| Problem Chosen | 2023 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| Y | | 2330869 |

# Sail Far, Discover More

## Summary

The value of sailboats varies with their own aging and changes in market conditions. In order to better understand the sailboat market, we have established the following three models: Forecasting Model Based On Random Forest, Finite Mixture Model Using Nonparametric Methods, K-means Clustering Subset Selection

For problem1, through exploring the sailboat data, we found that there is a strong complexity between the feature variables and the target variable. Combining the characteristics of multi-feature variables in our dataset, we chose to establish a **random forest** algorithm model using **K-Fold Cross-Validation**. It can achieve high **prediction** accuracy and effectively avoid model underfitting and overfitting. We obtained the two features that have the greatest impact on the price of used sailboats through the model: sail area and displacement.

For problem 2: During the data exploration process, we found significant differences in average sailboat prices in different regions, as shown in Figure 6. In order to study the specific impact of regions on prices, we considered three regional characteristics: average cargo throughput, GDP, and coastline length, and established a **finite mixture model** using **nonparametric methods** based on the **EM algorithm**. Unlike traditional models, it can effectively correct the bias caused by regional heterogeneity in the data. Our conclusion for problem 2 is: GDP is the main cause of regional effects.

For problem 3: We used an improved **K-means** model to obtain a subset of sailboats with high information content. Then, when discussing the regional impact of Hong Kong on sailboat prices, we continued to use model 1. We found that Hong Kong has different effects on monohull sailboats and multi-hull sailboats.

For problem 4, since we not only considered multiple factors related to the second-hand sailboats themselves when modeling, but also discussed aspects such as geography, policy, and demand, our model has strong adaptability. It can be used for regions outside the scope of our study.

In addition, we conducted sensitivity analysis on the two main models used, and the results objectively showed that our models are not sensitive to parameter changes.

Regarding problem 5, we summarized the conclusions obtained through the models and prepared a report on the pricing of second-hand sailboats for brokers in Hong Kong. The report is easy to understand, with a logical structure, and includes visual aids that facilitate comprehension.

**Keywords:** Predict**;** Random Forest; K-Fold Cross-Validation; K-means; Finite Mixture Model; EM Algorithm; Nonparametric Methods

# Contents

# 1 Introduction

Sailing is a diverse sport that encompasses several competitive formats, which are regulated by different sailing federations and yacht clubs. These racing disciplines involve matches within a fleet of sailing crafts, between pairs, or among teams.

This helps to ensure a level playing field and to create fair and exciting races for all participants.

Sailboats can be considered a luxury items because they are expensive to purchase and maintain and usually require a lot off financial resources and time. Some of the larger sailboats can cost up to millions of dollars, and maintenance and upkeep can be very expensive. In addition, sailboats require specialized knowledge and skills to operate, so it may be necessary to hire a crew or pay for training. As a result, sailboat ownership is often seen as a luxury that only the more affluent can afford. However, there are smaller and less expensive sailboats available for the average consumer to purchase and enjoy.

Buying a used sailboat can be a more affordable option, as used boats are often much less expensive than brand-new boats. Buying a used sailboat can also be a more environmentally friendly option, as it reduces the resources and energy needed to build a new boat.

## 1.1 Problem Background

We focus on a more practical problem of how to set a pricing strategy in used sailboats selling, which is in a highly competitive marketplace.

For now, our reality is that the market for selling used sailboats is very complex and requires consideration of both the boat and the area where it is being sold. In terms of boats, we need to consider three factors: cost, history of the boat, and hull condition. In terms of the area of sale, there are three factors to consider: regional population, latitude, and economic development.

## 1.2 Restatement of the Problem

In this issue, we were given data on the sale of selected sailboats sold in Europe, the Caribbean and the United States. In order to better understand the sailing market and to price used sailing boats, in the following paper we will:

1. Gather information to expand useful predictors and sample size for the provided dataset.
2. Build a suitable regression model to predict the listing price of each sailboat in the table based on the expanded data set and assess the accuracy of the model.
3. Modeling to explore and explain the impact of region on the listing price of sailing boats, discussing the consistency of the regional effect in the context of practical and statistical significance.
4. Rely on some method to select an informative subset and collect a comparable selling data in the Hong Kong market from which the regional impact of Hong Kong can be obtained.
5. Discover more interesting and informative inferences or conclusions with the help of

the obtained model results.

6. An easy-to-understand report on pricing used sailing boats for Hong Kong sailing brokers.

## 1.3 Our Work

The title imposes multiple requirements on us. Our work mainly includes the following:

1 Based on our expanded sailboat data, we built a random forest model to achieve accurate prediction of used sailboat prices.

2 A finite mixture model based on the EM algorithm in the non-parametric case was developed to analyze the effect of regional effects on prices.

3 A subset of sailboats containing information is selected by an improved K-means algorithm, and model one is followed for specific analysis.

4 We tried to find other reliable information to explain our model results and found some interesting conclusions and inferences.

5 We integrated the conclusions obtained through the model and prepared a report for sailboat brokers in Hong Kong.

To show our workflow more visually, the flow chart is shown in Figure 3.

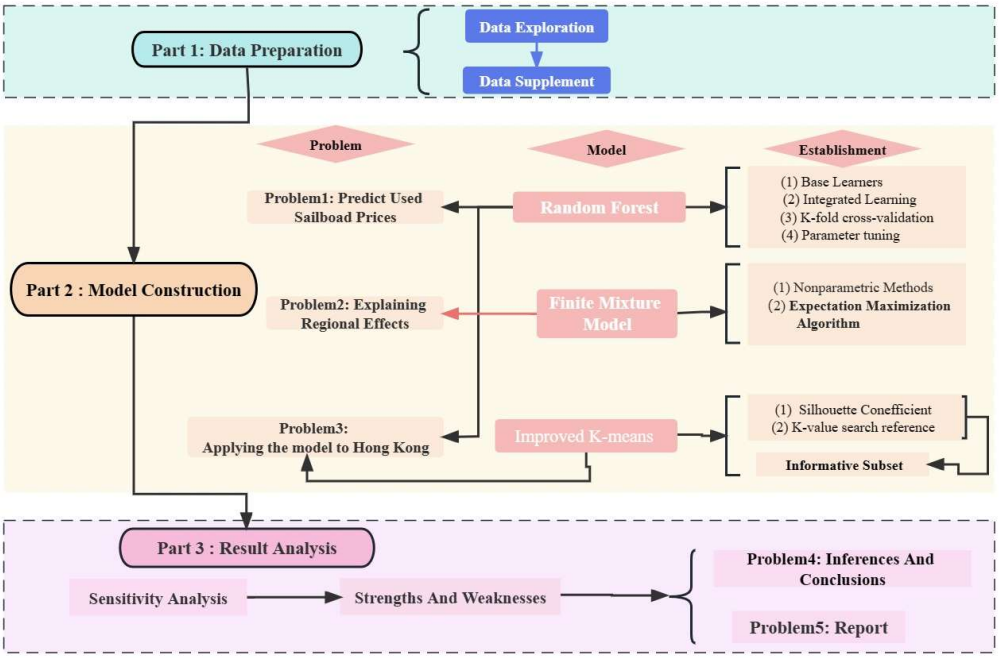


**Figure 1: The Workflow**

# 2 Assumptions and Justifications

1. **The distributor is attentive, reliable and non-deceptive**

   Careful means that the dealer will carefully inspect the health status of the boat when purchasing from others and set a reasonable selling price based on the condition. No deception means that the dealer will not hide the true condition of the boat when selling it to the consumer. Being reliable means that the dealer will provide after-sales

service to the consumer if a problem arises after the purchase of the boat. When a dealer satisfies these three characteristics, it means that the consumer's purchase of a boat is a completely commercial transaction, without involving the game of human nature. This assumption is essential to our model, which is based entirely on psychology.

2. **Differences between regions, considering only two factors: economy and average cargo throughput**

   Generally speaking, the low and middle latitudes are better locations for sailing, due to their moderate temperatures. Besides being often extreme, the main characteristic of the weather in high latitudes is its extreme volubility. As a luxury item, sailing is extremely expensive to purchase and maintain, so economic factors must be considered when selling sailboats. The Population is often closely related to the economy of a region. Generally speaking, areas with a large population have a relatively prosperous overall economy. These three factors are certainly three essential factors when considering regional differences.

# 3 Notations

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

**Table 1: Notations used in this paper**

| Symbol | Description |
|---|---|
| $g_\theta(x_i)$ | density function of stochastic semiparametric EM algorithm |
| $K(.)$ | kernel density function |
| $h_{jl}$ | bandwidth for component and block density estimate |
| $f_{jb_k}(.)$ | density function |
| $\Phi_{ij}$ | an n × m matrix |
| $H(D)$ | information entropy |
| $H(D\|A)$ | conditional entropy |
| $g(D,A)$ | information gain |
| $S(i)$ | Silhouette Coefficient |

# 4 Data Description

We supplemented the parameter columns, Beam(ft), Draft(ft), Displacement(lbs.), Sail Area(sq ft), Average cargo throughput(tons), GDP(USD billion), GDP per capita(USD), Engine Hours, Coastline(km), through multiple data sources and supplemented and modified missing and problematic data in the dataset through website information.

