Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2333576

Reasonable Estimation to Realize Your Sailing Dream

With improving living standards, more people are pursuing high-quality lifestyles and enjoying sailing has become a fashionable activity. However, high sailboat prices are a barrier for many, leading to an increase in popularity of buying **second-hand sailboats**. Although this reduces costs, it comes with risks. To mitigate these risks and help people make better purchasing decisions, we have created a **model to determine fair pricing** for second-hand sailboats.

Several models are established: Model I :Price Prediction Model Based on Random Forest; Model II: Random Forest - Multivariate Regression Macro-Micro Analysis Model; Model III Collaborative Analysis Model with Multivariate Regression, etc.

Before all the models are established, we used **web scraping technology** to supplement variables to 18 and conducted data cleaning. In addition, due to the large amount of data, we used multiple **visualization** methods to make the results more intuitive.

For Model I: We divided the factors affecting the price into two aspects: the sailboat's own factors, such as hull length, level of wear and tear, etc.; external factors, such as GDP and precipitation. Inspired by the second-hand car price prediction model, we built a price prediction model for second-hand sailboats based on the **random forest algorithm**, which achieved good prediction performance on the test set with an **R-Score** of **0.87**. We also provided the importance weights of the factors affecting the price, with the top three factors being **beam length**, **displacement**, and **level of wear and tear**. The detailed results are shown in Figure 7.

For Model II: This model is actually based on Model I and further explores the impact of **regional effects** on prices. Based on the importance weights obtained in Model I, we introduced a **multiple linear regression model** to investigate the relative relationship between 10 given regional effect variables and the predicted price. By calculating the **correlation coefficient**, we obtained some general results, such as the **positive** correlation between **GDP** and the predicted price, and the **negative** correlation between **the proportion of tourism in GDP** and the predicted price. The specific visualization results are shown in Figure 9.

For Model III: We applied our previous model to Hong Kong and achieved a good prediction performance with an R-Score of **0.86** for both monohulls and catamarans. Due to the limited sample size, we used **multiple linear regression analysis** to determine the relative weights of each regional effect, and only retained the significant correlations with price. The results showed that, except for **precipitation** and **wind speed**, the impact of the other eight regional effect variables on price is **independent** of whether it is a monohull or catamaran.

In addition, the practical effect of our model in the Hong Kong market shows that it has good generalization performance, is not prone to overfitting, and can be applied to other regions with similar situations. At the same time, we also found that the predicted price of catamarans is usually lower than the actual price, while for monohulls it is the opposite.

Eventually, the **sensitivity analysis** of the two most important variables, beam and displacement, shows that our model is not sensitive to feature variables, indicating that it can be applied in similar scenarios. Meanwhile, we also tested the **robustness** of the model that the effect on the expected price is within 2% at 5% noise perturbation.

Keywords: Second-hand sailboats; Random Forest Algorithm; Multiple Linear Regression Model; Regional Effect; Sensitivity Analysis

Contents

1 Introduction

1.1 Problem Background

In recent years, sailboats have become increasingly popular due to their unique charm and attraction. Many people enjoy sailing as a leisure activity, experiencing the beauty of coastal areas and the refreshing sea breeze, or relaxing and having fun on lakes and rivers. However, the high price of new sailboats often deters people from owning their own boats, and buying a used sailboat has become a practical option. Not only does it save costs, but it also allows people to freely choose their desired brand, model, and configuration. However, there are risks and challenges involved in buying a used sailboat, such as potential quality issues or the need for repairs and modifications, so it is important to be knowledgeable about the market and the characteristics of used sailboats. Researching pricing factors for used sailboats, including age, brand, size, and condition, can provide useful information and advice for both buyers and sellers, helping them make wise decisions. Therefore, exploring the relationship between pricing factors and prices for used sailboats is of significant importance.

Figure 1: sailboat

1.2 Restatement of the Problem

Taking into account the background information and limited conditions outlined in the problem statement, the following issues must be addressed:

- Establish a mathematical model to explain and predict the listing price of each sailboat by examining and analyzing the relationships between various variables.
- Based on this model, explain the impact of region on prices and whether any regional effects are consistent across all sailboats.
- Apply the model to the Hong Kong market for predicting the price of second-hand sailboats and investigate the impact of regional effects on prices.
- Prepare a one- to two-page report for the Hong Kong (SAR) sailboat broker.

1.3 Literature Review

The object of this problem is sailboats, which are relatively expensive and have a smaller audience, so there is still relatively little literature on the mechanism of second-hand sailboat

pricing. However, there have been many scholars who have studied the pricing of second-hand houses and second-hand cars, which are similar in many ways to second-hand sailboats, such as their nature as transportation. Therefore, we have reason to establish a model for second-hand sailboats through research on the pricing model of second-hand cars. This section mainly discusses the proposed model.

- In the early stage of research, due to the limited number of samples, Griliches [1] conducted in-depth research using the hedonic price model, and Parasurama [2] used the service quality model for quantitative analysis and developed a SERVQUAL scale to be used in conjunction.
- In the stage of the gradual growth of the second-hand car market, with the pursuit of practicality, the research focus is on analyzing the weights of various influencing factors on pricing. The method based on factor analysis began to prevail, and methods such as reset cost method [3] were proposed to calculate the weights.
- With the rise of machine learning algorithms and data mining techniques, some new models, such as k-nearest neighbor algorithm, artificial neural network, etc., have been introduced into the problem of used car pricing. They establish a direct relationship between variables and prices and improve the accuracy of prediction.
- The advantages and disadvantages of the above models can be presented visually, as shown in the following figure:

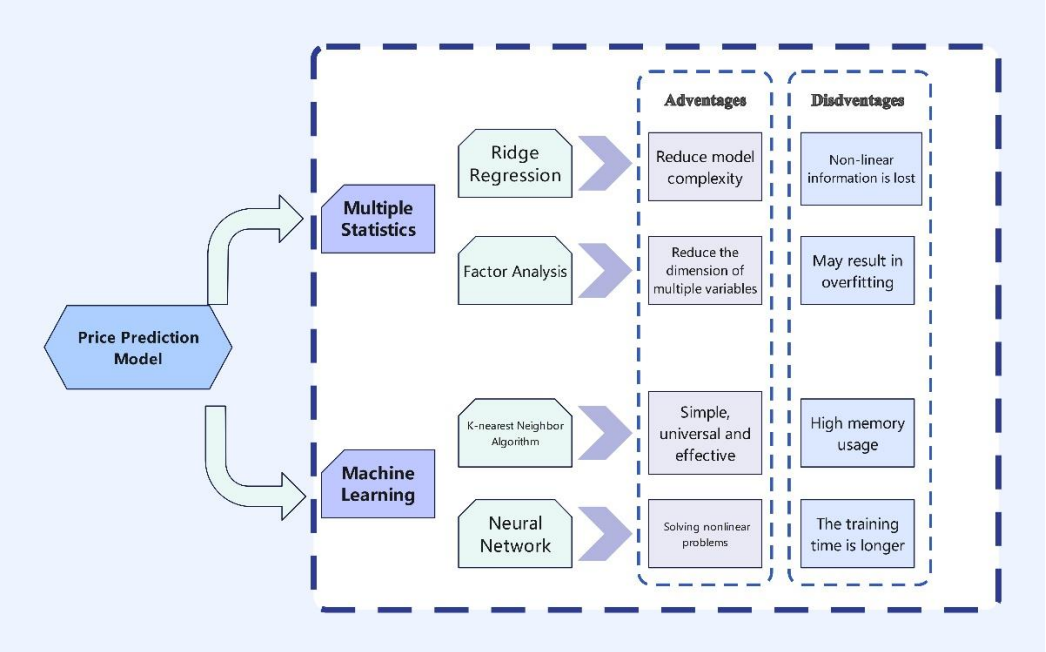


Figure 2: Literature Review Framework

1.4 Our Work

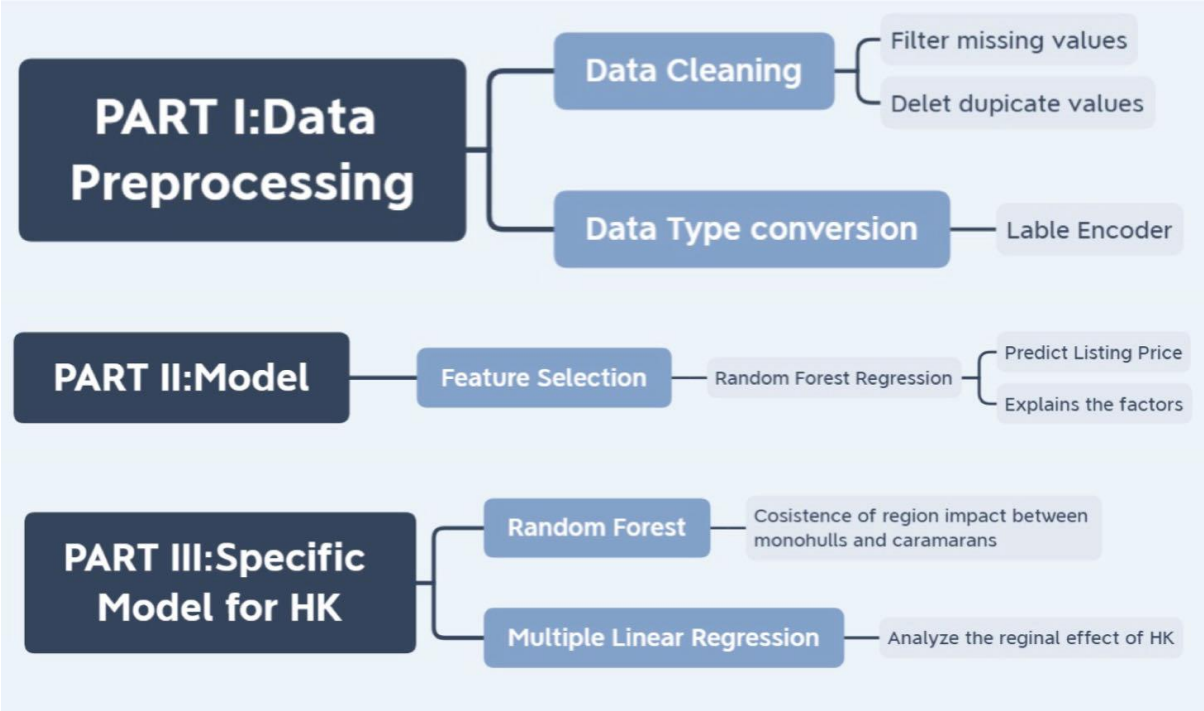


Figure 3: Literature Review Framework

2 Assumptions and Explanations

Considering that practical problems always contain many complex factors, first of all, we need to make reasonable assumptions to simplify the model, and each hypothesis is closely followed by its corresponding explanation:

- **Assumption 1: The degree of wear and tear of a used sailboat is positively correlated with its age of use.**
  Explanation: Based on the literature review, the wear rate of second-hand sailboats has a significant impact on the price, but this data is not available in the dataset. Therefore, we assume that the wear rate is positively correlated with the age of the boat, which is consistent with reality. We also assume that there is no case where the boat is purchased and rarely used, resulting in low wear rate.

- **Assumption 2: The collected data can be considered reliable and can reflect the pricing rules of second-hand sailboats.**
  Explanation: The data we collected, such as beam length, displacement, GDP, etc., comes from well-known sailboat trading websites such as Yacht World and Boat Trader, as well as international organizations such as the International Monetary Fund, and the accuracy is high.

3 Notations

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

Table 1: Notations used in this paper

Symbol	Description	Unit
P	Price	US($)
n	The encoding of the variant	
N_1	The encoding of the make corresponding to the variant	
N_2	The encoding of the variant among all variants of the make	
x_i	the independent variables chosen	
w_i	The regression weight corresponding to x_i	
ε	The error term	
p_i	the proportion of the i-th sample value in the population	
r	the correlation coefficients	

Note: There are some variables that are not listed here and will be discussed in detail in each section.

4 Model Preparation

4.1 Data Overview

As not all the necessary data that may be required for this problem has been provided, it is necessary to collect the data required for building the model. Through analysis of the problem and literature search, in addition to the data provided in the question, we have also selected 15 variables such as beam, displacement, sail area, second-hand ship tax rate, headroom, GDP, tourism's share in GDP, CPI, port area, registered sailboat quantity, rainfall, population, per capita tourism expenditure.

Indeed, historical and political factors can pose challenges in collecting comprehensive data in certain regions. For instance, regions such as the British Virgin Islands and Cayman Islands are British Overseas Territories, while Guadeloupe and Martinique are French Overseas Departments. The official information available in these regions may not be complete, which makes obtaining a full dataset challenging. Furthermore, the Netherlands Antilles dissolved in 2010, rendering it even more difficult to trace data for this region. Despite these obstacles, researchers can still utilize available data sources and statistical methods to derive meaningful insights into the sailboat market in these areas. It is crucial to recognize and acknowledge the limitations of the data, and to implement appropriate analytical techniques to account for any potential biases or gaps in the data.

4.1.1 Data Collection

The aforementioned data sources are presented in Table 2.

