Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2333441

A Price Model of Used Sailboats Based on Random Forest and Mulitiple Linear Regression

Abstract

Used sailboats are expensive, often listed in several hundred thousand dollars. A variety of factors could affect the price. A **Radom-Forest-Multiple-Linear-Regression (RF-MLR) Model** is established to explain the relationship between the listing price of used sailboats and the factors, find out possible regional effects, and forecast the Hong Kong market.

Firstly, preprocess data and use **Random Forest algorithm** to train model for monohulled sailboats and catamarans. Then the top 1/3 variables in importance order are selected for model II fitting. **Multiple Linear Regression Model** is used to fit the features that have the greatest influence on the selection of two types of sailboats. In this process, the categorical variable manufacturer's brand is converted into **dummy variable** to participate in regression. Evaluate the prediction accuracy of each variation. For monohulled sailboats and catamarans. The deviation within 10% accounted for 96% and 94% respectively.

Thirdly, analyze the regional effect of listing price of sailboats. Compare the average listing price of sailboats in Caribbean, Europe and USA, which results **USA >Global≈Europe >Caribbean.** This is consistent with the RF-MLR model that the listing price of monohulled sailboats is positively correlated with regional per capita consumption expenditure in 2020. However, for each variant, there is the possibility that isn't consistent with regional effects. Finally, the statistical and practical significance are analyzed.

Next, we add actual listing price to Multiple Linear Regression models for modification, and establish a new **regional effect evaluation function RE**. Select the same observations, and then substitute it into RE. The listing price in Hong Kong is higher than that in the other three geographical regions, no matter for which type of sailboats

There are two interesting facts. First, in terms of brand, Germany and France have the tendency of local brands, domestic brands accounting for 83.2% and 75%. In terms of hull choice, Caribbean tends to choose catamarans. Second, the displacement of the hull has a moderate strength correlation with the region.

Sensitivity analysis of Multiple Linear Regression models was also performed. When we remove one of the arguments, R squared drops by more than 5%, and the magnitude of the drop is positively correlated with the magnitude of its importance. On the other hand, the larger the coefficient change, the greater the deviation is.

Finally, we provided a report to the sailboat broker in Hong Kong, in which we briefly describe the principles of the RF-MLR model and sour conclusions we.

Keywords: Random Forest, Multiple Linear Regression, pricing model, sailboat

Contents

1 Introduction **2**
1.1 Problem Background . . . 2
1.2 Restatement of the Problem . . . 2
1.3 Our Work . . . 2

2 Assumptions and Justifications **3**

3 Notations **3**

4 Model Preparation **4**
4.1 Data preparation . . . 4
4.2 Variable selection . . . 5

5 Modeling **6**
5.1 Idea analysis . . . 6
5.2 Random forest model establishment . . . 7
5.2.1 Model principle . . . 7
5.2.2 Model building . . . 7
5.2.3 Model calculation . . . 8
5.3 Multiple linear regression model prediction . . . 11
5.3.1 Model building . . . 11
5.3.2 Precision of each variant . . . 12

6 An explanation of regional effects using our Model **13**
6.1 Global regional effect and interpretation . . . 13
6.2 Regional effect consistency test for different variants . . . 15
6.3 Meaning . . . 15

7 Model Applying in Hong Kong **16**
7.1 How to apply our models . . . 16
7.2 Hong Kong second-hand sailboat data search . . . 17
7.3 Regional effect in Hong Kong . . . 17

8 Interesting discovery **18**
8.1 Trading tendency . . . 18
8.2 Correlation between region and displacement . . . 19

9 Conclusion **20**
9.1 Sensitivity Analysis . . . 20
9.2 Evaluation of models . . . 21

10 A report on the Pricing of Used sailboats **23**

1 Introduction

1.1 Problem Background

Common sailboats can be divided into monohulled sailboats and catamarans according to the number of hulls. Like many luxury goods, sailboats vary in value as they age and market conditions change. Unlike luxury goods, the listed price of a used sailboat is not only related to the make, but also to the hull specification (variant).Examples include length, beam, draft, displacement, rigging, sail area, hull material, engine hours, sleeping capacity, headroom and electronics.

At the same time, depending on the use nature of sailing, whether the listed area is near the sea, whether the economy is developed and whether the tourism industry is developed will also affect the price level. For example, the per capita GDP and per capita consumption expenditure of region A and Region B, which are also near the sea, are higher than those of region B. Under normal circumstances, affected by the market environment, the listed price of the same brand, model and years of yachts in region A is likely to be higher than that in region B.

The countries and regions given in the attachment are all near the sea. Therefore, to accurately predict the listing price of second-hand sailing boats, it is also necessary to analyze the economic level and tourism development level of each country and region.

The above factors have different impacts on the listed price of sailing ships. Mathematical modeling can be used to analyze the effects of each factor and rank them according to their importance, so as to comprehensively consider the impact of each factor on the listed price and then accurately predict.

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to establish a mathematical model that explains the listing price of each of the sailboats in the provided spreadsheet, and complete the following tasks using the model:

- Draw on other sources to understand additional features of a given sailboat, such as beam, draft, displacement and sail area. Remember to identify and describe all sources of data used.
- Discuss the precision of the estimate for each sailboat variant's price. And use the model to explain the effect of region on listing price.
- Discuss how the modeling of the given geographic regions can be useful in Hong Kong (SAR) market.
- Find and discuss interesting and informative inferences from the data.
- Provide a one-to two-page report including a few graphics for Hong Kong (SAR) sailboat broker.

1.3 Our Work

The problem requires us to collect and process data on sailboats. And then develop a mathematical model about influencing factors of listing prices and use it to predict the listing price of sailboats in Hong Kong (SAR) market. Therefore, our work includes the following:

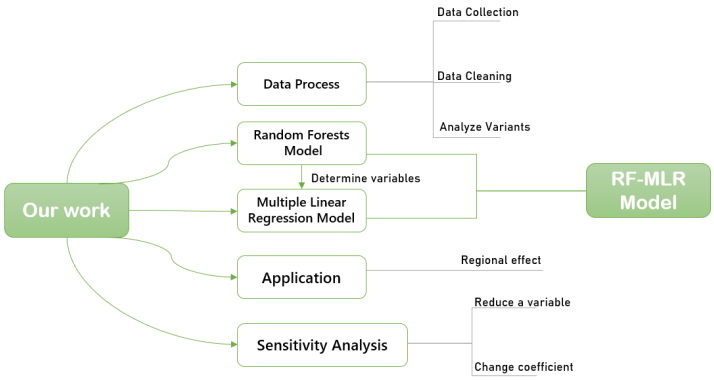


Figure 1: Random forest algorithm framework

2 Assumptions and Justifications

Assumption : The model constructed in this paper does not take into account the actual transaction taxes, commissions, agency fees, etc. We only consider the impact of the hull itself and the economy and consumption level of the transaction place on the listing price.

Justification : Taxes, commissions and agency fees are related to the listing price and the policies of the intermediary company. Compared with the listing price, the values are small and fluctuate greatly. For the sake of simplification, they are not considered for the moment.

Assumption : The model constructed in this paper does not consider the impact of COVID-19 on the listing price.

Justification : Considering that the data given in the attachment is December 2020, at which time the covid-19 epidemic has been alleviated globally, the impact of the COVID-19 on the listing price is not taken into account.

Assumption : Per capita GDP and per capita consumption expenditure are selected as indicators to describe the regional economic level.

Justification : Among the given regions and countries, there are many islands with small area and small population, as well as countries with large area and large population such as France and Mexico. If the total GDP is used to reflect the economic environment, it will lead to large errors. Therefore, it is more accurate to use per capita to reflect the regional economic environment.

Assumption : Make is independent of other variables.

Justification : The difference between makes is mainly reflected in the make effect, which directly affects the price. It has nothing to do with the specifications of the sailboat itself, for example: length, beam, sail area, draft, displacement, etc.

3 Notations

Notations are shown in the table.

