Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2302376

Data Analysis and Prediction of Used Sailboats Prise Based on Random Forest Model

Summary

Sailboats are important commodities in the luxury goods market and their values are influenced by various features. In order to help the broker to better understand the sailboat market, it is necessary to analysis the data and study the features effect. In this paper, we focus on the data of used sailboats and develop totally five models to study the features of the data, and make some predictions.

For **Model I**, the features of used sailboats are quantified and normalized. We firstly find the features that are possibly related to the price and then we use **Pearson Product-Moment Correlation Coefficient (PPMCC)** to judge the relevance between each feature and price.

For **Model II**, all the data is processed and we build **Random Forest Model** to predict the price with the features. The model training uses large amounts of data and we use two metrics **R-Squared** and **Mean Absolute Error (MAE)** to evaluate the prediction accuracy.

For **Model III**, due to the effect of Make and Region are significant, we reassess the quantification and use **Discrete Processing** to quantify them. Then, the relevance between them and price is judge by **Spearman's Rank Correlation Coefficient (SRCC)**.

For **Model IV**, we use **Random Forest Model** with **Discrete Variables** to predict and evaluate the region effect of the price of used sailboats. Then, we analyse the result and make some conclusions.

For **Model V**, we choose an informative subset of used sailboats and compare it to the data from the Hong Kong (SAR) market. Then, the region effect is studied by **Random Forest Model**. The data set is split between monohulls and catamarans, and we consider the difference of region effect between them.

Some other inferences of the data are also simply mentioned. At last, we reflect on the strengths and weaknesses of our models. We hope our prediction model to be useful for the sailboat broker in Hong Kong (SAR).

Keywords: Used Sailboat Random Forest Model Pearson Product-moment Correlation Coefficient Spearman's Rank Correlation Coefficient Discrete Processing Region Effect

Contents

1 Introduction

1.1 Problem Background

Sailboats are important commodities in the luxury goods market and their values are influenced by various features such as age and market trends. Accurately predicting the listing price of used sailboats is critical for both sailboat brokers and enthusiasts, as it helps to make informed buying and selling decisions. While features such as make, variant, length, and year of manufacture have traditionally been used to determine sailboat pricing, other features such as geographic region and economic data by year and region can also play a role in determining sailboat pricing. Therefore, the accurate modeling and analysis of these features can provide important insights into the sailboat market and inform pricing decisions.

1.2 Restatement of the Problem

In order to develop practical mathematical models through the analysis of data and complete the report, the Hong Kong (SAR) sailboat broker puts forward several problems to be solved. The specific requirements are as follows.

- Develop a mathematical model which can explain and predict the price of any used sailboat.
- Discuss the regional effect on the price and address the practical and statistical significance of any regional effects noted.
- Explain how modeling sailboat prices in specific regions can be useful in the Hong Kong (SAR) market, using a subset of sailboats from a provided spreadsheet.
- Compare listing price data for this subset in Hong Kong, and model the regional effect on sailboat prices in Hong Kong for both monohull and catamarans.
- Identify and discuss any other interesting and informative inferences from the data.
- Prepare a report for the Hong Kong (SAR) sailboat broker to help understanding.

1.3 Our Approach

This topic requires us to build mathematical models for the Hong Kong (SAR) sailboat broker to better understand the features of used sailboat which influence the price. **Figure 1** mainly shows our approach.

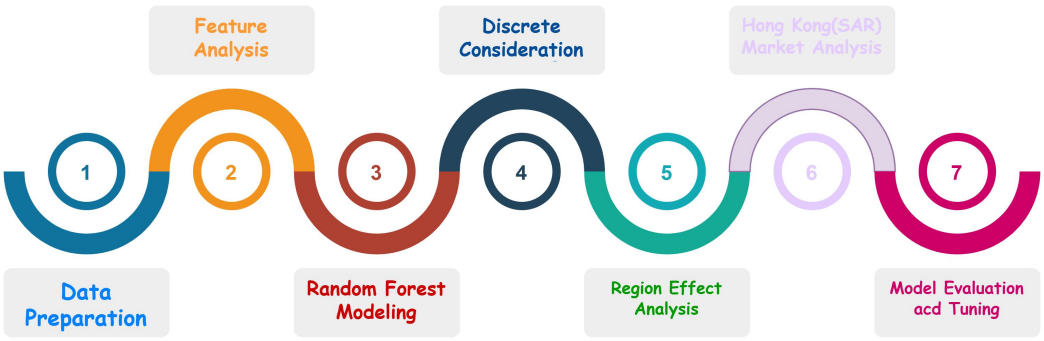


Figure 1: Our Approach

2 Assumptions

To simplify the problem, the basic assumptions are as follows, each of which is properly justified.

- **Assumption 1:**
  The ignored features of used sailboats have subtle impact on the price.

  Justification:
  The ignored features are not shown in the data source, which means they are not important to the customers. Therefore, they are highly likely unrelated to the price.

- **Assumption 2:**
  The economic data can be represented by GDP per capita and total GDP.

  Justification:
  GDP per capita and total GDP are widely used to represent the economy of a region.

- **Assumption 3:**
  The geographic region of the boat's location is mainly one of Caribbean, Europe and USA.

  Justification:
  Although the geographic region can be others, but most of regions in the data can be thought in these three geographic region.

3 Model Preparation

3.1 Data Collection

The data we used include Make, Variant, Length, Geographic Region, Country/Region/State, Listing Price, Year. The data sources are listed in **Table 1**.

Here, the second database file is obtained from Yachtworld website.[1]The third database file is obtained from the Bureau of Economic Analysis (BEA).[2] The forth and fifth database file is downloaded from International Monetary Fund website.[3]

Table 1: Data Source

Database File	Data Type
2023_MCM_Problem_Y_Boats.xlsx	Data set
Used_Sailboats_Yachtworld.csv	Data set
GDP per capita of each states of America.xlsx	Data set
GDP per capita.xls	Data set
GDP.xls	Data set

3.2 Notations

Table 2 shows symbols and notations used in this paper. Note that symbols used only once are not included and will be defined later.

