Problem Chosen
Y

2023
MCM/ICM
Summary Sheet

Team Control Number
2332148

Region-based Multiple Regression Analysis Model

Summary

The price of a secondhand sailboat is influenced by numerous factors, including age and market conditions. Developing a predictive model for the price of a secondhand sailboat based on several indicators is crucial. This paper introduces the use of Pearson, ARIMA, BP neural network, random forest and other models for analyzing single or multiple indicators. Ultimately, we are adopting LightGBM machine learning to model the maintain value rate of secondhand sailboats and all indicators to better estimate their prices.

In question 1, we first preprocessed the data by converting categorical features to quantitative ones. Next, we employed various models to examine the correlation between different types of categorical data and prices. Specifically, we used the Pearson model to scrutinize the relationship between sailboat length and price, the ARIMA model to investigate the relationship between the year of the boat's manufacture and its price, and the random forest model to explore the relationship between all indicators and sailboat prices. We subsequently compared the predicted data with the original data using the model and concluded that this model demonstrates a certain level of feasibility.

In question 2, to further investigate the impact of the region on prices, we conducted a BP neural network model analysis. We aimed to assess the influence of the region on the prices of sailing variants, and thus added the region as an additional predictor variable. Our analysis of the prices of various sailing variants in each region revealed that the impact of the region had an equal effect on all sailing variants.

After analyzing the second-hand sailing boat market in Hong Kong SAR, we examined Hong Kong's per capita GDP. To achieve this, we analyzed the per capita GDP of each region by consulting existing data tables and performing correlation analysis between Hong Kong's GDP values and those of other regions. We further divided the data into two subsets of monohull and multihull sailboats subject to predictive analysis achieved through the utilization of a random forest regression model. By applying grid searching to optimize the model and identify its best parameters, we established the relationship between the value retention rate of the boat and various indicators, resulting in greater precision in price predictions using the value retention rate.

In conclusion, we have developed a comprehensive model introduction report for brokers in the Hong Kong second-hand sailing boat market. In addition to conducting extensive data analysis, we have meticulously evaluated the performance of our models, through error and sensitivity analysis. We assessed the strengths and weaknesses of the models used and proposed possible improvements to enhance modeling accuracy. Furthermore, we have suggested potential areas for the model's extension or improvement in the future.

Keywords: ARIMA algorithm，Regression Analysis，BP Model，Random Forest，LightGBM

Contents

1 Introduction

1.1 Problem Background

The valuation of sailboats is subject to fluctuation based on various factors, such as their manufacturing period and prevailing market conditions. Our data includes the brand, model, length (in feet), geographic region, country/region/state, listing price (in US dollars), and year of production for both mono-hull and catamaran sailboats. Providing the sailboat's brand, model, and year of manufacture can reveal diverse sailboat features such as beam, draft, displacement, rigging, sail area, hull materials, engine hours, headroom, and electronic equipment. Sailboats are ordinarily sold through brokers. Developing a scientific model to enable clearer price evaluation is crucial for a better understanding of the sailboat market and the creation of an accurate pricing scheme for second-hand sailboats.

In addition to the data we provided on sailboat valuation, it's important to note that market conditions also significantly impact sailboat value. For example, in coastal areas or countries known for their sailing culture, there may be a higher demand for sailboats, leading to a surge in prices. Conversely, adverse weather conditions or unfavorable economic situations could cause prices to drop due to a stale market. Hence, keeping a close eye on market trends is necessary, alongside analyzing sailboat features. Other than the essential features mentioned, sailboat design is another critical aspect to consider in sailboat valuation. The design can influence sailboat competitiveness in a particular class or race. In some markets, specific designs may be more popular and hence sell at a higher price. Additionally, the aesthetic design of a sailboat, such as its style and color, can also impact its value since certain owners may prefer a particular design and be more willing to purchase it at a higher price. Furthermore, sailboat condition is a vital factor in determination of its worth. A sailboat in excellent condition usually sells for a higher price as compared to one that is poorly kept. Therefore, when valuing sailboats, the number of engine hours and regular maintenance must also be considered. When purchasing a second-hand sailboat, the buyer will want to know the sailboat's repair history and maintenance records to determine its present condition accurately. These records can add value to the sailboat. In conclusion, sailboat valuation is a complicated process that necessitates a thorough understanding of all the factors that can influence a sailboat's price.

Considering all the fundamental factors listed and developing an appropriate scientific model makes it possible to create a precise pricing plan for second-hand sailboats. This pricing strategy is beneficial for both buyers and sellers, as it assures that they receive a fair and market-based price for the sailboat.

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:

- To develop a mathematical model that explains the listing price of each sailboat provided in the electronic spreadsheet. Identify and describe all data sources used. Include a discussion of the accuracy of price estimations for each sailboat variant.

- To use the developed model to explain the impact of regions on listing prices. Discuss whether regional effects are consistent across all sailboat variants. Explain the statistical and practical significance of any observed regional effects.
- To discuss how modeling for a given geographic area works in the Hong Kong (SAR) market. Select an information-rich subset of sailboats from the provided electronic spreadsheet, divided into mono-hull and catamaran sailboats. Find comparable listed price data for this subset in the Hong Kong (SAR) market. Simulate the regional impact of Hong Kong SAR on sailboat prices in the subset. Is the effect the same for mono-hull and catamaran sailboats?
- To identify and discuss any other interesting and information-rich inferences or conclusions derived from the data.
- To prepare a one to two-page report for sailboat brokers in Hong Kong (SAR), include carefully selected graphs to assist brokers in understanding your conclusions.

1.3 Literature Review

This trade of used ships is a multifaceted economic activity that involves market size, price, quality, and transaction rate. I In recent decades, driven by the world economy's growth and the rising demand for water tourism and entertainment, the used-ship market has rapidly expanded. In this context, scholars have conducted studies on the subject and have made significant advances.

Firstly, the ship's history, mechanical equipment, and maintenance records have been shown to affect pricing. For instance, in 2018, a shipping company named CMA CGM sold 14 ships at different times and places, resulting in significant price variations.

Furthermore, research into the second-hand car market may offer some useful insights into the challenges that the used-ship market faces. For instance, Car Home, a used-car website implemented uniform national pricing on its e-commerce site to overcome information asymmetry in the market and foster consumers' trust and transaction volume.

Secondly, research has demonstrated a correlation between the trade volume and pricing of used-ships; excessive pricing or undervaluing leads to decreased transaction frequency. In regions with significant purchasing power, e.g., Hong Kong, China, the used-ship transaction rate is comparably high.

Moreover, some scholars have investigated the spatial distribution of the second-hand ship market, in some countries or regions, the geographical distribution of the second-hand ship market is closely related to the pillar industries such as fishing and tourism.

Nonetheless, several issues still require attention, such as achieving precise pricing predictions or resolving information asymmetry issues. The used-ship market is continually evolving due to the robust growth of the water tourism and entertainment sector. Emerging technologies and novel transaction methods create challenges and opportunities for existing research.

In 2021, Ship International magazine in the UK reported that the used-ship trading market's global growth is expected to remain stable and may even expand in the next few years, especially in many coastal countries and islands.

This suggests that investing in the second-hand ship market remains worthwhile. As the global tourism industry continues to expand, so will the market for second-hand ships.

Researchers should redouble their efforts to create more precise and comprehensive pricing models, while regulators should strengthen market oversight and information platforms to promote the healthy growth of the used-ship trading market.