Variable name	Description	Unit
Make	the manufacture of the sailboat	\
beam	the width of a boat at its widest point	ft
displacement	the weight of the volume of water displaced by a boat	kg
draft	the min depth of the water required to boat	ft
fuel	the tank capacity of the sailboat	L
lengthft	the length of the boat in feet	ft
sail	the total surface area of the sails when fully raised	ft^2
Year	the year the boat was manufactured	\
GDPP	Gross Domestic Product per capita	billion dollar
PC	Per capita consumption expenditure	dollar
TRV	regional tourism income in the past five years	million dollar

4 Model Preparation

The listing price of a sailboat may depend on the qualities of the hull, its age and the region in which it is located.

Hull characteristics include brand, model and related parameters. Brand effect will affect the price, model and parameters and determine the production and use of sailboats.It will also affect the price to a large extent. The sale of second-hand sailboats also takes into account the value of wear and tear on the sailboat, which is related to longevity and therefore the year in which the sailboat was produced. At the same time, regional economic level and market environment will affect the supply and demand relationship and purchasing power, as well as the listed price of sailboats, so it should also be taken into account.

4.1 Data preparation

(1)Data acquisition

Data are mainly obtained by retrieving and downloading relevant files from the official website and Python crawler. The data in this paper mainly comes from the following websites.

Data	Website
Listing price of Hong Kong sailboats	www.boats.com www.simpsonmarine.com
Hull data of sailing ships	www.sailboatdata.com www.boat-specs.com www.sailmagazine.com
Regional economic data	www.ceicdata.com www.databank.worldbank.org www.apps.bea.gov www.ivanstat.com www.censtatd.gov.hk

In the process of obtaining the listing price of sailboats in Hong Kong, we searched the website and set the keywords as ”Hong Kong” and ”used” to obtain the listing price data set of second-hand sailboats

in Hong Kong in the recent 15 years, and then selected specific brands and models and summarized them into a table.

The ship hull data can be obtained mainly by entering the model and obtaining relevant data on the official website.Regional economic indicators can be accessed directly from the World Bank, the government Statistics website, or other reliable databases such as CEIC.

(2)Data process

Step1: Complete the required data type

There are many indicators used to measure the hull, such as sail area/displacement (used to measure the power of the boat), boat displacement/length (used to measure the minimum power needed to drive the boat to its nominal hull speed) and so on.Add new categories to the existing table by combining the data obtained from the various websites and the analysis in 4.1

Step2: Complete missing values of individual data

For hull data of different brands of sailboats and catamarans, use Excel to retrieve missing items. If more than three data are missing, the data will be deleted. The rest of the data was filled in using the average of similar hull data of the same make and variant.

For variables describing the economic circumstances of different regions, Cork is located in Ireland, so Cork is filled in with Irish data; The geographical proximity of Gibraltar to Spain made it up with Spanish data; In recent five years, tourism income data in the Caribbean region is missing a lot. Considering that this region as a whole is a tourist resort, the services it provides, the scenery it enjoys and the projects it experiences are also similar, so a random value between the maximum value and the minimum value is used to fill in. For other missing values, the mean value of the geographical region is used to fill in.

4.2 Variable selection

It is impossible to determine the influence size of various possible factors only through theoretical analysis, so we decide to use one-way analysis of variance to analyze possible factors one by one, so as to determine variables according to the size of influencing factors.[1]

Taking the monohulled sailboat as an example, the single factor analysis of variance is carried out for the make. The prerequisite assumption is equal variance, that is, it is assumed that different brands have the same level of influence on the listing price. Import all the data into the data editor in Stata software, and then use one-way analysis of variance to test whether different brands have the same influence factors on the price.[2]

The graph shows a list of the results of the analysis of variance.As reflected in the table, the horizontal term's deviation square and SSA are 2.4528e+13, 69 degrees of freedom, 3.5548e+11, the error term's deviation square sum is 2.6694e+13, 2276 degrees of freedom, the intra-group mean square SSE is 1.1729e+19, the statistic F is 30.31, and the p value is 0.000.

The result of variance analysis of monohulled sailboats is chi2(65) = 515.7459, Prob>chi2 = 0.000, which means that small probability events occur, and the null hypothesis needs to be rejected, that is, different makes in sailboats have significant differences in the influence level of listing price.

Hull specifications include a range of specifications for the hull, including year of manufacture, beam, draft, displacement, sail area, hull material, engine hours, sleeping capacity, headroom, etc. These indicators are closely related to the manufacture and use of sailboats, so whether the impact of these factors on the listed price is significant should be considered separately.The method is the same as above. Single factor contrast analysis is used in Stata software respectively to test whether specific

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	2.4528e+13	69	3.5548e+11	30.31	0.0000
Within groups	2.6694e+13	2276	1.1729e+10		
Total	5.1222e+13	2345	2.1843e+10		

Bartlett's test for equal variances: chi2(65) = 515.7459 Prob>chi2 = 0.000

Figure 2: One-way analysis of variance of Makes

hull indicators have the same influence level on the listing price. The results are shown in the table below.

Hull index	Results of one-way analysis of variance		difference
lengthft	chi2(19) = 922.6559	Prob>chi2 = 0.000	Obvious
beam	chi2(150) = 1.8e+03	Prob>chi2 = 0.000	Obvious
displacement	chi2(182) = 1.6e+03	Prob>chi2 = 0.000	Obvious
draft	chi2(116) = 1.9e+03	Prob>chi2 = 0.000	Obvious
sail	chi2(221) = 1.4e+03	Prob>chi2 = 0.000	Obvious
fuel	chi2(104) = 1.2e+03	Prob>chi2 = 0.000	Obvious
headroom	chi2(20) = 116.0215	Prob>chi2 = 0.000	Obvious
Year	chi2(14) = 381.8227	Prob>chi2 = 0.000	Obvious
rigging	chi2(3) = 1.7139	Prob>chi2 = 0.711	Not obvious
hull material	chi2(3) = 0.7683	Prob>chi2 = 0.842	Not obvious
engine hours	chi2(3) = 0.2587	Prob>chi2 = 0.676	Not obvious
sleeping capacity	chi2(3) = 0.9741	Prob>chi2 = 0.815	Not obvious
electronics	chi2(3) = 2.7659	Prob>chi2 = 0.743	Not obvious
mast	chi2(3) = 0.4656	Prob>chi2 = 0.926	Not obvious

Different countries have different economic levels, GDPP and per capita consumption, which directly affects the market environment and people's purchasing power. One-way variance analysis was used to determine whether different brands of sailboats had the same influence level on prices. The results of variance analysis showed that chi2(60) = 602.1072, Prob>chi2 = 0.000, indicating significant differences.

5 Modeling

5.1 Idea analysis

Based on the results of the data preprocessing section, we know that some features are strongly correlated and numerous. In order to better explain the listing price of second-hand sailboats, it is necessary to screen the characteristics to achieve dimensionality reduction without losing most of the

information. Considering that there are both numerical and categorical variables in the 11 features, and the target listing price to be predicted is a continuous variable, which is suitable for the random forest model, the random forest algorithm is used to fit the data. The importance of each feature is then sorted and the top third of features that have the greatest impact on the listing price are finally selected and performed with SPSS.

5.2 Random forest model establishment

5.2.1 Model principle

(1)Decision Tree

Decision tree algorithms abstract the structure of real-life trees and construct an algorithm for decision-making. The tree structure generally contains parts such as stumps, branches and leaves, which correspond to abstract concepts one-to-one, It can be thought of as the stump as the complete set of samples; Where the branches fork are internal nodes, which are manifested as splitting different decision-making paths according to certain characteristics through internal nodes; The leaves are the final output. A schematic representation of a decision path from the stump to the leaf node is shown.