**Table 1: Data, Database Website and Data Type**

| Database Names | Database Website | Data Type |
|---|---|---|

| | | |
|---|---|---|
| Sailboat Data | https://sailboatdata.com/ | Sailboat |
| Yacht World | https://www.yachtworld.com/ | Sailboat |
| Saffron Marine | https://saffron-marina.com/ | Sailboat |
| United States Census Bureau | https://data.census.gov/ | Region |
| International Federation of Forwarders Associations | https://fiata.org/ | Region |
| The World Bank | https://data.worldbank.org/ | Region |
| World Economic Forum | https://cn.weforum.org/ | Region |
| Kaggle | https://www.kaggle.com/ | Region |
| Census and Statistic Department in HK | https://www.censtatd.gov.hk/ | Region |
| Yacht World | https://www.yachtworld.com/ | Region |

**Table 2: Data Description**

| | Year | Listing Price | Length | Beam | Draft | Displace-ment | Sail Area | Average Cargo Throughput | GDP | GDP per capita | Engine Hours | Coastline |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Mean** | 2010.38 | 226305.17 | 45.27 | 13.99 | 6.78 | 26476.9 | 1053.76 | 3.5E+07 | 1061.8 | 33338.61 | 9.62 | 5175.42 |
| **STD** | 4.05 | 144641.8 | 4.77 | 1.08 | 0.87 | 7956.23 | 267.99 | 5.2E+07 | 1078.41 | 20499.77 | 4.05 | 4360.76 |
| **Min** | 2005 | 45000 | 36 | 9.5 | 3.94 | 6393 | 516 | 5.5E+04 | 0.8 | 4871 | 1 | 0 |
| **25%** | 2007 | 139000 | 40.25 | 13.08 | 6.33 | 19621 | 861 | 1.9E+06 | 57.8 | 13933 | 6 | 121 |
| **50%** | 2009 | 190303.5 | 45 | 13.94 | 6.75 | 25353 | 1032 | 2.0E+07 | 650 | 30459 | 11 | 4964 |
| **75%** | 2014 | 267054.25 | 49 | 14.73 | 7.22 | 31085 | 1191 | 5.2E+07 | 2005 | 44494 | 13 | 7600 |
| **Max** | 2019 | 1885229 | 56 | 16.73 | 11.58 | 63900 | 2314 | 3.0E+08 | 3861 | 93944 | 15 | 25148 |

Finding and handling missing values was a time-consuming process. We used a matrix plot in Figure 1 to describe the general shape of data completeness. From the figure, it can be seen that cargo throughput, GDP, and GDP per capita are severely missing, so we removed rows with missing values. During the analysis, we found outliers in the listing price of some sailboats. Therefore, we classified the ships with the same attributes and took the average, reducing the impact of outliers. After these two steps, 2363 sets of raw data were reduced to 1839 sets of processed data, providing a basis for our subsequent analysis. Our data description is shown in Table 2.

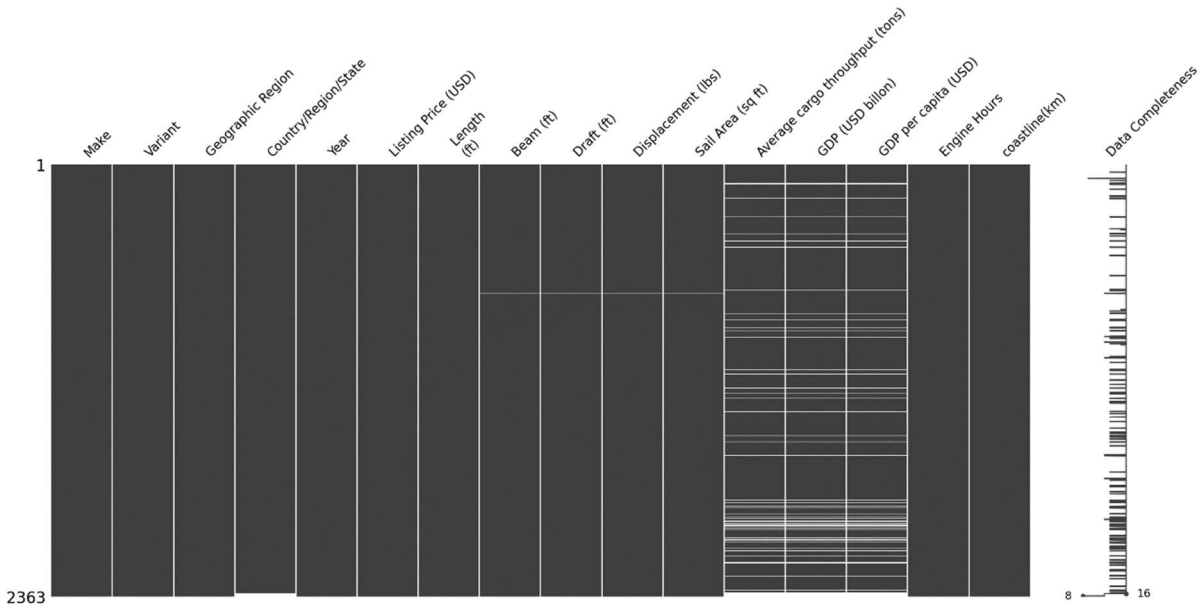


**Figure 2: Missing Values Processing**

After preprocessing the data, a histogram shown in Figure 2 was plotted with the listing price, and the fitted curve approximates a normal distribution, indicating the correctness of the data preprocessing process. Meanwhile, as shown in Figure 3, prices have a regional effect. In Figure 4, it can be seen that the frequency of sailboat sales varies with length, selling price, and year.

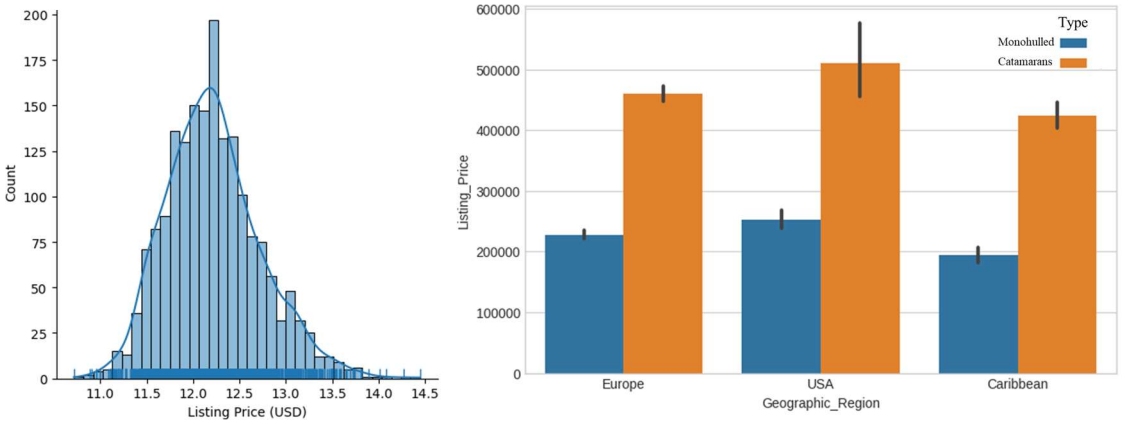


**Figure 3: Histogram and Fitted Curve(left) and Regional Differences(right)**

Overall, many factors have an impact on the sale price of sailboats. Therefore, we use a heatmap in Figure 5 to visually show the correlation between different factors and prices. From this, it can be concluded that Displacement, Sail Area, Beam, Draft, Length, Make, GDP, and price have a relatively strong correlation. Therefore, we further analyze these factors.

