Table 2：KMO and Bartlett's Test

KMO Measure		.813
Bartlett's Test	Asymptotic Chi Test	21207.854
	Degree Of Freedom	180
	Significance	.000

We calculated the eigenvalues of the correlation coefficient matrix R, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{539} \geq 0$, and the corresponding eigenvectors, $\mu_1, \mu_2, \cdots, \mu_{539}$, and for μ_j, $\mu_j = (\mu_{1j}, \mu_{2j}, \cdots \mu_{Nj})^T$

The primary loading matrix：

$$A = \left[\sqrt{\lambda_1}\mu_1 \sqrt{\lambda_2}\mu_2 \cdots \sqrt{\lambda_{539}}\mu_{539}\right] \quad (1)$$

By combining the scree plot and the total variance explained plot, we found that the contributions reached 60%, 85%, 22 components. After considering the trade-offs, we chose 22 components to ensure minimal information loss.

We conducted factor analysis on the dataset using SPSS software, and according to the scree plot and the sum of squared loadings (ESS) table after rotation, we found that the contribution reached 70.233% at the 161st component.

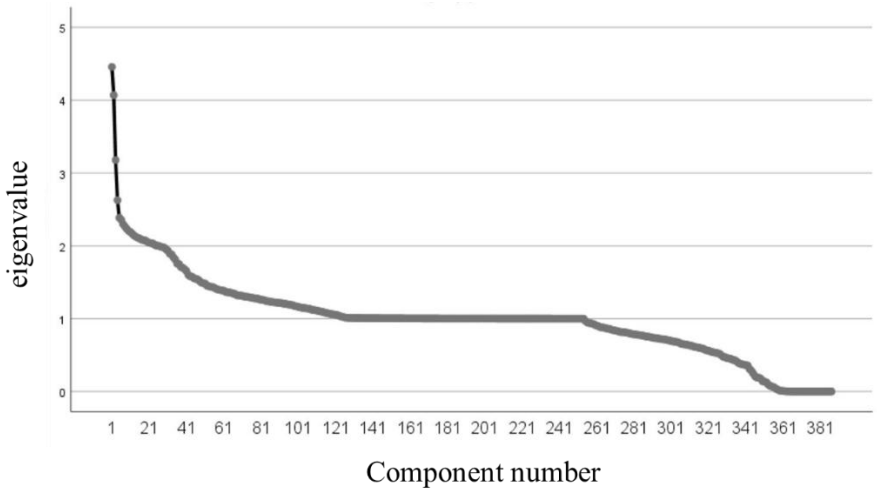


Figure2：Gravel Fig

Table 3：Extraction Sums of Rotating Squared Loadings

Component	Initial Eigenvalues of Variance	Cumulative
1	8.057	8.057
2	3.690	11.747
3	2.957	14.704
4	2.456	17.160
5	2.257	19.417
6	1.991	21.409
7	1.808	23.216
8	1.574	24.790
9	1.482	26.272
10	1.381	27.653
…	…	…
158	0.261	69.451
159	0.261	69.712
160	0.261	69.973
161	0.260	70.233

4.3 Correlation Analysis and Chi-square Test

4.3.1 Analysis Process

Correlation analysis is a statistical method used to measure the degree of relationship between two or more variables. Pearson correlation coefficient is applicable for measuring the linear correlation between two continuous variables. Spearman's rank correlation coefficient is applicable for measuring the non-linear relationship between two variables or data that generally does not satisfy the normal distribution assumption.

Correlation analysis can help us understand the relationship between variables and which variables may have a significant impact on the target variable of interest in model building or prediction. Correlation analysis can also be used to identify redundant variables, which are highly correlated with other variables, to reduce model complexity.

We conducted correlation analysis on quantitative variables using the SPSS software and used chi-square test to analyze the correlation between qualitative variables and price.

4.3.2 Result

The integrated results are shown in the following table:

Table 4: Correlation ranking of factors influencing the prices in the used sailboat market.

Factor	Initial Eigenvalues of Variance	Cumulative
M_t	0.7623	2
Y	0.9120	1
M_a	0.7625	3
R_g	0.7289	5
D_v	0.7313	4
T_p	0.6812	6

5 Price Prediction Model Based on BP Neural Network Regression.

5.1 The Foundation of Model

Back Propagation Neural Network (BP Neural Network) is a common artificial neural network with strong learning ability and adaptability. [2]Its basic principle is to iteratively calculate the connection weights between neurons by using input and output data, in order to

achieve approximation and prediction of nonlinear patterns.

The basic structure of BP neural network consists of input layer, hidden layer and output layer. The input layer accepts external input data, the hidden layer is an intermediate layer composed of multiple neurons, and the output layer outputs the final result. During the training process, the network uses forward propagation to pass the input data from the input layer to the output layer, and then calculates the output error using the back propagation algorithm, and adjusts the connection weights between neurons according to the error. This process is repeated many times until the network error reaches the preset threshold or the maximum number of iterations is reached.

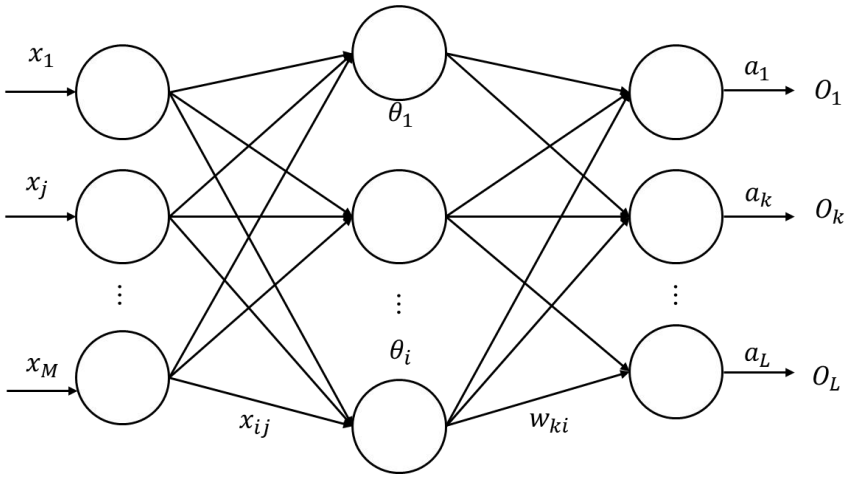


Figure 3: Three-layer Neural Network Structure

5.2 Algorithms used in the model

We used the Levenberg-Marquardt (LM) algorithm to train the neural network. LM algorithm (Levenberg-Marquardt algorithm) is a commonly used nonlinear least squares optimization algorithm, mainly used in the training process of neural networks.[3]

The LM algorithm is based on the following objective function:

$$\phi(x) = \frac{1}{2}||f(x) - p||^2 = \frac{1}{2}\sum_{i=1}^{m}[f_i(x) - p_i]^2 \quad (1)$$

Where $|\cdot|$ denotes the Euclidean norm, m is the number of residuals, and $f_i(x)$ is the predicted value of the i^{th} input of function f at x.

