Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2330197

Used Sailboats Pricing Model Based on Decision Tree Regression with AdaBoost

Summary

Like many luxury goods, sailboats vary in value as they age and as market conditions change. In order to gain a better understanding of the sailboat market, we have studied the listing price pattern of used sailboats.

In this paper, we first process the data provided by the boating enthusiast. On this basis, we establish Model I using **Decision Tree Regression with AdaBoost** and find the optimal parameters through grid search. Then, we use the **R-Square** score to evaluate the model goodness, and use the **Shapley value** in Model I to analyze the regional effects on listing price. Next, we remove the regional and establish Model II and analyze the difference between Model I and Model II, explain regional effects significance. And we use Model I to predict the Hong Kong (HK) data we searched and discuss the discrepancy between the predictions and the reality, we also establish Model III after the HK data are added and analyze the difference of the predictions between Model I and Model III. Finally, we do **sensitivity analysis** to assess the stability and adaptability of our model.

For Model I: Through data processing, the data after removal of null values and encoding are obtained, then the five features "Make", "variant", "Length", "Country/Region/State", "Year" are used to train machine learning model Decision Tree Regression with AdaBoost for two ship types. Our model can predict the price of used sailboats of certain kind **accurately**, and gain a high score of **0.852** for Monohulled Sailboat and **0.828** for Catamarans. On this basis, different variants with relatively complete data are predicted, and the error is basically **within 0.1**, which means accuracy is high. In addition, based on the idea of **"Shapley"** value, a parameter is introduced to judge the significance of the characteristic of the region on the listing price, indicating that "Country/Region/State" is important for predicting the outcome. And by predicting the impact of different regions of the three variants on pricing, it can be seen that the regional effect has a consistent effect on the overall price trend in each variant, but due to factors such as regional preferences, a certain variant shows different trends in some regions.

For Model II: For Model II: Compared to Model I, the model scores of the two types of ships were **0.829** and **0.776**, respectively, which is slightly lower than that of Model I.

Combined with Model I and Model II, and referring to the statistical data from 2006 to 2008, concluded that the regional impact on second-hand prices is significant, and the practical and statistical significance is explained from many aspects of regional differences, for example, regional economy, market, population, regional preferences and other factors may cause price differences.

For Model III: In order to compare the impact of HK data, we first use Model I to predict HK data and find the actual value of the monohull is basically similar to the predicted value, some are slightly larger, and the actual value of the catamaran is larger than the predicted value. Since HK 's economy is developed, the results are consistent with the analysis of the region in this paper. Then we add HK data to train the model, using the five features of "Make", "variant", "Length", "Geographic Region", "Year", the model scores are 0.853 and 0.866, respectively. After re-predicting the subset data, the predictions increase by **0.009** and **0.019** respectively, comparing to Model I. That is, after the training concentrates on adding data from economically developed regions, the overall price of the model will increase slightly, which is in line with the previous judgment of regional effects, indicating that the adaptability of the model is strong.

Finally, we do the **sensitivity analysis** on Model I by adding the noise to the data set, and the predicted stability rates of the two ship models are **0.966** and **0.887**, respectively, which have high stability, indicating that the model has good stability.

Keywords: Used Sailboats Pricing; AdaBoost; Decision Tree Regression; R-Square

Content

1 Introduction

1.1 Background

Shipping is the most important carrier of global trade, according to the United Nations Association for the Promotion of Trade statistics, maritime trade accounts for 90% of the total global trade volume. From the current fleet age distribution, 16-20 years old ship capacity accounted for 16.5%, 20 years old ship capacity accounted for 7.1%, ship aging trend is obvious, the number to be dismantled gradually rise, new ship capacity supplement demand gradually upward. [1] Considering the replenishment of sailing vessels, their value cannot be ignored. Second-hand sailing boats are also considered by many sailing brokers as their prices are more affordable compared to new boats. As with many luxury items, the value of a sailing boat varies with age and market conditions (Make, Variant, Length, Geographic Region, Country/Region/State.).

Sailboats are often sold through brokers. To gain a better understanding of the sailboat market, our team was commissioned by a sailboat broker in Hong Kong, China (SAR) to prepare a report on the pricing of used sailboats.

Based on data provided by a boating enthusiast, we developed a mathematical model to assess the relationship between the pricing of used sailing boats and various sailing characteristics, studied the impact of various characteristics and presented a report here.