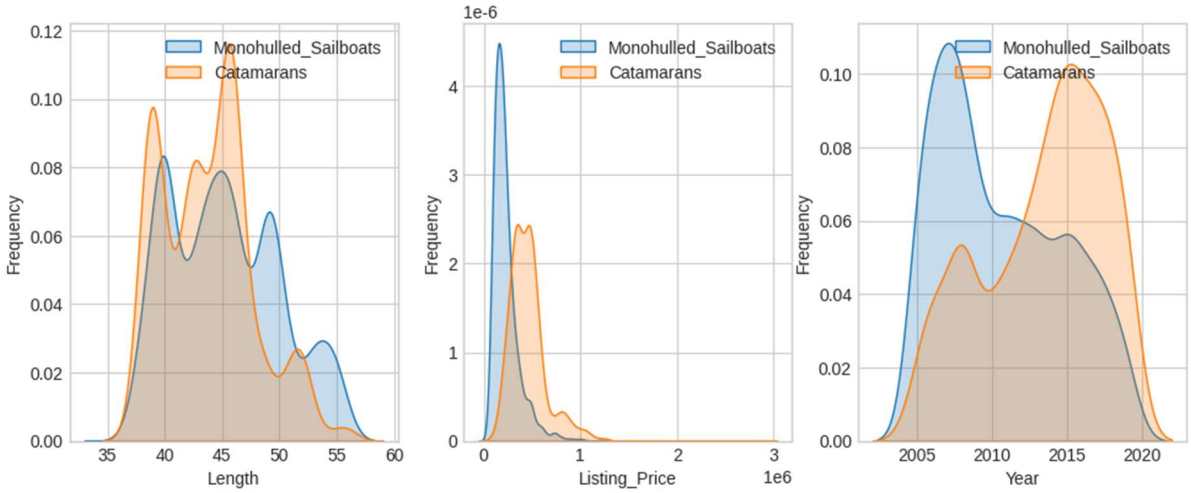


**Figure 4: Sales Status**

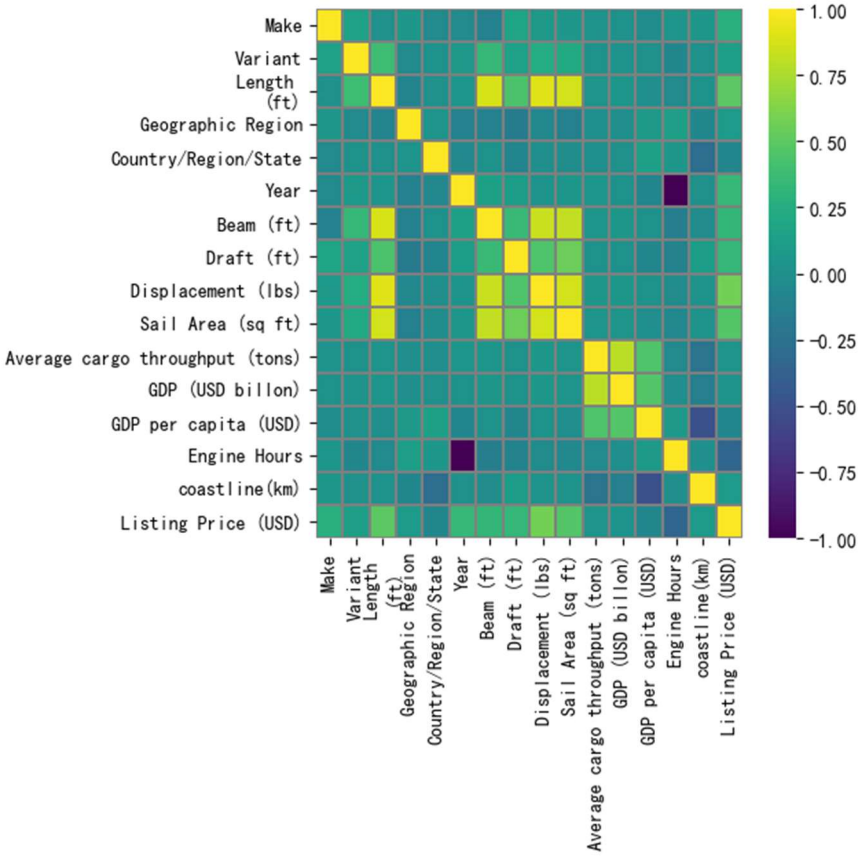


**Figure 5: Correlativity**

Before studying regional effects, we visualized the average sailboat prices for different regions in the United States separately. As can be seen from Figure 6, there are significant differences in sailboat prices in different regions of the East Coast of the United States. Based on this, we preliminarily believe that sailboat price data has regional heterogeneity, which needs to be taken into account when establishing the model.

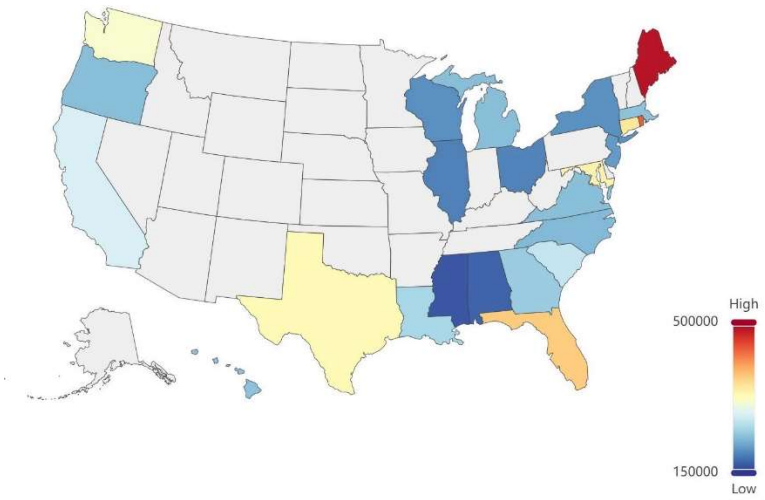


**Figure 6: Mean Price of Sailboats (USA)**

# 5 Forecasting Model Based On Random Forest

In accordance with the requirements of Problem 1, all sailboat characteristic variables need to be selected and a quantitative prediction model of the main characteristics of sailboats on the listing price needs to be constructed accordingly. The model is then used to predict the listing price of sailing boats for a selected sample of the test set and the accuracy of the model is evaluated by comparing the true values of the sample using an appropriate metric. The challenge is to choose the forecasting model algorithm and to improve the optimal parameters of the model to achieve the best accuracy.

The first is the selection of the forecasting model. In the data description section, we found that each characteristic variable has diversity and complexity, the phenomenon of non-linearity between characteristic variables, and the relationship between the characteristic variables and the target variable (the listing price of the sailboat) is more complex, therefore, the use of a simple linear model will certainly not converge and cannot achieve high forecasting accuracy, we give priority to non-linear forecasting algorithms modeling. At the same time, considering that the number of feature variables in this question is large for statistical models and small for deep learning, which is prone to under- or over-fitting of the model accuracy, we believe that it is more appropriate to choose the algorithm modeling traditional machine learning.

In traditional machine learning algorithms for regression problems, single regression models still struggle to handle complex variable relationships, and in turn, it is easy to think of ensemble learning algorithms - Random Forests[1] for building regression models. Random Forests is a supervised machine learning algorithm based on ensemble learning, which can incorporate different types of algorithms or the same algorithm multiple times, and could not be more appropriate for this problem given the characteristics of the data set.

## 5.1 The Establishment of the Model

For this problem, the base learner of the random forest does not need to be too complex,

otherwise, it is prone to overfitting, and a decision tree model can be chosen, which also facilitates the interpretation of the model results in the light of their practical implications.

The decision tree model is presented as a tree structure and the process is similar to what we do in real life, where we ask a series of questions about the data before arriving at a final decision. The subtlety of the decision tree is how to get the answer with the least number of questions, i.e. requiring each decision to make the greatest contribution to the solution of the final decision, and the decision tree model does this by selecting attributes through an information gain criterion as follows:

1. The information entropy H(D) of the characteristic variable Ci of the data set D, i.e. the uncertainty of the information of the characteristic variable Ci. This can be expressed by equation 1:

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \tag{1}$$

2. Calculate the conditional entropy H(D|A) of the characteristic variable A on the data set D, i.e. the information uncertainty conditional on the determination of the characteristic variable A. This can be expressed by equation 2:

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|} \tag{2}$$

3. Calculate the information gain g(D, A), i.e. the reduction in information uncertainty after learning the characteristic variable A. This can be expressed by equation 3:

$$g(D, A) = H(D) - H(D|A) \tag{3}$$

4. Select the feature variable with the greatest information gain as the child node, then recursively call the above method to obtain all child nodes

It can be seen that decision trees are in fact greedy algorithms, i.e. they prioritize the feature variables with greater information gain.

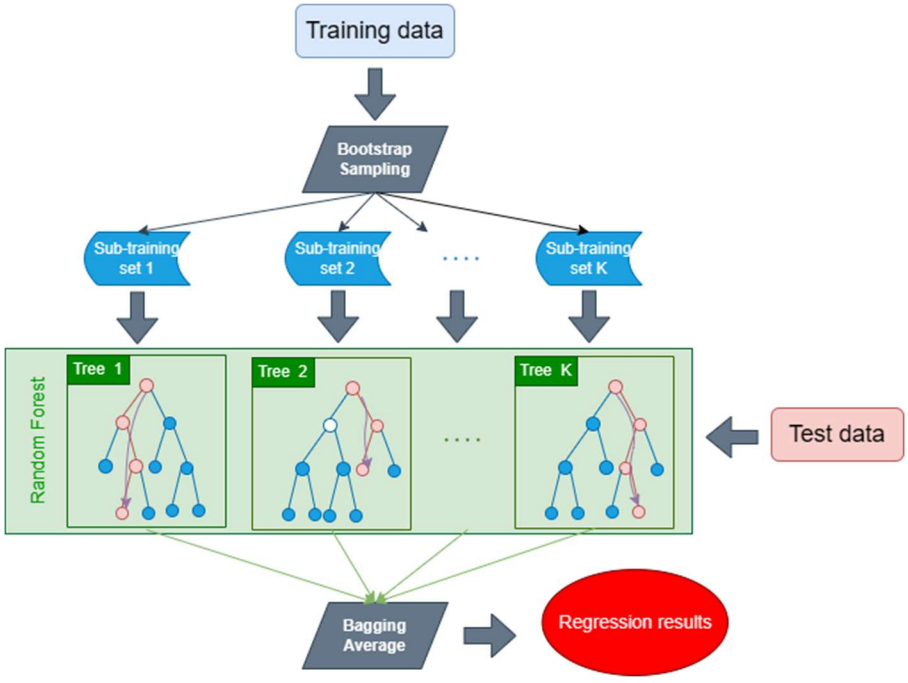


**Figure 7: Schematic Diagram of the Random Forest Algorithm**

Random forests integrate multiple decision trees to get results, similar to a real-life scenario where multiple people vote on decisions. The basic principle is to first select a subset of N samples from a finite data set by repeatedly sampling the data using Bootstrapping, and then construct a decision tree based on these N samples. The regression is based on the small variance of the results of the decision tree, so that h1(x) represents the predicted value of one of the decision trees, and then the predicted value of the random forest regression is obtained by averaging the predicted values of the decision trees. The algorithm works as shown in Figure 7.