Table 2: Notations

Symbol	Definition
P	Price of the used sailboat
T	Time since manufacture
L	Length of the used sailboat
D	Maximum Draft of the used sailboat
B	Beam of the used sailboat
E	Engine hours of the used sailboat
H	Headroom number of the used sailboat
G	GDP per capita
M	Market Share of the manufacturer
r	Pearson product-moment correlation coefficient (PPMCC)
ρ	Spearman's rank correlation coefficient (SRCC)
γ	Confidence coefficient
R^2	R-Squared

3.3 Data Clean

Since the data sets may have missing data or other issues, it is necessary to do data cleaning prior to analysis. Therefore, all the data sets are processed by Python. Firstly, the data that has missing features are deleted. Then, features are transformed into correct format that is easy to use in our model.

4 Model I: Feature Analysis

4.1 Description

In this model, discussed which features of the used sailboats affect the price significantly. According to the statistics method, the Pearson product-moment correlation coefficient and the Spearman's rank correlation coefficient can be calculated by Python, which means the relevance to the price of each feature is determined.

4.2 Implement

4.2.1 Features of the Used Sailboats

First, it is necessary to determine the features of the used sailboats. According to the data from **Table 1**, there are several main features as **Figure 2** shows

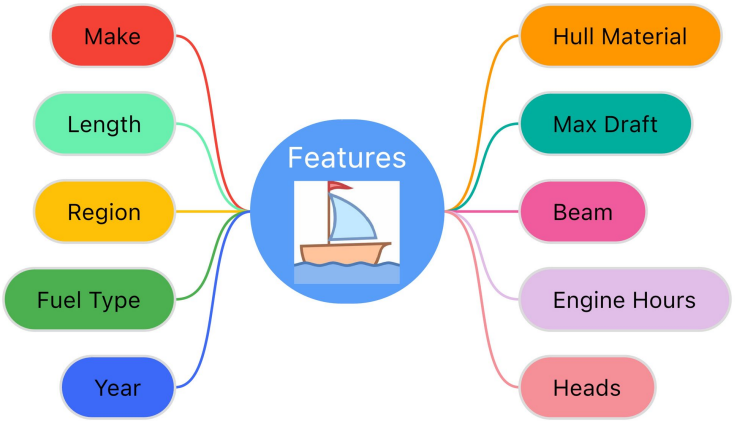


Figure 2: Features

- **Year:** The year the boat was manufactured.
- **Make:** The name of the manufacturer of the boat.
- **Length:** The length of the boat in feet.
- **Region:** The region of the boat's location
- **Fuel Type:** The type of the fuel that the boat uses.
- **Hull Material:** The materials of which a boat's hull is made.
- **Max Draft:** The maximum depth of water that can float a boat.
- **Beam:** The width of a boat at its widest point.
- **Engine Hours:** The number of hours the boat's engine(s) have run since new.

- **Heads:** The height available to stand up in the cabin.

After the primary analysis of the data, the fuel type and the hull material should not be considered since almost all sailboats use diesel fuel and the hull material data is few.

Therefore, the other eight features can be the factors which may affect the price of the used sailboats. In order to build a mathematical model to explain and predict the price, the data should be quantified and normalized first.

4.2.2 Quantification and Normalization

According to the qualitative data, the quantification are as follows

- **Make:** Use market share $M = \frac{Num_M}{Num_{total}}$ to represent.
- **Region:** Primarily use GDP per capita G to represent.

Furthermore, in order to make the result better, the normalization is also necessary.

- **Year:** $T = 2023 - Manufactured\,Time.$

4.2.3 Pearson Product-moment Correlation Coefficient (PPMCC)

Pearson Product-Moment Correlation Coefficient (PPMCC) is a coefficient to describe the relevance between two variables. It is convenience to judge the relevance between the attributes of the word and the difficulty of the word by calculating PPMCC, denoted by r.

According to the definition of r, there is

$$r(X,Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\sqrt{E(Y^2) - E(Y)^2}} \tag{1}$$

Here, X and Y represent two variables. $E(S) = \bar{S} = \frac{1}{n}\sum_{i=1}^{n} S_i (S = X, Y, XY, X^2, Y^2)$, which means the average value of S.

Then, according to **Equation 1**, PPMCC r and its confidence coefficient γ of each feature can be calculated. The results of the calculation are as follows

Table 3: PPMCC and Confidence Coefficient of Different Features

Features	T	L	D	B
r	-0.361	0.626	0.237	0.690
γ	1.14×10^{-16}	3.95×10^{-55}	9.78×10^{-8}	5.11×10^{-71}
Features	E	H	G	M
r	0.060	0.670	-0.071	0.038
γ	0.182	1.33×10^{-65}	0.113	0.402

In general, if the confidence coefficient γ is smaller than the significance level α, then the two variables have significant correlations.

According to the values of γ in **Table 3** and the default value $\alpha = 0.05$, it is clear that engine hours, GDP and the market share of manufacturers make an insignificant affect to the price.

4.3 Conclusion

As **Table 3** shows, the result is that the price of used sailboats has the factors including Year, Length, Max Draft, Beam, Heads and Region. In addition, the region effect is not only caused by the economy because G is a unrelated factor to the price. Therefore, the price can be denoted as $P = P(T, L, D, B, H)$ without the effect of the region. The expression of P will be discussed in **Model II**.

5 Model II: Random Forest Price Prediction

5.1 Description

In this model, the variables from **Model I** are utilized with high correlation to the used sailboat prices to predict their prices, using the *random forest model*, which is an ensemble learning model based on multiple decision trees. Specifically, there is a set of relevant features that affect sailboat prices that discussed in **Model I**. Then, the random forest model is trained on these variables and can predict the price of each sailboat. The estimation will also be discussed below.

5.2 Implement

5.2.1 Model Training

In the random forest model, each decision tree is built based on the following algorithm, where N represents the number of training samples and M represents the number of features. An input feature number m is selected to determine the decision at each node of the decision tree, where m should be much smaller than M. From the N training samples, a training set is formed by bootstrap sampling with replacement, and the samples not selected are used for prediction and error evaluation. For each node, m features are randomly selected, and the decision at each node is based on these features. Based on these m features, the optimal split is calculated. Each tree is fully grown without pruning, which may be used after building a normal tree classifier.