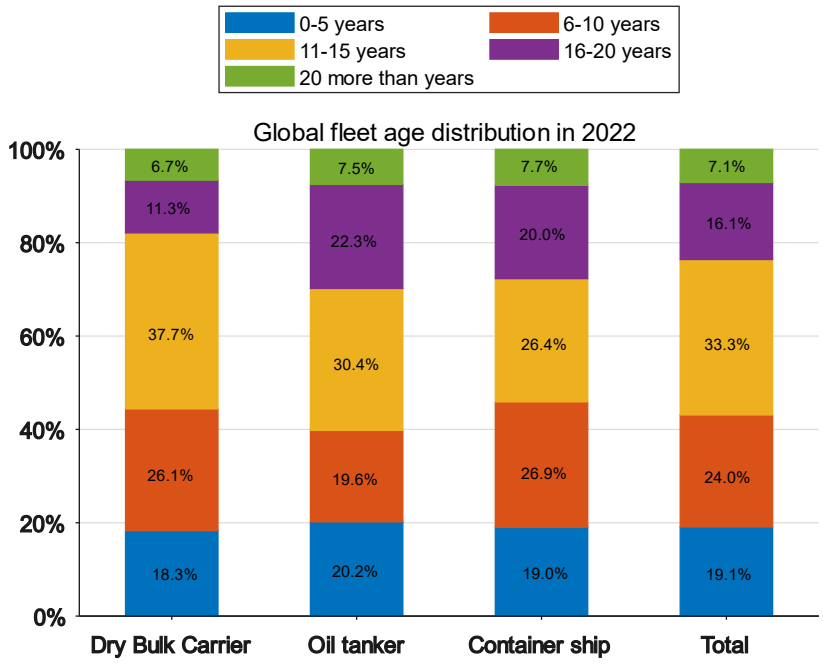


Figure 1 Global fleet age distribution in 2022

Figure 1: Global fleet age distribution in 2022. Data from www.huaon.com.

1.2 Restatement of the Problem

Through in-depth analysis and research on the background of the problem, combined with the specific constraints given, the restate of the problem can be expressed as follows:

- Develop a mathematical model that explains the listing price of each of the sailboats in the provided spreadsheet. Include Make, Variant, Length, Geographic Region, Country/Region/State and Age(Year). Include a discussion of the precision of your estimate for each sailboat variant's price.
- Based on the model you have established, explain the effect of region on listing prices. Discuss whether any regional effect is consistent across all sailboat variants. Address the practical and statistical significance of any regional effects noted.
- Discuss how our modeling of the given geographic regions can be useful in the Hong Kong (SAR) market. Find comparable listing price data for that subset from the Hong Kong (SAR) market. Model what the regional effect of Hong Kong (SAR) would be, if there is one, on each of the sailboat prices for the sailboats in your subset. Whether the effect is the same for both Catamarans and monohull sailboats.
- Prepare a one- to two-page report for the Hong Kong (SAR) sailboat broker. Include a few well-chosen graphics to help the broker understand your conclusions.