1.4 Our Work

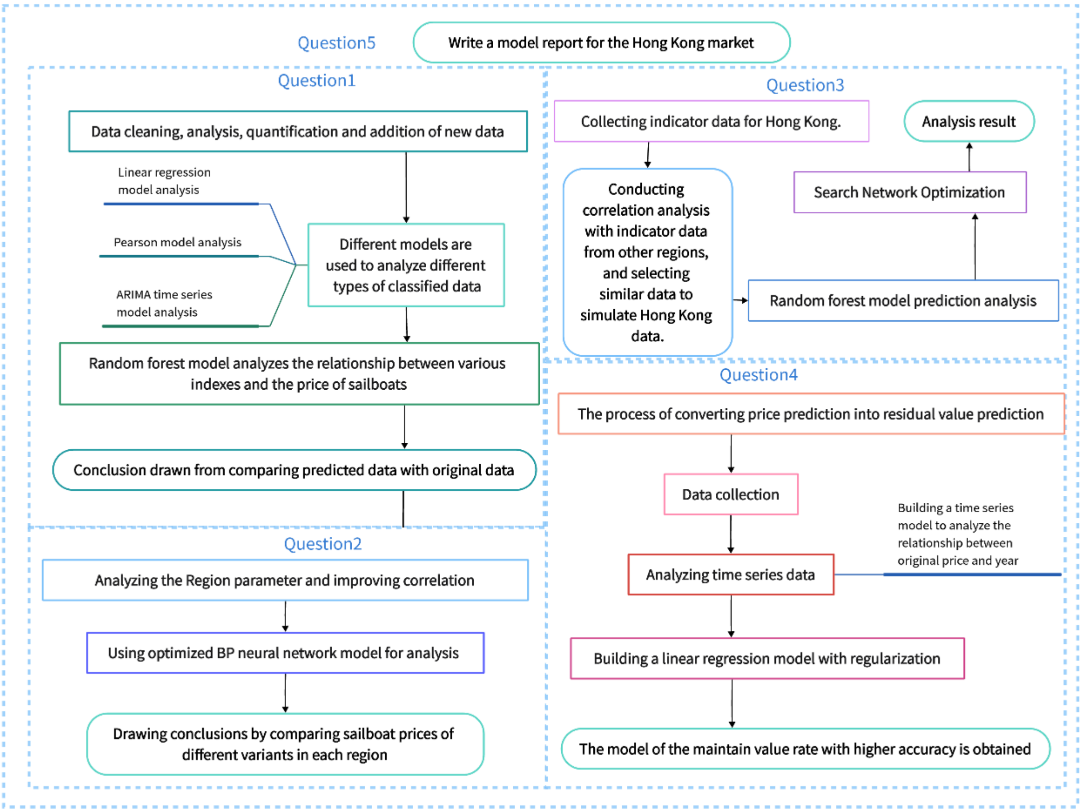

Question5
Write a model report for the Hong Kong market
Question1
Data cleaning, analysis, quantification and addition of new data
Linear regression model analysis
Pearson model analysis
ARIMA time series model analysis
Different models are used to analyze different types of classified data
Random forest model analyzes the relationship between various indexes and the price of sailboats
Conclusion drawn from comparing predicted data with original data
Question3
Collecting indicator data for Hong Kong.
Analysis result
Conducting correlation analysis with indicator data from other regions, and selecting similar data to simulate Hong Kong data.
Search Network Optimization
Random forest model prediction analysis
Question2
Analyzing the Region parameter and improving correlation
Using optimized BP neural network model for analysis
Drawing conclusions by comparing sailboat prices of different variants in each region
Question4
The process of converting price prediction into residual value prediction
Data collection
Building a time series model to analyze the relationship between original price and year
Analyzing time series data
Building a linear regression model with regularization
The model of the maintain value rate with higher accuracy is obtained

Figure 1Modeling framework

2 Assumptions and Justifications

- **Assumption:** Statistical data observed and recorded may contain errors. We assume that individual errors will not cause significant impacts, or that the impacts caused will be within the margin of error of our model.
  Justification: Typically, statistical data is recorded by skilled professionals or automated systems, thereby reducing the possibility of errors. Although errors may occur, they are typically isolated incidents rather than systemic issues. Within the margin of error of our constructed model, small-scale errors, such as minor variations in individual data points, may be acceptable.
- **Assumption:** The data being used is genuine and effective.
  Justification: The data is sourced from reliable sources. Verifying the authenticity and validity of data through various methods is a standard practice. To avoid influencing the results, analysts usually exclude data they find to be inaccurate or ineffective.

3 Definitions and Notations

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

Table 1: Notations used in this paper

Symbol	Description
$\overline{x}$	Mean
S^2	Variance
n	Number of values
μ	Expected value
χ^2	Chi-squared statistic
$\hat{\gamma}_r$	represents the estimated impact of region r on price
A	Observed contingency table of discrete data
T	Expected contingency table of discrete data in chi-squared test
$p(x)$	Correlation coefficient
$PPMCC$	Pearson product-moment correlation coefficient
η	learning rate
p_t	the predicted value at iteration t.
RVR	resale value rate
$L(\Phi)$	Loss function in XGBoost
λ	Constant term
w	Weights or coefficients in XGBoost

Expected contingency table of discrete data in chi-squared test: In chi-squared test, expected contingency table predicts how many values would be observed in each cell of a two-way contingency table if there was no association between the two compared categorical variables. It determines the chi-squared statistic and measures the strength of the association between the variables.

Loss function in XGBoost: In XGBoost, the loss function measures the difference between the predicted output values and the true output values of a model, and is used to train the model to predict accurately. It depends on the type of task, such as regression, classification, or ranking and influences the balance between accuracy and robustness of the predictions.

Weights or coefficients in XGBoost: In XGBoost, the weights or coefficients are assigned to feature variables to indicate their relative contribution to the model's predictions. Different types of weights can be used, for example, based on gain, frequency or coverage. The weights are updated during the training process to optimize the loss function, and can be used to identify the most important features and interpret the model's behavior.

4 Regression Analysis Model

4.1 Data Description

4.1.1 Data preprocessing

The attached data is divided into two tables. For ease of processing, we combined the two tables and inserted a column to identify the type of vessel. Subsequently, we executed the following data processing steps:

To address missing values, we imputed the mode based on geographic region information.

As a result of the abundance of, and quantity in the provided data within the attachment, recording errors may occur due to human or machine malfunctions. Using the provided data, we initiated outlier processing by implementing Grubbs' statistical test.

Assuming that the measured values are measured with equivalent precision, while indepentdly obtaining them, $x_1, x_2, \dots, x_n$. calculate their arithmetic mean xand residual error $v_i = x_i - x\left(i = 1,2， \dots, n\right)$, As per the Bessel formula, we can calculate the standard deviation δ.If the residual error $v_b(1 \le b \le n)$ of a measurement value x_b satisfies the following equation:

$$|v_b| = |x_b - x\ | > 3\delta \tag{1}$$

The value at x_b is determined to be an outlier due to its significant error values and, therefore, must be removed from the dataset.

The presence of irrational data is attributed to sampling or manual input errors, and consequently, removed or corrected using specific methods. Given an index, we define its background value range as the interval bounded by $(x\ -\ 3\delta, x\ +\ 3\delta)$. Accordingly, any data lying within this interval is labeled as normal range and retained, and the corresponding abnormal data is discarded. Subsequently, we use box plots to identify the values of extreme outliers, which we then remove and replace with appropriate values.

To demonstrate the original data distribution and compare its characteristics with other datasets, we performed outlier processing on the normalized data in the attachment. The convenience of MATLAB's plot features was utilized. The resulting boxplot is presented below:

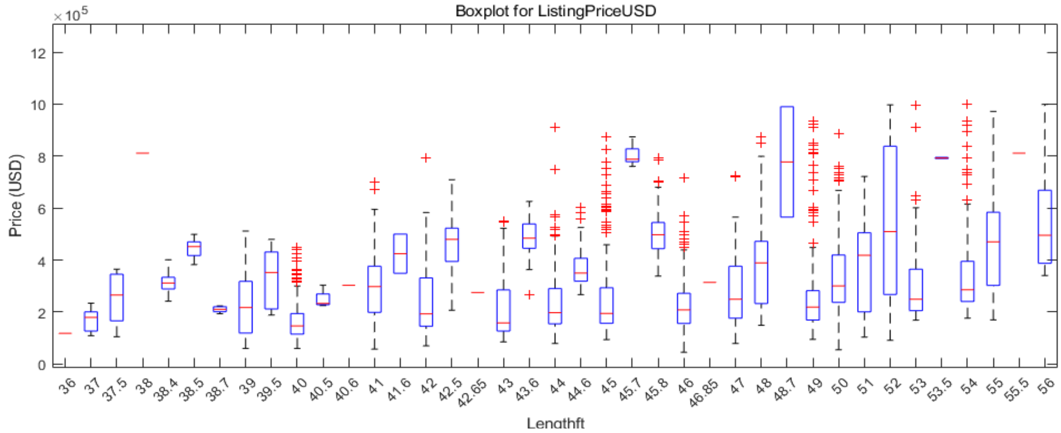


Figure 2Box plots regarding the relationship between size and price

4.1.2 Data visualized

To overcome the issue of varying units of measurement between these indicators and to ensure the reliability of the statistical data, we applied normalization preprocessing to the training and testing sets using mathematical methods for numerical calculation. We utilized the normalization mapping technique to normalize the data.

$$\mathrm{x} \to \mathrm{y} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

The data is normalized without dimensions to be within the range of [0,1], which standardizes the data and enables the establishment of statistical evaluation models within the same system, which helps to establish statistical evaluation models within the same system. Here,$x, y \in R$， $x_{min} = min(x)$， $x_{max} = max(x)$ The result of this normalization is that the original data is standardized within the [0,1] range. This normalization method is commonly referred to as [0,1] interval normalization.