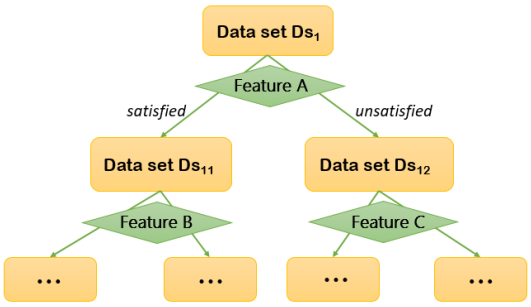


Figure 3: Decision tree schematic

(2)Random forest

Obviously, it is highly likely that a single decision tree will have some random properties, which will bring instability to the results. To solve this problem, the natural idea is to use the results of multiple decision trees for combined decisions to effectively reduce the bias and variance of the prediction. The basic principle of random forest is to combine the results of multiple independent decision trees for prediction, generally giving the same weight to the measurement results of each decision tree in the random forest.[3] For the regression task involved in this question, the final prediction result is the average of multiple decision trees in the forest. Its model structure is shown in the figure.

5.2.2 Model building

In the decision-making process of a decision tree, the choice of attributes is critical each time. As the division process progresses, in general, the ”purity” of the samples contained in the branch nodes increases. In this problem, the division principle we use is the minimum mean squared error (MSE), That is, for arbitrarily divided feature A, the data sets D1 and D2 divided on both sides of

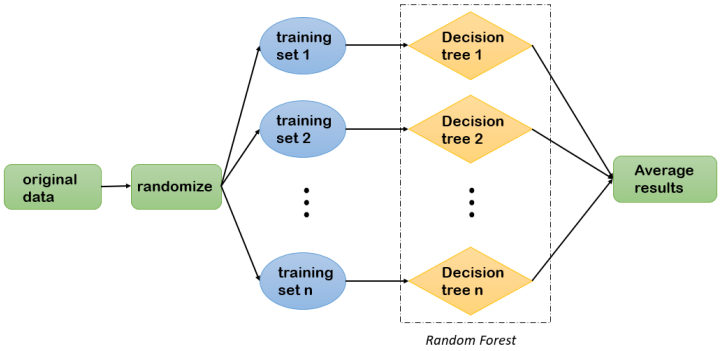


Figure 4: Random forest algorithm framework

the corresponding arbitrary division point s are found to minimize the mean square deviation of the respective sets of D1 and D2, and the corresponding feature and feature value division points are the smallest sum of the mean square deviations of D1 and D2.[4] The expression is:

$$min_{(A,s)}\left[min_{(c_1)}\sum_{x_i\in D_1(A,s)}(y_i-c_1)^2+min_{(c_2)}\sum_{x_i\in D_2(A,s)}(y_i-c_2)^2\right]$$

(c1 is the sample output mean of dataset D1, and c2 is the sample output mean of dataset D2.)

Select a different number of leaves in the random forest model, set a larger number of trees, and compare the MSE of the training results. The smaller the effect, the better. Then determine the optimal number of leaves and the minimum number of trees that can make the results stable, so as to reduce the amount of computation as much as possible without affecting the final effect, and shorten the time required for model training. Then, the pre-processed data set is randomly divided into the training set and the test set at a ratio of 1:4, and the training set is substituted into the model. Then, the trained random forest model is applied to the test set, and the correlation coefficient r is selected to reflect the accuracy of random forest fitting estimation. The calculation formula is as follows:

$$r(TrueV,PrdcV)=\frac{Cov(TrueV,PrdcV)}{\sqrt{Var[TrueV]Var[PrdcV]}}$$

After training the Random Forest model, we needed to use it to find the top third of features that had the most impact on the listing price of a used sailboat. In order to measure the influence degree of the feature, namely the importance, for each feature, the value of the feature is arranged among each observation value in the data set, and how bad the MSE becomes after the arrangement. Average the MSE increments of all trees and divide by the standard deviation. The bigger the result, the more important the variable.

$$ImPt=\frac{\sum_{i=1}^{ntree}\triangle MSE}{\sqrt{Var[x]}}\text{ ,for each feature}$$

Where ntree represents the best tree tree determined earlier, and x is the set of MSE increments of each tree.

5.2.3 Model calculation

The treebaggar function in MATLAB is used to train and predict the random forest model. Considering the different numbers of propellers, fuel tanks and other equipment of monohulled sailboats and

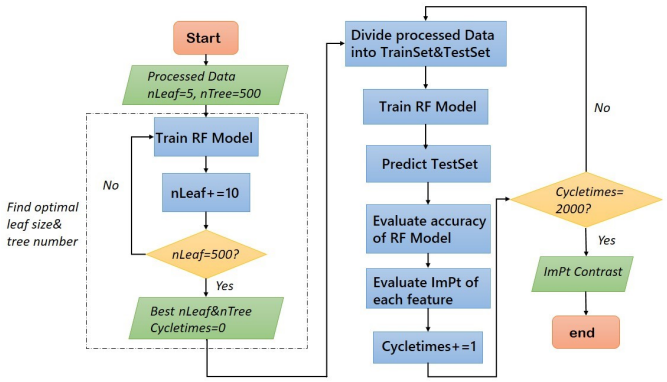


Figure 5: Random forest algorithm framework

catamarans, two random forest regression models are trained for monohulled sailboats and catamarans respectively. The calculation results are as follows:

As shown in the figure, the comparison between the real value and the predicted value of monohulled sailboats and catamarans is shown respectively. It can be seen that the predicted value of the two models is in good agreement with the real value in most cases. In addition, the correlation coefficient r values of the random forest model trained by monohulled sailboats and catamarans are 0.8975 and 0.8497, respectively, which further indicates that the model has a good fitting effect from a quantitative perspective.

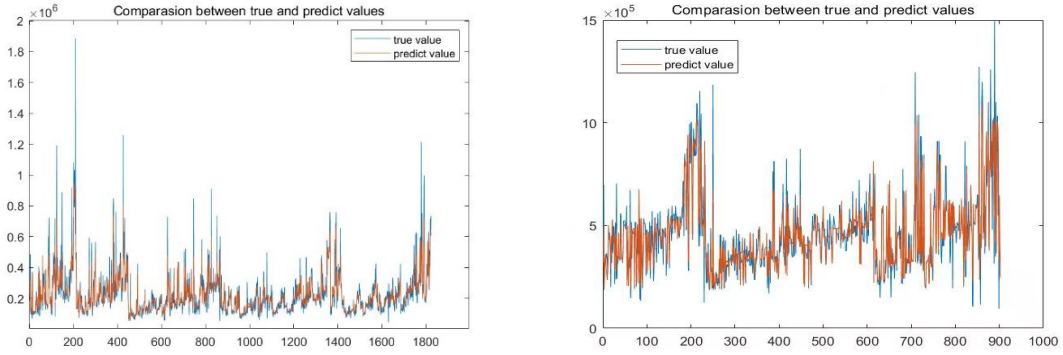


Figure 6: comparison between the real value and the predicted value

As shown in the figure, the curves of both models are lowest when the number of leaves is 5, which means the MSE is the smallest and the effect is the best. Therefore, the number of leaves is 5 for both models. When the random forest model training of monohulled sailboats and catamarans tends to be stable, the number of decision trees is about 200 and 70 respectively. For the sake of safety, the number of decision trees is increased slightly to avoid getting results before convergence. Therefore, 205 decision trees and 75 decision trees were selected for the random forest model of monohulled sailboats and catamarans in the subsequent training.

As shown in the figure, for monohulled sailboats, the top four influential factors are year of manufacture(year), Make, 2020 per capita consumption expenditure (PC) in sales region and displacement.