1.3 Our Work

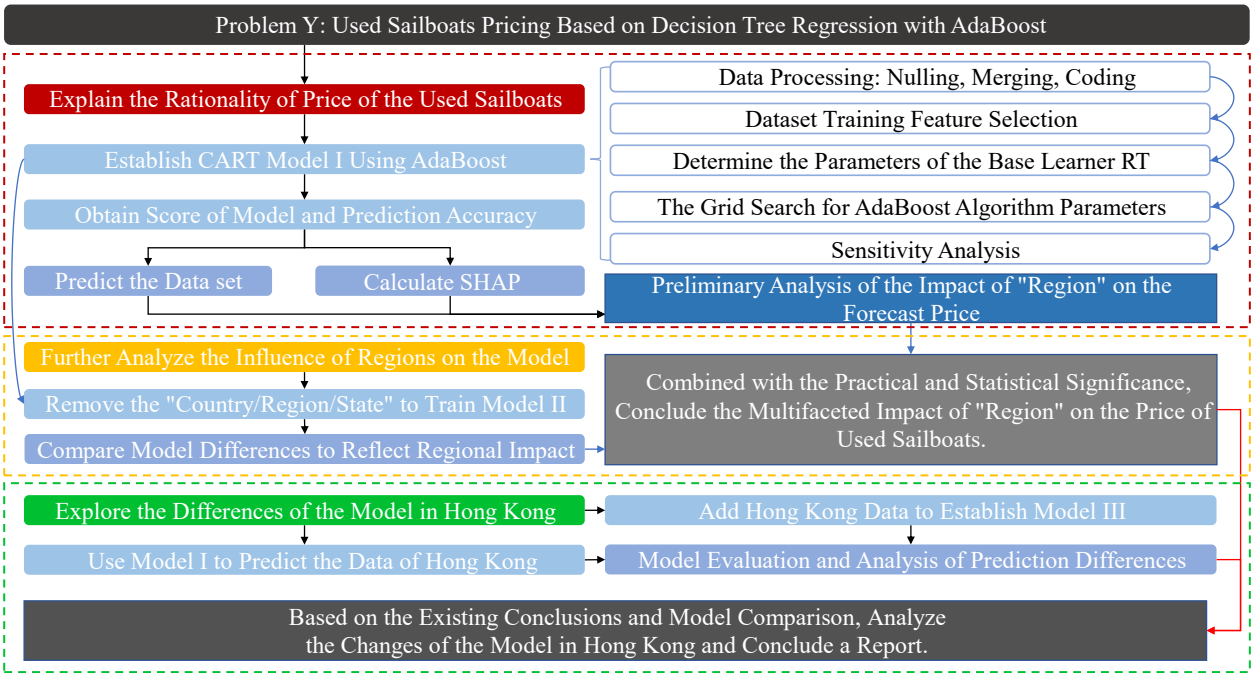


Figure 2 Our Work

2 Assumptions and Justifications

- **The data provided is accurate.** The data is provided by a boating enthusiasts, we consider him or her is familiar with sailboat markets and is accessible to accurate data. Besides, we do some data cleaning work including eliminating the null values and the empty space to make sure the data is effective.
- **During our modeling based on the provided data, we don't consider Geographic Region as a factor.** Since Country/Region/State subordinates to Geographic Region, we consider Geographic Region into our model and abandon Country/Region/State factors to lower the dimensionality under the premise of ensuring the accuracy of our model.
- **For the regions provided, we think they are all coastal.** Sailboats are only widely used in coastal areas, we think the regions provided are all accessible to the sea and don't discuss this in regional effects in our paper.
- **There is a linear positive correlation between listing price and year.** With time goes by, the used ship will wear out. The degree of breakage of the boat is only related to the time of use, and the rate of depreciation is constant.

3 Notations

The primary notations used in this paper are listed in Table 1.

Table 1 Notations

Symbol	Description
$\boldsymbol{x_i}$	the i^{th} independent vector
n	number of characteristics of the independent variable
N	sample capacity
e/E	error(rate)
$Dist_t(\boldsymbol{x_i})$	error of the i^{th}variable used for the t^{th} training
h_t	weak learner for the t^{th} training
MS	Monohulled Sailboat
CM	Catamarans

4 Model I: Decision Tree Regression with AdaBoost Based on the Data Provided

We only use the data provided, no more search for other feature vectors.

4.1 Data Visualization and Processing

4.1.1 Date Visualization

Since most of the analysis of the results is for variant, we make a heat map of listed prices for

the dataset with year and variant as independent variables (averaged over other types) to show the dataset distribution.