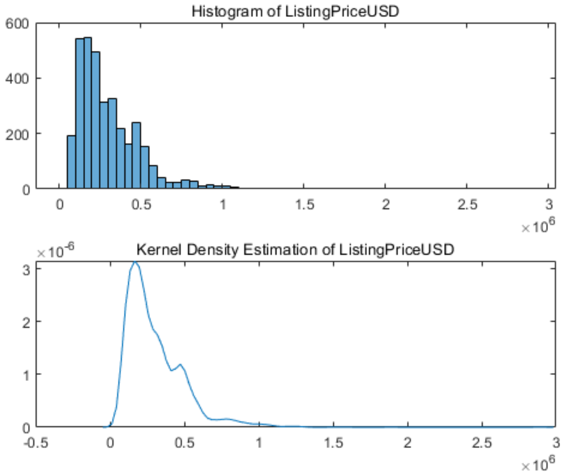


Figure 3 distribution of the price (before changes)

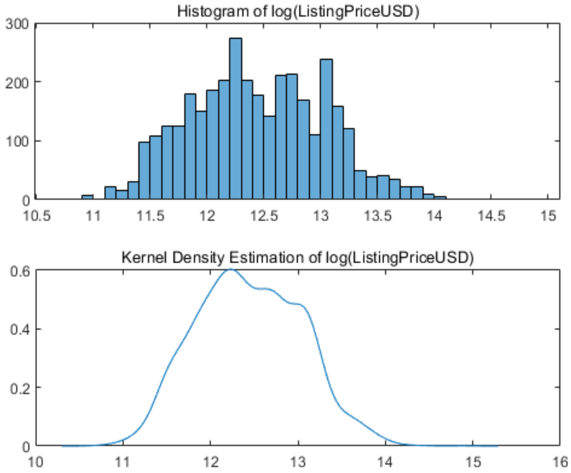


Figure 4distribution of the price (after changes)

The sailboat pricing results were visualized using size and time as variables for fitting. The resulting visualization is presented in the figure above. Size tends to have a normally distributed influence on price. Smaller sizes tend to result in lower prices. To quantify the extent of influence, we calculated the Pearson correlation coefficient using mathematical statistics, and present the results below.

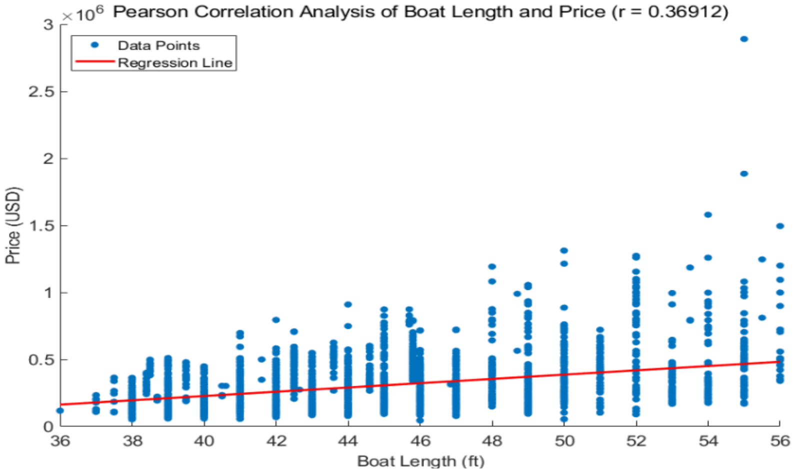
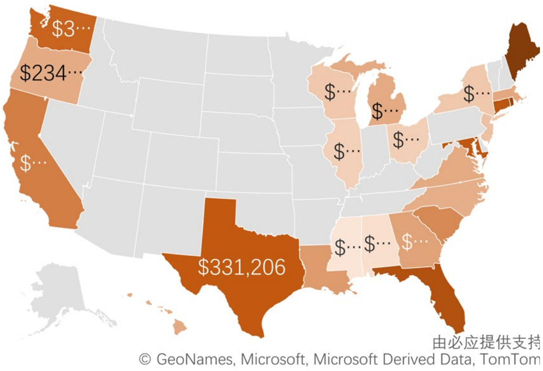


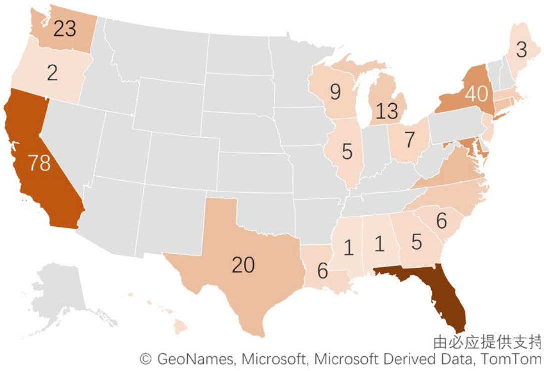
Figure 5Pearson correlation coefficient

MATLAB was utilized to perform statistical analysis and operations on the data provided in the attachment. Boat sales volume was computed for each region and average price was calculated. The data was mapped and six visualizations were created after processing. These visualizations are displayed below.



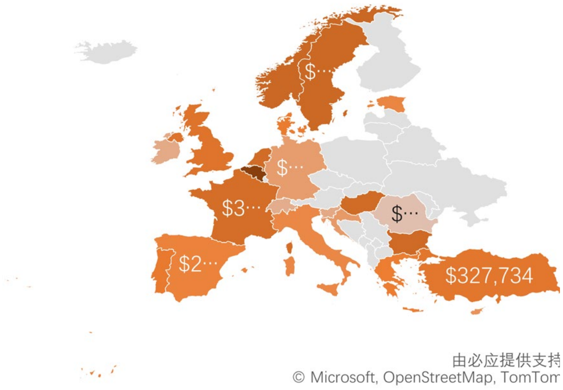
由必应提供支持
© GeoNames, Microsoft, Microsoft Derived Data, TomTom

Figure 6USA-ListingPriceUSD



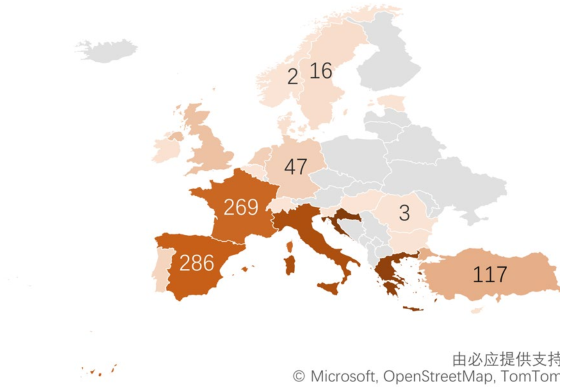
由必应提供支持
© GeoNames, Microsoft, Microsoft Derived Data, TomTom

Figure 7USA-GroupCount



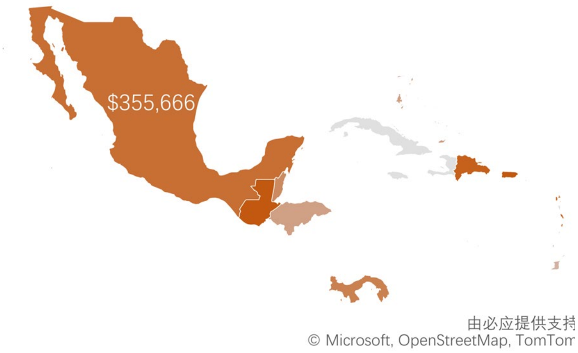
由必应提供支持
© Microsoft, OpenStreetMap, TomTom

Figure 8 Europe-ListingPriceUSD



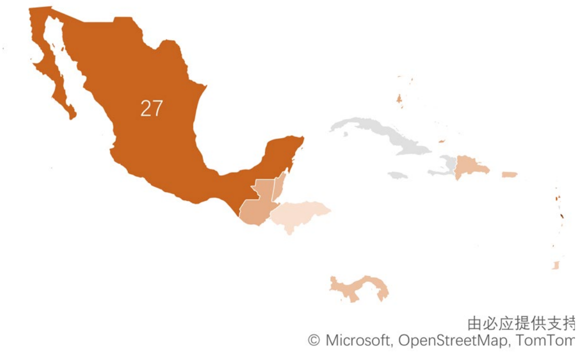
由必应提供支持
© Microsoft, OpenStreetMap, TomTom

Figure 9 Europe-GroupCount



由必应提供支持
© Microsoft, OpenStreetMap, TomTom

Figure 10 Caribbean-ListingPriceUSD



由必应提供支持
© Microsoft, OpenStreetMap, TomTom

Figure 11 Caribbean-GroupCount

The relationship between length and price was further investigated through data visualization, taking into account the influence of Geographic Region. We analyzed the relationship by creating a scatter plot and regression line, represented in the figure below.