## 5.2 The Solution of the Model

### 5.2.1 Analysis of the Train/Test Sets

**Table 3: Training Set Test Set Data Comparison**

| Characteristic variables | Mean | | Std | |
|---|---|---|---|---|
| | train | test | train | test |
| Length(ft) | 45.3 | 45.25 | 4.78 | 4.85 |
| Geographic Regin | 1.10 | 1.14 | 0.50 | 0.53 |
| Beam(ft) | 13.97 | 14.00 | 1.08 | 1.11 |
| Draft(ft) | 6.8 | 6.73 | 0.91 | 0.88 |
| Displacement(lbs.) | 26519.13 | 26713.65 | 8005.34 | 8321.34 |
| Sail Area(sq ft) | 1054.58 | 1055.24 | 274.07 | 261.60 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| GDP(USD billon) | 1073.29 | 1136.12 | 1075.56 | 1143.99 |

The question asked to determine the estimated accuracy of each sailboat variety, so the sample selection required the inclusion of each sailboat variety. Based on the relationship between the total number of samples and the number of sailboat categories, the test set size was approximately 80% of the total sample size, which is in line with the most important 80/20 law of only approximately 20%.

Considering that we will use the established random forest prediction model to make predictions on the test set, we first conducted statistics on the mean and variance of the feature data of the training and test sets, and the results are shown in Table 3. It can be found that the data distributions of the test and training sets are basically the same, indicating that the test set can represent the features of each sailboat well.

### 5.2.2 K-Fold Cross-Validation

Due to the large number of feature variables and large sample size in the dataset, it will be found during model tuning that better prediction results can only be obtained when the values of parameters (such as the number of base learners, maximum number of features, maximum number of leaf nodes, etc.) are large, but at the same time, large models are also prone to overfitting. In order to improve the robustness and accuracy of the model, the validation set is used to solve this problem:

However, simply dividing the dataset into three parts: training set, validation set and test set, actually only makes the model more convergent to the validation set (a smaller subset relative to the original training set), which in turn leads to overfitting of the model and does not achieve the idea of crossover. This in turn leads to over-fitting of the model and does not achieve cross-validation. So we add the idea of randomness to the division of the set.

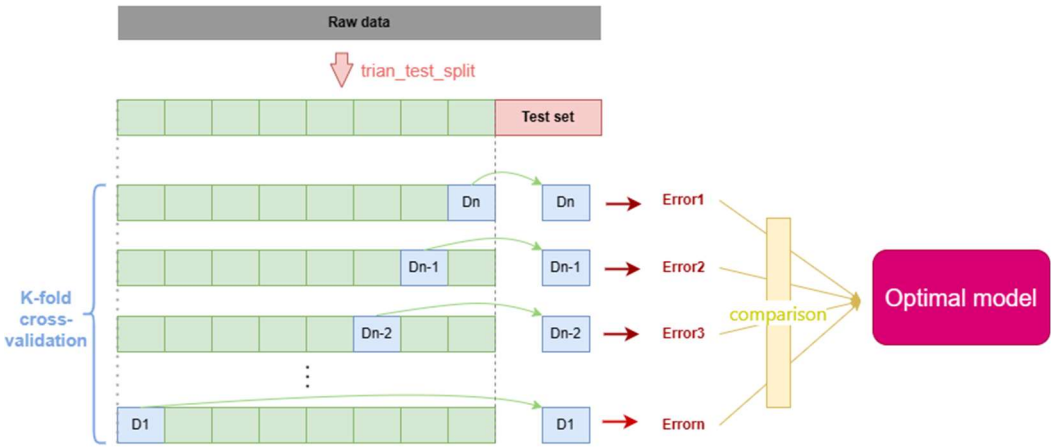


**Figure 8: Data Division Process**

K-fold cross-validation is shown in Figure 8. Firstly, all the data sets are divided into N copies, and then one of them is taken as the validation set without repeating each time, while the other n-1 data are used as the training set for training the model and tuning the reference, so we need to train N models, and different training and validation sets are used for each training, and finally the evaluation metrics of the N models are calculated and the best model is selected accordingly.

### 5.2.3 Parameter Tuning and Test Results

We use a combination of randomized search and individual search to automatically superparameter tune the problem.

At first, we are not sure of the approximate location of the parameters, so we can first take random values throughout the space to get the approximate location of the optimal hyperparameters, and then take a smaller range around that number of hyperparameters to search the grid one by one. Adjust the parameters to achieve the best model. The model results show that the weights of each characteristic variable which are displayed in Figure 9.

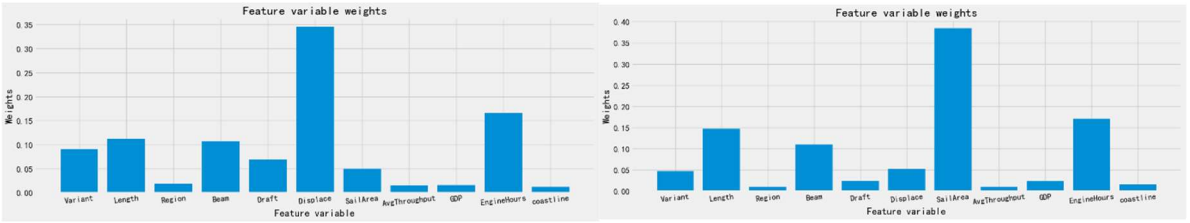


**Figure 9: Feature Weights of Monohulled(left) and Catamarans(right)**

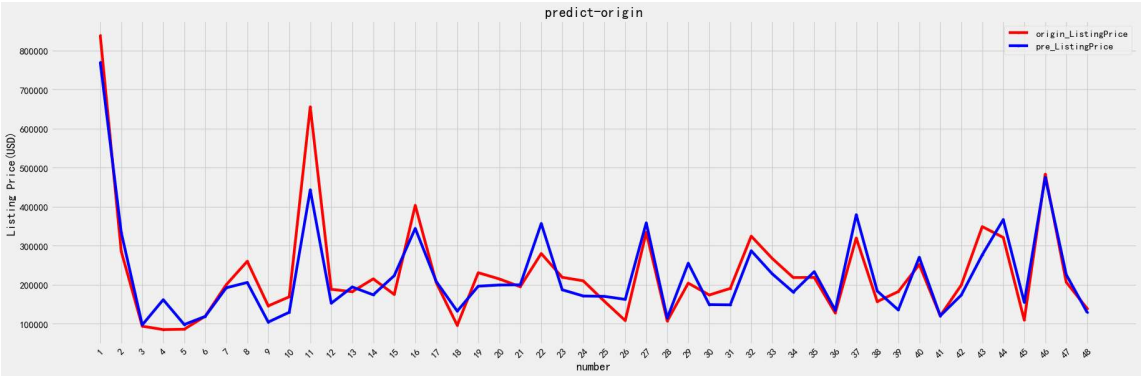


**Figure 10: Chart of Real Results versus Predicted Results**

The previously divided training set was fed into the above trained model, and the prediction results were obtained with an accuracy of over 80%. Figure 10 shows the comparison between the target variables and the predicted target variables in the original test set.

# 6 Finite Mixture Model Using Nonparametric Methods

Since the data we used came from different countries and regions, even the same product would have different transaction prices due to regional and economic differences, making our data discrete. We originally planned to use a Poisson model for data analysis, but after researching, we learned that traditional models often fail to consider the heterogeneity of data in their assumptions, resulting in model bias. Therefore, after comprehensive consideration, we ultimately established finite mixture model using nonparametric methods for solving the problem.