The pseudocode of this algorithm is as follows:

Algorithm 1: Random Forest Training

Input: Training dataset $D = (x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$; Number of trees T; Number of features to consider at each split k; Maximum depth d_{max}.

Output: Ensemble of decision trees $F = f_1, f_2, \cdots, f_T$.

for $t = 1$ *to* T **do**
  Select a random subset of features $S \subseteq \{1, 2, \cdots, d\}$ with $|S| = k$;
  Sample a bootstrap dataset D_t from D;
  Grow a decision tree f_t from D_t by recursively splitting the nodes;
  while $d_{tree}! = d_{max}$ **do**
    if *the node contains samples belonging to only one class* **then**
      break;
    end
    Select the best feature $j \in S$ and the best split point s to maximize information gain;
    Add the tree f_t to the ensemble F;
  end
end
return F

5.2.2 Model Predicting

The prediction result of Random Forest Regression (RFR) is obtained by averaging the predictions of all internal binary decision trees. The prediction process of a binary decision tree mainly consists of the following steps:

1. For a given input sample, starting from the root node of the binary decision tree, determine whether the current node is a leaf node. If it is a leaf node, return the predicted value of the leaf (i.e., the average of the target variable of the samples in the current leaf). If not, proceed to the next step.

2. Compare the value of the corresponding variable in the sample with the splitting value of the current node, based on the splitting variable and value of the current node. If the sample variable value is less than or equal to the splitting value of the current node, then visit the left child node of the current node. If the sample variable value is greater than the splitting value of the current node, then visit the right child node of the current node.

3. Repeat step 2 until a leaf node is visited, and return the predicted value of the leaf node.

The pseudocode of this algorithm is as follows:

Algorithm 2: Random Forest Prediction

Input: An unseen sample x; An ensemble of decision trees $F = f_1, f_2, \cdots, f_T$.
Output: A predicted class label $\hat{y}$.

```
for t = 1 to T do
    for node in f_t do
        predict the class label of the input sample x;
    end
    Record the predicted class label y_t;
end
Compute the mode of the predicted class labels y_1, y_2, ..., y_T as the final prediction ŷ;
return ŷ
```

For the used sailboat price prediction, the result is as **Figure 3** and **Figure 4** show

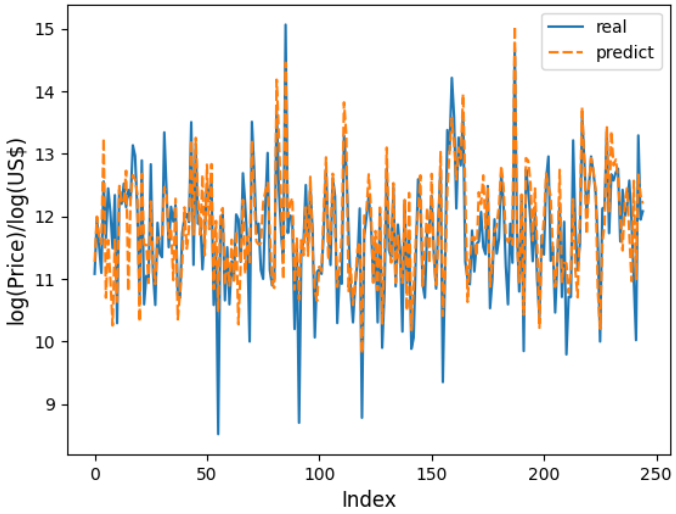


Figure 3: Price Prediction I

5.2.3 Feature Importance

The feature importance can be calculated again as we used the RFR to predict the price.In sklearn's internal tree models, feature importance is calculated using either the Gini Importance or the Mean Decrease Impurity (MDI) method.

For Gini Importance, the calculation formula is:

$$Gini\ Importance = \sum_{j=1}^{n} w_j \times (p(1-p))j - wj' \times (p'(1-p'))_j \tag{2}$$

where n is the number of features, w is the sample weight, p represents the proportion of samples in the original set, and p' represents the proportion of samples in the subset after splitting. It can be

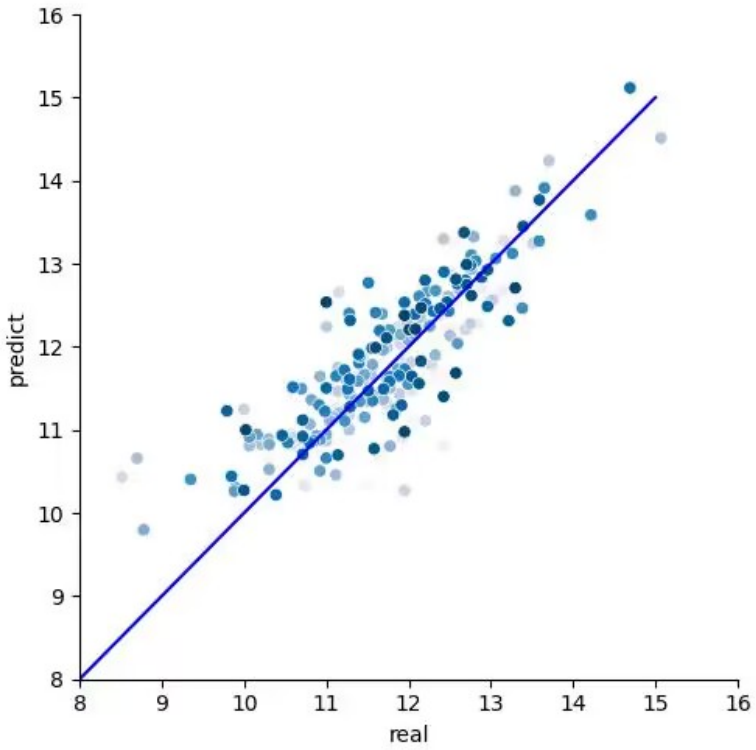


Figure 4: Price Prediction II

seen that for each feature, its Gini Importance is calculated in all trees, and the average is taken as the importance score.