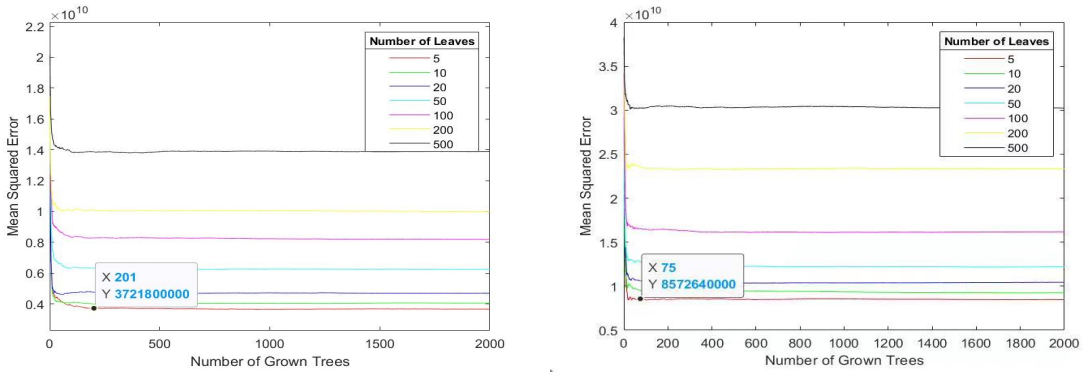


Figure 7: the curves of both models

For catamarans, the difference between the fourth and fifth most influential factor is small, so the top five are chosen for later models. They are year of sailboat manufacture (year), sailboat length (lengthft), Make, beam and the tourism income in the last five years (TRV).

In a practical sense, people prefer new items, even on the second-hand market. In addition to the impact of collectible value, the less old the item, the lower the discount. In addition, the displacement and beam width of a sailboat can reflect the size and magnitude of the sailboat to some extent. The larger the size of the sailboat is generally the higher the manufacturing cost, so the more expensive the price, which is in line with people's intuitive cognition and life experience. As for Make, according to the hypothesis in the beginning, we believe that this feature will mainly affect the listing price of sailboats in terms of brand effect, manufacturing process and manufacturing materials. In real life, we also naturally know that the quality of the same product produced by different manufacturers is different, so this feature is of high importance and easy to be recognized. All these indicate that the random forest model we trained can reflect the reality better.

Among the important influence characteristics of monohulled sailboats and catamarans, the per capita consumption expenditure of the sales region in 2020 has a greater influence on the listing price of monohulled sailboats. One guess we can make is that catamarans may not be as widely used and popular as monohulled sailboats, and may only be popular in certain small areas. Therefore, the consumption power of the sale area of second-hand catamarans has a lower impact on the listing price of catamarans than on the listing price ofmonohulled sailboats.

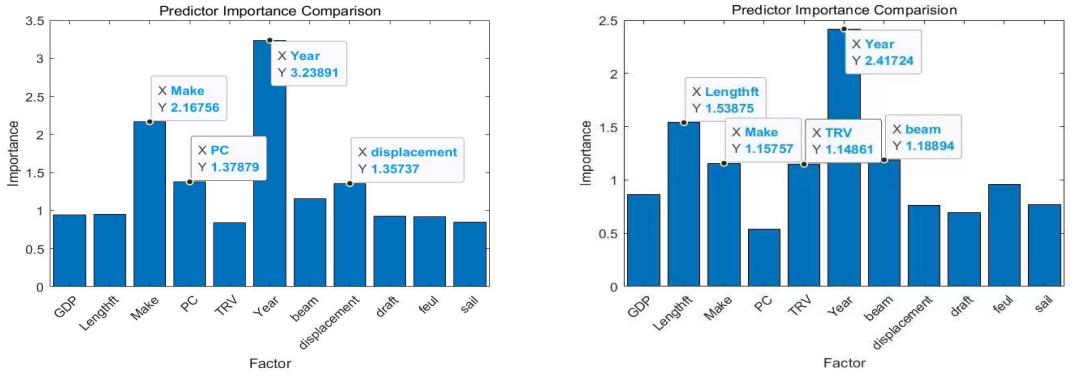


Figure 8: the precision of our model's estimate for each sailboat variant's price

5.3 Multiple linear regression model prediction

5.3.1 Model building

According to Numeric Predictor Association Estimates, the four or five factors that have the greatest influence on mono/catamaran in random forest regression model can be regarded as independent of each other.All the conditions of establishing multiple regression linear model are satisfied. We tried to use SPSS to establish a simpler multiple linear regression model with major influencial factors as independent variables without cross terms:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \mu, (n = 4, 5)$$

In order to establish multiple linear regression model, it is necessary to process the categorical variable Make first. Make variable was converted into dummy variable, one of which was selected as the reference variable, and all remaining variables were analyzed with the other variables by using the linear regression analysis function of SPSS. The coefficient and R of each model variable are obtained.

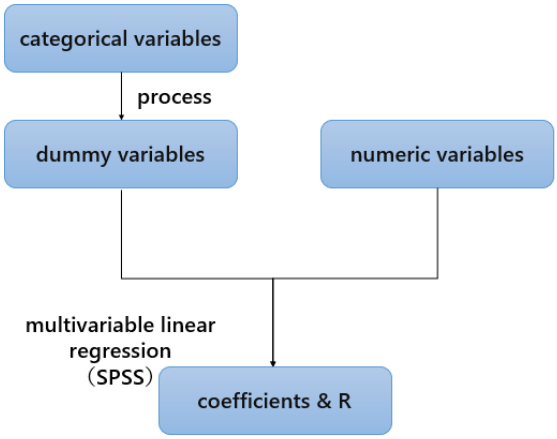


Figure 9: The flow chart

For the monohulled sailboats, we select Bavaria, which has a large brand and many variants, as the reference variable to get the regression model. (μ is related to variable Make,R_2=0.86)

$$y = displacement * 22.739 + 12947.703 * Year + 1.856 * PC - 26169390 + \mu$$

For catamarans, we randomly select HH-catamarans as the reference variable to get the regression model.(μ is related to variable Make,R_2=0.82)

$$y = -48402929.1 + 31552.654 * lengthft + 9938.56 * beam + 23453.05 * Year + 0.836 * TRV + \mu$$

It can be seen from the R value that the fitting effect of the obtained multiple linear regression model is better.

5.3.2 Precision of each variant

Set the error rate as w (parameter), and use the obtained multivariate linear formula to predict the price *yi* (parameter) of each variation of monohulled sailboats and catamarans.The average error rate can be calculated by the following formula.

$$w = \frac{1}{n} \sum \frac{y_i - y}{y}$$

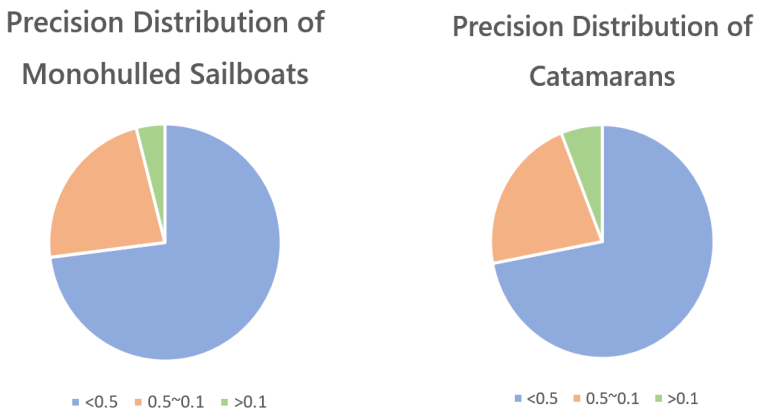
The following is part of the data, and the rest is in the appendix.