In the LM algorithm, we first use the Gauss-Newton method to obtain an initial value for x. Then, we calculate the gradient $g(x)$ and the Hessian matrix $H(x)$ at x. Next, we use the following formula to calculate the update Δx for x:

$$(H(x) + \lambda I) \Delta x = - g(x) \tag{2}$$

where I is the identity matrix and λ is a parameter that controls the step size. If λ is small, the update Δx is close to the Gauss-Newton method, and if λ is large, the update Δx is close to the gradient descent method. Therefore, the value of λ should be balanced between the two.

Finally, update the value of x:

$$x \leftarrow x + \Delta x \tag{3}$$

This process iterates until the error reaches the predetermined threshold or the maximum number of iterations is reached.

5.3 Training Process

1. We used the dimensionally reduced data with strong correlations obtained from factor analysis and correlation analysis as the column variables of the input matrix, and the number of sailboat species for different brands and variants as the row variables, to form the input matrix X. We also used the sailboats for different brands and variants as the row variables and the given prices as the column variables of the input matrix Y.
2. We selected 161 features with strong correlation to sailboat prices based on the conclusion of correlation analysis, and divided the additional 32,984 data into a 70% training set, a 15% validation set, and a 15% test set, then constructed a neural network with one input layer, ten hidden layers, and one output layer.
3. We use MATLAB software to program and implement the above operations. Training will continue until a pre-defined stopping condition is met (Epoch greater than or equal to 1000 or the validation error of the iteration continues to increase).

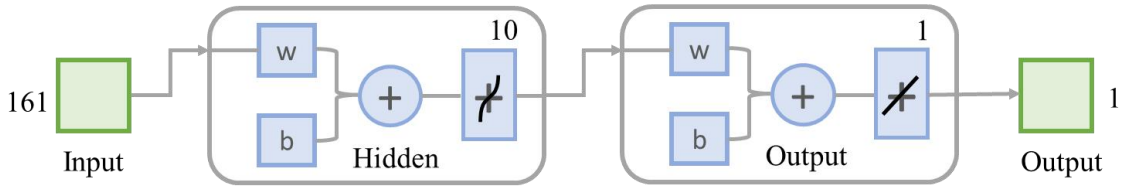


Figure 4: Neural Network Construction

5.4 Model Accuracy Analysis

Linear regression plots were generated to display the corresponding (target) network predictions (output) for the training set, validation set, and test set. If the fit is perfect, the data should fall along a 45 degree line where the network output is equal to the response. It

can be seen from the figure below that the dataset fits very well. To obtain more accurate results, the dataset can be trained again with different initial weights and biases for the network, and the network can be improved after retraining.

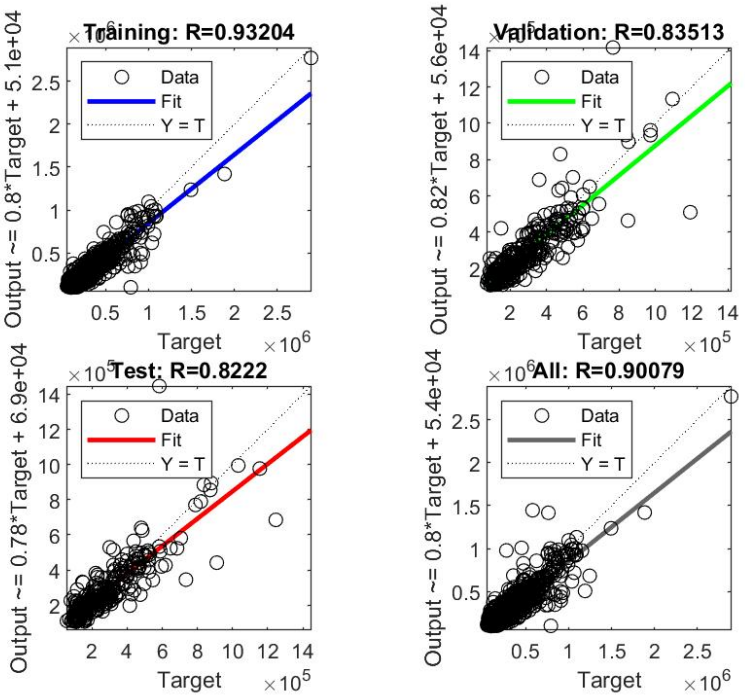


Figure 5: Regression Plot

The results of the mean square error (MSE) and R-value indicate that the network has high accuracy in predicting the prices of used sailboats, reflecting the efficiency of the network model.

Table 5: Parameter Index Chart

	Observations	MSE	R
Training	23088	1.8087e+10	0.7237
Validation	4948	1.7412e+10	0.6873
Test	4948	2.3298e+10	0.6736

This training used conditional stop training, which stopped at 6 epochs.

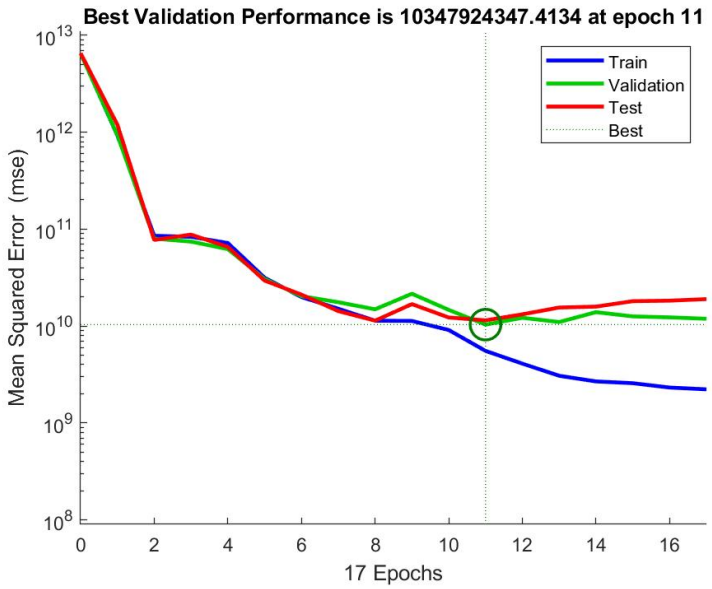


Figure 6: Optimal State Chart

View the error histogram for additional verification of network performance.