For MDI, the calculation process is as follows:

For each tree, calculate the Impurity Decrease (ID) of each feature at each node during splitting;

For each feature, take the average ID across all trees as its importance score. Therefore, the following formula can be used to calculate the MDI of feature j:

$$MDI_j = \frac{\sum_{t=1}^{T} 1(j, s_t)}{N_t} \times ID(j, s_t) \tag{3}$$

where T is the number of trees, $1(j, s_t)$ represents whether feature j is used in the split of tree t, N_t represents the number of nodes in tree t, and ID(j, s_t) represents the impurity decrease of feature j at node s_t of tree t.For the price prediction, the importance of each feature is as follows

Table 4: Feature Importance

Features	T	L	D	B	H
$Importance$	0.1439	0.1505	0.0750	0.2899	0.3407

Feature importance indicates the degree of influence that a feature has on the prediction results. A higher feature importance value indicates a greater impact on the prediction results, while a lower

value indicates a smaller impact.

5.2.4 Regression Evaluation (R-Squared and MAE)

Goodness of Fit refers to the degree to which a regression line fits the observed values. The statistical measure that quantifies the goodness of fit is the Coefficient of Determination, denoted as R^2.

For m samples, $(\overrightarrow{x_1}, y_1), (\overrightarrow{x_2}, y_2), ..., (\overrightarrow{x_m}, y_m)$, the estimated values of a certain model are $(\overrightarrow{x_1}, \widehat{y}_1), (\overrightarrow{x_2}, \widehat{y}_2), ..., (\overrightarrow{x_m}, \widehat{y}_m)$.

The calculation of the Total Sum of Squares (TSS) for a sample is below:

$$TSS = \sum_{i=1}^{m} (y_i - \bar{y})^2 \tag{4}$$

The calculation of Residual Sum of Squares (RSS) is given by the formula:

$$RSS = \sum_{i=1}^{m} (\widehat{y}_i - y)^2 \tag{5}$$

Then, the definition of R^2 is

$$R^2 = 1 - \frac{RSS}{TSS} \tag{6}$$

The coefficient R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. The value of R-squared ranges from 0 to 1, where a higher R-squared value indicates a better fit of the regression model to the data.

Another regression evaluation is the MAE, or Mean Absolute Error. It is a statistical measure that calculates the average of the absolute differences between the predicted and actual values in a dataset. The formula for calculating MAE is:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |(y_i - \widehat{y_i})| \tag{7}$$

The resulting value of MSE is always non-negative, with a value of 0 indicating a perfect fit between the predicted and actual values.

For the price prediction,the R-squared is

$$R^2 = 0.7555824269519621$$

the MAE is

$$MAE = 81006.06441805104$$

5.3 Conclusion

Based on the evaluation results, the random forest model can perform well in predicting prices given the features provided, and provides feature importance rankings. It should be noted that the feature set does not include "Region" and its impact will be discussed in detail in the next model.

6 Model III: Discrete Processing

6.1 Description

In this model, the Make feature and Region feature will be reassessed. In **Model I**, the Make feature is considered as the market share and the region is considered to be represented by GDP per capita G, but the correlations are both poor. However, according to **Figure 5** and **Figure 6**, the effect of Make and Region are significant. Therefore, the effect of make and region should be reconsidered in other ways.

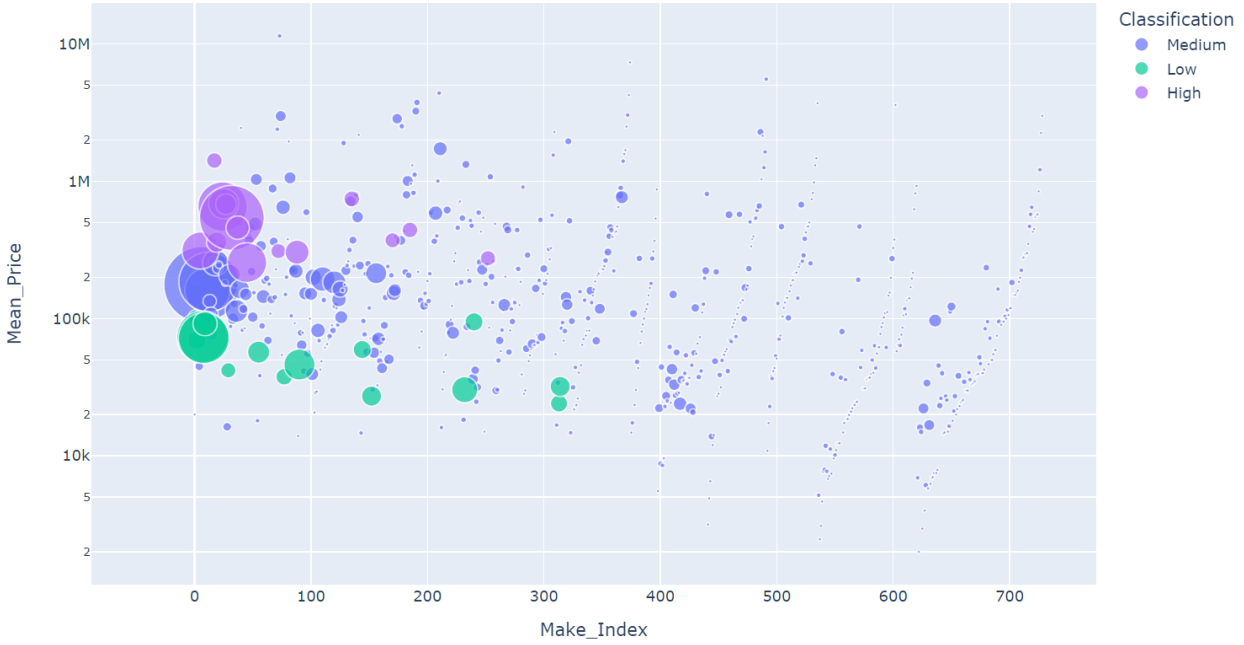


Figure 5: Mean Price of Different Makes

6.2 Implement

6.2.1 Make Reassessment

According to the data in **Table 1** and **Figure 5**, the price distributions of different Makes are different significantly. Therefore, the quantification of Make is reconsidered as discrete variable. Different value represents different Make.