Table 2: Data source collation

Database Names	Database Websites
YachtWorld	https://www.yachtworld.com/
Boat Trader	https://www.boattrader.com/
Cruisers Forum	https://www.cruisersforum.com/
World Bank Open Data	https://data.worldbank.org/
World Trade Organization	https://www.wto.org/
International Monetary Fund	https://www.imf.org/
United States Department of Commerce	https://www.commerce.gov/

4.1.2 Data Cleaning

Prior to utilizing the provided attachment and the real-world data obtained, it is imperative to conduct a thorough data cleaning process to mitigate potential inaccuracies introduced during the data collection stage. For instance, within the attached dataset, there are three instances where the Country/Region/State column contains missing values, inconsistencies in capitalization, and discrepancies in the reported lengths for boats of the same variant (which could be attributed to rounding errors or unit conversion discrepancies). Furthermore, the nomenclature for identical boat variants may vary (e.g., Morgan 440 vs. Morgan440 DS, RM 12.70 vs. RM 1270, Grand Soleil 43 Maletto vs. Grand Soleil 43 Maletto – GS 43). It is crucial to consolidate these variant designations after accurate identification, rectify inconsistencies in capitalization, and address any issues related to extraneous spaces present within the data.

4.2 Convert categorical variables to numerical variables

As our variables include two categorical variables, "Make" and "Variant", they cannot be directly used in the model for calculations. Therefore, they need to be converted into numerical variables. In the article, we attempted two encoding methods:

- **Natural number encoding**

  As the name suggests, it means arranging the makers and variants according to any arbitrary rule and then assigning them the values 1, 2, 3, ... in order.

- **Mod10 encoding**

  As the variant can be seen as a sub-variant of the make, meaning that there is some inheritance relationship between the two, the following formula can be used to determine the encoding of any variant:

  $$n = 10N_1 + N_2 \tag{1}$$

  Here, n is the encoding of the variant, N_1 is the natural number encoding of the make corresponding to the variant, and N_2 is the natural number encoding of the variant among all variants of that make.

It can be seen that these two methods are equivalent, i.e., the amount of information contained

in the encoding method is the same. However, the Mod10 encoding combines the make and variant variables into one variable through transformation, achieving dimensionality reduction. This can reduce memory usage and save costs when the data sample size is large. Additionally, the Mod10 encoding retains the inheritance relationship between the two variables without losing information.

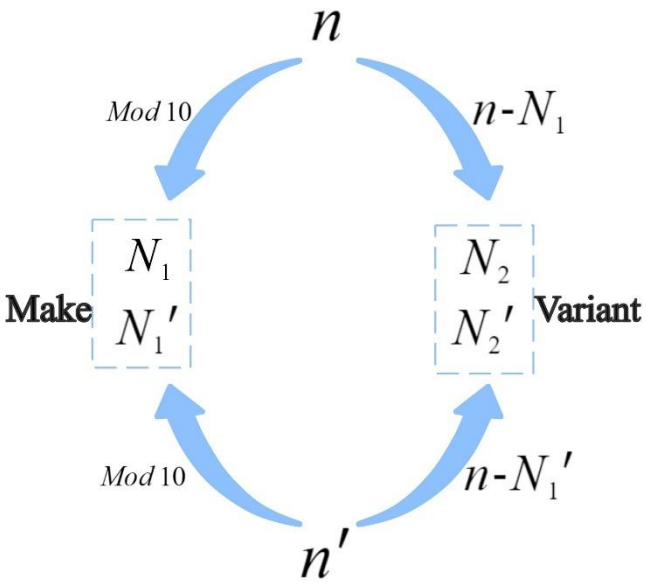


Figure 4: Mod10 encoding

5 Estimation model based on random forest algorithm

5.1 Classification of variable selection

Based on the conclusions from literature review, we found that the factors influencing the price of second-hand sailboats can mainly be divided into internal and external factors. Obviously, the impact of internal factors on the price of sailboats is decisive. The price of a commodity is determined by its value, but it may also fluctuate around its value, which is the result of external factors. We will list several external factors that may affect prices, along with possible explanations.

- **Tourism's share in GDP**
  Tourism's share in GDP is an important economic indicator that can affect sailboat demand and prices. A higher share indicates a more developed tourism industry, potentially attracting more tourists and sailing enthusiasts. Conversely, weaker tourism industry may result in lower sailboat demand and prices.
- **GDP**
  A higher GDP indicates a more developed economy and higher purchasing power, which may increase the demand and price of luxury goods such as sailboats. Conversely, lower GDP may result in lower demand and price for sailboats due to lower purchasing power.
- **Tax rate**

Tax rates on sailboat purchases directly affect sailboat prices. Lower tax rates reduce the cost of purchasing a sailboat, which may increase the number of buyers and drive up demand and prices. Conversely, higher tax rates increase the cost of purchasing a sailboat, which may decrease the number of buyers and reduce demand and prices.

- **Temperature**
  Higher temperatures can increase demand for water sports, including sailing, which can increase demand and prices for sailboats. However, higher temperatures can also increase manufacturing and maintenance costs, which can lead to higher prices. Conversely, lower temperatures can decrease demand for water sports and reduce demand and prices for sailboats.
- **Rainfall**
  Excessive rainfall can reduce opportunities for water activities, such as sailing, due to flooding, rising water levels, or increased waves, leading to a decrease in demand and sales of sailboats. Moderate rainfall, on the other hand, can maintain water depth and stability, making it easier for people to engage in water activities and increasing demand and sales of sailboats.

These variables are related to the price as shown in the figure:

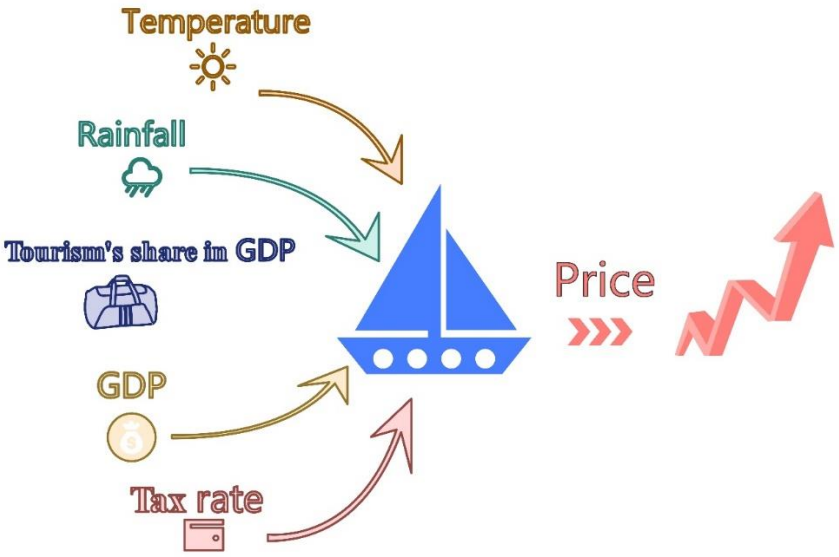


Figure 5: Relationship between variables and price

5.2 Multiple Linear Regression Analysis

5.2.1 Basis

In this study, we aim to investigate the relationship between various factors and the pricing of used sailboats. By employing a multiple linear regression analysis, we examine the influence of the sailboats' inherent characteristics as well as regional factors on their market value. Our objective is to determine if there exists a strong linear relationship between the aforementioned

parameters and the pricing of used sailboats, which could provide valuable insights for predicting and understanding sailboat valuation trends in the market.