## 6.1 The Establishment of the Model

This section focuses on nonparametric multivariate finite mixture models, which are an extension of the stochastic semiparametric EM algorithm[2] presented in equation 4, allowing for different distributions for each component and coordinate of $Xi$. Notably, if the density function $f_{jk}(\cdot)$ does not depend on $k$, the $Xi$ are not only conditionally independent but also identically distributed. We also assume that the coordinates of Xi are conditionally independent, with blocks of coordinates being identically distributed. Specifically, we can denote the block to which the kth coordinate belongs as $b_k$, where $1 \leq b_k \leq$ B and B is the total number of such blocks. With this notation, the equation can be expressed as follows:

$$g_\theta(x_i) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jb_k}(x_{ik}) \tag{4}$$

Throughout this section, we will use the indices $i$, $j$, $k$, and $l$ to refer to a generic individual, component (subpopulation), coordinate (repeated measurement), and block, respectively. Thus, we always have $1 \leq i \leq$ n, $1 \leq j \leq$ m, $1 \leq k \leq$ r, and $1 \leq l \leq$ B. To estimate model (7), we employ the EM algorithm, which involves an E-step and an M-step. This yields a weighted nonparametric kernel density estimate expressed as follows:

$$f_{jl}^{t+1} = \frac{1}{nh_{jl}C_l\lambda_j^{t+1}} \sum_{k=1}^{r} \sum_{i=1}^{n} p_{ij}^{(t)} I\{b_k = l\} K\left(\frac{u - x_{ik}}{h_{jl}}\right) \tag{5}$$

where $K(\cdot)$ is a kernel density function, $h_{jl}$ is the bandwidth for the $j$ th component and $l$ th block density estimate, and $C_l$ is the number of coordinates in the *l*th block. For any real $u$, define for each component $j \in \{1, \cdots, m\}$ and each block $l \in \{1, \cdots, B\}$

We have certain instances of equation 4[3] where certain $f_{jb_k}(\cdot)$ densities are believed to be identical except for a change in location and scale. Such situations are referred to as semi-parametric since the estimation of each $f_{jb_k}(\cdot)$ entails determining an unknown density along with numerous location and scale parameters. To illustrate, equation (6) of establishes this idea.

$$f_{jl}^{t+1} = \frac{1}{nh_{jl}C_l\lambda_j^{t+1}} \sum_{k=1}^{r} \sum_{i=1}^{n} p_{ij}^{(t)} I\{b_k = l\} K\left(\frac{u - x_{ik}}{h_{jl}}\right) \tag{6}$$

where $l = b_k$ for a generic *k*.

To fit equation 6, the mixtools package provides an algorithm in the spEM function. The npEM function also requires updating the values of $f_{jb_k}(x_{ik})$ for all $i$, $j$, and $k$. The spEM algorithm updates an n × m matrix called Φ, which stores the current values of $f_{jb_k}(x_{ik})$.

$$\Phi_{ij} \equiv \phi_j(Xi) = \prod_{k=1}^{r} f_{jb_k(x_{ik})} \tag{7}$$

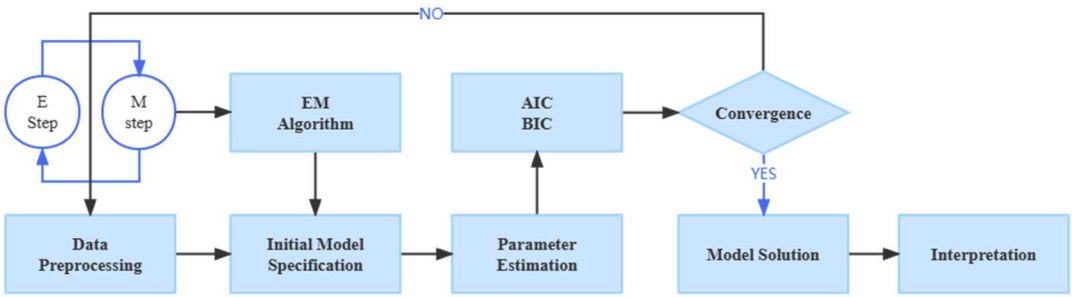


**Figure 11: Process of Finite Mixture Model**

## 6.2 Explaining Regional Effects

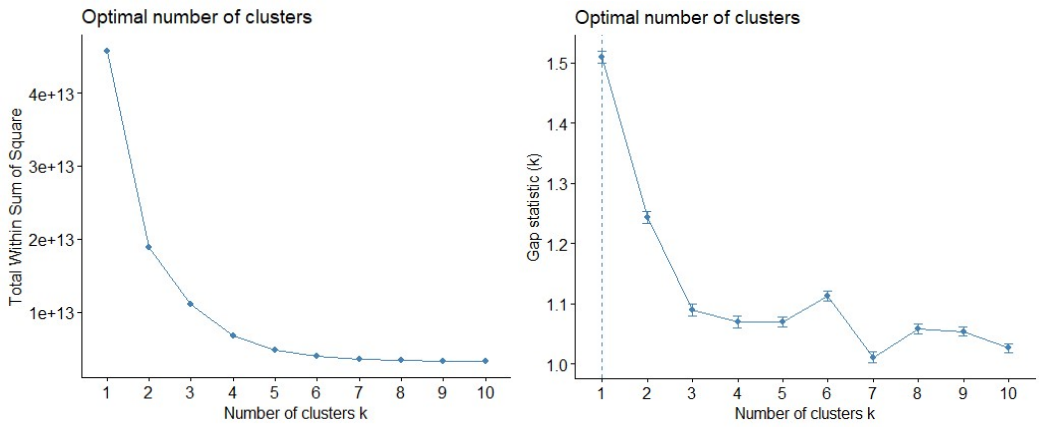


**Figure 12: Finding the Optimal Number of Clusters k**

In the data preprocessing section, we obtained several variables that have a strong correlation with price. Since this model is designed to explain the effect of regional factors on price, we introduce three variables: Listing Price, GDP, and coastlines. We fit them using a finite

mixture model and determine the statistical significance of regional effect.

First, we perform K-Means clustering and plot a graph to select the value of k in Figure 12 (left). We search for the point on the curve that corresponds to the value of the total sum of squares for a certain k, where the relationship between the number of clusters and the total sum of squares starts to curve or level off. When the graph shows an elbow shape, it usually indicates the ideal number of clusters. It is evident from the graph that the elbow shape appears at k = 4. Therefore, we choose k = 4. At the same time, we estimate the value of k again using a different method, by comparing the relationship between the number of clusters and the statistical disparity. From the graph, it can be seen that the gap statistic is largest when k = 1, which contradicts the results from the previous graph. By using statistical methods, we ultimately determine k = 4 as the number of clusters, and the model constructed at this value can better explain our results.

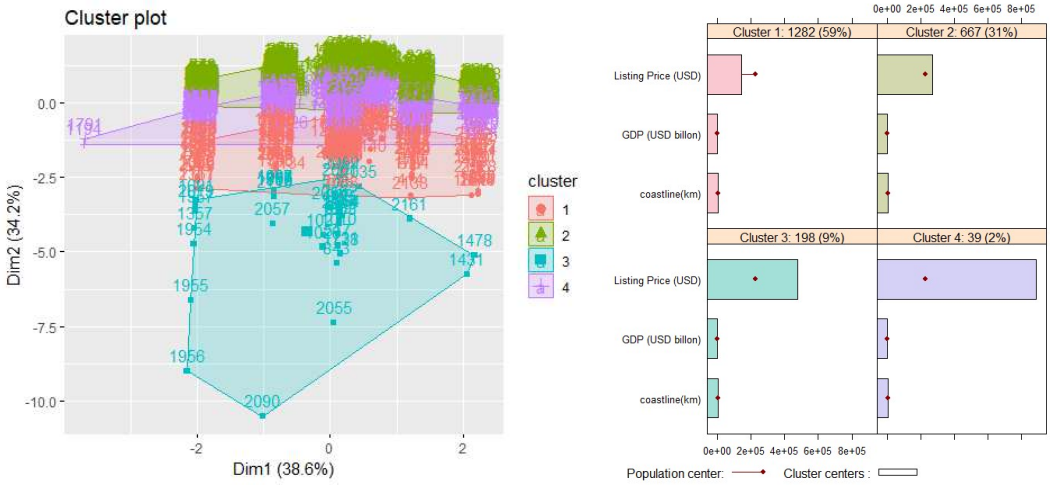


**Figure 13: Clustering Result**

**Table 4: The Mean of Variables in Each Cluster**

| Cluster | Listing Price | GDP | Coastline |
|---|---|---|---|
| 1 | 135684.6 | 987.6517 | 5111.487 |
| 2 | 880757.9 | 1063.6125 | 6302.950 |
| 3 | 246913.3 | 1139.8484 | 5046.368 |
| 4 | 449605.0 | 1135.6186 | 5681.206 |