6.2.2 Region Reassessment

According to the data in **Table 1** and **Figure 6**, the price distributions in different cities are different significantly. In order to reassess the region effect, the quantification of the region is reconsidered by the following methods:

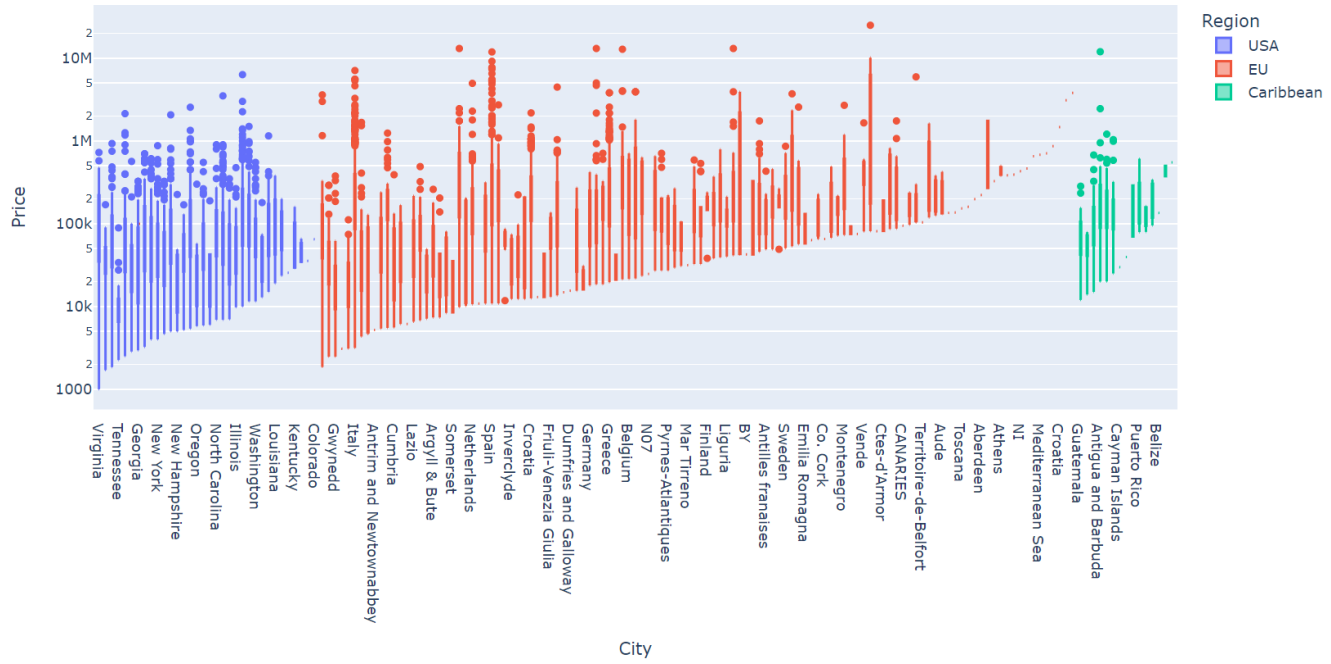


Figure 6: Price in Different Cities

- **Method I: Geographic Discrete Transform**
  The geographic region includes USA, Europe, Caribbean and other region. Therefore, the region can be transformed into the discrete variable, which has four possible value.

- **Method II: Total GDP Discrete Transform**
  As mentioned, GDP per capita is not a relative factor. But the total GDP considers both the population and economic situation, which may be a relative factor.

- **Method III: Mean Price Discrete Transform**
  As **Figure 6** shows, the mean price of different countries have difference, to quantify the difference, the countries can be transformed into three discrete variable: -1,0,1, representing the mean price is high, medium or low.

- **Method IV: Country Discrete Transform**
  Another methods to quantify the country difference is to transform each country into discrete variable as features to predict the price.

Then, for each method, the relevance can be both represented by the Spearman's rank correlation coefficient (SRCC) and the feature importance in the random forest prediction.

6.2.3 Spearman's Rank Correlation Coefficient (SRCC)

The Spearman's rank correlation coefficient is defined as the Pearson Product-Moment Correlation Coefficient between the rank variables. While Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). The relevance between the region and the price can be judged by the Spearman's rank correlation coefficient.

The expression is shown below

$$\rho_{X,Y} = r_{R(X),R(Y)} = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \tag{8}$$

Here, X, Y are converted to ranks $R(X), R(Y)$. r denotes the PPMCC, but applied to the rank variables. $cov(R(X), R(Y))$ is the covariance of the rank variables. $\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are the standard deviations of the rank variables.

Then, according to **Equation 8**, SRCC ρ for each method can be calculated respectively. The result is as follows

Table 5: SRCC and Confidence Coefficient of Different Methods

Method	I	II	III	IV
ρ	0.260	-0.355	0.455	-0.355
γ	2.15×10^{-88}	1.97×10^{-167}	3.97×10^{-288}	1.97×10^{-167}

The result shown in **Table 5** means that each method is reasonable.

6.2.4 Feature Importance in The Random Forest Prediction

The Feature importance has been introduced in **Model II**.

For each method, the Feature importance is shown as follows

Table 6: Feature Importance of Different Methods

Methods	I	II	III	IV
$Importance$	0.0011	0.0290	0.0036	0.0175
$R-squared$	0.8238	0.8079	0.8372	0.8234

6.3 Conclusion

According to **Table 6**, the prediction accuracy is improved since two features are discrete considered. Then, the region effect will be discussed in **Model IV** by using this model.

7 Model IV: Region Effect

7.1 Description

In this model, the region effect would be considered by **Model III**. The price prediction is considered with the region feature and without the region feature respectively. Then the region effect is shown by the difference of two price prediction results.

7.2 Implement

The process is similar to **Model II**, only Make feature and Region feature are reassessed. Since different methods of Region feature lead to similar model accuracy, here taking **Method III** to quantify the region feature.