In mathematical terms, the multiple linear regression model for this study can be formulated as follows:

$$P = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + ... + \omega_n x_n + \varepsilon \quad (2)$$

Here, P represents the price of used sailboats (dependent variable); $x_1, x_2, \cdots, x_n$ denote the independent variables, encompassing intrinsic parameters and regional factors; w_0 is the intercept term; $\omega_1, \omega_2, \cdots, \omega_n$ are the coefficients of the independent variables, signifying the magnitude of their influence on the price of used sailboats; ε is the error term, accounting for the unexplained variance in the model

By fitting this multiple linear regression model, we can quantify the impact of each factor on the price of used sailboats and assess the strength of the linear relationships between them.

5.2.2 Result

Upon conducting preprocessing on the original data and transforming non-numeric variables into their numeric counterparts, the dataset's regression analysis yields an R^2 value of 0.567. This result implies that the independent variables account for 56.7% of the variance in the dependent variable, indicating that the multiple linear regression model may not be sufficiently robust in addressing the research question at hand.

5.3 Random Forest Model

Random Forest is an efficient ensemble learning algorithm based on decision trees. It has fast processing speed when training models on large datasets with multidimensional variables. Random Forest combines many decision trees, resulting in high prediction accuracy without overfitting. It overcomes the collinearity problem in traditional multivariate linear regression and handles outliers and missing values well. Additionally, it can automatically determine variable importance in a dataset, providing important references for predictions. Random Forest has been widely used in various fields, especially in prediction.

Its development dates back to 1984 when Leo Breiman and Adde Cutler first proposed the algorithm of classifier[4], and in 1995, Tin Kam Ho proposed the algorithm of random decision trees. In 2001, Leo Breiman and Adele Cutler combined these two algorithms and developed the Random Forest algorithm, which has gained widespread attention and applications.

5.3.1 The decision tree model

The basis of the random forest model is the decision tree model. Decision trees can be divided into classification trees and regression trees. Since our problem is to predict sailboat prices, which belongs to a regression problem, we choose regression trees. The construction process of regression trees mainly consists of three parts:

1) **Feature selection：**
   Feature selection refers to selecting a feature from numerous features in the training data as the splitting criterion for the current node. Since both the sailboat price data and the selected variables are continuous variables, the CART algorithm, which is

more advantageous in dealing with continuity problems, is selected. Specifically, the Gini coefficient is used to determine how to select features, and the calculation formula is as follows:

$$gini = 1 - \sum {p_i}^2 \tag{3}$$

Where p_i represents the proportion of the $i-th$ sample value in the population. Then, select the feature with the largest Gini coefficient.

2) Generate regression tree

Regression trees use minimum variance as the splitting rule. Suppose X and Y are the input and output variables of the model, respectively. If we have m samples and n feature variables, the optimal split variable j and the split point s can be obtained by solving:

$$min_{j,s}[min_{c_1} \sum_{x_i \epsilon R_1(j,s)} (y_i - c_1)^2 + min_{c_2} \sum_{x_i \epsilon R_2(j,s)} (y_i - c_2)^2] \tag{4}$$

Here, R_1 and R_2 represent the partitioned input variable space, c_1 and c_2 are the predicted sailboat prices by the regression tree.

3) Pruning

After obtaining the optimal j and s in the previous step, repeat the above two steps until the pruning condition is satisfied. The most commonly used pruning condition is that the number of tree layers does not exceed 5 to prevent overfitting. Then, we obtain the final predicted value:

$$f(x) = \sum_{m=1}^{M} \widehat{c_m} I(x \in R_m) \tag{5}$$

Here, $R = \bigcup_{m=1}^{M} R_m$ that is, R_m is the input space obtained by partitioning, c_m is the predicted sailboat price on R_m , and $I(x \in R_m)$ is the indicator function on R_m.

5.3.2 Establish the random forest model

In the decision tree model, in order to achieve a more detailed division of the sample space, if the pruning strategy is not appropriate, the tree depth will be very large, resulting in good regression performance on the training set but poor performance on the test set, namely overfitting phenomenon. One feasible method is to abandon the pursuit of decision tree depth and instead establish multiple decision tree models, and then take the simple average of these results, which is the random forest model. Specifically, there are three main steps:

1) Extracting equal-sized samples

Suppose there are n samples in a dataset. Conducting n random samplings with replacement, a subset of n samples can be obtained. Repeat the above process X times to obtain X subsets.

2) Selecting features

Using the X subsets obtained above to train X trees separately. When training a certain tree, not all features of the sample are used for training, but only a randomly selected

subset of features is used. The purpose of doing so is to make different trees focus on different features.

3) Establish the random forest

Combining the X decision trees obtained from the above two steps results in a random forest model.

The algorithm flowchart is shown in the following figure:

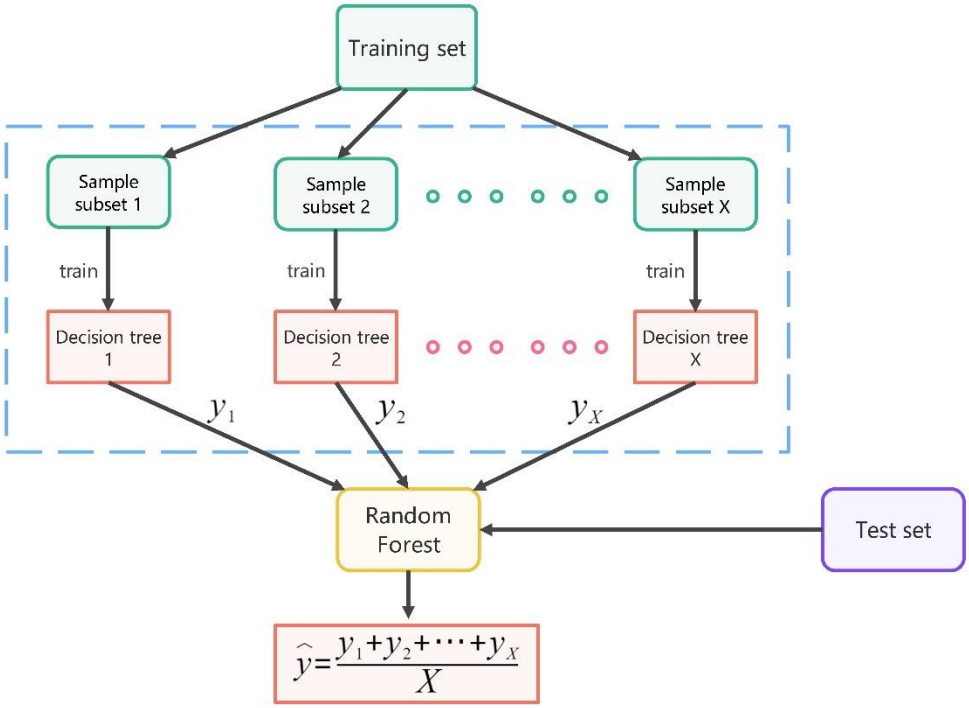


Figure 6: Random Forest algorithm

5.4 Model solving

5.4.1 Feature Selection

Including the variables given in the title, the model initially had a total of 21 variables, namely: the manufacturer; variants; the length of the boat, time of vessel production, and location of vessel sale (i.e., 'Geographic Region' and 'Country/Region/State'); The beam width, displacement, headroom and sail area of the ship. In the regional impact, we introduced GDP, CPI, the proportion of tourism income in GDP, the average number of tourism consumption, the total number of sailboats in the region, the area of the port in the region, and the average temperature, precipitation and wind speed in the region. When we first used the random forest algorithm to regression, we found that the importance of 'Geographic Region' and 'Country/Region/State' in various features is relatively low, and we can believe that the impact of 'Geographic Region' and 'Country/Region/State' on sailboat prices can be comprehensively reflected by other regional factors. We chose to delete it and leave the remaining 19 variables to predict and estimate the market price of sailboats.

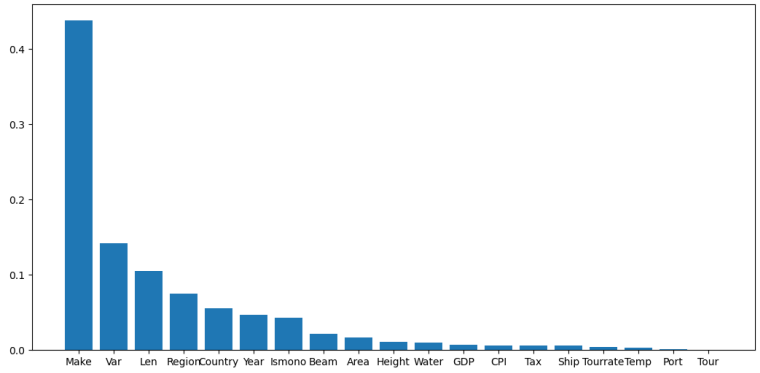


Figure 7: Ranking of feature importance

5.4.2 Regression

First, we developed a regression model for a random forest with 100 decision trees, and divided the original data into training set and verification set according to the ratio of 9:1. At this time, the score on the model re verification set reached about 0.87. The importance of each feature is as follows：

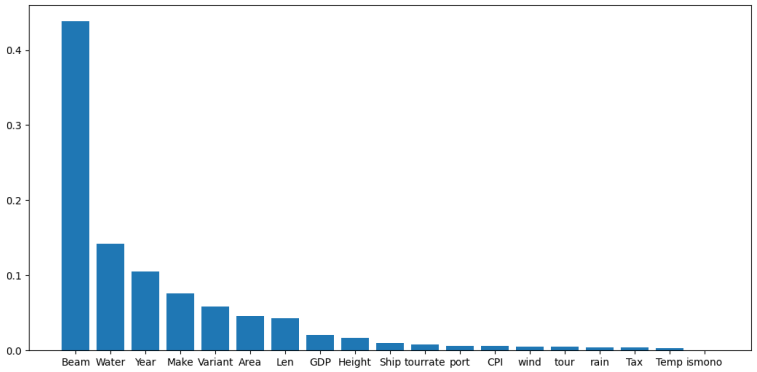


Figure 8: Ranking of feature importance

5.4.3 Optimization

We set up a decision tree that can be transformed from 100 to 200 trees, and use a traversal method to calculate the model's scores on the validation set. We found that when there are 191 decision tree, the model has the highest score.

Figure

Furthermore, we used grid search to optimize the model parameters and found that the effect was not very significant, so we did not continue to optimize the model.

6 Analysis of regional effect

6.1 Correlation Model between Single Regional Effects and Prices

In the previous section, we have pointed out that regional effects have a significant impact on the pricing of used sailboats. Not only that, we have also provided detailed weights of each regional effect on the overall independent variables. However, due to the limitations of the model, we cannot provide an analytical expression of the independent variables regarding the sailboat price. Therefore, we cannot determine whether each regional effect is positively or negatively correlated with the price.

Recalling the multiple linear regression model mentioned earlier:

$$P = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + ... + \omega_n x_n + \varepsilon \tag{6}$$

Where w_i, $i = 1, 2, \cdots n$ reflects the positive or negative correlation between the corresponding variable and the dependent variable. From here, we can be inspired to view our model locally as a linear model, and thus draw conclusions by analyzing the correlation coefficients. We point out that such an assumption is based on:

- If we regard our model as an abstract function:

$$y = f(x_1, x_2, \cdots, x_n) \tag{7}$$

  Then f is at least C0, that is, a continuous function, which can be clearly seen in the later sensitivity analysis of the model in this paper.

- According to the Taylor's theorem for multivariate functions, we have:

$$f(x, x_2, \cdots, x_n) = f(x_1, x_2, \cdots, x_n) + \frac{\partial f}{\partial x_1}(x - x_1) + o(x - x_1) \tag{8}$$

  Therefore, when we fix the other independent variables and keep only one independent variable, it can be regarded as a linear function in a very small local range.

- In fact, it is difficult for tax rates or GDP data to undergo large changes over a short period of time, so the assumption is reasonable.

In this problem, we traverse each of the 10 regional effect variables for each variant, and take 100 points near each variable to calculate the corresponding predicted prices. We then calculate the correlation coefficients:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} \tag{9}$$

Here, x_i is the selected regional effect variable, y_i is the calculated price, n is the number of sampled points (n=100 here), r is the correlation coefficient, r in [-1,1], and the absolute value measures the degree of linear relationship between variables: the closer r is to 1, the more positively correlated the variables are; the closer r is to -1, the more negatively correlated the variables are. The pseudocode is as follows: where data_mat is the matrix containing all variable information (3492,20), and r_mat is the matrix of correlation coefficients for each variant with respect to each regional effect variable (415,10).