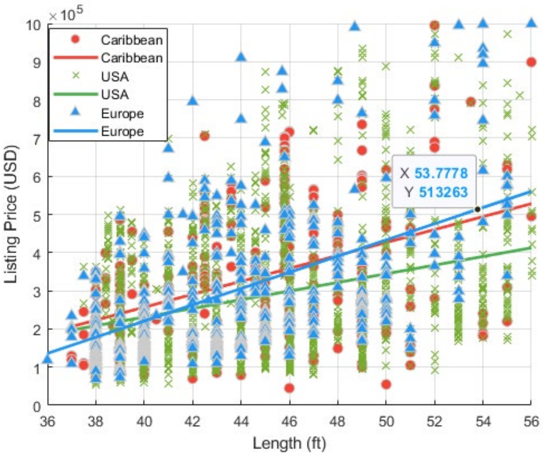


Figure 12GeographicRegion& length-ListingPriceUSD

4.2 The Establishment of Regression Analysis Model

Cleaning and increasing data indicators, the model has the following parameters: Make, Variant, original price, Length, GeographicRegion, Country_Region_State, Listing-Price, Year, whether is it a monohull sailboat,, average annual port throughput, per capita GDP and residual value rate.

In the model, all parameters are used to maintain the formula :

$$RVR = \frac{ListingPrice}{original\,Price} \tag{3}$$

Linear regression is used to analyze the relationship between ListingPrice and Make, Variant, Length, ListingPrice, GeographicRegion, and Country_Region_State.

$$ListingPrice = \beta_0 + \beta_1 Make + \beta_2 Variant + \cdots + \beta_5 Country\_Region\_State + \Omega \tag{4}$$

ARIMA time series model is used to analyze the relationship between original price and Country_Region_State on the time series of year.

$$ARIMA(Country\_Region\_State, year) \tag{5}$$

Finally,

$$RVR = \frac{\beta_0 + \beta_1 Make + \beta_2 Variant + \ldots + \beta_5 Country\_Region\_State + \Omega}{ARIMA(Country\_Region\_State, year)} \tag{6}$$

Searching for and processing GDP-related data from the World Bank. The visualization results are as follows：

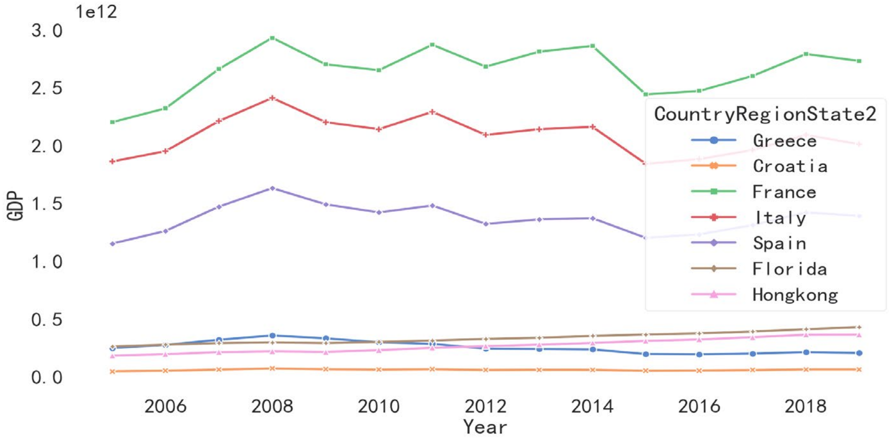


Figure 13Fluctuations in GDP

As the majority of the data used in the analysis consists of categorical and discrete variables, tree-based models, such as random forest, are preferred for modeling and predictive analysis.

Analyze the importance of variable features in a random regression model.

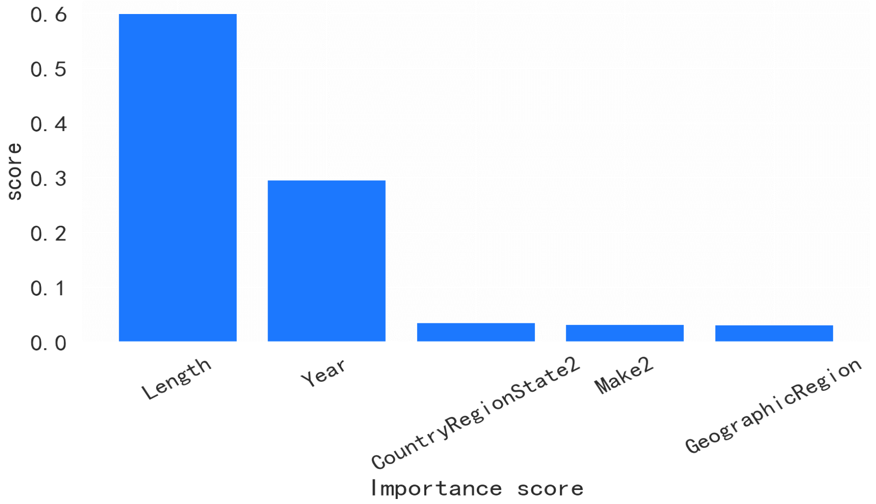


Figure 14Importance of regression characteristics of random forest

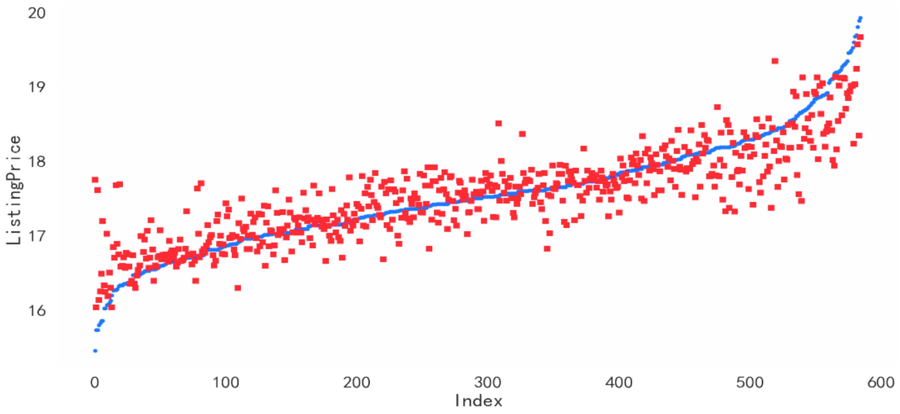


Figure 15Random forest regression model (validation set)

Establish an RBF kernel Support Vector Machine regression model to explore the effectiveness of the regression model.

The prediction effectiveness of the visualization model on the validation set:

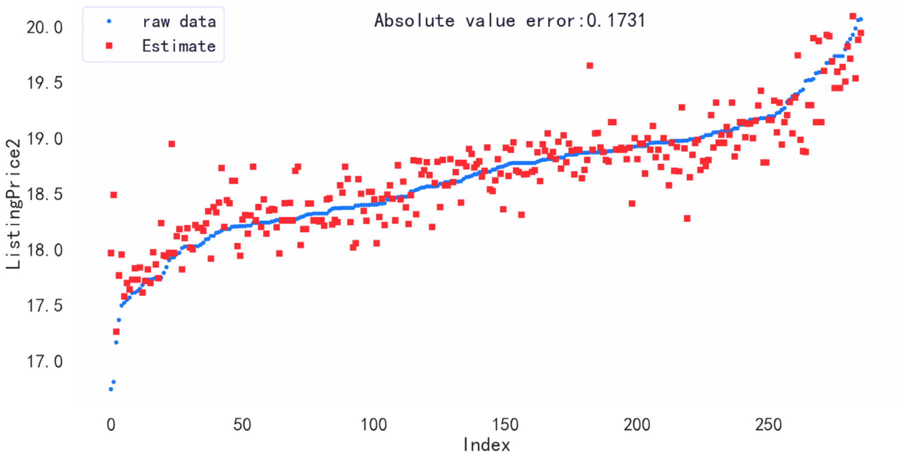


Figure 16Support Vector Machine Regression Model (Validation Set)

4.3 The Solution of Regression Analysis Model

The first step in data analysis is to convert categorical data into continuous data using a specific methodology.