Figure 7 shows the price difference between the prediction value and real value.

Figure 8 shows the price difference which is caused by region effect.

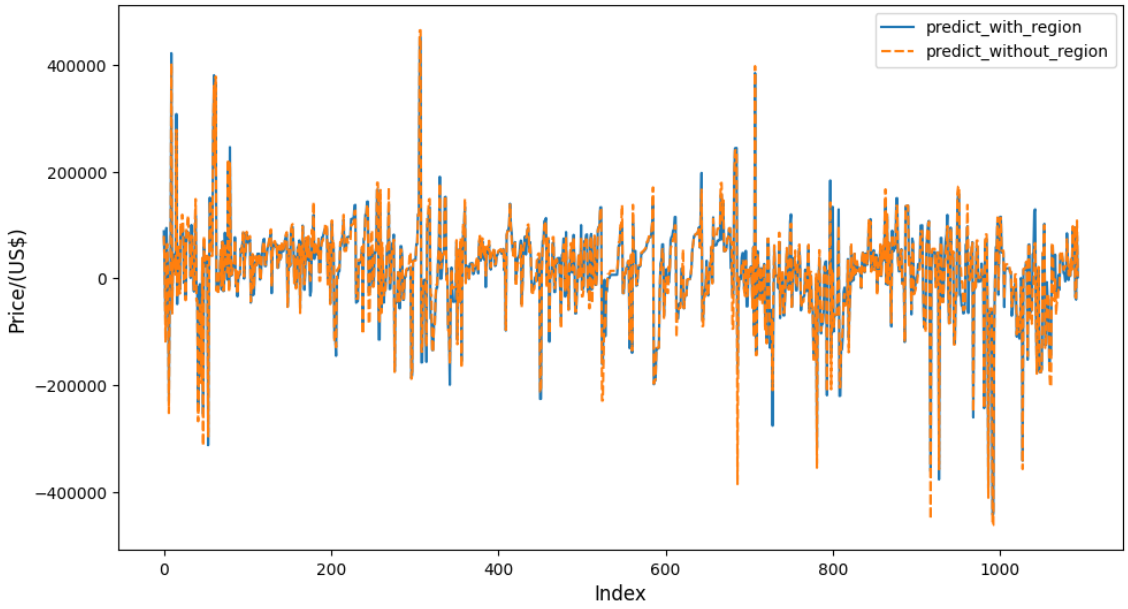


Figure 7: Price Difference I

According to the two figures, there are some obvious features:

- The prediction is accurate for the most data, but there are still some data that have a large error.
- The region effect is small compared to the difference between the prediction value and real value.

7.3 Conclusion

Although the region effect is relative to the price as **Table 5** shows, but it is subtle compared to the price of used sailboats. The result is expected since the importance in random forest is small shown in **Table 6**.

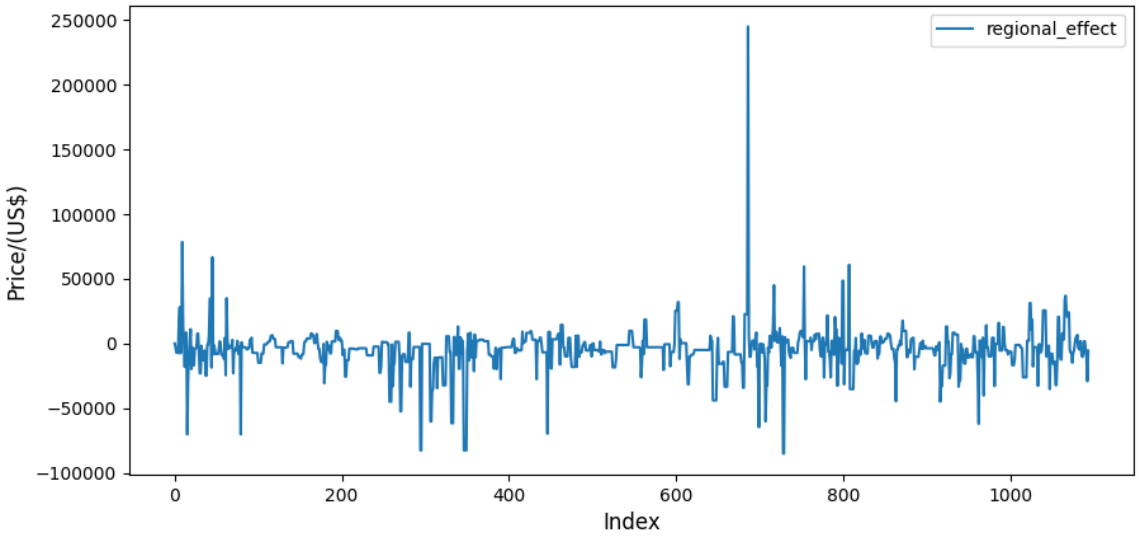


Figure 8: Price Difference II

8 Model V: Hong Kong (SAR) Market Analysis

8.1 Description

In this model, the region effect of Hong Kong (SAR) would be considered. According to the data, the catamarans and monohull sailboats will be considered respectively.

8.2 Implement

8.2.1 Subset Chosen

To compare with the data from the Hong Kong (SAR) market, the informative subset of sailboats is chosen only the manufactures which have more than 10 data and the data is complete and reasonable. Then, the subset is split between monohulls and catamarans.

8.2.2 Monohull Sailboats Result

The prediction result is as follows

Table 7: Monohull Sailboats Result

Features	$Length$	$Year$	$Make$	$Region$
$Importance(With)$	0.370	0.305	0.143	0.182
$Importance(Without)$	0.467	0.373	0.167	/

R-Squared Value: $R^2(With) = 0.7324, R^2(Without) = 0.7217.$

8.2.3 Catamaran Sailboats Result

The prediction result is as follows

Table 8: Catamaran Sailboats Result

Features	$Length$	$Year$	$Make$	$Region$
$Importance(With)$	0.423	0.342	0.128	0.107
$Importance(Without)$	0.461	0.385	0.154	/

R-Squared Value: $R^2(With) = 0.6833, R^2(Without) = 0.6649$.