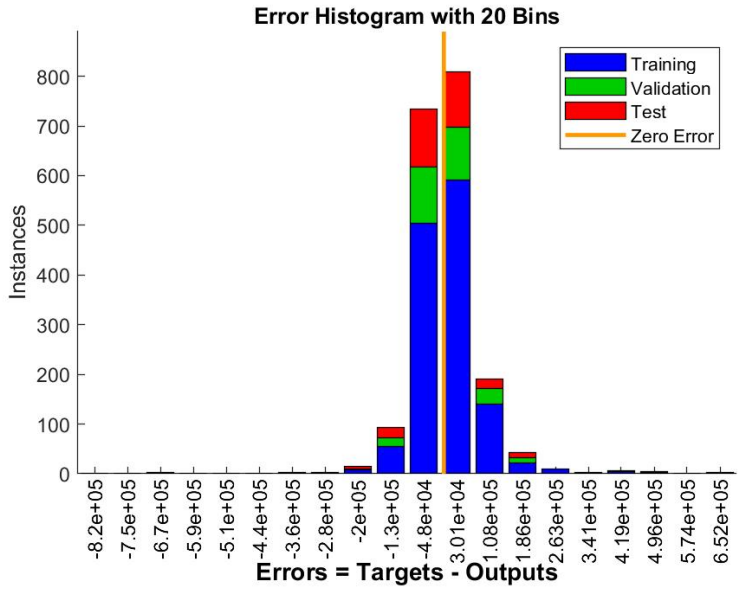


Figure 7: Error Histogram

The blue bars indicate training data, the green bars indicate validation data, and the red bars indicate test data. The histogram can indicate outliers, which are data points that fit significantly worse than the majority of the data. The outliers are examined to determine if the data are poor within the error range or if any of these data points differ from the rest of the data set, and the data points with valid outliers are interpolated using the network.

5.5 Result Analysis

5.5.1 Task 1: The Accuracy of Price Estimates for Each Variant of Sailboat.

We used a BP neural network model to predict prices for the brands and variants in the original table provided in the question. The following image is a visualization of the deviation between our estimated prices and the actual prices (due to the large number of variants, only a screenshot with limited length is displayed, the full image is available in the appendix 2), measured by the R-value and MSE.

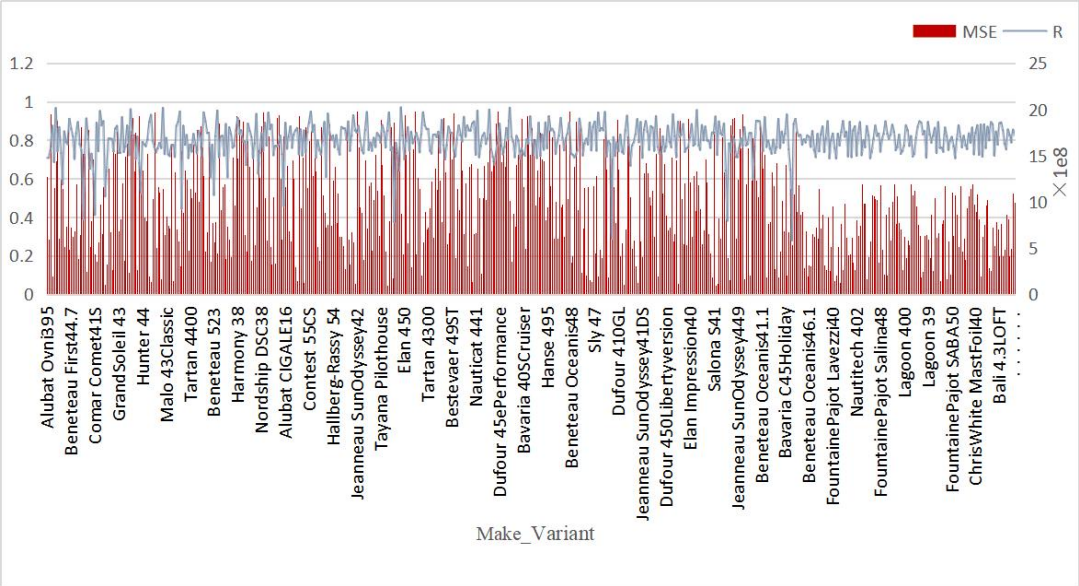


Figure 8: Estimated MSE and R-Squared Variance for Each Variant

The above figure shows that the R-value deviation between the model-predicted price and the actual price fluctuates around 0.7, and the MSE is controlled at the level of tens of millions. Therefore, the model's predictions are reasonably accurate. However, it is observed that for some variants, there is a large deviation between the model's predicted price and the actual price. Further investigation revealed that these variants were private custom sailboat brands, such as Baltic 67, Oyster Yachts 545, and some sailboats whose usage and age were unknown. Overall, the model provides reasonably accurate price predictions for different variants within an acceptable range of error. But for some specific variants, the accuracy deviation is significant.

5.5.2 Task2: The Impact of Region on Prices and Its Practical Significance

1. **Data set segmentation and processing:**

   We filtered out variants that are sold in all three major regions from the supplemented data set, and averaged the prices of multiple variants in the same region. This resulted in a total of 12,861 data points. Then, we segmented and integrated the three data points of each variant according to the region. After factor analysis and dimensionality reduction, we obtained 127 columns of factors as input variables for the three input matrices. Thus, the construction of the three input matrices is completed.

2. **Build three neural networks for each of the three major regions:**

Train a BP neural network on the processed data, constructing a neural network model with 1 input layer, 10 hidden layers, and 1 output layer. Divide the dataset randomly into training set, validation data, and test data, with percentages of 70%, 15%, and 15% respectively. The range of hidden layer nodes is calculated by Formula (1), where m represents the number of input layer nodes, n represents the number of output layer nodes, and a takes values between 1 and 10. Based on Formula (1), 10 selectable hidden layer node numbers can be obtained. Set these 10 hidden layer node numbers for the BP network in turn and calculate the mean squared error (MSE) of the training set. Finally, the hidden layer node corresponding to the minimum error is the optimal hidden layer node.

$$x = \sqrt{m+n} + a \tag{1}$$

After debugging and training, the neural network models trained on the three regions can predict prices that are close to the actual values.