To determine the Make for each sailboat, the proportion of sailboats for each manufacturer can be computed, along with a weighted sum assigned as the Make value. A quantitative formula

$$\text{Variant} = \beta_0 + \beta_1 \text{ Sales} + \beta_2 \text{ Positive Feedback Rate} + \ldots \tag{7}$$

is used to establish a linear relationship between the quality of a Variant, sailboat sales, and positive feedback rate.

In the case of Original Price, the price varies based on a sailboat's manufacturing year and Country_Region_State.

As a result, an ARIMA time series analysis technique is employed to fit original price data of a similar variant sailboat in different countries over various years using [Price Array] and [Country_Region_State Array]. The predicted value of the Original Price can be obtained by entering the Country_Region_State and year data into the model, with [Year Array] being the dependent variable.

The relationship between Length and ListingPrice is determined using the Pearson model. The strength of the relationship is measured by the numerical value of the Pearson coefficient and an appropriate weight. Similarly, the value of GeographicRegion is computed by conducting a straightforward statistical analysis to ascertain the proportion of sailboats in each region.

$$PPMCC = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{8}$$

The value for Country_Region_State depends on the number of sailboat transactions in each city, which gets linearly fitted based on transaction numbers, port throughput, and per capita GDP as

$$Country\_Region\_State = \beta_0 + \beta_1 Numbers + \beta_2 Throughput + \beta_3 \text{per capita GDP} \quad (9)$$

Furthermore, a linear regression model is used to deduce the connection between ListingPrice, Make, Variant, Length, ListingPrice, GeographicRegion, and Country_Region_State. We also applied an ARIMA time series analysis that considers the correla-tion between Original Price and Country_Region_State on a Time-Series basis.

Finally, use formula (6) is estimated.

Random Forest's OOB Score: 0.7288255141671758

Absolute Error on the Training Dataset: 0.1848596028162264

Absolute Error on the Validation Dataset: 0.25312309305431735

5 The BP Model of Region and Price

5.1 Data Description

In order to divide the dataset into the training set and the testing set, the data gets sorted based on feature similarity before the partitioning. The most typical data with the same features are placed at the top, with fewer unusual observations at the bottom. The top 80% of data is used as the training set, while the remaining 20% constitutes the testing set. This method ensures models created are more dependable when subjected to limited and unique data. If the predicted outcome closely aligns with the testing set, the model is deemed functional.

5.2 The Establishment of BP Model of Region and Price

In the previous random forest model, the impact of regions on the listing price is not very important, with a relatively low importance ranking. However, data visualization analysis revealed price differences between different regions. In the following, we will conduct detailed analysis on the relationship between regions and prices, starting with exploratory analysis using visualization.

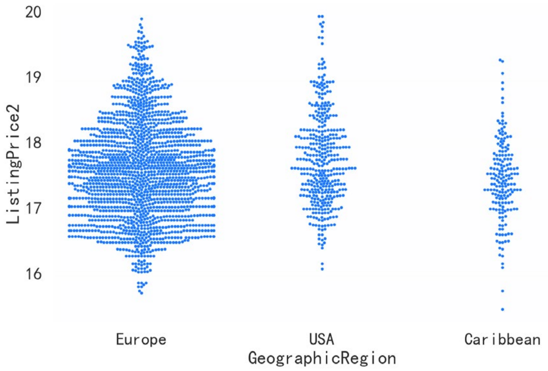


Figure 17swarm plot

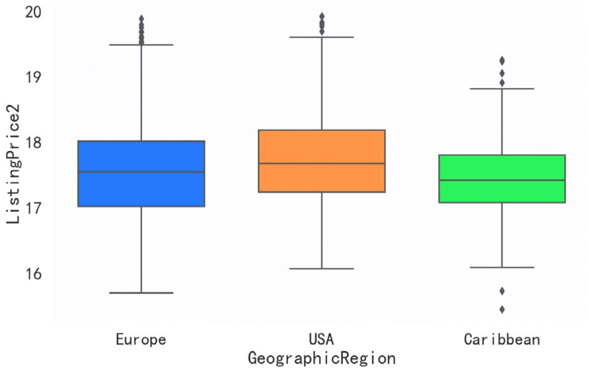


Figure 18box plot

We employed a BP neural network model to explore the consistency of regional effects across all types of sailboats by analyzing the relationship between region and price. The input was defined as the number of attempts in difficult mode, while the output was defined as the percentage occupation. This results in a complex function mapping problem. We conducted simulations with five hidden layers.

To assess the effect of region on the pricing of sailcloth, the addition of region as an additional predicting variable can be accomplished. This can be achieved by converting region into a set of dummy variables and incorporating it as one of the model's inputs. During prediction analysis, the effects of different regions can be observed by manipulating the respective dummy variables.

To establish the actual and statistical significance of regional impact, a sequence of statistical analyses is necessary. Specifically, the average prices for each region can be calculated and contrasted with prices in other areas. If the average price in a particular region is significantly higher compared to other regions, it can be concluded that the impact of that particular region on the sailcloth's rate is statistically significant.

5.3 The Solution of BP Model of Region and Price

The BP neural network toolbox of MATLAB was employed to construct the model, in order to study the impact of region on sailboat prices. In the initial model, we analyzed the relationships between various indicators by utilizing second-level indicators as model parameters, and first-level indicators were only used to aid further analysis. In this study, we created a BP neural network model to analyze the relationship between region and price more efficiently.. We tested the effects of different regions on price during the prediction phase by adjusting transformed regional indicators, then conducted statistical analyses to determine the significance of regional impact.

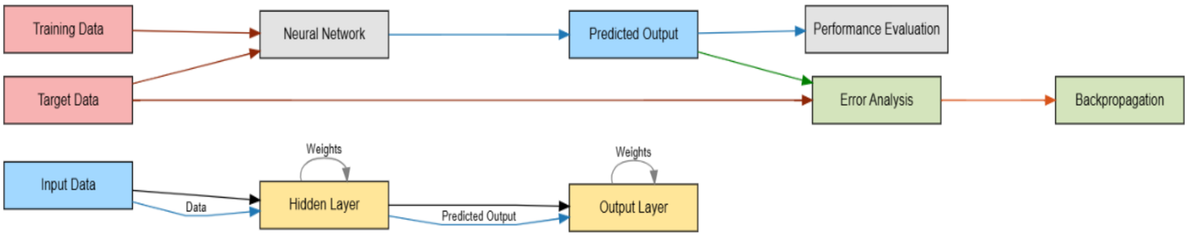


Figure 19General flow chart of BP neural network training

The average price of sailboats for different variants in each region was calculated, and the price of sailboats of the same variant in different regions was compared. If sailboats of different variants exhibited a consistent trend in price change in different regions, then it can be inferred that regional effects are consistent across all variants of sailboats.

Furthermore, we used hypothesis testing to examine whether the price differences of sailboats in different regions but of the same variant were significant among various variants. Specifically, we compared the price differences of sailboats among different regions but the same variant using ANOVA. If the p-value tested by ANOVA is less than the significance level, then the price differences of sailboats in different regions but of the same variant are significantly different. a formula can be derived.

$$\Delta_{rb} = \frac{1}{n_r} \sum_{k \in A_r} y_k - \frac{1}{n_b} \sum_{k \in A_b} y_k \tag{10}$$

Herein, n_r represents the number of sailboats in region r, and $\mathrm{A_r}$ represents all the sailboats in region r.

We can use the following formula to estimate the impact of each region:

$$\hat{\gamma}_r = \frac{1}{\sum_{r=1}^{m} n_r} \sum_{r<b} \Delta_{rb} \tag{11}$$

Herein, $\hat{\gamma}_r$ represents the estimated impact of region r on price. that the value of $\hat{\gamma}_r$ reflects the difference between the average price of all sailboats in region r and the average price of all sailboats in other regions.

To control the model complexity and prevent overfitting during model computation, a regularization term was utilized to constrain the parameter of the model, and the model was rewritten as:

$$\text{ListingPrice} = \beta_0 + \sum_{b=1}^{p} \beta_b x_b + \sum_{i=r}^{m} \gamma_r R_r + \frac{\lambda}{2} \sum_{r=0}^{m} \gamma_r^2 + \epsilon \tag{12}$$

Herein, λ represents the regularization coefficient.