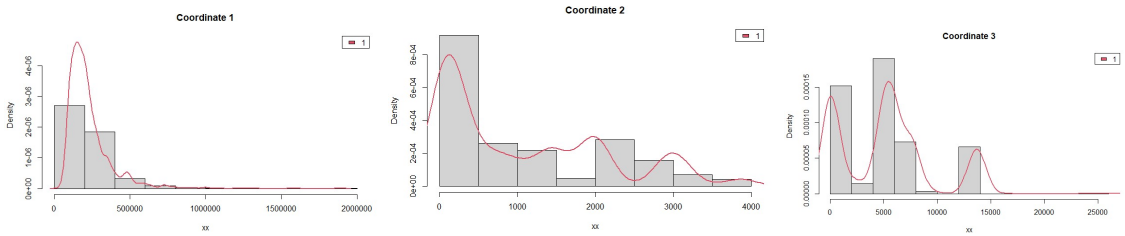


**Figure 14: Estimated Component Densities**

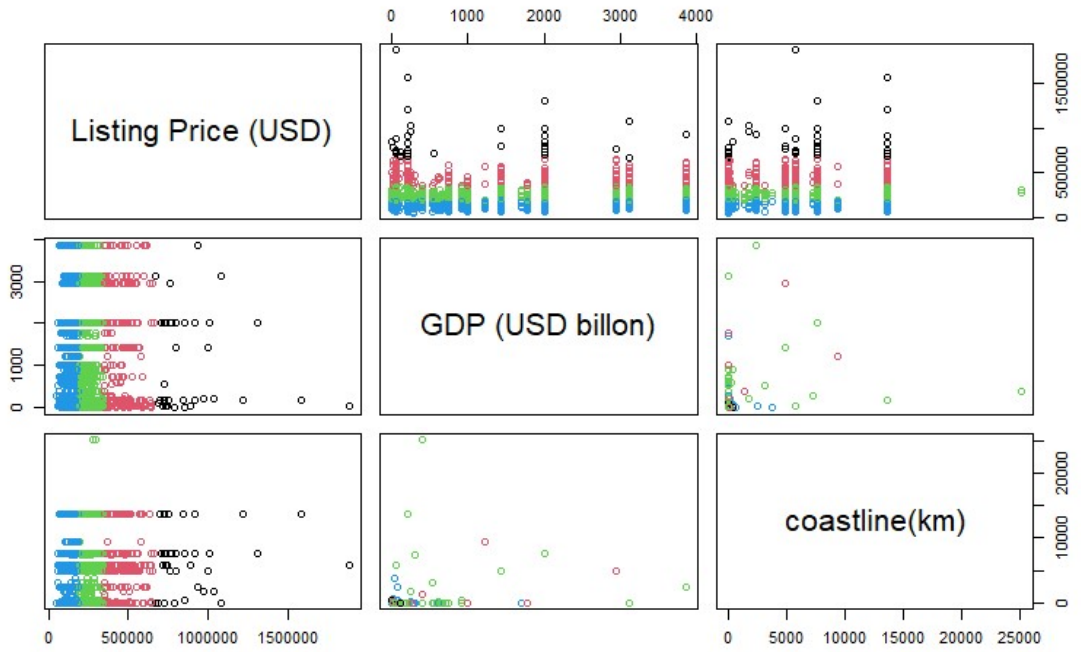


**Figure 15: Pair Plots**

# 7 K-means Clustering Subset Selection

## 7.1 The Role of Models in Hong Kong

For the random forest model established in Problem 1, Figure 16 shows the decisions after pruning of a representative decision tree for the single hull ship dataset, and we analyses the model details further.

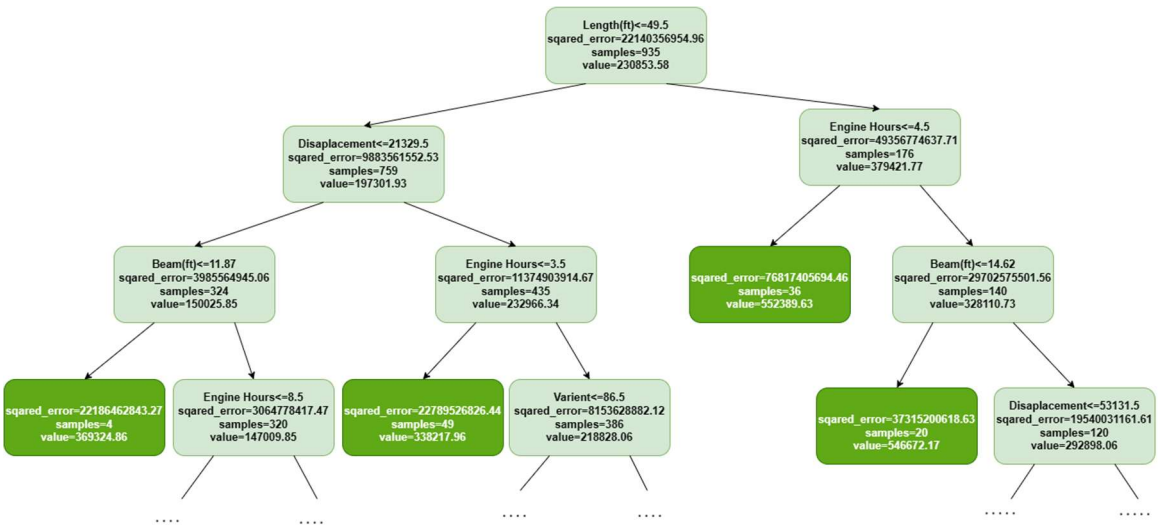


**Figure 16: Partial Results Graph of a Representative Decision Tree for a Single Hull**

As can be seen from the results in the Figure 15, for the listed prices of monohull vessels, the Displacement, Length, and Engine Hours of the vessel are more deterministic; correspondingly, in the random forest generated from the catamaran dataset, the Sail Area, Length, and

Engine Hours of the vessel are more deterministic. In conclusion, for both monohull and catamaran vessels, the factors determining their prices were less correlated with the area-related characteristic variables. Also, in the analysis of Question 2, we have a similar conclusion that Displacement, Sail Area play a decisive role in the listing price of monohull and catamaran respectively. Our modelling of the given geographical areas is therefore of some relevance as a guide to the Hong Kong market regarding the pricing of sailing vessels.

In order to make the sailboat subset as representative of the whole dataset as possible, we tried a number of methods (Improved K-means Algorithm, Ordering Points to Identify Clustering Structure (OPTI CS), Density Peak Cluster (DP C), etc.). Algorithm (OPTI CS), Density Peak Cluster (DP C), etc.) and finally chose the Improved K-means Algorithm model, which is specific in that it self-searches for the best clustering K values and uses a fast aggregation method based on optimizing the initial clustering centers and silhouette coefficients.

We then used the results of the clustering algorithm as a comparable subset of the Hong Kong listing price data for the same vessels in Hong Kong from the websites in the table in the dataset above.

## 7.2 The Establishment of the Model

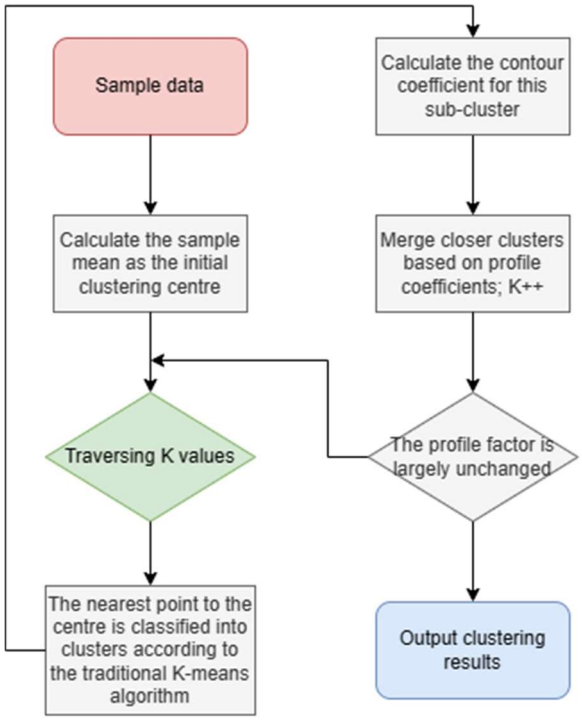


**Figure 17: K-means Clustering Algorithm based on Initial Clustering Centers and silhouette coefficients**

The K-means clustering algorithm belongs to the category of unsupervised learning and is an iterative repositioning algorithm. The k in the algorithm represents the clustering of k clusters, and the means represents the mean value of the data values in each cluster as the center of the cluster, i.e. the cluster is described by the center of mass of each of the classes. The basic idea is to find a kind of central division of the k clusters by iterating so that the loss function corresponding to the clustering result is minimized, where the loss function can be defined as the sum of the distances between any point in the cluster and the centroid.

The core step in the algorithm is the search for the number of cluster centers K. The initial cluster centers of the traditional K-means clustering algorithm are chosen randomly, and the number of clusters K is also determined artificially, once the initial cluster centers and K are determined, the whole iterative process of the clustering algorithm is calculated by the algorithm itself, but there is also a problem with traditional K-means clustering, that is, each time the clustering results in a different outcome due to the different values of the initial cluster centers and K. In order to solve this problem, we use the silhouette coefficient method and the K-value optimal search method. In order to solve this problem, the silhouette coefficient method and the K-optimal referencing method are used, and we believe that the mean point of all data should be closest to the cluster center, so the mean point is used to initialize the cluster center. The algorithm workflow can also be represented by Figure 17.