Table 6: Model error in the Caribbean

	Observations	MSE	R
Training	9003	8.2267e+09	0.9130
Validation	1929	1.3856e+10	0.8035
Test	1929	1.0305e+10	0.8431

Table 7: Model error in the USA

	Observations	MSE	R
Training	9003	1.3269e+09	0.9815
Validation	1929	4.5275e+09	0.9252
Test	1929	1.9210e+10	0.8466

Table 8: Model error in the Europe

	Observations	MSE	R
Training	9003	1.3099e+09	0.9886
Validation	1929	9.5413e+09	0.8328
Test	1929	8.7025e+09	0.8427

After obtaining the three neural network models mentioned above, we removed the regional variable from the original data and performed dimensionality reduction using factor analysis, resulting in 127 factors. We used the matrix of row and column variables

as input and used the three regional neural networks we established to predict prices, resulting in three price prediction datasets:U_{pred}, E_{pred} and C_{pred}.

3. Analysis of F-test results

Draw boxplots for the three obtained datasets for overall analysis:

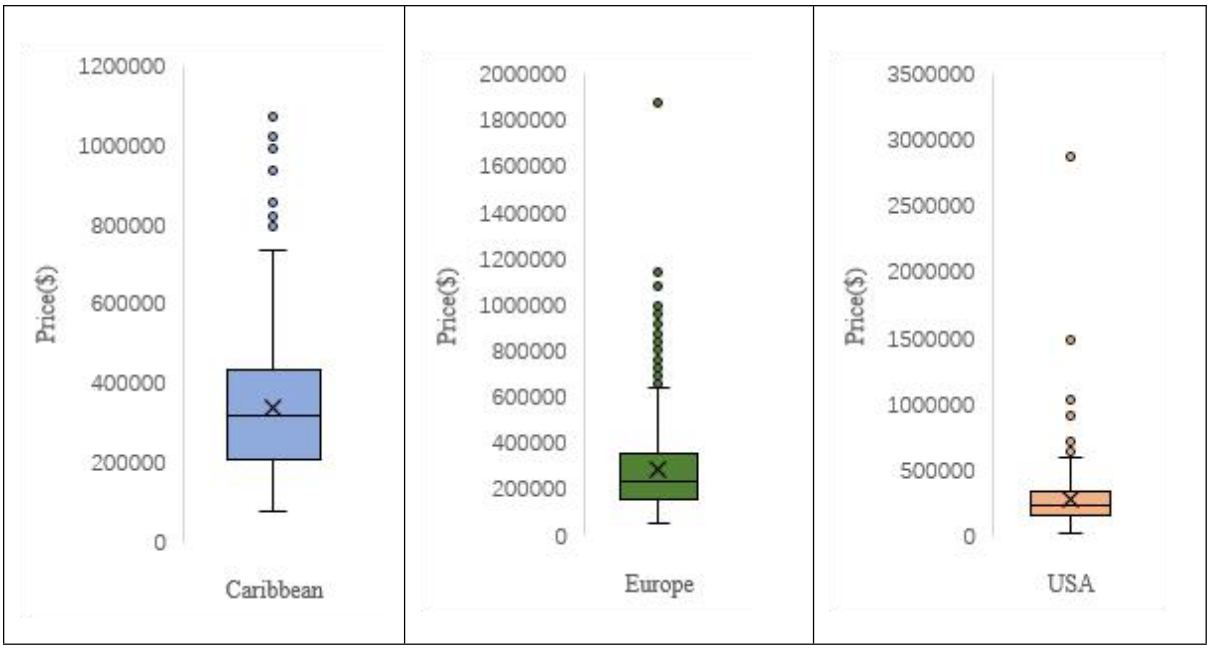


Figure 9: Histograms of prices in Caribbean, Europe and USA

And plot the price-variant curves for the three major regions with sailboat variants as the horizontal axis and prices as the vertical axis:

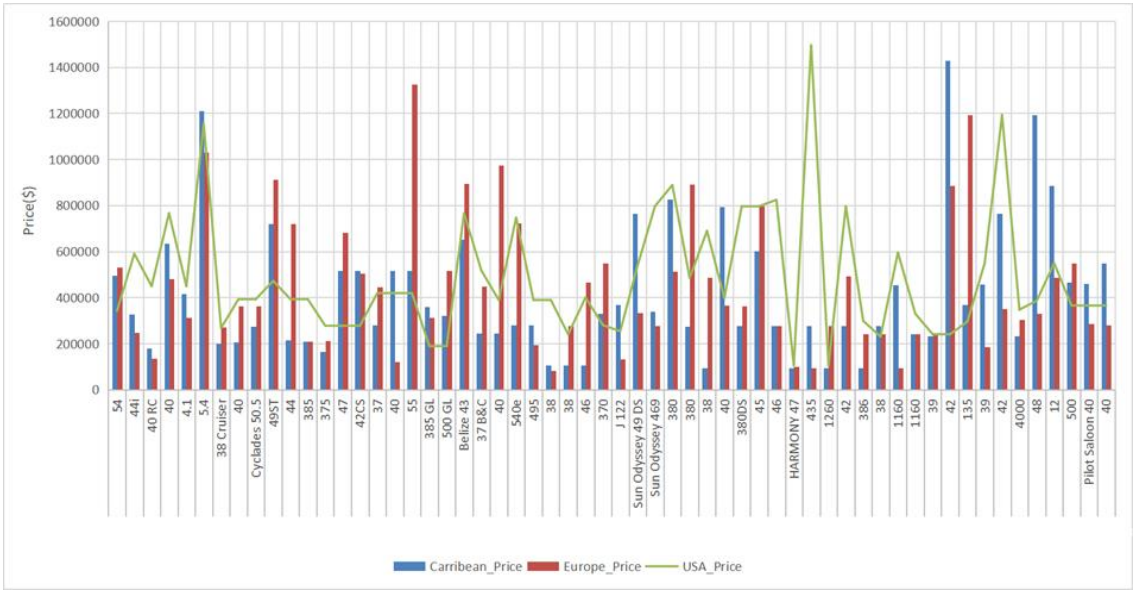


Figure 10: Price Prediction for California, Europe, and the USA

Principle:

To prove that the region has a consistent impact on each variant, it is only necessary to examine whether the price distribution of each variant in the three regions is consistent. [4]Therefore, an F-test is used to analyze whether the variance of the data in the three regions

is equal.

$$F = \frac{\sum_{n_i} (\overline{x_i} - \overline{x})^2/(k-1)}{\sum\sum (x_i - \overline{x_i})^2/(n-k)} \quad (2)$$

Where n_i represents the sample size of the i^{th} group, x^i represents the mean value of the i^{th} group, k represents the number of groups, n represents the total sample size, x_i represents the observed value in the i^{th} group, and $\overline{x}$ represents the mean value of all samples.