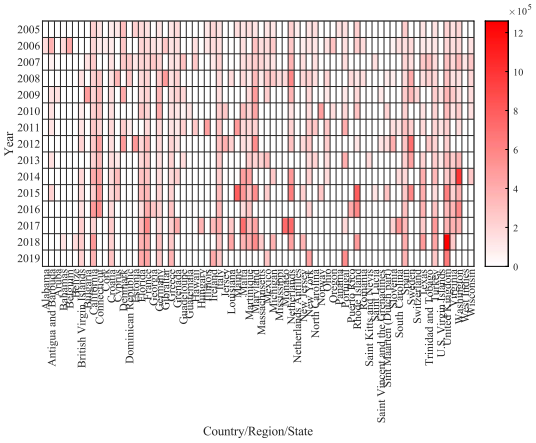


Figure 3(a): Monohulled Sailboat

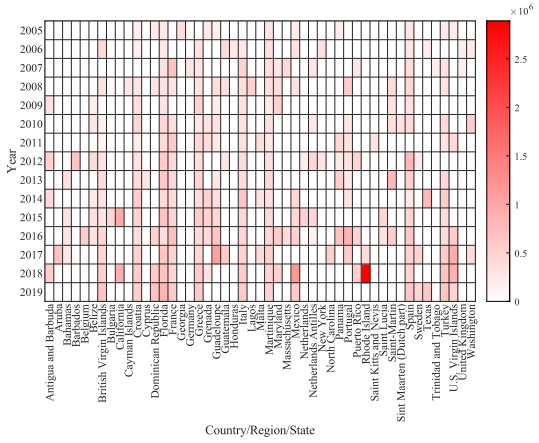


Figure 3(b): Catamarans

Figure 3: Heat map of price intensity with year and variant as variables using MATLAB.

As can be seen from the figure, the distribution of the data set is scattered and the intensity shows an increase with each year, so we consider the elimination of the year dimension when performing some results presentation, that is, projecting them to the same year.

4.1.2 Data Processing

We first remove three null values. Then we find that some strings have the numbers and letters reversed and some strings have blanks behind it which seriously impact our category coding since the computer will consider these as different categories. To solve the problem, we use strip function to delete the blanks behind the strings and rename the variant using Python.

Next, we number the category value according to the character sequence. The steps are as follows, first compare the initial character, alphabet is prior to number, numbers are sorted by sizes, alphabets are sorted based on the alphabetical order. If the initial number is the same, we can use the second character and so on to compare. Lastly, we code them sequentially.

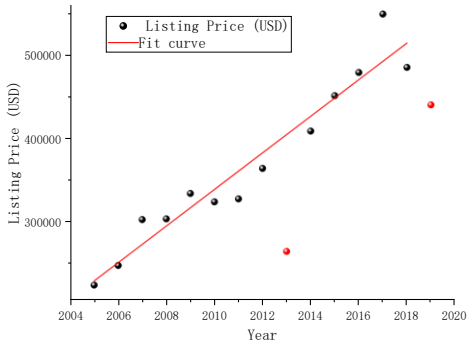
After data processing, the number of possible values for each of its features is shown in the Table 2.

Table 2 Number of Features

feature	Make	Variant	Length	Geographic Region	Country/ Region/State	Year
Monohulled Sail-boat	62	362	57	3	72	15
Catamarans	20	92	57	3	49	15

4.1.3 Fitting of Year

We assume that there is a linear positive correlation between listing price and year, to prove our assumption, we use origin 2023b to calculate the average over current year data and then fit and analyze, the outcome is shown in figure 4.



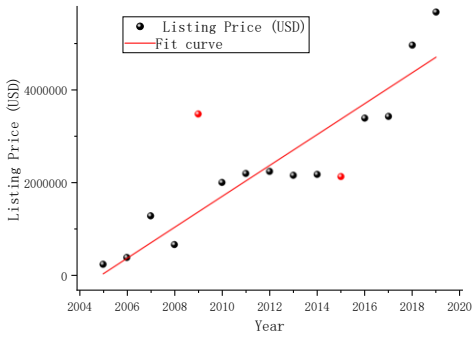


Figure 4(a): Year Fitted Curve of MS **Figure 4(b): Year Fitted Curve of MC**

Figure 4: in Origin Pro fitted curve，the red data point are considered as outliers and should be removed.

The curve parameters we got are as follows:

Table 3(a) Year Fitting Parameters for Monohulled Sailboat

Parameter name	Intercept distance/E6		Slope/E4	
	Value	Standard deviation	Value	Standard deviation
Parameter size	-43.9	3.5	2.2	0.17
	R-Square = 0.936			

Table 3(b) Year Fitting Parameters for Catamarans

Parameter name	Intercept distance/E7		Slope/E4	
	Value	Standard deviation	Value	Standard deviation
Parameter size	-66.5	6.9	33.2	3.4
	R-Square = 0.894			

According to this year fitted curve, we can transfer the year of the sailboats to the same year.

4.2 Establishment of Model

4.2.1 Decision Tree Regression Principle

The decision tree is a basic classification and regression method, we use the regression part in our model. The regression decision tree mainly refers to the CART (Classification and Regression Tree) algorithm, where the internal node features take the values of "true" and "false", and is a binary tree structure. [2]

By regression, we mean that the corresponding output values are determined based on the feature vectors. A regression tree will divide the feature space into some cells, each of which has a specific output. [3] Since each node is a "true" or "false" judgment, the boundary of the division is parallel to the coordinate axis. For test data, we can simply assign it to a cell according to the

features and get the corresponding output value.