### 7.2.1 Silhouette Coefficient

The silhouette coefficient is a way to evaluate how good the clustering is, and it includes the degree of cohesion and the degree of separation. Suppose a tuple Xi in the data set belongs to cluster Ci, then the cohesion ai is calculated as the average distance between Xi and other points in the same cluster, representing the degree of dissimilarity between Xi and other points in the same cluster; the separation bi is calculated as the average distance between Xi and another cluster C nearest to Xi , and all points in cluster C, representing the degree of dissimilarity between Xi and the most adjacent cluster C. The final profile coefficient is expressed as equation 8:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{8}$$

The silhouette coefficients range from [-1,1], with the closer to 1 representing a relatively good degree of both cohesion and separation.

### 7.2.2 K-value Search Reference

The range of k values is [2,n] , n represents the number of data points. Since the value of n is too large from this question, we can use a large step to determine a small range in a large range, and then use a smaller step to determine a smaller range in a small range to find the best k value, the evaluation function of k values is the silhouette coefficient, Figure 18 shows the process of our search for parameters.

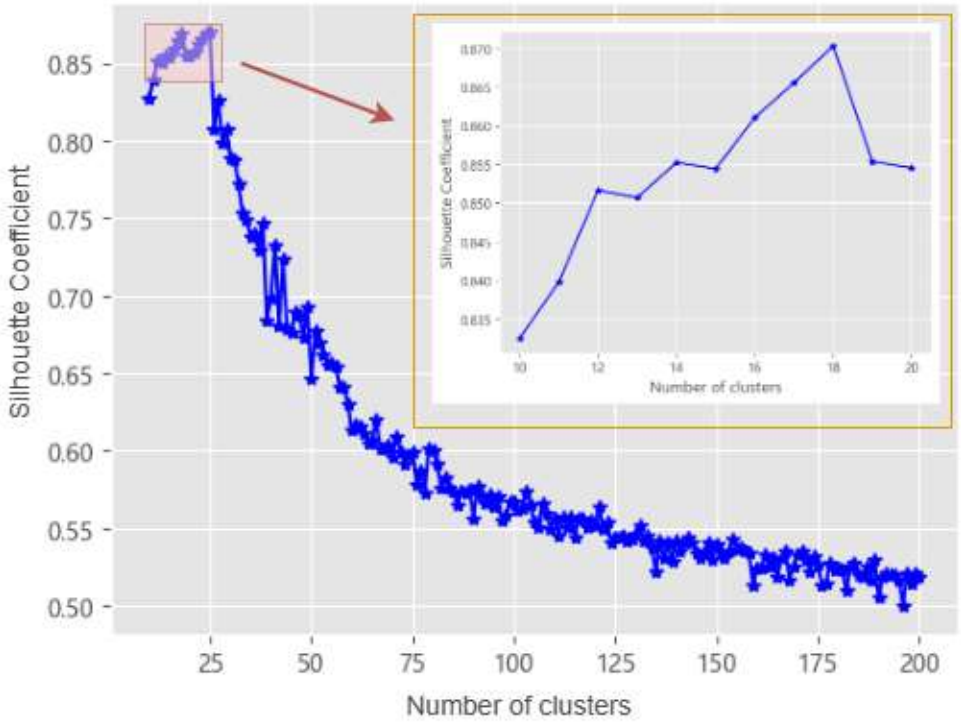


**Figure 18: Number of Cluster**

We then obtained a super informative 18 figures and accordingly found the listing price of the corresponding vessel in Hong Kong on the website.

### 7.2.3 Regional impact

Preliminary forecasts were made for Hong Kong using the two models developed and a comparison of the listing prices of the same sailing boats in Hong Kong and other regions was carried out, resulting in Table 5:

**Table 5: List Price Comparison Table**

| Listing Price in Hong Kong | | Catamarans | |
|---|---|---|---|
| Price (Hong Kong) | Price (Elsewhere) | Price (Hong Kong) | Price (Elsewhere) |
| 240000 | 95961 | 685000 | 633646 |
| 543375 | 479824 | 560000 | 539753 |
| 1770000 | 1311872 | 538500 | 534784 |
| ⋮ | ⋮ | ⋮ | ⋮ |

It can be seen that for catamarans, listing prices are relatively stable across regions, but for monohulls prices are relatively volatile and the Hong Kong market can adjust the market's inventory and prices for monohulls and catamarans accordingly.

Why is there such a regional effect for monohulls only in the Hong Kong region? We have found, after searching for information, that this price stability difference is in fact related to issues concerning oil spills from vessels and marine environmental protection in recent years. Some regions have adopted strict regulations on the use of monohulls because of the higher risk of oil spills and environmental pollution in the event of an accident. This has led to an increase in demand for catamarans globally. The strict regulations on monohulls have created some risk in the monohull trade and this has contributed to the unstable prices on the market

for monohulls.

In addition, we have found from many of the "Reports on the Study of Marine Issues in Hong Kong" that the quality of the marine environmental protection team in Hong Kong is relatively strong, and that scientific research and consultation on marine environmental protection is more solid. The Bunker Oil Pollution[4], provides that if pollution damage is caused in Hong Kong as a result of an accident, the owner of the ship concerned shall be liable for that damage. This also enhances the regional effect of monohull vessels in Hong Kong to a certain extent.

# 8 Sensitivity Analysis

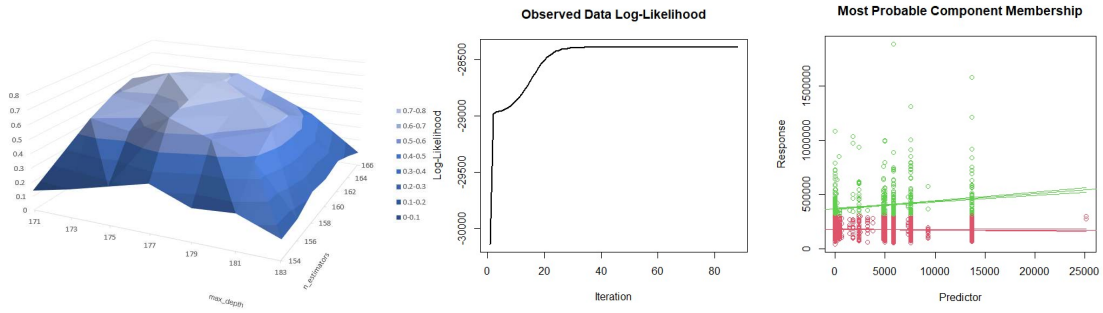


**Figure 19: Figures for Sensitivity Analysis**

The EM algorithm is an iterative algorithm used to estimate the parameters of a finite mixture model. In each iteration, the algorithm updates the estimates of the component parameters and the probabilities of belonging to each component given the data. When the number of iteration is 87, convergence is reached. The curve starts at a lower value and steadily increases, thus, it is a good fit for the data. In Figure 18, for the observed values corresponding to the green line, since some of the observed values are in the region of higher probability and many others are in the region of lower probability, it indicates that the mixture model can effectively assign different data points to different groups, demonstrating a good model fit. From the graph, it can be observed that when the number of iterations is around 25, the curve changes from steep to nearly horizontal, indicating that the fitted models after more than 25 iterations are extremely close to convergence. Therefore, it can be concluded that the finite mixture model we built is very insensitive and has strong universality, representing outstanding model performance.

Figure 19 shows the trend of the random forest model predicting the listing price accuracy with n_estimators and max_depth. From the figure we can see that: listing price accuracy can be achieved better when n_estimators=160 and max_depth=175. In addition, the effect of the model's accuracy in predicting listing prices is minimal when n_estimators and max_depth spread in all directions from 160 and 175 respectively (this is evident from the flatter top of the hill in the middle panel of the figure). The model is therefore able to be applied to different types of data sets and has a strong generalization capability.

# 9 Model Evaluation and Further Discussion

## 9.1 Strengths

1. Random forest algorithms are insensitive to noise in the dataset, which facilitates a robust model, and the use of a set of uncorrelated decision trees can effectively prevent overfitting of the model.
2. We used a finite mixture model based on the EM algorithm, which takes into account the regional heterogeneity of the data, resulting in our model having better predictive performance.
3. The K-means algorithm is relatively scalable and efficient, and we have improved the model to obtain a globally optimal model with excellent clustering results.

## 9.2 Weaknesses

1. Random forest models are complex and they require more time to train than other similar algorithms.
2. Due to limited time and dataset, we lacked exploration of more dimensions related to the regions. If we could obtain more feature data related to the regions, our model would be more reliable and have stronger interpretability.

## 9.3 Further Discussion

With the random forest model we can construct decision trees using a dataset selection method that is more specific to the dataset. The base learner can be joined with other learners to make the model more robust

# 10 Inferences

Our model shows that displacement and sail area have the most significant effect on the price of used sailboats. In single variant sailboats, displacement has the highest weight of influence on their higher or lower price. While in double variant sailboats, the highest weight of influence is on sail area. By reviewing a large amount of literature, we found that the two sailboat characteristics mentioned above correspond exactly to the two important parameters that affect the speed of a sailboat.