Statistical significance:

Assuming H_0: there is no significant difference in variance; H1: there is a difference in variance. At a significance level of α=0.05 and with 180 degrees of freedom, the calculated F-test result is 1.054 $< F_{0.05}(180) = 2.118$. Therefore, we reject the null hypothesis and conclude that there is no significant difference in variance. Overall, the impact of the region on all variants is consistent.

However, from the above combination chart, we observe that for some variants, such as 435, there are significant differences in the predicted prices among the three regions. This may be related to cultural factors, the number of ports, and other factors.

Practical significance:

The above conclusion indicates that the BP neural network price prediction model established based on the data of the three major regions can provide some reference for Hong Kong brokers to predict the price situation in the Hong Kong market. However, for non-general cases, such as sailboat variants or custom-made variants with regional preferences, further analysis and discussion are needed.

5.5.3 Task 3: The role of Hong Kong region and its impact on monohull and catamaran boats.

Firstly, we selected records with a variation number greater than or equal to 3 from the electronic spreadsheet provided by the organizer as a subset of the data, which was divided into monohull and catamaran boats. Next, we obtained the important characteristic parameters such as length and beam of Hong Kong's sailboat variants, as well as the corresponding market prices from relevant websites such as https://www.asia-boating.com/ (see Appendix 1 for details). We processed missing values, outliers, and other issues before inputting these parameters into the BP neural network trained earlier to obtain predicted market prices for various sailboats in the Hong Kong region

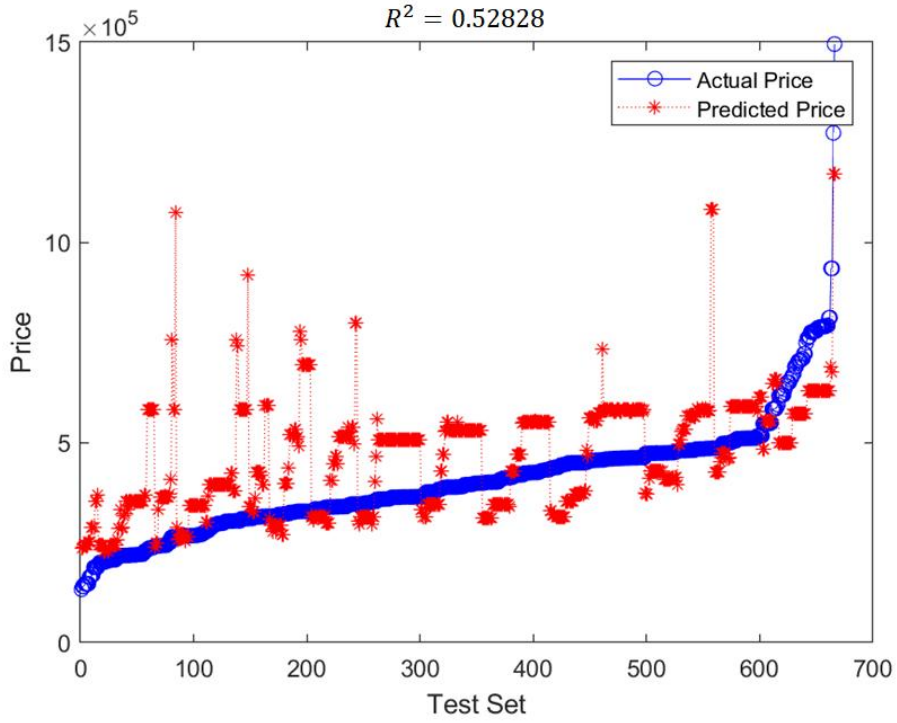


Figure 11: Visualization of Results Comparison For Monohull Sailboats

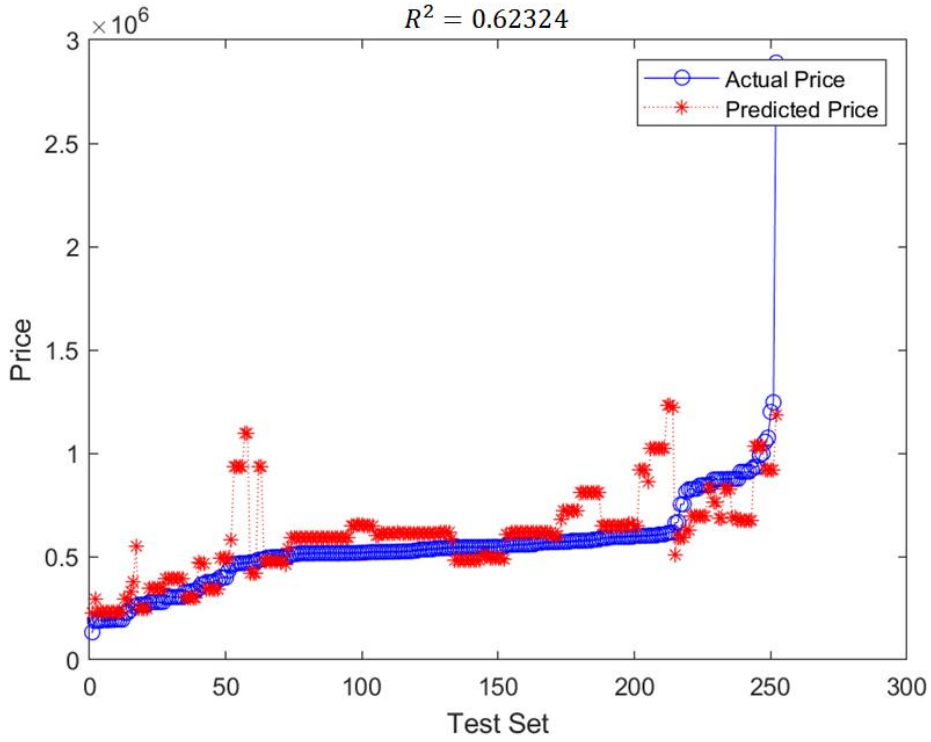


Figure 12: Visualization of Results Comparison For Catamaran Sailboats

As shown in the figure, the predicted data has some deviation from the actual data, but the overall fitting effect is good, and the error is within an acceptable range.