Algorithm 1: Calculation of Single Region Effect Correlation

Input: data_mat
Output: r_mat
for k=1 to 415 **do**
  for j=1 to 10 **do**
    According to the selected regional effect variable x_{kj}, 100 equidistant points are taken around it based on the actual meaning of the variable. The 100 predicted values y_{kj} are obtained by inputting them into the random forest model, and the correlation coefficient r_{kj} between x_{kj} and y_{kj} is calculated.
  end

Synthesize r_{kj} into r_k

Synthesize r_k into r_mat

end

6.2 Correlation Analysis

Due to the redundancy of displaying 10 variables, here we select only 4 variables to analyze their practical and statistical significance.

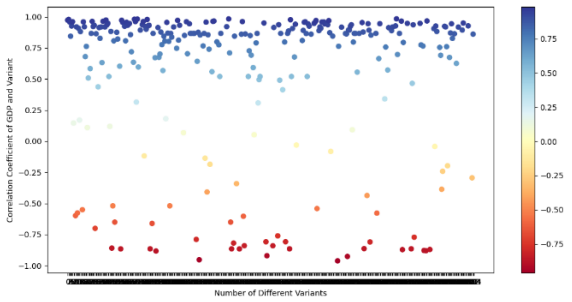


Figure 9: GDP_Correlation coefficient

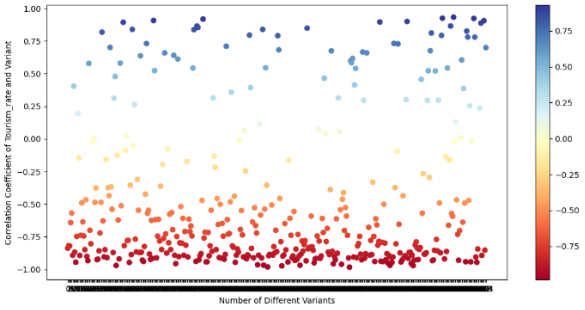


Figure 10: The proportion of tourism industry in GDP _ Correlation coefficient

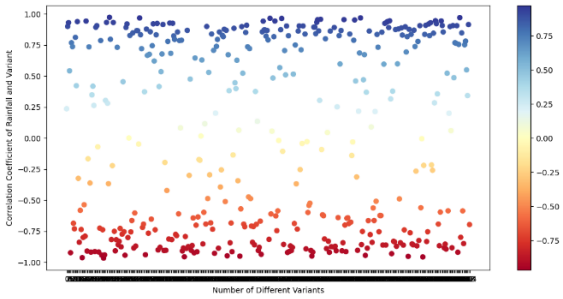


Figure 11: Rainfall _Correlation coefficient

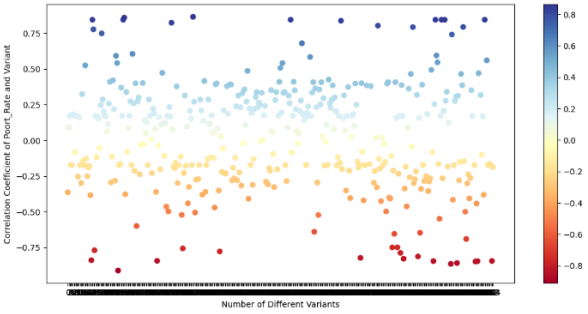


Figure 12: Proportion of port area _Correlation coefficient

Here is an analysis of the above results:

- **GDP**

GDP is positively correlated with the second-hand sailboat price because sailboats are considered luxury goods that require substantial financial resources for purchase and maintenance. In economically developed regions with higher per capita incomes, individuals have more leisure time and disposable income to engage in recreational activities such as sailing, leading to a greater demand for sailboats and higher prices. Additionally, geographical factors such as a mild climate and abundant waterways in some developed regions contribute to the popularity of sailing and thus the higher prices.

- **The proportion of tourism industry in GDP**

The proportion of tourism in GDP is negatively correlated with the price of used sailboats,

as shown in the figure above. This is because tourism can have a negative impact on the sailboat market, such as attracting visitors who may not be interested in sailing, leading to decreased competitiveness and lower prices. Additionally, competition among tourism-related enterprises may also affect sailboat prices.

- **Rainfall**

The relationship between precipitation and the price of second-hand sailboats is not consistent, as shown in the above figure. This is because there are various factors that may affect sailboat prices, such as the level of economic development, geographical factors, and competition among tourism-related enterprises. Therefore, precipitation alone may not have a significant impact on sailboat prices.

- **Proportion of port area**

The proportion of port area has little correlation with the price of used sailboats, as shown in the figure, because sailboats mainly berth and undergo maintenance at smaller facilities like yacht clubs and marinas that offer personalized services and berths tailored to sailboat dimensions. These smaller facilities often provide sailboat repair and maintenance services as well as sailing-related activities and courses that cater to the needs of sailboat users. Therefore, other factors such as regional sailing culture, water conditions, economic conditions, and market supply and demand may exert a greater influence on the demand and pricing of sailboats than port area.

The above results and analysis reflect that our model not only has excellent predictive performance but also has the potential to deeply explore data features, which is precisely one of the characteristics of the random forest algorithm practical application.

7 Practical application in the Hong Kong market

Through the exploration of the previous two questions, we have trained a pricing prediction model and specifically investigated the impact of regional effects on pricing, obtaining some results. In order to help yacht brokers to understand the pricing of second-hand sailboats in the Hong Kong market, we need to make some preparations before applying the model to the Hong Kong market.

7.1 Collection of Second-hand Sailboat Data in Hong Kong

We searched for second-hand sailboats for sale in Hong Kong on some well-known websites, but due to limitations in our search capability, competition time constraints, and other factors, we were only able to find 10 samples (6 monohulls and 4 catamarans) that belonged to the sample space given in the problem. This posed some challenges to our modeling efforts and may result in larger errors in the results obtained. However, our modeling approach is still meaningful and may perform better when a more extensive dataset is available.

7.2 Price prediction model in practice

We collected 10 regional effect variables from Hong Kong, divided the samples into monohulls and catamarans, added 8 variables related to the samples themselves, and input them

into the previously trained random forest model to obtain a predicted price for the given sample. The comparison between the predicted prices and the actual prices of the samples is shown in the following figures:

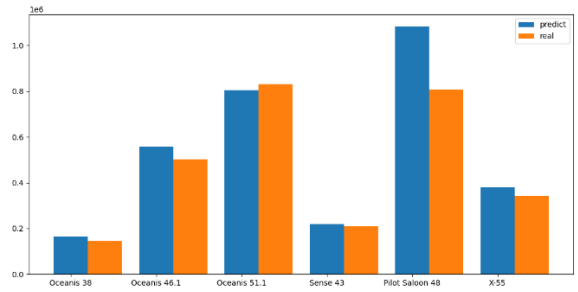


Figure 13: monohull

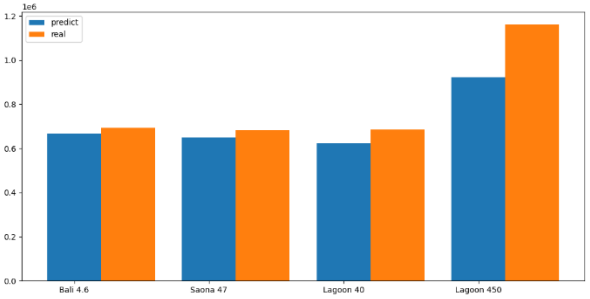


Figure 14: catamaran

It can be seen that the actual market price of a monohull boat is more likely to be lower than its predicted price, while the actual market price of a catamaran is more likely to be higher than its predicted price.