The process of dividing is also the process of building a tree, and with each division, the output corresponding to the divided cell is subsequently determined, which means one more node. When the division is terminated according to the stopping condition, the final output of each cell is also determined, which is the leaf node.

Its cut-off point is selected by least squares methods, and the output value is the mean value within the cell.

Suppose X and Y are respectively the input and output variables, and Y is a continuous variable. Given a training data set $D = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \dots, (\boldsymbol{x_N}, y_N)\}$, where $\boldsymbol{x_i} = \left(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)}\right)$is the input instance (feature vector), n represents the number of features, $\mathrm{i} = 1,2, \dots, \mathrm{N}, \mathrm{N}$ is the sample capacity.

The partitioning of the feature space is done heuristically, and each partition examines all the values of all the features in the current set one by one, and selects the best one as the cut point according to the squared error minimization criterion. For example, the j^{th} feature variable $x^{(j)}$ and its value s are used as cut variables and cut points. Define two regions $R_1(j,s) = \{x \mid x^{(j)} \leqslant s\}$ and $R_2(j,s) = \{x \mid x^{(j)} > s\}$. By solving the formula below, we can find the optimal j and s.

$$\min_{j,s}\left[\min_{c1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2\right] \tag{1}$$

That is, to find the j and s that minimize the sum of squared errors of the two regions to be divided.

Where c_1, c_2 are the fixed output values in the two regions after division, the two min in square brackets means that the optimal c_1 and c_2 are used, that is, the c_1 and c_2 that minimize the squared error in their respective regions. The two optimal output values are the mean values of Y in their respective corresponding regions, so the above formula can be written as

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

in which

$$\hat{c_1} = \frac{1}{N_1}\sum_{x_i \in R_1(j,s)} y_i,\ \hat{c_2} = \frac{1}{N_2}\sum_{x_i \in R_2(j,s)} y_i \tag{3}$$

After finding the optimal cut point (j,s), the input space is divided into two regions in turn, and then the above division process is repeated for each region until the stopping condition is satisfied.

Finally, the input space is divided into M regions $R_1, R_2, \dots, R_M$, and the decision tree is generated:

$$H(x) = \sum_{m=1}^{M} \hat{c_m} I(x \in R_m) \tag{4}$$

In (4), I represents instruction function, $I = \begin{cases} 1 & \text{if } (x \in R_m) \\ 0 & \text{if } (x \notin R_m) \end{cases}$

Finally, a regression tree is generated, and such a regression tree is usually called a least squares regression tree.

4.2.2 AdaBoost Principle

Boosting, also known as augmented learning or boosting, is an important integrated learning technique that can augment a weak learner with prediction accuracy only slightly higher than random guesses into a strong learner with high prediction accuracy, which provides an effective new idea and method for the design of learning algorithms when it is very difficult to construct strong learners directly. [4]

The AdaBoost algorithm can be used not only for classification tasks, but also for regression tasks. Since the results obtained from regression prediction are continuous values, such as stock prices. Since the stock price trend curve is continuous, stock prices have very many possible values in the real number range, unlike the category labels in classification tasks which only have a number of fixed integer values.

AdaBoost generates strong learners by giving a high weight to learners with low error rates and a low weight to learners with high error rates, combining each learner and the corresponding weights. [5]

Suppose there is a dataset consists of m samples: $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)\}$.

In D, $\boldsymbol{x}_i \in \mathbb{R}^n$ (each sample data has n features), y_i is the target value of sample $\boldsymbol{x}_i$. AdaBoost regression algorithm are as follows:

First, initialize the weights. Let the sample distribution of the data set in the initial state be $Dist_1$.The weights are initialized to $1/N$ for each sample $\boldsymbol{x}_i$, then $Dist_1(\boldsymbol{x}_i) = 1/N$. $Dist_1$ distribution is used to train the first weak classifier, $Dist_t$ is used to train the t weak classifier h_t, and same for the rest.