Based on this, we infer that:

1. Our modeling of regional influence is reasonable and can provide a good reference for predicting prices in the Hong Kong market. This can help sellers price their boats reasonably, assist buyers in planning their sailing expenses, and contribute to stabilizing market prices.
2. The regional impact on monohull and catamaran boats is different, and the impact on catamarans is more stable. Through literature review, we believe that compared to monohull boats of the same tonnage, catamarans have a larger deck area and cabin capacity, and are generally used for cargo ships. Due to the high saturation of the freight market, the demand for catamarans is often close to saturation, so the sensitivity of prices to changes in demand is low, and the price stability is strong. Moreover, catamarans connect two boats horizontally, making them less likely to capsize at sea and able to withstand larger waves and winds.

6 Analysis and Inference Based on Exploratory Data Analysis

6.1 Analysis Principle

Exploratory data analysis (EDA) refers to the process of using visualization, statistical methods, and other techniques to understand the characteristics of data, examine hypotheses, and discover patterns and trends in data.[5] This helps to better understand the data, validate previous assumptions, or discover new conclusions. In order to explore and discover potential unknown relationships between variables and to generate insightful conclusions, we conducted exploratory data analysis.

6.2 Analysis Process

"We used visualization and statistical tools such as Excel and SPSS to perform exploratory data analysis from various aspects. For example, we used bar charts to summarize the total listing price, average listing price, and average price growth rate for each year after classifying the data by year. We also used line charts to analyze the listing prices in the three regions after classifying the data by region. We used box plots and scatter plots to analyze the distribution of the data, identified outliers, and conducted further in-depth analysis.

6.3 Inference and Conclusion

6.3.1 Sailing Fever.

The market price of used sailboats is affected by cultural background factors such as sailboat races and the "sailing fever." For example, a series of sailboat races such as the America's Cup and the San Francisco Big Sailboat Race were held in the United States

from 2007 to 2008. These sailboat competitions promoted the emergence of the "sailing fever" phenomenon and increased demand for sailboats, which led to a certain upward trend in the listing prices of used sailboats during that year.

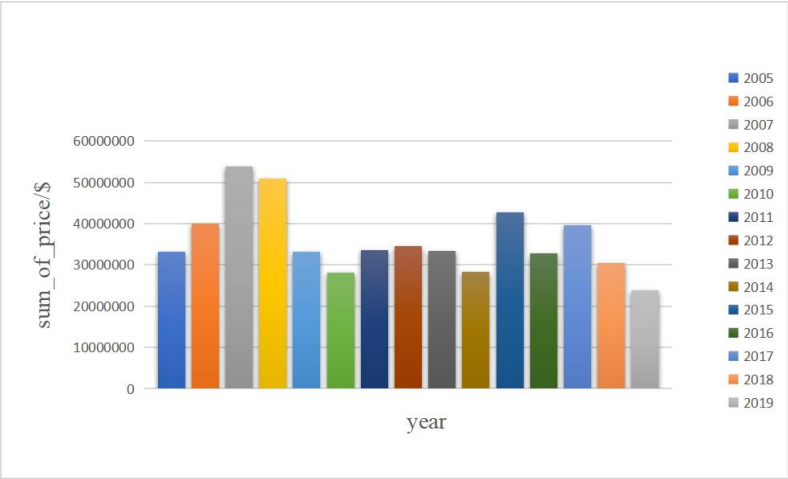


Figure 13: Total Price from 2005 to 2019

6.3.2 Depreciation .

As the years pass, the average listing price shows a clear upward trend. The following bar chart shows the average listing price for each year between 2005 and 2019. First, the regulations regarding the years of use for sailboats lead to the problem of depreciation in the resale market. The more remaining years of use a sailboat has, the higher its value in the used market. Second, the working time of the sailboat engine decreases gradually with the production time. The number of engine hours on a sailboat can significantly affect its value. The more engine hours a sailboat has, the less it is worth.

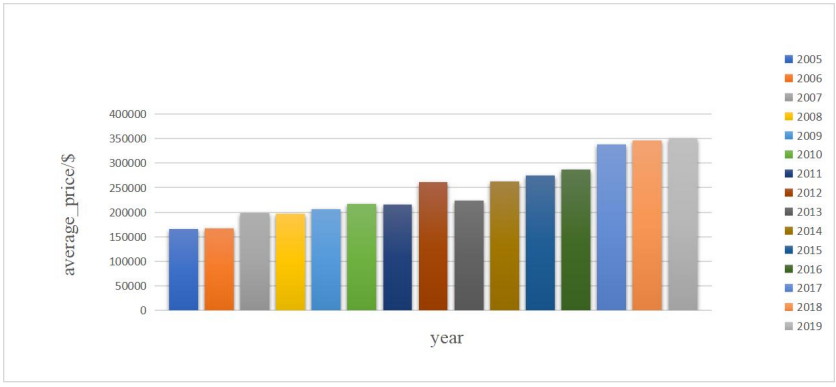


Figure 14:Average Price from 2005 to 2019.

6.3.3 Custom-made.

Compared to sailboats of the same model, some personally customized sailboats have a significantly higher price. In most cases, personally customized models are

tailored to the user's specific needs, with less engine time and better hull materials, while also meeting certain personalized requirements of the user. The manufacturing process is more complex, making the higher price justifiable. Interestingly, we previously mentioned the HH55 sailboat during the analysis of outliers in data processing. It is a limited edition luxury sailboat with excellent performance, luxurious interior decoration, and advanced navigation equipment, providing the perfect choice for sailing enthusiasts seeking a high-end travel experience. Such products naturally come with a surprisingly high price tag, sometimes one to two times higher than that of other sailboats.

6.3.4 Brand Effect.

Well-known brands have a "brand effect" and their listed prices are higher compared to other brands. According to the data, well-known brands such as Bavaria, Beneteau, Jeanneau, and Catalina have an average listed price for monohull boats that is about 15%-25% higher than other brands, and an average listed price for catamaran boats that is about 20%-30% higher than other brands.

6.3.5 Country of Origin Effect(COO).

The listed prices of used sailboats in their production regions are lower compared to the listed prices in other regions. Data shows that sailboats produced in Europe have a listed price in Europe that is about 15% lower than the listed price in the United States and the Caribbean region. Sailboats produced in the United States and the Caribbean region have a listed price in those areas that is about 20% lower than the listed price in Europe.