The first is the Displacement-Length Ratio (D/L) of the sailboat, which is calculated as follows:

$$D/L = \frac{(Displacement/2240)}{(0.01 \times LWL)^{3}} \tag{9}$$

Equation 9 illustrates that for different models of sailboats, Equation 9 illustrates that, for different models of sailboats, displacement is a better determinant of parameter size than LWL(the length of the waterline) . The significance of the displacement-length ratio is that the lighter the sailboat is relative to its waterline length, the higher its speed potential of the sailboat, especially in the displacement mode[5]. At the same time, the lower the D/L, the more uncomfortable the boat will be in the channel and the more sensitive the boat will be to overload. Therefore, buyers who have a need to race with a single variant sailboat, then displacement

would be an important consideration, which explains its prominence.

The next is the Sail Area-Displacement Ratio (SA/D) for sailboats, which is calculated as follows:

$$SA/D = \frac{Sail\ Area}{(Displacement/64)^{2/3}} \tag{10}$$

While there is no single number that sums up the performance of every different type of sailboat, SA/D, provides a relatively robust way to discuss the relationship between sail power and weight that determines many aspects of a boat's acceleration, maneuverability and performance capabilities[6]. SA/D is also commonly used to measure how easy it is for a sailboat to reach maximum speed. Within this parameter, sail area is more determinative of the parameter than displacement. Generally speaking, a boat with a sail area-displacement ratio below 15 would be considered under-canvased; values above 15 would indicate reasonably good performance.

Due to time cost and other problems, we do not have information about the speed of sailing boats, but we can get the following **inference** based on the above two parameter formulas: **buyers mainly consider the speed of sailing boats when buying sailing boats, and the speed of single variant sailing boats is more influenced by the displacement, while the speed of double variant sailing boats is more influenced by the sail area.**

# References

[1] Cap. 605 Bunker Oil Pollution (Liability and Compensation) (2023) s. 1(5).

[2] Benaglia T , Chauveau D , Hunter D R , et al. mixtools: An R Package for Analyzing Mixture Models[J]. Journal of statistical software, 2009, 32.

[3] Wang Q, Nguyen T. T., Huang J. Z., Nguyen T. T. An efficient random forests algorithm for high dimensional data classification[J]. Advances in Data Analysis and Classification, 2018, 12(4): 953-972.

[4] McLachlan G J, Lee S X, Rathnayake S I. Finite mixture models[J]. Annual review of statistics and its application, 2019, 6: 355-378.

[5] BoatQuest. https://www.sailmagazine.com/boats/comparing-design-ratios. Accessed on April 2, 2023.

[6] Life Of Sailing. https://www.lifeofsailing.com/post/what-is-sail-area-displacement-

## Report

Firstly, based on the data we have collected, sailboat sales are very strong in Hong Kong, with a relatively high compound growth rate. “Hong Kong is, for its size, probably the number one destination for yacht deliveries globally”, said by Charles Massey[3 https://www.sevenstar-yacht-transport.com/news/hong-kong-is-top-destination-for-yacht-deliveries], a sales in Hongkong. Therefore, selling second-hand sailboats in Hong Kong is a wise and forward-looking choice. According to our analysis, we found that the sale price of second-hand sailboats is related to two factors, namely the boat's characteristics and the regional effect. We will describe our analysis results in detail for you.

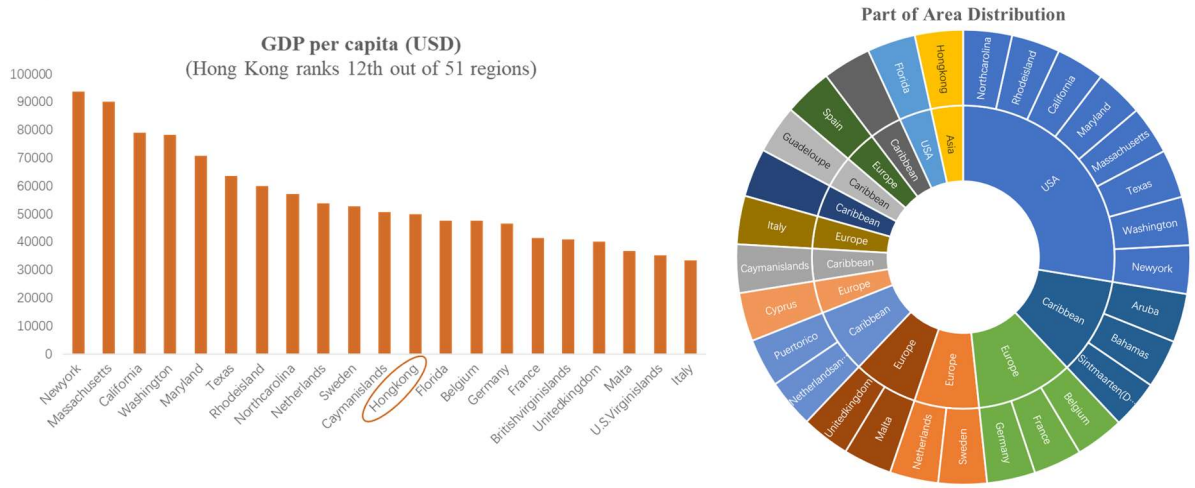


**Figure 19: Market Situation in Hong Kong and Graph of Area Distribution**

Regional Factors: Identify the major factors that can impact sailboat prices in the Hong Kong region, such as economic conditions, market demand, and availability of sailboats.

Based on the data we have collected, we divided the regional effect into three categories: GDP, coastline length, and average cargo throughput. However, in our analysis, we found that GDP is the primary factor influencing the sales volume of used sailboats. We have learned that the GDP of Hong Kong is..., ranking... among all regions. Therefore, there is a large market for second-hand sailboats in Hong Kong.

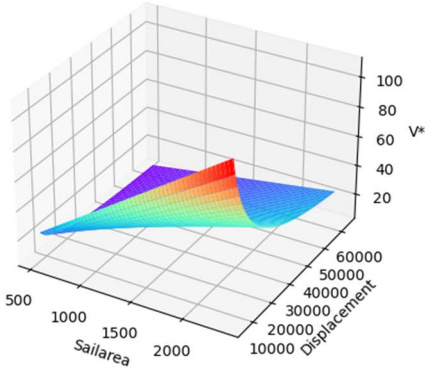


**Figure 18: 3D Relationship between Speed, Sail Area and Displacement**

In terms of boat characteristics, through our research, we found that Sail Area and Displacement are the main factors affecting sailboat prices. These two factors are directly related to the speed of the sailboat, so we speculate that the main consideration for consumers when purchasing sailboats is speed. The secondary factor that affects sailboat prices is Lengths, which is related to the deck area of the sailboat, indicating that consumers not only consider speed when purchasing sailboats but also consider the deck size related to comfort. Therefore, in pricing, we recommend that the selling price should fully consider the comprehensive influence of the three factors, Sail Area, Displacement, and Lengths.

Another point that cannot be ignored is that in our analysis, we found that the prediction of monohull sailboats is unstable and easily affected by various factors, while catamarans are very stable in predictions. This is due to environmental and policy factors. Once a monohull sailboat is damaged, it will have a relatively serious impact on the environment, while for catamarans, this impact is much smaller. Therefore, some governments will use policies to reduce the use of monohull sailboats. Therefore, we recommend focusing on selling catamarans in sales.

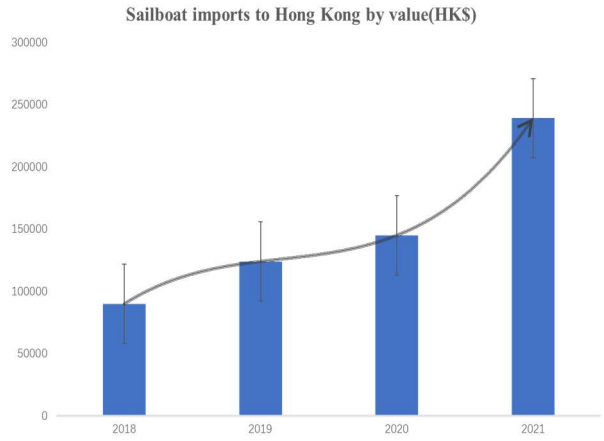


**Figure 19: Sailboat Import in Hong Kong**

The new data shows that in 2021, Hong Kong imports of motorboats for pleasure use (not including outboard boats) rose over 45% year on year, from nearly HK$2 billion in 2020 to HK$2.9 billion. The import value of sailboats rose 56% from 2020 to 2021, from HK$147 million to HK$230 million. The value of imports in both categories far outstripped any previous year over the past decade.

Based on the analysis, we come to the conclusion that Hong Kong is a market with great potential for selling second-hand sailboats, which is a wise decision. When determining the selling price, the sail area, displacement and lengths factors should be fully considered, and pricing should be logical and reasonable. In order to maximize profits, it is recommended to focus on selling catamarans as the main product. However, as a seller, it is important to also focus on providing good customer service, including after-sales service, to enhance the customer experience.