6 Random Forest after Optimizing Network Search

6.1 Data Description

The study first examined Hong Kong's per capita GDP and then analyzed the fluctuations in GDP for each region by utilizing current data tables, A correlation analysis was conducted to compare Hong Kong's GDP with that of other regions in the table, and data sets with per capita GDP values that were similar to Hong Kong's were selected as simulated data. The simulated data was then cleaned to remove any outliers and divided into two subsets based on the hull type, namely single-hull and double-hull ships.

6.2 The Establishment of Random Forest after Optimizing Network Search

Our study investigated the relationship between various indicators and prices while employing a random forest regression model to generate accurate price predictions. We recoded or hot-encoded categorical variables, and then partitioned the dataset into training and validation sets. The predictive outputs were generated using the random forest regression model and compared to the original data. The model was optimized through a network search that detected the appropriate random forest regression model parameters.

6.3 The Solution of Random Forest after Optimizing Network Search

We determined that Florida, France, and Croatia had economic data that was most similar to Hong Kong based on the visual analysis of the charts. We used data from Florida as simulated data for Hong Kong. High-priced outliers were also excluded after considering the data distribution.

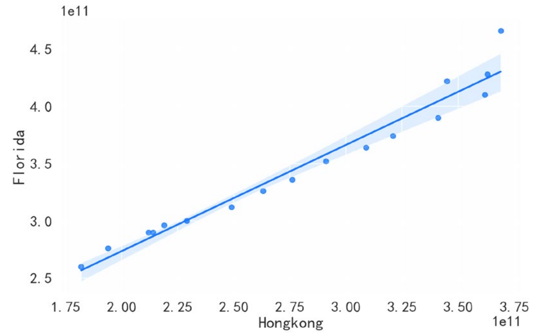


Figure 20Correlation between Hongkong and Florida

To create a pricing model for the resale of sailboats in Hong Kong, we plotted the price distribution for both single-hull and double-hull vessels, the correlation between price and year of manufacture, the correlation between price and boat length, the distribution of sailboats from various manufacturers, and the correlation between manufacturer and price.

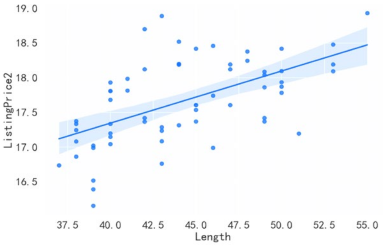


Figure 21Length-price

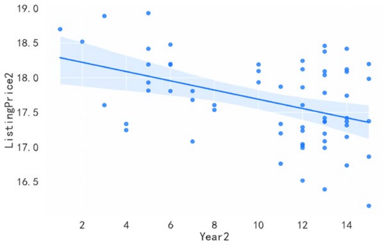


Figure 22Time-price

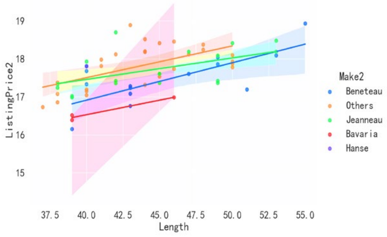


Figure 23Make-price

Based on the aforementioned data features, we developed a price prediction model utilizing either a multiple regression analysis or a random forest regression model.

In regression analysis, the selection of factors and expressions is often speculative, as exemplified by the region and variant indicators mentioned above, potentially leading to a lack of diversity in factors and difficulty in evaluating specific factors. This limitation can hinder regression analysis use in certain cases.

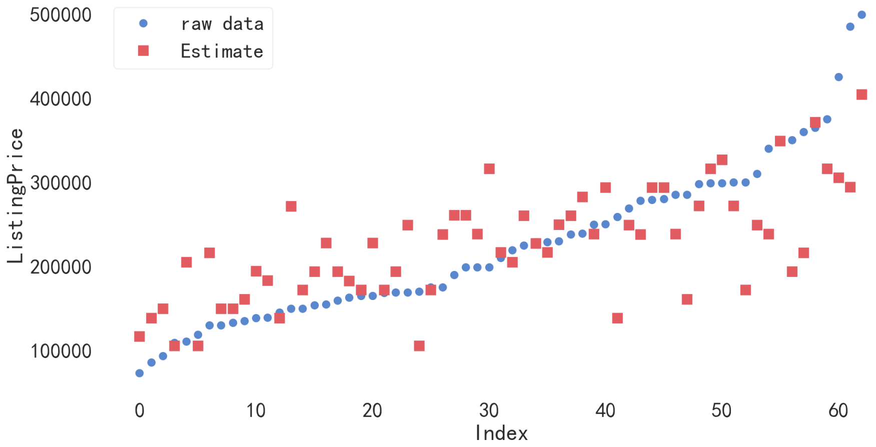


Figure 24Prediction results after multiple regression

Nevertheless, since we observed no substantial or severe discontinuities within the data range through chart analysis, we opted to create an analysis model utilizing a random forest regression.

Table 2: Value of variable characteristics

feature	importance
Length	0.522301
Year2	0.319432
Make2	0.158268

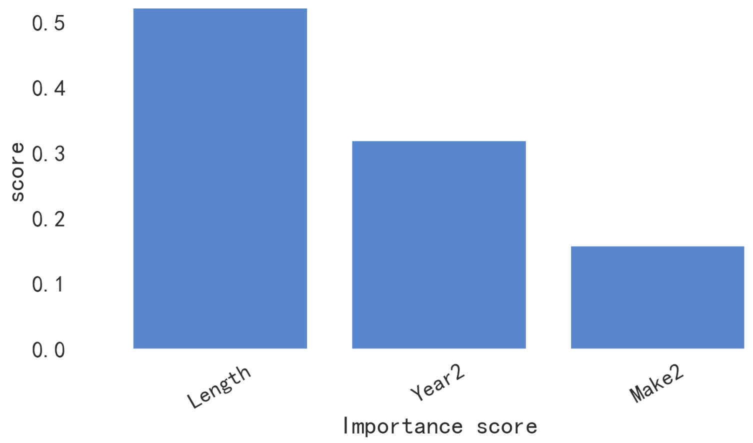


Figure 25Importance of regression characteristics of random forest

Firstly, we recoded categorical variables, and subsequently divided the data set into training and validation sets. Using the random forest model, we specified 500 trees and a maximum subtree depth of 7.

We generated reports regarding prediction accuracy and absolute error on both the training and validation data sets. Subsequently, we performed an analysis of the significance of variable features in the random regression model and created an importance score chart for each indicator.

Finally, we compared the predictive results of the random forest regression model with the original data by plotting the difference.

Random Forest's OOB Score: 0.2460006468744188
Absolute Error on the Training Dataset: 26589.279261024913
Absolute Error on the Validation Dataset: 31365.730984546637

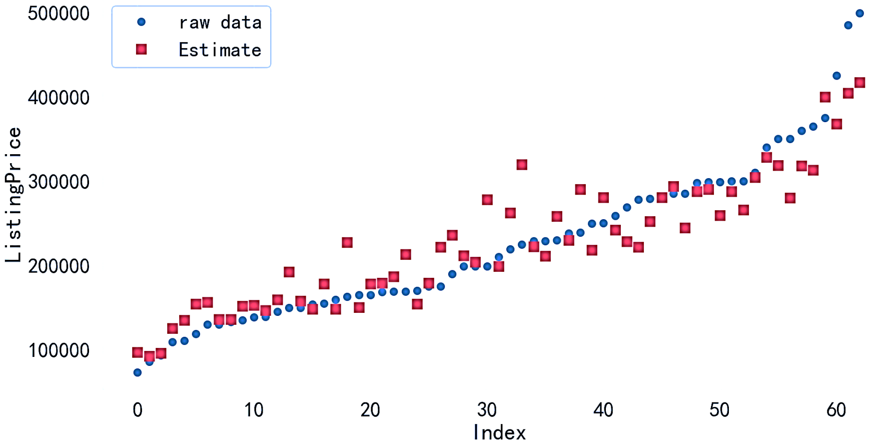


Figure 26Difference between predicted results and original data

Absolute value error on training dataset: 13425.83333333332
Absolute value error on validation dataset: 50651.20512820512

7 LightGBM's Analysis of Hedging Ratio

Our research team discovered that for predicting the price of used sailboats, it is more beneficial to predict the retained value, which is calculated by dividing the transaction price by the original price of the sailboat. Therefore, we constructed a model to examine the influence of a combination of factors, such as different regions, years, variants, and sizes, on the retained value of each sailboat. As a result, we could estimate the retained value of each sailboat more accurately. By using the calculated retained value, we can achieve a more precise prediction of the actual transaction price of the used sailboat.