Deviation of Prediction			
variant	w	variant	w
44	0.10	First 44.7	0.02
AZZURO 53	0.09	First 45	0.12
44	0.09	First 47.7	0.01
CIGALE 16	0.07	Moorings 42.3	0.04
Ovni 395	0.15	Cyclades 50.5	0.31
54	0.03	First 45	0.02
40 RC	0.02	115	0.12
40	0.27	121	0.03
46	0.08	375	0.08

Deviation of Prediction			
variant	w	variant	w
44GS	0.12	DH550	0.22
44i	0.29	Astrea 42	0.09
4.1	0.01	Athena 38	0.04
4.3	0.15	Bahia 46	0.11
4.5	0.03	Helia 44	0.01
385	0.12	Lavezzi 40	0.18
415	0.42	Lipari 41	0.10
47 Ocean Class	0.05	SABA 50	0.01
47	0.07	Salina 48	0.08

Figure 10: The left table is monohulled sailboats and the right table is catamarans

The following figure shows the distribution of error rate within the same hull type



By comparing the error rate of different hull types, it can be found that the more data used for fitting, the more accurate the price predicted by the model. By comparing the accuracy of different variants of the same hull type, it can be found that the error rate predicted by the model is lower when there are more data of the same variant, while the error rate is larger when there are fewer variants (especially only one). We believe that the reasons for the above phenomena are as follows: when there is a small amount of variant data, the factors we regard as having a low influence but not being used as the independent variables of the model may have a greater impact on the data, and thus produce errors, which are accidental.

We also compared the error rate of the multiple linear regression model and the random forest model, and obtained the following scatter plot.

As can be seen from the figure, the error rate of random forest model is more concentrated and close to 0, and the prediction accuracy is higher.

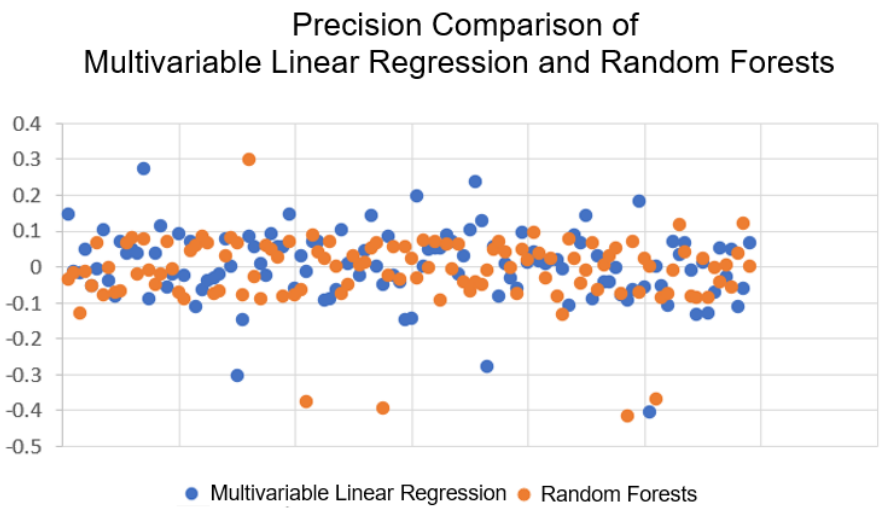


Figure 11: The scatter plot

6 An explanation of regional effects using our Model

6.1 Global regional effect and interpretation

(1)Monohulled sailboats
Step1: Regional effect

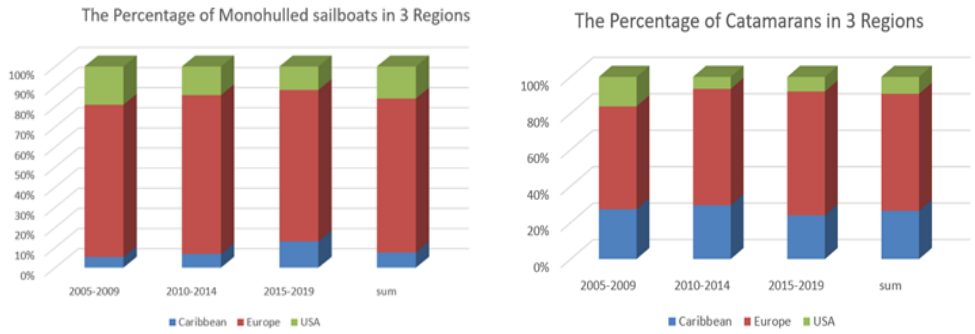


Figure 12: the data of the Year of manufacture of sailboats

As shown in the figure above, the data of the Year of manufacture of sailboats are evenly distributed among regions. Therefore, when finding out whether there is regional effect in the listing price of second-hand sailboats, we do not need to consider the most influential feature: the year of manufacture. Calculate the average listing price of the Caribbean, Europe and USA monohulled sailboats in the three geographical regions by using the EXCEL average formula.

Take the average listing price of all second-hand monohulled sailboats as the standard and make comparison by using the percentage of deviation, that is, calculate:

$$Deviation = \frac{(RegionAver - TotalAver)}{TotalAver} * 100\%$$

This formula can effectively avoid the influence of the absolute value of data on the comparison. The calculation results are shown in the following table:

Averaged Listing Price of Total/dollar	Geographic Region	Averaged Listing Price/dollar	Deviation
226928	Caribbean	193691	-14.65%
	Europe	226727	-0.09%
	USA	243642	7.37%

Among the three regions, only Europe has a mean deviation percentage within 5%, which indicates that the listing price of second-hand monomer sailboats has a relatively significant regional effect, that is, the listing price level varies greatly among different regions.

Specifically, the price level of USA>Europe>Caribbean, and the listing price of monohulled sailboats in USA is higher than the world average level, the listing price of monohulled sailboats in Europe is the same as the world average level, and the listing price of monohulled sailboats in Caribbean is lower than the world average level.

Step2: Explain using our Model

According to the correlation analysis in 4, per capita personal consumption expenditure (defined as PC), per capita GDP (defined as GDPP) and regional tourism income (defined as TRV) of a region all have a significant impact on the listing price of second-hand sailing boats. According to the ranking diagram of the importance of variables obtained by the random forest model of single ship in 5.3, the influence degree of these three features is ranked as PC>GDPP>TRV.

Similarly, the average values of these three features in the Caribbean, Europe and USA are calculated, and compared with the average values of all data, it can be found that the regional differences are consistent with the listing price: That is, USA >Europe> Caribbean, and compared with the overall average, the levels of the three regions are respectively low, flat and high.These three characteristics describe the regional economic environment from the aspects of consumption, production and tourism respectively, so it can be believed that the reason why listing price shows regional effect is related to the difference of economic conditions among regions.

	PC/dollar	GDPP/billion dollar	TRV /million dollar
Average of Caribbean	15903	134	1476.084
Average of Europe	20625.671	1071.353	18382.635
Average of USA	44504.044	1429.561	51242.617
Average of total	24072.624	1057.693	22338.965

According to the model in 5.3, when other conditions are completely consistent, the listing price of second-hand sailboats presents a positive linear correlation with the regional per capita consumption expenditure in 2020, with a coefficient of 1.856 (when both units are in US dollars).

In other words, the higher the regional per capita consumption expenditure, the listing price of second-hand monomer sailboats will also rise. It can be seen from the data in the second column of the above table that the regional average PC presents a relationship of USA >Total ≈ Europe > Caribbean, so the listing price of second-hand mono sailboats also presents the same size relationship.

(2)Catamarans

Step1: Regional effect

Similarly, the average listing price of catamarans in different regions is compared with the overall average. The results are shown in the table below:

Averaged Listing Price of Total/dollar	Geographic Region	Averaged Listing Price/dollar	Deviation
454532	Caribbean	421572	-7.26%
	Europe	460098	1.22%
	USA	510065	12.21%

Although the specific values are different, the listing price of catamaran shows the same regional effect as that of mono ship: USA>Total≈Europe> Caribbean.

Step2: Explain using our Model

Similar to the model explanation of monomer ships, the correlation analysis in 3 shows that GDPP, PC and TRV all have significant influence on the listing price of second-hand sailboats. According to the ranking diagram of the importance of variables obtained by the Catamaran Random forest model in 5.3, the ranking of the influence degree of these three features is TRV > GDPP>PC, and the importance of TRV is in the top 50% of all variables.