Second, perform T-round iterations circularly. The numbering of the weak classifier in each iteration round is denoted as t and $t \in \{1,2,3, \dots, T\}$. Take step two as a loop body, the steps in the loop body are as follows:

(1) Based on a sample distribution of $Dist_t(\boldsymbol{x})$, train weak classifier h_t in dataset D;

(2) Calculate the max error E_t of classifier h_t on training set D, the calculation formula is

$$E_t = \max|y_i - h_t(\boldsymbol{x}_i)|, \ i = 1,2, \dots, N \tag{5}$$

In (5), $h_t(\boldsymbol{x}_i)$ shows the prediction outcomes of weak classifier h_t for sample $\boldsymbol{x}_i$, y_i shows the target value of sample $\boldsymbol{x}_i$;

(3) Based on the maximum error E_t of h_t obtained above, calculate the relative error of h_t for each sample. There are many ways to calculate it, here is ana example of squared error:

$$e_{ti} = \frac{\left(y_i - h_t(\boldsymbol{x}_i)\right)^2}{E_t^2},\ i = 1,2,\dots,N \tag{6}$$

(4) Based on the relative error e_{ti} obtained above, calculate the error rate of current weak classifier:

$$e_t = \sum_{i=1}^{N} Dist_t(\boldsymbol{x}_i) e_{ti} \tag{7}$$

which is the sum of the products of the weights and errors for all samples in the data set;

(5) Update the weights of the current weak classifier h_t, the calculation formula is as follows:

$$w_t = \frac{e_t}{1 - e_t} \tag{8}$$

(6) Update the weight distribution of the data set, for sample $\boldsymbol{x}_i$, the update weights are calculated by:

$$Dist_{t+1}(\boldsymbol{x}_i) = \frac{Dist_t(\boldsymbol{x}_i)}{Z_t} w_t^{1-e_{ti}} \tag{9}$$

in which, Z_t is the normalization factor，the calculation formula is as follow:

$$Z_t = \sum_{i=1}^{N} Dist_t(\boldsymbol{x}_i) w_t^{1-e_{ti}} \tag{10}$$

(7) Make $t := t + 1$，then go back to step one in the loop body.

Third, end t-round iterations, the strong learner obtained is as follows:

$$H(\boldsymbol{x}) = \sum_{i=1}^{N} \ln\left(\frac{1}{w_t}\right) f(\boldsymbol{x}) = \left[\sum_{i=1}^{N} \ln\left(\frac{1}{w_t}\right)\right] f(\boldsymbol{x}) \tag{11}$$

$f(\boldsymbol{x})$ is the median of $w_t h_t(\boldsymbol{x})(t = 1,2,\dots,T)$，that is, the median of the weighted output results of all weak learners.

4.2.3 Parameter Selection

The parameters of this model include: object type(base learner) and its base parameters, maximum number of iterations of the base learner, learning rate (weight reduction factor for each base learner), error function.

As for the base learner, we choose Decision Tree Regression Model, the reasons are as follows:

• For decision trees, the preparation of data is often simple or unnecessary. In this model, there is no abnormal data as well as blank data (although we have eliminated them), and it is more appropriate to use decision trees.

• Ability to handle both numerical and categorial data. Other models tend to require a single data attribute and generally require a data type.

• It is possible to produce feasible and effective results for large data sources in a relatively short period of time. The size of our data is not small, and using a decision tree model can help us

get outcomes quickly.

• It can handle uncorrelated feature data. We consider our data to be uncorrelated in our hypothesis, and regression prediction of this dataset can be achieved with this model.

As for AdaBoost, we use grid search to solve.

4.2.4 Model Solving

We use Python for model solving and train a machine learning model with variable x (contains Make, variant, Length, Country/Region/State, Year). The model can output a direct prediction of the price of a sailboat in a certain situation.

We performed a grid search for this model using Python and found that the model fits better at a maximum depth of 20 and a minimum number of classifications of 2. Therefore, we use this base learner in AdaBoost.

Then we use grid search for optimal parameter fitting of the AdaBoost algorithm again. It can be concluded that the maximum number of iterations of the best learner is 300, the learning rate is 0.6 and the error function is ‘square’, which represents square loss function.

For Catamarans, the maximum number of iterations of the base learner is 100, the learning rate is 0.8, and the error function is 'square'.

Accurate sailboat price forecasts can be obtained by using the model's prediction function.