8.3 Conclusion

The above results show the region effect of Hong Kong (SAR) market. The effect of catamarans is larger than monohulls, but both are small. The prediction is not very accurate since the data is few.

9 Other Inferences

Upon reexamining the data table, some noteworthy conclusions were drawn as follows

- **The Impact of Age on the Price of Used Boats**
  The model suggests that as the age of the boat becomes newer, the used boat prices in each region are gradually increasing and the correlation is relatively significant, as shown in the **Figure 9** below. Therefore, age is an important factor to consider when predicting the price of a used boat.

- **Other Significant Features**
  When other features are added to the given sailboats, it is noted that two other features, as the **Table 4** shows, are significant in predicting the price of the used sailboat.

  - **Head Room**
    The number of head rooms is positively correlated with price and has a significant impact, according to our model.
  - **Beam**
    The model shows that there is a positive correlation between the beam and the price of used boats, and the effect is significant.

  When replacing the generic monohulled sailboats and catamarans categories with the two variables mentioned above, head room and beam, more accurate predictive results were obtained.

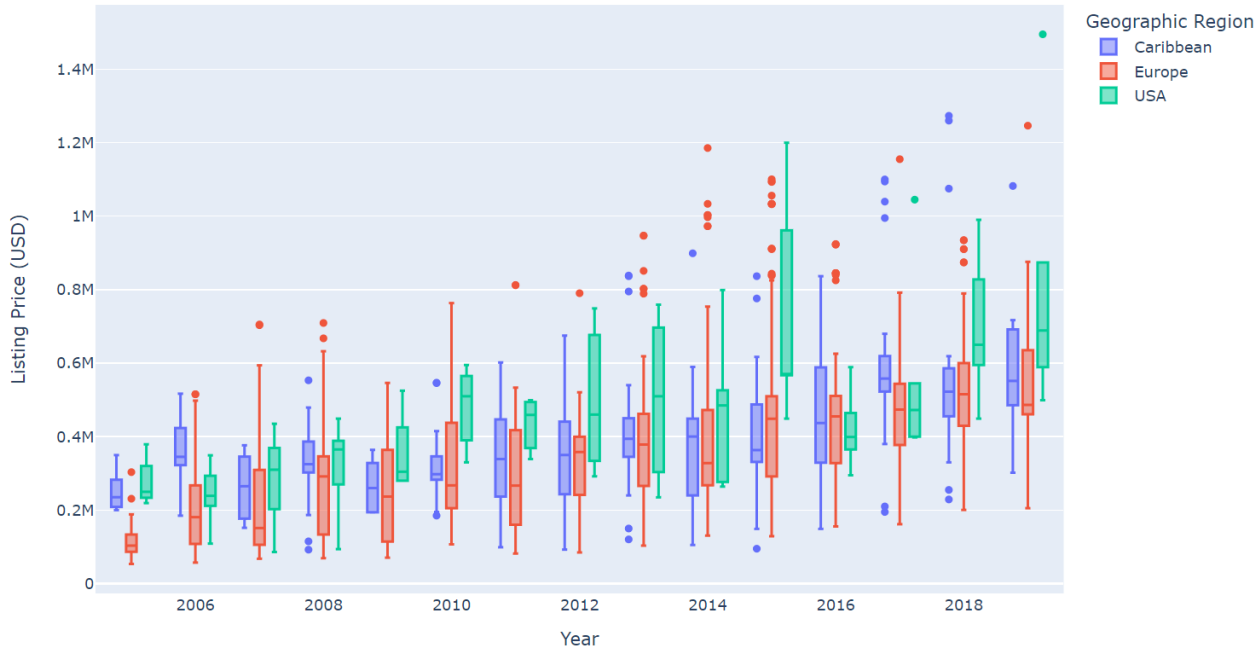


Figure 9: Price by Year

10 Model Evaluation and Discussion

10.1 Strength and Weakness

10.1.1 Strength

- **Good Accuracy**
  The Random Forest Algorithm we uesd is a strong model that integrates decision tree model and the bagging method. As a result, it's more accurate in the regression process.

- **Strong Generalization Ability**
  When creating random forests, unbiased estimation is used for generlization errors, and the model generalization ability is strong.

10.1.2 Weakness

- **Lack of data**
  The data for the model is not adequate enough and more features need to be considered to accurately predict.

- **Overfitting**
  The Random Forest Algorithm is prone to overfitting when the data is noisy.

10.2 Possible Improvements

- **More Features Considered**
  The considered features are limited. If we consider more features, the model can be more accurate and persuasive.
- **More Credible Data Needed**
  The data of used sailboats with copious features are limited. If we could collect more credible data, we can obtain more accurate models and results.
- **More Professional Analysis Methods**
  The data can be further analysed by more professional methods.

References

[1] https://www.yachtworld.com/boats-for-sale/condition-used/

[2] https://www.bea.gov/

[3] https://www.imf.org/en/Home

Appendices

A Random Forest Main Code

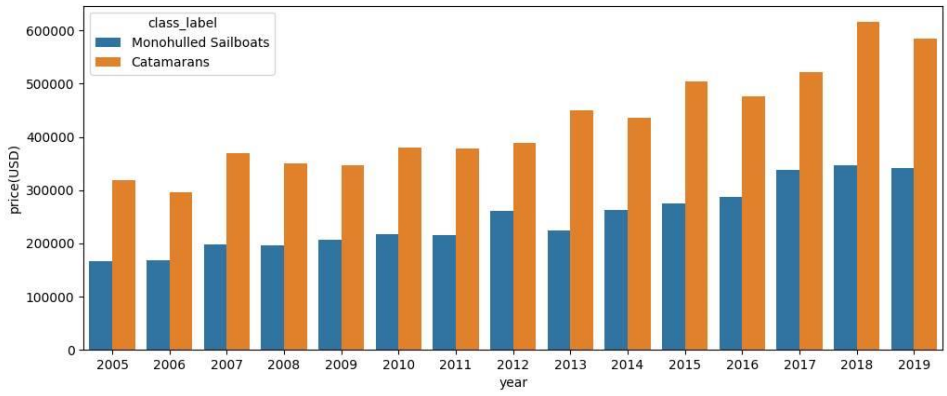
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from sklearn import ensemble
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score
from pprint import pprint
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
random_seed=44
random_forest_seed=np.random.randint(low=1,high=230)
df_train=pd.read_excel('2023_MCM_Problem_Y_Boats_ver6_1.xls')
df_test=pd.read_csv('sailboats_hongkong_1.csv')
print(df_train.columns)
print(df_test.columns)
y_test=df_test['Price'].tolist()
length_test=df_test['Length'].tolist()
year_test=df_test['Year'].tolist()
```