On the other hand, according to the model in 4.3, when other conditions are completely consistent, the listing price of second-hand monohulled sailboats presents a positive linear correlation with regional per capita consumption expenditure in 2020, with a coefficient of 0.836 (when both units are in US dollars).The mean value of TRV in different regions was calculated, and the results were as follows: $1333, $16,680, $54,504. The average TRV of all data is $16,132. acctually has the same comparison result with the listing price between different regions, which indicates that both the linear regression model and the stochastic forest model established previously can explain the regional effect of listing price well.

6.2 Regional effect consistency test for different variants

(1)Monohulled sailboats

For each variation, ten sets of data are captured in the table below to show the average of listed prices in different regions, the average of the population, and whether they are consistent with regional effects. The rest of the data is in the appendix.

(2)Catamarans

The number of variant groups in which the monohull maintained consistency with the regional effect was 14, and the number of variant groups in which the monohull did not maintain consistency was 21. The number of catamaran variants consistent with regional effects was 6, and the number of catamaran variants not consistent with regional effects was 15.

6.3 Meaning

From the statistical point of view, it can be seen that the region has a significant impact on the listing price of second-hand monohulled sailboats. In a practical sense, when selling and buying second-hand monohulled sailboats, both sides of the transaction should fully take into account the regional factors. It is not sensible to compare the lowest price across regions, but it should also be understood that

Variants	Caribbean	Europe	USA	Total	Consistence
38	105000	199259	110173.3	125354.5	N
39	256165.5	198368.2	149950	200452.4	N
50	251591	356418.8	298724	339833.6	N
54	495000	697156.9	767000	694033.4	Y
380	99950	233156	109000	147368.7	N
385	72908	164182	215725	206106	Y
415	187166.7	169292.4	244000	182828.3	N
445	166566	195056	306050	213456	Y
455	279450	241636	329500	247086	N
465	330000	319756	450000	366585.3	N

Variants	Caribbean	Europe	USA	Total	Consistence
39	300443.5	338045.7	319250	325513.4	N
40	320142.7	346144.5	227250	327062.6	N
42	522976	484933.7	567000	495061.4	N
44	336499.9	320869	531666.7	371085.8	N
46	434556	339514	438000	426085	N
400	272655	297176	429000	294957	N
420	321079	333876	352129	334326	Y
440	333810	345345	353878	343884	Y
450	498464	470390	484975	476787	N
500	487883	518082	395000	504533	N

regional effects are a holistic estimate of all sailboat sales in the region and do not necessarily apply to specific variants.

7 Model Applying in Hong Kong

7.1 How to apply our models

Regional effect refers to the overall difference in the listing price of second-hand sailboats in different geographical regions, which has nothing to do with the nature of the hull, such as manufacturer, specifications, equipment intelligence degree and other factors, but more to do with the economic and social conditions and natural geographical conditions of the specific region. According to this idea, in the multiple linear regression model we established before, there are both independent variables that can measure the nature of the ship itself and indicators that can represent regional differences. At this time, only the variables related to the region are removed, and the deleted model only represents the influence of different hull characteristics on the listing price of the ship, and then the real value is compared with the value obtained by the new model. The influence of the region on the listing price can be obtained.

7.2 Hong Kong second-hand sailboat data search

Relevant variants, parameters, listing price and other data of monohulled sailboats and catamarans are found respectively on the Hong Kong second-hand sailing websites, such as www.boat-specs.com. Then manufacturers and variants that coincide with the original documents are selected as the observation set. The statistical table is as follows:

(1)Monohulled sailboats

Make	Variant	Length (ft)	Listing Price (USD)	Year	Beam	Draft	Displacement	Sail area	Fuel
Beneteau	Oceanis 38	37	165000	2014	13.09	6.8224	6580	707.19	130
Beneteau	Oceanis 51.1	51	545315	2018	15.744	7.7408	13930	1528.49	200
Beneteau	Sense 43	43	220000	2012	14.53	6.724	11800	1068	401
Beneteau	Oceanis 46.1	47	555747	2023	14.76	8.692	10697	1231.4	200
Beneteau	Oceanis Clipper 473	47	165615	2003	14.137	6.888	11010	1226.02	237
Beneteau	43	43	139500	2007	12.825	5.084	6280	678.024	201
Nautor	Swan 54	54	1770000	2019	15.7	8	26000	1538	495
Hanse	400	40	$94,276	2007	13.4	6.8	8400	1137	140

(2)Catamarans

Make	Variant	Length (ft)	Listing Price (USD)	Year	Beam	Draft	Displacement	Sail area	fuel
Fountaine Pajot	Saona 47	46	1163395	2020	25.3	4.2	807	13800	939
Fountaine Pajot	Saona 47	46	1066920	2020	25.3	4.2	807	13800	939
Lagoon	450	46	685000	2017	25.62	4.27	15003	1398	1041
Lagoon	450	46	560000	2016	25.62	4.27	15003	1398	1041
Lagoon	450	46	538500	2014	25.62	4.27	15003	1398	1041
Lagoon	450	46	535066	2013	25.62	4.27	15003	1398	1041
Lagoon	380	38	280272	2009	21.42	3.75	17600	506	200
Lagoon	39	39	318491	2012	22.28	4.17	11672	817	401

7.3 Regional effect in Hong Kong

(1)Model building

According to the ideas in 7.1, we can simply transform the previous multiple linear regression model, so as to achieve the effect of predicting the regional effect of second-hand sailing boats in Hong Kong.The mathematical model is established as follows

Define the impact of the region on the listing price:

(1)Monohulled sailboats

$$ReEffc = TrueV - (displacement * 22.739 + 12947.703 * Year - 26169390 + \mu)$$

(2)Catamarans

$$ReEffc = TrueV - (-48402929.1 + 31552.654 * lengthft + 9938.56 * beam + 23453.05 * Year + \mu)$$

Considering that a single result is not representative, and the listing price of sailboats will be affected by some random factors, in order to make the model more representative, the average value of all observations in the geographic area to be evaluated can be taken as the basis for the evaluation of regional effects. The formula is as follows:(Q is the whole observation of second-hand sailboats in the area to be evaluated)

$$RE = \frac{\sum_{a \in Q} ReEffc(a)}{|Q|}$$

If the value of RE is greater than 0, it indicates that the regional effect will increase the listing price on the basis of the sale of sailboats themselves in the region. If the value of RE is approximately 0, it means that the price of sailboats in the region is not particularly affected, but is only related to the characteristics of sailboats. If the value of RE is less than 0, the regional effect will reduce the listing price on the basis of the sailboat itself.

(2)Apply the model to Hong Kong

The above model is used to calculate the sorted data in 7.2, and the results are as follows:

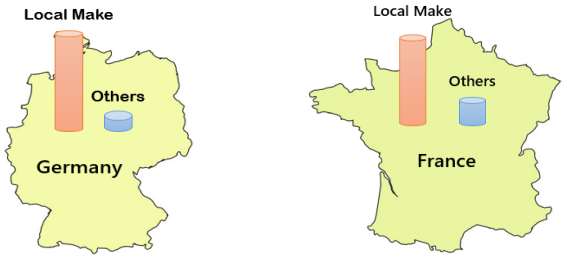
RE	Caribbean	Europe	USA	Hong Kong
Monohulled Sailboats	-125800	269.23	9830.22	221286.6
Catamarans	-99781.49	4275.46	96128.71	169466.3

The results of this table are consistent with the analysis of regional effects in 5, that is, the overall listing price level of sailboats in the Caribbean is lower than the average level, and the listing price level of sailboats in the United States is higher than that in Europe, which indicates the correctness and effectiveness of the model.Compared with the data of Hong Kong, the listing price of sailboats here is higher than that of the other three regions. The regional effects of monohulled ships and catamarans in Hong Kong are consistent.