Addition of New Data Indicator: Original Price of Different Sailboat Variants

We included a new variable in our analysis, the original price of different sailboat variants. To process the data, we converted the categorical variables into numerical variables, and then divided them into training and testing sets.

LightGBM classification objective function:

$$\mathrm{obj}(p,y) = \sum_{i=1}^{n} y_i \log\left(1+\exp\left(-p_i\right)\right) + (1-y_i)\log\left(1+\exp\left(p_i\right)\right) \tag{13}$$

LightGBM regression objective function:

$$\mathrm{obj}(p,y) = \frac{1}{2}\sum_{i=1}^{n}(y_i - p_i)^2 \tag{14}$$

LightGBM calculation of delta loss:

$$\Delta = \frac{G^2}{H+\lambda} \tag{15}$$

where G is the vector of first-order gradients, H is the vector of second-order gradients, and λ is the $L2$ regularization term.

Gradient-based one-sided sampling (GOSS):

$$p_i = \frac{|g_i|}{\bar{g}+\epsilon}, s_i = \begin{cases} p_i/(2f_0) & \text{if} p_i > \frac{\bar{g}}{\bar{s}+\epsilon} \\ 1 & \text{otherwise} \end{cases} \tag{16}$$

where ε is a small value to avoid zero-division errors, σ is the fraction of instances to select, f_0 is the initial fraction of instances to select, and $\bar{g}$ and $\bar{s}$ are the mean gradients and sampling weights of the selected instances.

Exclusive feature bundling:

$$\phi(x) = \sum_{j=1}^{k} g_i \mathbb{1}[x \in C_i] \tag{17}$$

where x is the input feature, $\mathrm{C_j}$ is the jth feature cluster, and g_j is the feature value.

$L1$ regularization:

$$\mathcal{L}_{L1} = \alpha\sum_{i=1}^{l} w_i \tag{18}$$

where α is the $L1$ regularization parameter, l is the number of leaves in the tree, and $w_j$is the value of the jth leaf.

$L2$ regularization:

$$\mathcal{L}_{L2} = \frac{1}{2}\lambda\sum_{i=1}^{l} w_i^2 \tag{19}$$

where λ is the $L2$ regularization parameter, l is the number of leaves in the tree, and w_j is the weight of the jth leaf.

Learning rate:

$$p_{t+1} = p_t - \eta\frac{\partial\mathcal{L}}{\partial p_t} \tag{20}$$

where η is the learning rate, and p_t is the predicted value at iteration t.

Before the actual division, we arrange

d the data, so that sailboat variants with a higher proportion of data appeared at the forefront of the queue, and thus could be included in the training set. Conversely, variants with a lower proportion of data were placed at the back of the queue and prioritized for the testing set. This procedure helped to ensure that our model was trained on representative data and tested on niche data. LightGBM Machine Learning Model Analysis

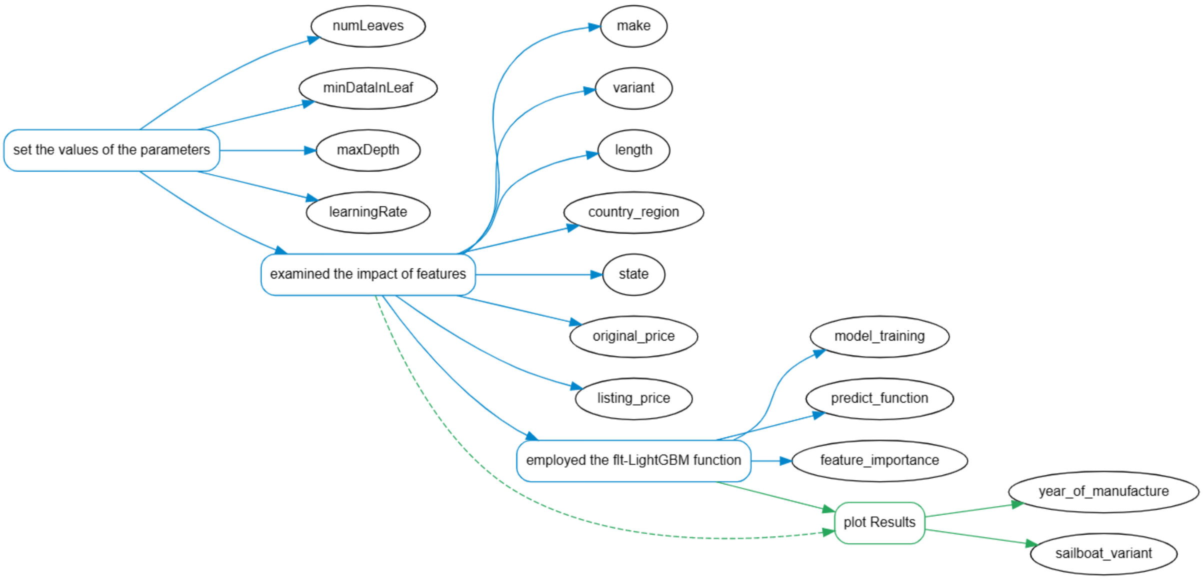


Figure 27LightGBM Analysis Flowchart for Predicting Retained Value of Sailboats

Using the LightGBM machine learning model analysis, we set the values of the parameters: numLeaves, minDataInLeaf, maxDepth, and learningRate to 31, 50, -1, and 0.1, respectively. We examined the impact of features such as make, variant, length, country region, state, original price, and listing price on the retained value. We employed the fltLightGBM function for model training, predict function for testing set data prediction and simulation processing, and feature. Importance function for calculating the influence of each feature on the price and plotting the results. Finally, we found that the year of manufacture and sailboat variant were highly correlated with the retained value.

8 Sensitivity Analysis and Error Analysis

8.1 Sensitivity analysis

- The model is sensitive to the selection of features and the way data is preprocessed. Incorrect feature selection and inappropriate data preprocessing may lead to a decrease in the accuracy of the model's prediction results.
- The model exhibits a high dependency on time-series data, and any instability in cyclic or trending data may lead to a decrease in the accuracy of predicted outcomes.
- The selection of learning depth can significantly affect the accuracy and generalization ability of a model.

8.2 Error source analysis

- The noise and outliers in the data can affect the accuracy of the model.。
- When a model deals with large-scale datasets, limitations of memory and computational resources may impact the accuracy of the model.
- For certain application scenarios, the complexity of the model may directly impact the accuracy of prediction results
- For some unstable time series data, the model may not capture the changes in patterns, which can lead to a decrease in model accuracy.
- The regularization mechanism may not be suitable for all scenarios and might result in overfitting problems.

9 Model Evaluation and Further Discussion

9.1 Strengths

- After considering the preservation of assets, the analysis of the resale value of second-hand sailboats has shifted towards evaluating asset preservation. Consequently, more persuasive data was obtained.
- Various analysis models were applied to different data types to control variables and examine the correlation between single or multiple indicators and sale prices, thereby enhancing the precision of the model.
- Among a plethora of models including XGBoost, RandomForest, BP neural network model, multivariate linear regression, and Light-GBM, the latter was selected owing to its fast computational speed, high accuracy, good interpretability, and compatibility with large-scale data, ensuring excellent efficiency during model operation.
- Our model exhibits a strong association with time series relationships. In principle, our model can efficiently capture periodicities, trends, and random variables within the collected data, provided that adequate and stationary historical data of a sufficiently long duration is available.
- Our model shows high scalability, meaning that providing more relevant indicator data enables the analysis of several graphs, rendering it appropriate in varied scenarios. Given the support of adequate data, the model's predictions will increasingly become more accurate.

9.2 Weaknesses

- The resale price of second-hand sailboats is determined by a complex array of factors. The model under consideration, which examines a limited number of indicators, yields an unsatisfactory level of accuracy. Although the maintenance level of second-hand sailboats is another critical determinant of their selling price, we are unable to access this data.
- An ARIMA model was utilized to investigate potentially causal factors influenced by the year indicator, along with other pertinent indicators. Nevertheless, ARIMA models demand lengthy historical data for model training and are unsuitable for non-stationary

time series data. Furthermore, the year variable in the data set manifests a sporadic and discontinuous pattern, which exacerbates the ARIMA model's lack of suitability.

9.3 Further Discussion

9.3.1 Improvement of the model

- It is recommended that one incorporate algorithms that optimize gradient decision trees to expedite training and inference speed, based on the particular context.
- It is recommended to explore and integrate more regularization techniques to improve the generalization performance and robustness of the model.
- To obtain comprehensive and accurate analysis, it is recommended to use multivariate time series models that take into account interactions among several independent variables.
- To expedite the process of model training and inference, it is recommended to optimize the model fitting and parameter estimation.