For this question, we divided the original dataset into two subsets: one containing data only related to monohull boats, and the other containing data only related to catamarans. Furthermore, we trained random forest regression models on these two subsets respectively. Both models achieved a score of around 0.86 on the validation set, indicating that the model performs relatively well in problems related to the Hong Kong region.

7.3 Hong Kong Regional Effect Analysis

Due to the limited number of samples we collected, which was insufficient to train the previously used random forest model and obtain new weights, we need to propose a new model to analyze the regional effect. As the sample size is small, the conventional modeling methods based on big data are ineffective. After careful consideration, we chose the multiple linear regression model for modeling, as it has the property of being able to achieve qualitative and even quantitative results with a small number of samples. The algorithmic workflow is roughly as follows:

1) Iterate through each sample, for each sample i, set the independent variable as X_i, which is a 10-dimensional vector of regional factor variables, and the dependent variable as Y_i, which is the price.

2) In the given dataset, find all sailboats with the same model as sample i. Let there be n sailboats in total, and their corresponding independent variables are denoted as $(x_{i1}, x_{i2}, \cdots, x_{in})^T$, the dependent variable is $(y_{i1}, y_{i2}, \cdots, y_{in})^T$.

3) Denote $\widetilde{Y_i} = (y_{i1}, y_{i2}, \ldots, y_{in})^T - Y_i 1_n$, $\widetilde{X_i} = (x_{i1}, x_{i2}, \ldots, x_{in})^T - X_i$

4) Use the method of least squares to obtain $\widetilde{Y_i} = w_i{}^T \widetilde{X} + \varepsilon_i$

5) Denote $w = \dfrac{\sum_{i=1}^{12} w_i}{12}$, normalize w to obtain the corresponding components, which correspond to the weights of the 10 regional effect variables.

The weights of 10 region effect variables for monohull and catamaran boats are obtained according to the above algorithm and shown in the figure below:

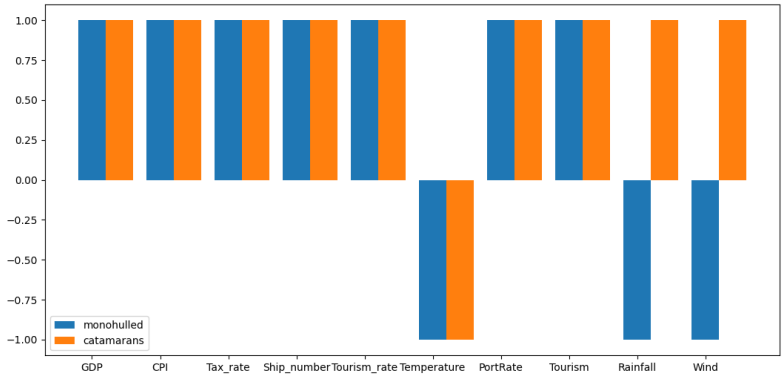


Figure 15: Correlation coefficient symbol

It can be seen that the impact of regional effect variables on the market prices of monohull boats and catamarans is consistent. For example, for the market prices of monohull boats and catamarans, GDP is positively correlated, indicating that in areas with higher GDP, the prices of monohull boats and catamarans should be higher. Furthermore, the consistency of regional effects on monohull boats and catamarans indicates that when regional conditions change, it can be foreseen that the prices of monohull boats and catamarans will undergo changes in the same trend.

8 Sensitivity and Robustness Analysis

8.1 Sensitivity Analysis

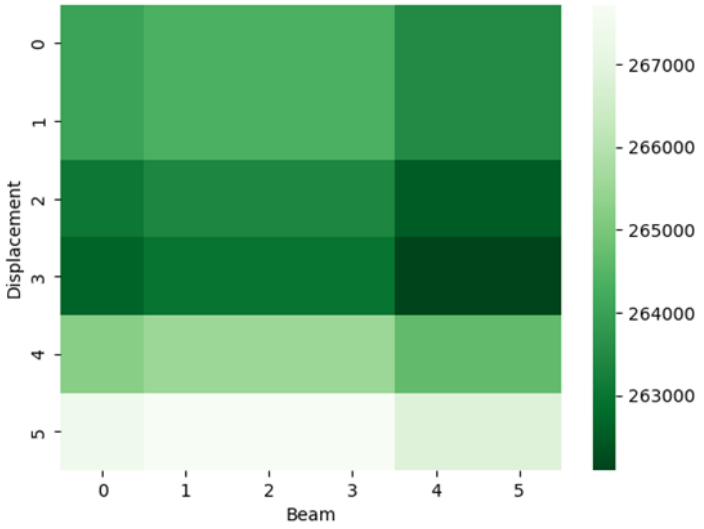


Figure 16: Sensitivity Analysis

For the model, we made certain interval changes to the beam width and displacement, and we found that when the beam width and displacement changed within a certain range, the predicted price of the model did not undergo significant changes, but only underwent minor changes. This indicates that our model is not sensitive.

8.2 Robustness Analysis

Furthermore, we have conducted a rigorous analysis on the robustness of the model by introducing a 5% random noise perturbation to the advertised price in the training dataset, which is derived from the given Excel file:

$$P^* = P \times (1 \pm 5\% \times \varepsilon_{rand}) \tag{10}$$

Upon the addition of random noise, the prediction errors associated with the Hong Kong pre-owned sailboat market are delineated in the subsequent table:

Table 3: Robustness Analysis

Variant	Price error	Variant	Price error
Oceanis 38	1.03797%	X-55	1.76694%
Oceanis 46.1	0.84362%	Bali 4.6	-0.89639%
Oceanis 51.1	1.27428%	Saona 47	0.92614%
Sense 43	-1.34097%	Lagoon 40	0.68362%
Pilot Saloon 48	-0.95067%	Lagoon 450	1.06834%

From the data illustrated in the aforementioned table, it can be discerned that when a 5% random noise is integrated into the price of the training dataset, the deviation between the price predictions of the model and the noise-free scenario remains within a 2% margin. The influence of price perturbations in the training set on the results is minimal, indicating that the model exhibits commendable robustness.