8 Interesting discovery

8.1 Trading tendency

(1)Brand orientation

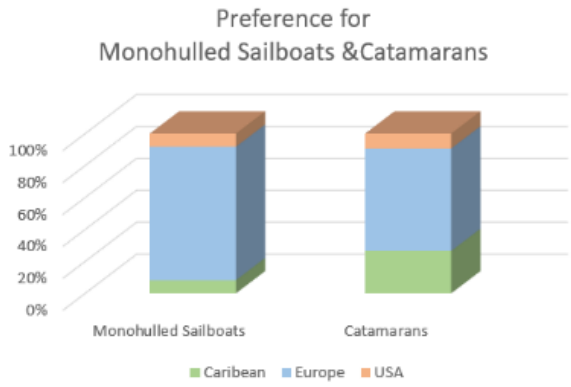


We find that 83.2% of the transactions in Germany are local German brands, such as Hanse, Bavaria, etc. and these brands are sold to all regions. This shows that,on one hand, sailboats made in Germany have strong competitiveness in the world market, which is consistent with the larger brand effect we get in the multiple linear regression model.

On the other hand, it also reflects to some extent that local people are more inclined to buy their own brands of sailboats when they choose to buy second-hand sailboats. For those who own these German brands, they can consider higher pricing and marketing, while non-German sailboats can consider selling outside Germany. France, home to many famous brands, has a similar regional bias, but it is slightly lower (75%) than Germany. The European market basically circulates internally and sells outwards.

However, there is no obvious regional bias in the United States, which also has well-known brands. Transactions in the United States involve various brands, and American brands are also sold to various regions.

(2)Hull inclination



Regionally, there is little difference in hull preference between Europe and the United States, while the Caribbean tends to buy catamarans. Therefore, the catamaran seller can deal for the Caribbean region, and this type of sailboat is easier to sell.

8.2 Correlation between region and displacement

To some extent, displacement can reflect the size and bearing capacity of the hull. After testing, we can use SPSS to conduct correlation analysis of region and displacement, and the following results are obtained:

ANOVA

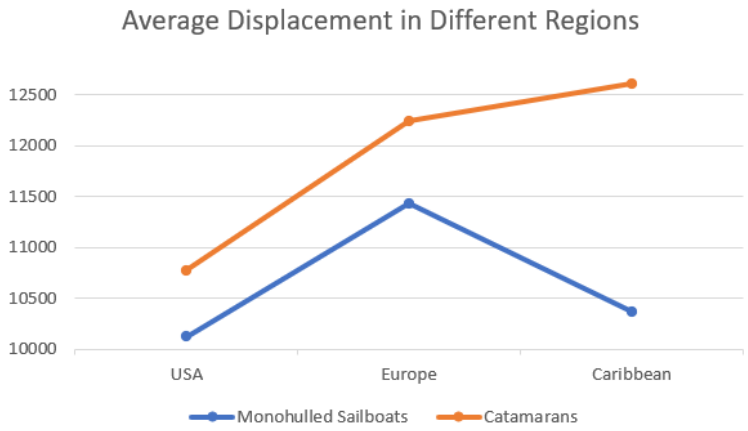
		Sum of Squares	df	Mean square	F	significance
displacement * Country/Region/State	between-group	1505447312.174	71	21203483.270	2.237	0.000
	with-group	20921479459.001	2207	9479601.024		
	Total	22426926771.175	2278			

Measurement of Association

	Eta	Eta^2
Displacement* Country/Region/State	0.259	0.067

It can be seen from the table that significance<0.05, there is a correlation between region and displacement. Now consider the degree of correlation. Eta2>0.06, with moderate correlation.

In terms of specific regions, it can be seen from the statistical results and tables that the displacement of the two types of sailboats traded in the United States is generally lower than that of the other two regions, and the displacement of catamarans increases less than that of monohulled sailboats. The

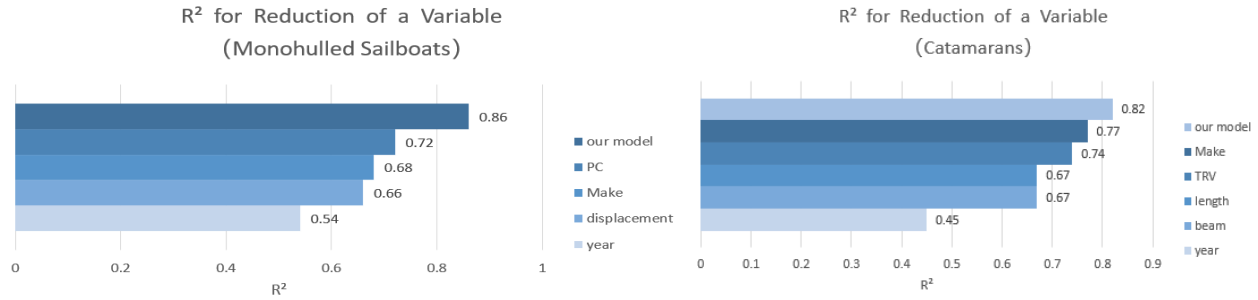


displacement of catamarans traded in the Caribbean is much higher than that of monomers, especially in the Virgin Islands, where charter trips explore uninhabited islands, where catamarans can be more than twice that of monomers. This may be due to the fact that the developed tourism industry, which is mainly based on sea traffic, has demanded more ships with larger displacement to increase economic efficiency.

9 Conclusion

9.1 Sensitivity Analysis

In our multiple regression model, we choose the factors of high importance in the random forest model as our independent variables for regression. Now we consider the fitting effect of the model after removing one of the independent variables respectively. The R^2 result obtained is entered as follows:



For both hulls, if you remove either of the independent variables, the R_2 decreases and the explanatory power of the model decreases. It can be seen from the results that the degree of influence of the removed variables on the fitting effect is roughly the same as the importance of the variables in the random forest.

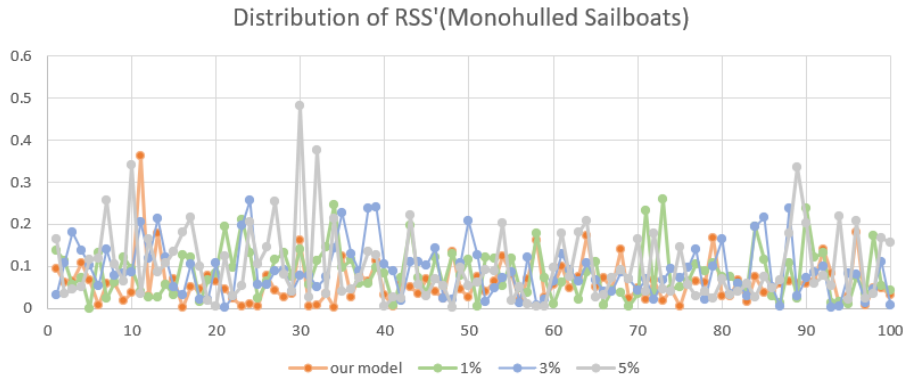
Then consider changing the regression coefficient according to the proportion of 1%, 3% and 5%, randomly extract 100 data into the changed model, and calculate the residual sum of squares.

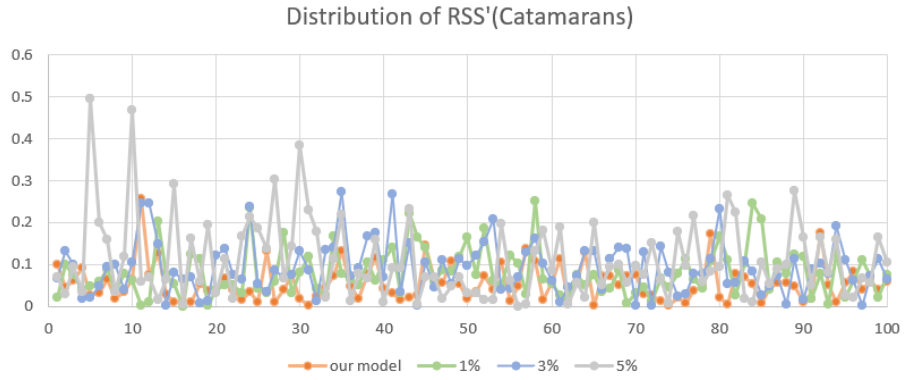
$$RSS = \sum_1^n (y_i - \hat{y}_i)^2$$

$$\text{Let } RSS' = \frac{RSS - min(RSS)}{max(RSS) - min(RSS)}$$

The closer the RSS ’is to 0, the better the fit. Take the coefficient of year as an example to make the scatter diagram of RSS ’

As can be seen from the figure, with the increase of coefficient variation, the dispersion degree and deviation of RSS ’increase greatly,and our model is sensitive to coefficient variation.