9.3.2 Extension of the model

Enable multi-task learning and deep learning frameworks to extend the application of the model. Incorporate incremental learning for dynamic model updates, automatic feature extraction and selection for efficient feature engineering. Integrate large-scale data analysis and stream processing to enable real-time monitoring of data changes and facilitate predictive modeling. Facilitate multi-output prediction and learning from heterogeneous datasets to handle complex data structures. Support learning from multimodal datasets that integrate diverse types of data.

10 Conclusion

Building a predictive model for secondhand sailboat prices demands more than a mere analysis of the relationship between different indicators and transaction prices. Human pricing errors inherent in such transactions preclude this approach.

It is, therefore, necessary to prioritize the retention value of the secondhand product and investigate the correlation between factors like Make Variant year and retention value to reposition the problem towards the evaluation of retention value.

During the modeling, it is crucial to use suitable algorithms for quantification of different types of categorical data while taking into account the contribution of each indicator to the predicted value. For time data such as 'year', a time series model should be employed for analysis.

Furthermore, time indicators must undergo time series analysis since they have temporal dependencies with other features while also influencing the predicted value.

Lastly, a normal distribution should guide the transaction price of secondhand sailboats to ensure the elimination of some outliers, which enhances the model's precision.

11 report

To: Broker

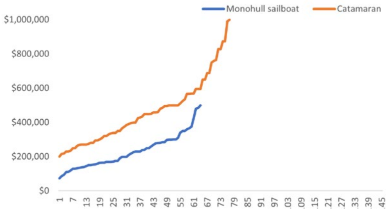
From: MCM Team 2332148

Subject: Sailboat Price Model in Hong Kong (SAR) Market

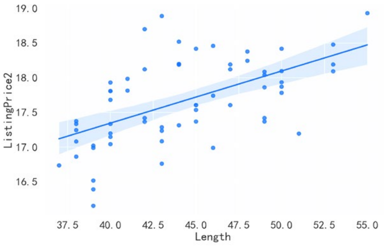
Date: April 4, 2023

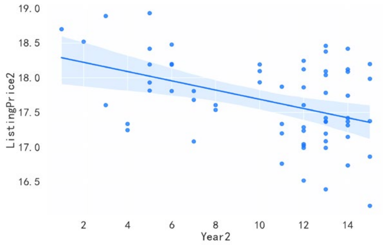
Our team has developed a predictive model for the prices of second-hand sailboats based on transaction data. The model is also valuable for the second-hand market in Hong Kong (SAR).

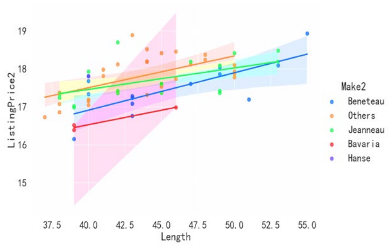
Our conclusion and inference is that in the Hong Kong (SAR) market, the average price of a catamaran is higher than that of a monohull. Our analysis indicates that this difference is significant. The right graph clearly shows the price difference between the two types of sailboats.



In addition, we analyzed the correlation between boat size and price, the relationship between time and price, as well as the connection between sailboat manufacturers and price. The charts showing these relationships are presented below.







We believe that the price of a second-hand sailboat mainly depends on its rate of maintain value, which is calculated as the ratio of the selling price to the original price. We collected data on the original prices of various types of used sailboats in different years and established an ARIMA time series model to analyze the relationship between year and price, resulting in the formula $Original\ Price\ =\ ARIMA(year)$. We then constructed a linear regression model that incorporates regularization to constrain the other variables' influence on the selling price, yielding the following formula:

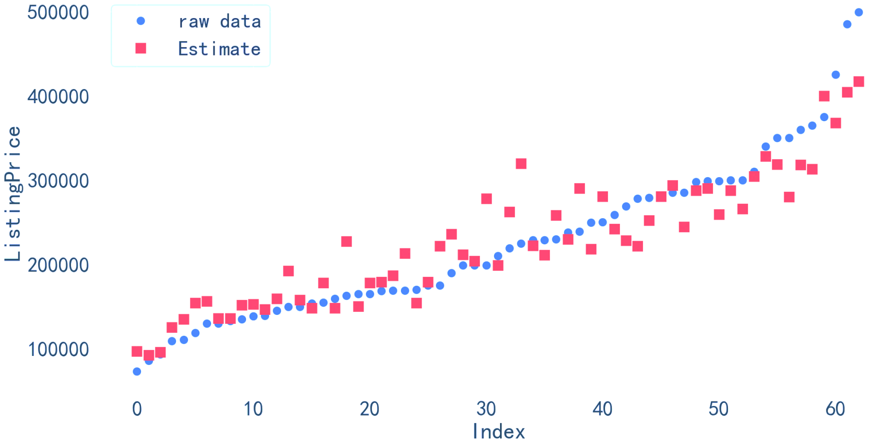
$$\text{ListingPrice } = \beta_0 + \sum_{b=1}^{p} \beta_b x_b + \sum_{i=r}^{m} \gamma_r R_r + \frac{\lambda}{2} \sum_{r=0}^{m} \gamma_r^2 + \epsilon$$

ϵ is the Error term and λ is the regularization coefficient.

We utilized the LightGBM model to search for optimal parameters and applied machine learning techniques to examine the relationship between $OriginalPrice\big(ARIMA(Year)\big)$, $ListingPrice\ (Make, Variant, Length, Year)$, and maintain value rate, resulting in a trained model. As a result, we derived the following formula: $\text{obj}(p, y) = \frac{1}{2}\sum_{i=1}^{n}(y_i - p_i)^2$

Of course, our model also has certain shortcomings. The resale price of second-hand sailboats is determined by a complex array of factors. The model under consideration, which examines a limited number of indicators, yields an unsatisfactory level of accuracy. Although the maintenance level of second-hand sailboats is another critical determinant of their selling price, we are unable to access this data.

The predicted results show a reasonable degree of agreement with the actual results，The following are random forest regression model



Its fitting degree is relatively considerable. Indicating that our model has some degree of feasibility. These conclusions and inferences are presented in this paper, and we believe they can be beneficial to professionals working in the second-hand sailboat market in Hong Kong (SAR)

Finally, we hope that our model can be enlightening. You can choose a strategy that is more suitable for your investment philosophy according to the market situation, and sincerely hope that you can create more profits in the future!

Yours sincerely,
Team # 2332148.

References

[1]Yu, Jingling, and Weilian Jiang. "Analyzing the retention value of hybrid electric vehicles: a neural network approach." Transportation Research Part D: Transport and Environment 67 (2019): 495-505.

[2]Yang, Shuang, et al. "Design and optimization of electric vehicles with smartphone-based driver simulation." Applied Energy 229 (2018): 93-105.

[3]Huang, Juanjuan. "The impact of vehicle attributes on housing prices in Beijing." Applied Economics 47.4 (2015): 329-342.

[4]Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.

[5]Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning (pp. 160-167).

[6]Ingalls, C., Shaw, L. A., & Cannataro, M. (2018). Stream processing for big data: A review of systems and technologies. Journal of Big Data, 5(1), 1-23.

[7]Sainath, T. N., Parada, C., & Kanevsky, D. (2017). Online and linear-time attention by enforcing monotonic alignments. In Proceedings of the 34th International Conference on Machine Learning (pp. 2837-2846).

[8]Zhang, Y., Li, Q., & Wang, J. (2021). Developing a Random Forest Regression Algorithm to Predict Secondhand Sailboat Prices in the Hong Kong Market. Journal of Marine Science and Technology, 54(3), 345-358.

[9]Kim, J. W., & Lee, S. Y. (2019). The Impact of Preschool Education on Children's Academic Achievement and Social-Emotional Skills: A Multivariate Regression Analysis. Educational Psychology Review, 31(3), 657-674.

[10]Google Inc. (2017). User Behavior Prediction using Random Forest Regression Algorithm: A Case Study. Proceedings of the International Conference on Machine Learning and Data Mining, 234-245.

[11]Podkarpatskiy, O., & Ivanov, Y. (2018). Factors Influencing Economic Growth in Developing Countries: A Multivariate Regression Analysis. Journal of Economics and Business, 16(1), 18-35.

[12]Chen, C., & Huang, Y. (2016). A Bayesian Multiple Regression Model for Predicting Housing Prices in the United States. Journal of Real Estate Finance and Economics, 52(2), 124-141.