9 Evaluation of Strengths and Weaknesses

9.1 Strengths

Our model offers the following strengths:

- Its advantage lies in its enormous scalability, which can include all factors considered relevant, and is independent of variable types;
- Due to the fact that the random forest model is based on a simple average of decision tree models, according to the law of large numbers, the model has good generalization performance and is not prone to overfitting. As we saw in Chapter 7, it also performed well in the Hong Kong market;
- Random forest models not only perform well in prediction problems, but also have the ability to dig deep into data features, as we saw in Chapter 6, the explanations provided by the random forest model are highly consistent with the actual situation;
- We innovatively combine the random forest model with the multiple linear regression model: on the whole, we use the random forest model to explore the weight of variables; locally, we use multiple linear regression to explore the correlation between variables;
- We have applied various visualization methods, such as the research methodology framework in the literature review, introducing the impact of external factors on sailboat pricing, and providing an illustrative explanation of the random forest algorithm, among

others. Boring data may be able to reflect the law, but not as intuitive as so many images;

- Through sensitivity analysis, the effectiveness of the model can be observed under different parameters. Therefore, the model can be applied to the prediction of sailboat prices in other regions.

9.2 Weaknesses and Further Improvements

Our model has the following limitations and related improvements:

- There is a significant degree of subjectivity in our selection of feature variables, and while the abundance of selected variables to some extent reduces the impact of this subjectivity on the results, it is still possible for the prediction performance to be compromised due to the loss of certain important variables.
- We abstracted the random forest model as a continuous function and performed Taylor expansion on it. However, this assumption may not hold at points where the properties are not good enough.
- In the modeling of the last question, our data collection was not sufficiently comprehensive, and the results obtained may deviate significantly from the true values.

References

[1] Griliches. Hedonic Price Indexes for Automobiles: an Econometric Analysis of Quality Change. Price Statistics ofthe Federal Government, 1961(73): 137—196.

[2] Parasuraman. Servqual: A multiplue—item scale for measuring consumer perspecions of service quality[J]. Journal ofrealailing, 1988, (64): 12-40.

[3] Chen Junyi, Wang Hongyan. Improved comprehensive adjustment method based on fuzzy evaluation for determining the value of second-hand cars. Shanghai Automotive, 2009(07): 18-23.

[4] Breiman L. Bagging Preditors[J]. Machine Leaming, 1996, 24(2).

Report to broker

Dear Broker,

We are excited to present our in-depth analysis report on used sailboat pricing. Our team has diligently studied the available data and developed a robust mathematical model to predict sailboat listing prices according to your requirements. Here is a summary of our findings and recommendations.

1. Model election, Scalability, and Flexibility: We opted for a combination of Random Forest and Multiple Linear Regression models to provide the most accurate predictions. Our model demonstrates excellent scalability, allowing the inclusion of various relevant factors, and flexibility to adapt to different market conditions.

2. Regional Impact, Sailboat Type Differences, and Economic Factors: Our analysis reveals that the region has a significant impact on sailboat prices, and there are differences in regional influences between monohulls and catamarans. Additionally, our model considers various economic factors, such as inflation and currency fluctuations, providing a more comprehensive understanding of their influence on sailboat prices.

3. Sensitivity Analysis, Validation, and Application: We performed a sensitivity analysis on our model, showcasing its effectiveness under different parameters. We also cross-referenced our predictions with actual listing prices to ensure the accuracy of our model. As a result, our model can be confidently applied to predict sailboat listing prices in other regions as well.

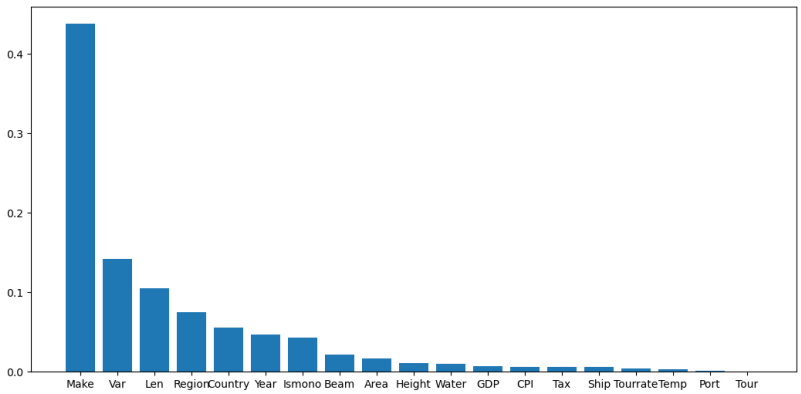
4. Market Trends and Decision-making: Our model can identify and analyze market trends, such as seasonal fluctuations and emerging preferences, enabling you to make well-informed decisions in a dynamic market environment.

Despite our thorough analysis, there are some limitations to our research, including subjectivity in selecting feature variables and potential gaps in data collection. We will continuously strive to enhance our model to ensure the most accurate predictions for you.

Please feel free to reach out to us if you have any questions or need further clarification on our analysis. We are more than happy to provide any additional support and insights you may require. We look forward to helping you navigate the sailboat market with confidence and success.

Warm regards,

Team #2333576



Appendices

Appendix 1

Introduce: python code

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn.pipeline import Pipeline
from sklearn.decomposition import PCA
from six import StringIO
from IPython.display import Image
from sklearn.tree import export_graphviz
import pydotplus
import os
from sklearn import metrics
import numpy as np
from torch.utils.data import Dataset, DataLoader, random_split
import torch

enc = LabelEncoder()
enc.fit(sailboat_df['Make'])
sailboat_df['Make'] = enc.transform(sailboat_df['Make'])
enc.fit(sailboat_df['Variant'])
sailboat_df['Variant'] = enc.transform(sailboat_df['Variant'])
enc.fit(sailboat_df['Geographic Region'])
sailboat_df['Geographic Region'] = enc.transform(sailboat_df['Geographic Region'])
enc.fit(sailboat_df['Country/Region/State '])
sailboat_df['Country/Region/State '] = enc.transform(sailboat_df['Country/Region/State '])
X = sailboat_df.iloc[:, :-1]
y = sailboat_df['Listing Price (USD)']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=0)
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
# regr = RandomForestRegressor()

regressor = RandomForestRegressor(n_estimators=100, random_state=0)
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)
# SSE = np.sum((y_pred-y_test)**2)
```

```
# mean = np.sum(y_test)/len(y_pred)
# SST = np.sum((y_test-mean)**2)
feat_labels = sailboat_df.columns
print(feat_labels)
score = regressor.score(X_test, y_test)
importances = regressor.feature_importances_
indices = np.argsort(importances)[::-1]
for f in range(X_train.shape[1]):
    print("%2d) %-*s %f" % \
            (f + 1, 30, feat_labels[indices[f]], importances[indices[f]]))

print('Score:', score)
print('MAE:', metrics.mean_absolute_error(y_test, y_pred))
print('MSE:', metrics.mean_squared_error(y_test, y_pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```