7 Sensitivity Analysis

The ROC (Receiver Operating Characteristic) curve is a common method used to evaluate the performance of a model. The horizontal axis of this curve represents the False Positive Rate (FPR), and the vertical axis represents the True Positive Rate (TPR).

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{1}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

Among them, TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives, and FN represents the number of false negatives. The ROC curve is plotted by changing the classification threshold of the classification model. For different classification thresholds, the corresponding TPR and FPR can be calculated, and the ROC curve can be obtained. Generally speaking, the larger the area

under the ROC curve (AUC), the better the performance of the classifier. The value of AUC ranges from 0 to 1.

Our ROC curve is shown below, with an AUC of 0.76>0.7. As the FPR increases, the slope of the TPR changes slightly, indicating that the results do not change much when making minor changes to the classification model. Based on this, we can infer that our model has good stability.

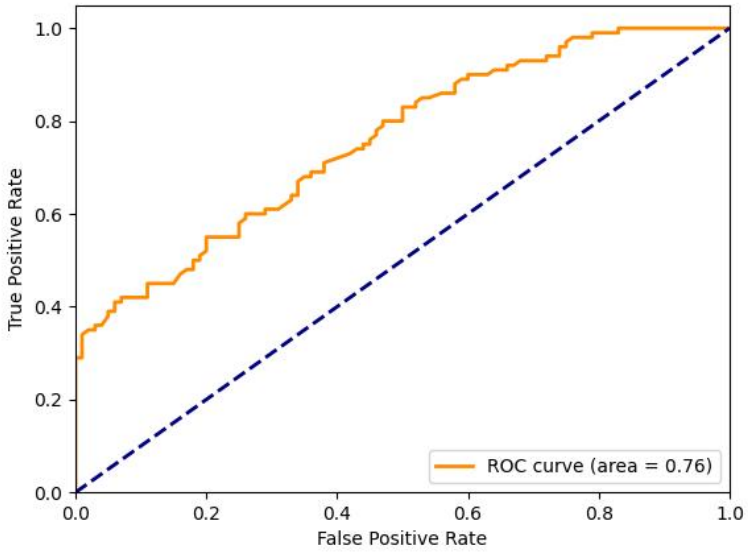


Figure 15: ROC Curve

8 Model Evaluation and Further Discussion

8.1 Strengths

- By using the LM algorithm to train our neural network, we have achieved a fast learning speed and strong generalization ability. Specifically, the network can complete the training of millions of data in a short period of time.
- With the assistance of factor analysis, we extract relevant features from the data while minimizing correlations between variables. This has resulted in a more accurate model.
- Our model exhibits high stability, as evidenced by sensitivity analysis that indicates errors remain within an acceptable range.

8.2 Weaknesses

- The model only considers the most important factors while neglecting weakly correlated factors such as cultural background and sailboat comfort, which results in a limited applicability range. Additionally, the model requires a large amount of data to support neural network training and is not suitable for small-scale data.

8.3 Further Discussion

We will improve the model by incorporating more variables and optimizing the model structure to enhance its predictive accuracy and applicability range. Additionally, we will investigate ways to enhance the model's performance under the condition of insufficient data, such as utilizing data augmentation techniques.

9 Conclusion

We have utilized factor analysis, correlation analysis, and BP neural network to determine the major factors that affect the prices of used sailboats, and successfully implemented the function of price prediction. We have also explained the impacts of regional factors on the prices and the difference between monohulled sailboats and catamarans, as well as discussed the application of our model in Hong Kong. We have drawn interesting conclusions and recommendations based on our analysis. Finally, we have prepared a report for sailboat brokers in Hong Kong to help them better understand the market and make informed decisions.

References

[1]Spearman, C. (1904). "General intelligence, objectively determined and measured". The American Journal of Psychology, vol. 15, no. 2, pp. 201-292.

[2] Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986). "Learning representations by back-propagating errors". Nature, vol. 323, pp. 533-536.

[3] Jelinek, F., Mercer, R.L. (1980). "Interpolated estimation of Markov source parameters from sparse data". Proceedings of the Workshop on Pattern Recognition in Practice, Elsevier North Holland, New York, pp. 381-397.

[4]Fisher, R.A. (1924). "The distribution of the partial correlation coefficient". Metron, vol. 3, pp. 329-332.

[5]Tukey, J.W. (1977). "Exploratory Data Analysis". Addison-Wesley

Appendices

Appendix 1

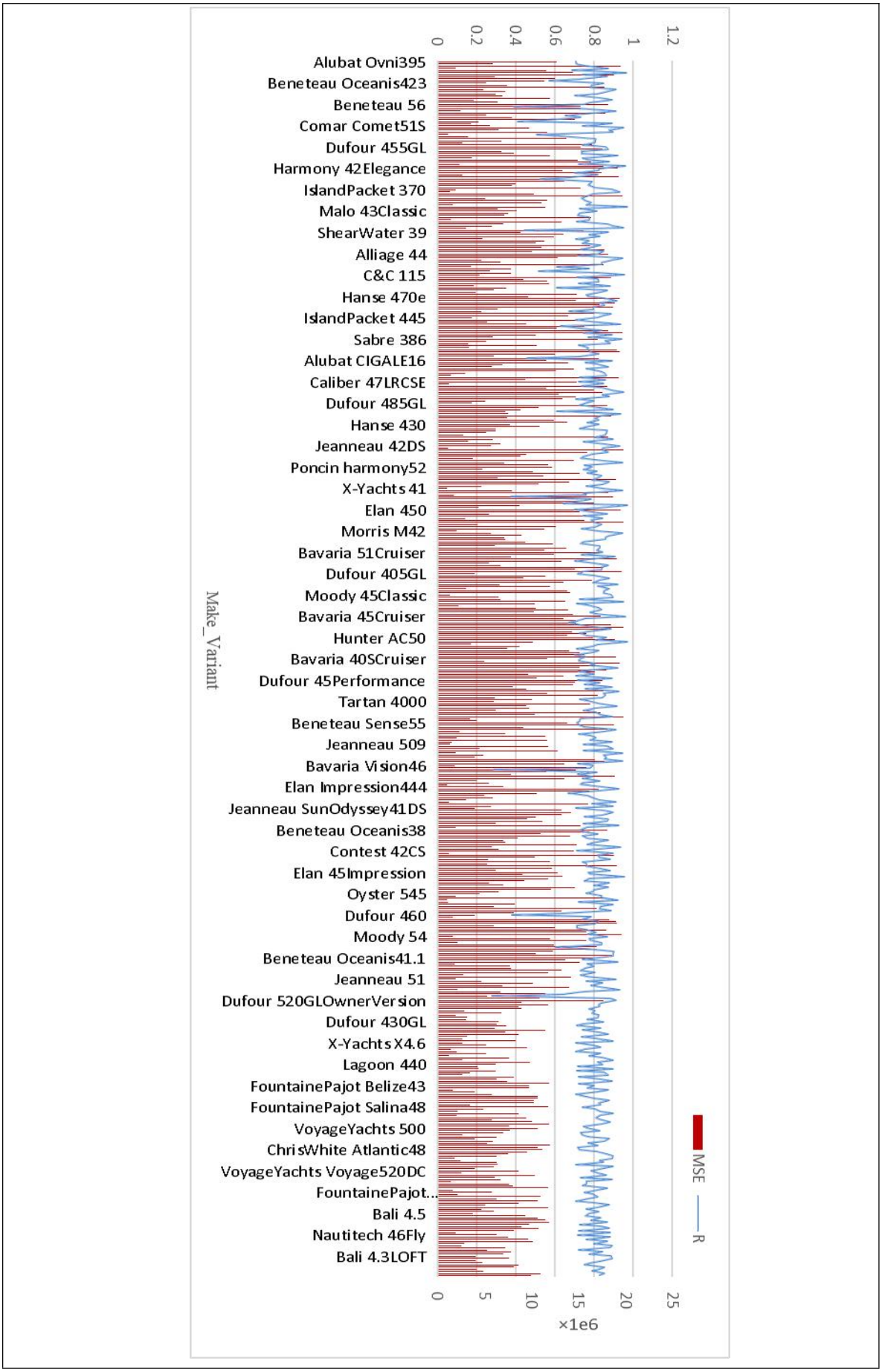
Introduce: Website addresses for data collection and their corresponding features.