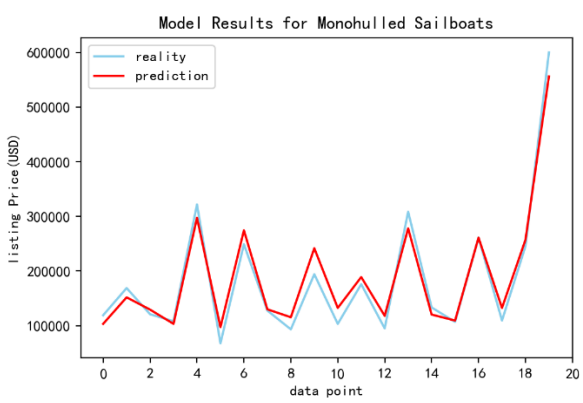
4.3 Result for Model I

4.3.1 Model Precision

For the models built by machine learning, we use model scores as a way to evaluate their accuracy.

The test set R-square score for the Monohulled Sailboat model is about 0.852 and the test set R-Square score for the Catamarans model is about 0.828.

We output 20 sets of comparison graphs of predicted data and real data respectively, which are shown in figure 5.



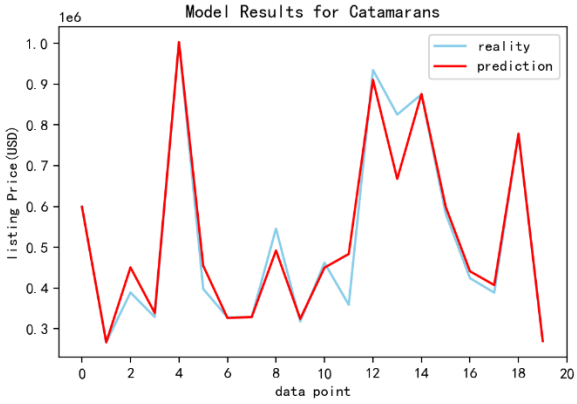


Figure 5(a): Model I Results for MS **Figure 5(b): Model I Results for CM**

Figure 5: The prediction results of Decision Tree Regression with AdaBoost trained by Python for the test set.

The graph shows that the model is solved accurately and is able to give a relatively accurate price prediction value for a certain characteristic sailboat.

4.3.2 A Discussion of the Precision of the Estimate for Each Sailboat Variant's Price

With the resulting model, to obtain the estimation accuracy of the variant, we project all independent variables in the dataset and then project them to 2012 using the year-fitted curve, and average the relative errors they obtain according to each variant to obtain the following figure. The x-axis represents the code of each model, y-axis represents the size of relative error.

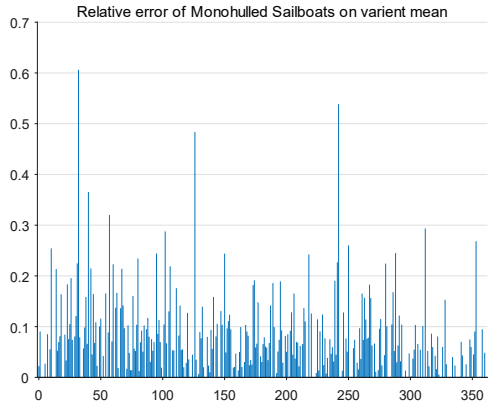


Figure 6(a): The average relative error of Model I for Monohulled Sailboat

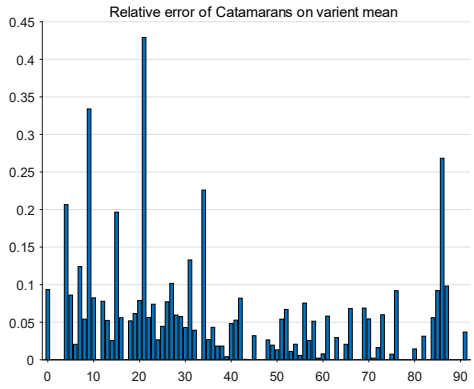


Figure 6(b): The average relative error of Model l for Catamarans

The graphs show that most of the error ranges are within 0.1, and the relative errors are larger for very few variants, probably due to the heterogeneous original data, but overall, the model is relatively accurate for variant estimation.