9.2 Evaluation of models

• Strengths

Our model takes the least coefficients into consideration as well as keeps most of effective information. According to the sensitivity analysis, when moving away one coefficient, the R2 reduces greatly.

Our model has a perfect complementarity.It can be applied into the situation of big data, for we have trained a Random Forest Model so that we can input the data set and get the results rapidly. However, because of the ”Black Box” feature of Random Forest,it lacks detailed mathematical explanation to its result.In comparison, the Multiple Linear Regression model has a solid math foundation, which can be accepted more easily. But its accuracy of prediction and the ability to deal big data is worse than the random forest model.We use both models so that we can combine their advantages and avoid disadvantages.

• Possible improvements

The model simply uses regional economic indicators to measureand ignores humanistic factor.For example,even if the two cities are both near the sea, there are differences in the way to travel.Some

cities' resident generally prefer sailboat for a short trip to the seabut there are also cities where the majority of people only want to enjoy the pleasure of speed.

The model describes regional effect roughly, just thinking of it in terms of the average listing prices of all sailboatswhich can only reflects the relative listing prices of sailboats in the area. However,regional effect should include the preference for different types of sailboats or other aspects.

• Further discussion

We can use more regression algorithms to compare which is better.For example,the artificial neutral networks is a algorithm that has wide application and bright prospect. It has the ability to carry out high precision regression and prediction.Otherwise, as for linear regression is too simple,it doesn't reflect the real world in most of times.So a nonlinear regression model with better applicability can be used.By comparing different models and learning from each other, we are able to gain a better understanding of the used sailboats market.

10 A report on the Pricing of Used sailboats

Dear broker,

We appreciate the trust you have placed on our team.In order to create a better understanding of only used sailboat market, we established a RF-MLR Model which can explain the relationship between the listing price and some features. You can also use our model to predict a price for any sailboat you want to buy/sell.Here is a brief introduction to the model and how it works, as well as some of the conclusions we draw from it.

• The RF-MLR Model

In order to better respond to your requirements, we combine the random forest model in machine learning with the classical multiple linear regression model to achieve a solid mathematical theoretical basis and efficient and accurate prediction ability.

• Random Forest Model

Random Forest Model belongs to integrated learning, a branch of machine learning. The principle is to train multiple learners through repeated sampling in place, and each learner is equivalent to an expert in a narrow field. These experts are gathered together to make decisions according to the principle of minimum relative error, so that a relatively accurate and comprehensive result can be obtained.

In our work,We selected enough factors influencing the listing price of sailboats. Then collect a lot of data through the official websites and other channels, eliminate outliers,make up for missing values,and then input the data into the Random Forest Model for training.

In this way, we have obtained a tool that can predict the listing price of a sailboat through some characteristics describing the sailboat itself and the economic and social situation of the region.The Random Forest Model not only has a highly accurate prediction ability, but also can rank the importance of the measurement indicators we input according to the degree of influence.We will explain this in detail below.

• Multiple Linear Regression Model

Generally speaking, multiple linear regression is a model that describes the correlation between one dependent variable and several independent variables. If the coefficient of the independent variable is positive, it means that the dependent variable increases with the increase of the independent variable. Otherwise, it means that the dependent variable decreases with the increase of the independent variable. Meanwhile, the greater the absolute value of the coefficient, the greater the influence of the independent variable on the dependent variable.

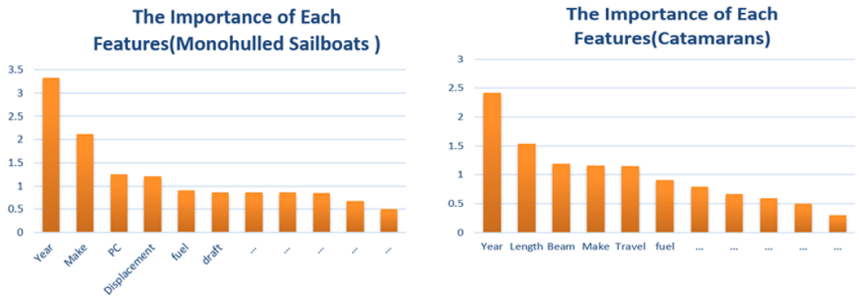
In our work, We selected the most important factors in the random forest model to carry out multiple linear regression on the listed price of sailboats,and obtained the following expression. You can use this formula to estimate the price.(μ is related to Make.)

$$y = displacement * 22.739 + 12947.703 * Year + 1.856 * PC - 26169390 + \mu$$
$$y = -48402929.1 + 31552.654 * lengthft + 9938.56 * beam + 23453.05 * Year + 0.836 * TRV + \mu$$

• The conclusions

Factors affecting listing price for monohulled sailboats, the top five are: (1) the year of the ship's manufacture, (2) Make,(3) the per capita consumption expenditure in the region of sales in 2020, (5)displacement; For catamarans, the top five are: (1) year of manufacture, (2)length of sail,(3)Make, (4)beam, and (5)average tourism revenue over the past five years in the region of sale.

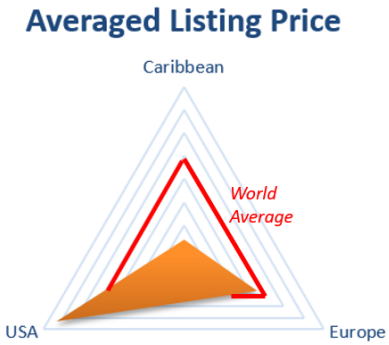
• Regional effect



(1)Global Market

In general, the average price of the United States is higher than the world average level, Europe is equal to the world level, and the average listing price of the Caribbean is the lowest among the three. It can be seen from our multiple linear regression model that the higher the per capita consumption expenditure, the higher the average price. The higher the regional tourism revenue, the higher the average price.

In addition, we also find an interesting conclusion that 83.2% of the transactions in Germany are local German brands, such as Hanse, Bavaria, etc., and these brands are sold to all regions. This indicates that, on the one hand, sailboats made in Germany are highly competitive in the world market; on the other hand, it also reflects to some extent that local people prefer to buy sailboats of their own brands when they choose to buy second-hand sailboats.



(2)Hong Kong Market

By screening the terms in the multiple linear regression model that are only related to the hull itself and calculating the difference with the real listing price of the obversation, the effect of the region on the price can be immediately obtained. Our conclusion is that the average listing price in Hong Kong is higher than that in the Caribbean market and the European market,even American market.Therefore, when buying and selling second-hand sailboats in Hong Kong, regional effects should be taken into account rather than using the lowest prices in other parts of the world to measure whether prices are inflated.

The above is a brief introduction of our study. We sincerely hope that it will provide you with useful information.

Yours sincerely
Team 2333441

References

[1] Enci Liu, Jie Li, Anni Zheng, Haoran Liu, and Tao Jiang. Research on the prediction model of the used car price in view of the pso-gra-bp neural network. *Sustainability*, 14(15):8993, 2022.

[2] James M Sallee, Sarah E West, and Wei Fan. Do consumers recognize the value of fuel economy? evidence from used car prices and gasoline price fluctuations. *Journal of Public Economics*, 135:61–73, 2016.

[3] Chuancan Chen, Lulu Hao, and Cong Xu. Comparative analysis of used car price evaluation models. In *AIP Conference Proceedings*, volume 1839, page 020165. AIP Publishing LLC, 2017.

[4] John R Brence. *Analysis of robust measures for random forest regression*. University of Virginia, 2004.