Table 9: Websites and Obtained Features

Websites	The obtained features
https://sailboatdata.com/	S.A.(reported), Draft (max), Displacement, Material, Beam etc.
https://www.boats.com/	Drive Type, Hull Shape, Number of Engines, Max Passengers etc.
https://m.jyacht.com/	Production status, Maximum Speed etc.
https://www.asia-boating.com	Length, beam, market prices of Hong Kong's sailboat variants, etc.
https://data.worldbank.org/indicator/NY.GDP.MKTP.CD	Hong Kong’s GDP

Appendix 2

Introduce: Estimated MSE and R-Squared Variance for each Variant


0
0.2
0.4
0.6
0.8
1
1.2
Alubat Ovni395
Beneteau Oceanis423
Beneteau 56
Comar Comet51S
Dufour 455GL
Harmony 42Elegance
IslandPacket 370
Malo 43Classic
ShearWater 39
Alliage 44
C&C 115
Hanse 470e
IslandPacket 445
Sabre 386
Alubat CIGALE16
Caliber 47LRCSE
Dufour 485GL
Hanse 430
Jeanneau 42DS
Poncin harmony52
X-Yachts 41
Elan 450
Morris M42
Bavaria 51Cruiser
Dufour 405GL
Moody 45Classic
Bavaria 45Cruiser
Hunter AC50
Bavaria 40SCruiser
Dufour 45Performance
Tartan 4000
Beneteau Sense55
Jeanneau 509
Bavaria Vision46
Elan Impression444
Jeanneau SunOdyssey41DS
Beneteau Oceanis38
Contest 42CS
Elan 45Impression
Oyster 545
Dufour 460
Moody 54
Beneteau Oceanis41.1
Jeanneau 51
Dufour 520GLOwnerVersion
Dufour 430GL
X-Yachts X4.6
Lagoon 440
FountainePajot Belize43
FountainePajot Salina48
VoyageYachts 500
ChrisWhite Atlantic48
VoyageYachts Voyage520DC
FountainePajot...
Bali 4.5
Nautitech 46Fly
Bali 4.3LOFT
Make_Variant
0
5
10
15
20
25
×1e6
MSE
R

Used Sailboat Price

Report to Broker

I am a researcher commissioned by you to investigate the pricing of used sailboats. In accordance with your requirements, we have conducted relevant research and modeling analysis. Now, please allow me to provide a report on the issue of pricing for used sailboats in Hong Kong.

I. Brief introduction of methods

The reason for conducting prediction is that it is the cornerstone of pricing for used sailboats. Firstly, we extensively consulted materials and collected market pricing data for used sailboats to identify factors that may have a significant impact on pricing. Then, we used factor analysis and chi-square test to further identify strongly correlated feature factors. Based on this, we constructed a BP neural network model to predict the prices of used sailboats. Through testing, our average prediction accuracy exceeds 71.23%, indicating that our model is relatively accurate. At the same time, based on the LM algorithm, the model can efficiently perform prediction analysis.

II. Result Analysis

1. Our research shows that the condition, material, category, and manufacturer of the sailboat are all key factors that affect its price. Among them, the better the condition, the better the material (such as carbon fiber, fiberglass), and the more famous the manufacturer, the higher the price. In addition, the average price of a catamaran is about twice that of a monohull. For example, in our data set, there is a sailboat called HH55 Catamaran, which is mainly made of carbon fiber and was produced by HH Catamarans in 2018. Its price is as high as 2,890,000$.

2. The influence of location factors on prices is significant. Factors such as GDP, the number of ports, marine environment, climate conditions, maritime culture, level of maritime industry development, tax situation, and transportation convenience will all affect prices. In addition, the effect of the place of origin is also significant in the used sailboat market. Sailboat manufacturers are mainly located in Europe, America, and Oceania, where sailboats can be cheaper by about 28 percentage points.

3. Regarding the Hong Kong region, we made price predictions for single and double-hulled boats separately and found that the prediction results for catamarans are more accurate, with an R-squared value of 0.60203 and a good fit, while the prediction accuracy for single-hulled boats is slightly lacking, with an R-squared value of 0.52828.

III. Pricing Suggestions:

- When pricing, the most important thing is to refer to our price prediction model to obtain a pricing benchmark.
- In some years when large-scale sailing competitions are held, a "sailing fever" may be triggered, during which the demand for sailboats will be unusually high, and prices can be raised appropriately.
- Price according to the seller's price expectations, which can greatly meet the seller's needs and also obtain more sources of goods in the future.
- As the years go by, the average price of sailboats is generally on the rise, with an average annual price growth rate of 5%. Therefore, prices can be raised appropriately as time goes by.