```
GDP_test=df_test['GDP'].tolist()
classlabel_test=df_test[' class_label'].tolist()
#countrylabel_train=df_train['country_label'].tolist()
#C_test=df_test['C'].tolist()
y_train=df_train['Listing Price (USD)'].tolist()
length=df_train['Length \n(ft)'].tolist()
year=df_train['Year'].tolist()
#R_train=df_train['R'].tolist()
GDP_train=df_train['GDP'].tolist()
classlabel_train=df_train['class_label'].tolist()
#countrylabel_test=df_test['country_label'].tolist()
#C_train=df_train['C'].tolist()
x_train1=[[x1,x2,x3,x4/1000] for x1,x2,x3,x4 in
    zip(length,year,classlabel_train,GDP_train)]
x_test1=[[x1,int(x2),x3,x4] for x1,x2,x3,x4 in
    zip(length_test,year_test,classlabel_test,GDP_test)]
x_train2=[[x1,x2,x3] for x1,x2,x3 in zip(length,year,classlabel_train)]
x_test2=[[x1,int(x2),x3] for x1,x2,x3 in zip(length_test,year_test,classlabel_test)]
random_forest_regressor1= ensemble.RandomForestRegressor(n_estimators=50)
random_forest_regressor1.fit(x_train1, y_train)
score = random_forest_regressor1.score(x_test1, y_test)
result1 = random_forest_regressor1.predict(x_test1)
print(score)
print(mean_absolute_error(y_test,result1))
print(r2_score(y_test,result1))
print(random_forest_regressor1.feature_importances_)
random_forest_regressor2= ensemble.RandomForestRegressor(n_estimators=50)
random_forest_regressor2.fit(x_train2, y_train)
score = random_forest_regressor2.score(x_test2, y_test)
result2 = random_forest_regressor2.predict(x_test2)
print(score)
print(mean_absolute_error(y_test,result2))
print(r2_score(y_test,result2))
print(random_forest_regressor2.feature_importances_)
dataset=[result1,y_test]
index=['predict_with_region','predict_without_region']
dfs=pd.DataFrame(data=dataset,index=index)
dfs=dfs.T
sns.lineplot(data=dfs)
plt.xlabel('Index',fontsize=12)
plt.ylabel('Price/(US$)',fontsize=12)
plt.show()
```

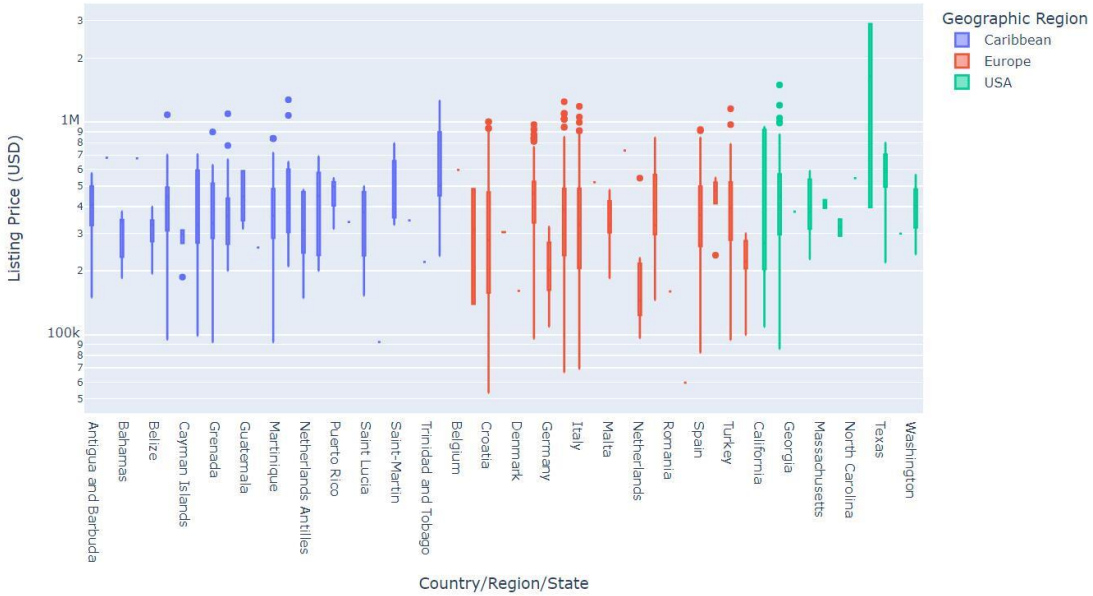
Dear Broker,

We have conducted a comprehensive analysis of the Hong Kong SAR sailing boat market based on the electronic spreadsheet data provided, as well as additional data obtained from other sources. Here are our conclusions.

1. In the Hong Kong SAR market as well as markets all around the world, catamarans have a higher average price than monohulls. Our data analysis shows that this difference is significant. The price distribution histogram plot below clearly shows the price differences between these two types of sailing boats.



2. The impact of region on sailing boat prices in Hong Kong SAR is not significant. We used a random forest model to estimate the impact of different regions, but our data analysis shows that the impact of region on sailing boat prices is small. This may be due to Hong Kong SAR being an international city, and the price differences between different regions are not significant. This conclusion also holds for other regions in the world.



3. We find that there are several important features that influence the listing price of the used sailboats, including the boat length, year, beams, and the number of headrooms. We found that there is a positive correlation between these factors and price.

These are our conclusions , and we believe they will be helpful in your work in the Hong Kong SAR sailing boat market.

Sincerely,
MCM 2302376 Team.