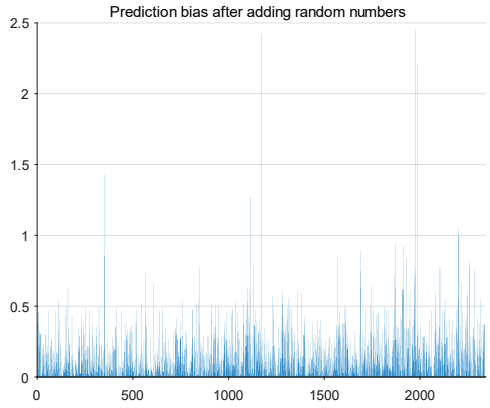
4.3.3 Discuss the Significance of Regional Effects on Listed Prices Using Model 1

Inspired by the problem of equitable distribution, we first consider an indicator "Shapley value" to describe the significance of the effect of a characteristic on the dependent variable:[7]

$$\phi_i(v) = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N|-|S|-1)!}{|N|!}\big(v(S \cup i) - v(S)\big) \tag{12}$$

In this problem, this means that we generate a new dataset by randomizing the codes corresponding to the feature of the region. This new dataset will contain all the features and corresponding labels of the original dataset, but replacing the values of the selected features with the values randomized in the test data. For the new dataset, the prediction is re-run and the difference between the prediction and the original prediction is calculated, and the average size of this value reflects the significance of the effect of the feature on the dependent variable.

Based on this idea, we obtained the relative errors between the new data set and the original data set as follows.



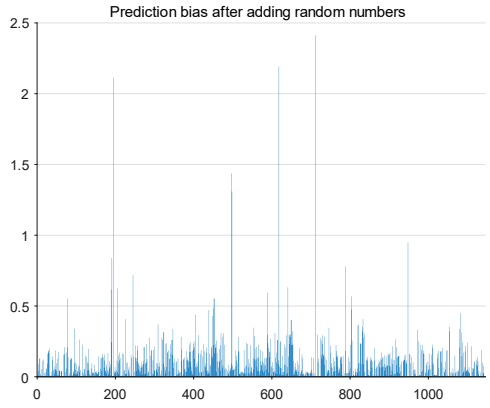


Figure 7(a): Relative Error for MS

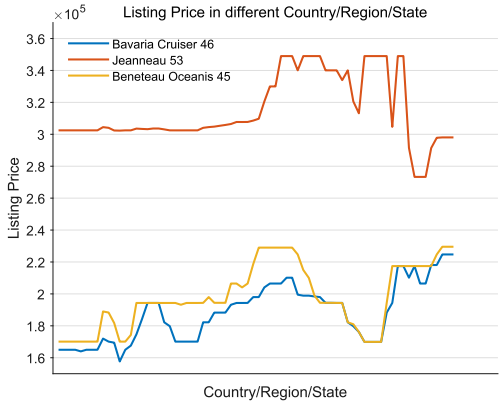
Figure 7(b): Relative Error for CM

On this basis, the deviations of the two predictions are averaged, and the "significance" of the region in model 1 is about 39097 for MS and 26215 for C. This is much greater than the "significance" of the other characteristics, so we believe that the influence of the regional characteristics on the listing price is quite significant.

4.3.4 Discuss Regional Effect Across All Sailboat Variants.

Many variants in the sample lack of data coming from many regions. In order to obtain the effect of region on all sailboat variants, we selected the three variants with the largest sample size to examine whether the region effect is consistent among the three variants, and if it is, then we can explain the effect of region in the variants.

The predicted values of different areas for three variants monohull (Bavaria Cruiser 46, Beneteau Oceanis 45, Jeanneau53) and three variants of multihull (Lagoon 420, Fountaine Pajot Lipari 41, Lagoon 450) were predicted by Model I and plotted using MATLAB as follows.



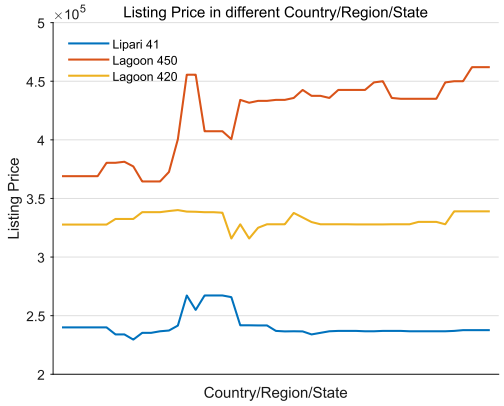


Figure 8(a): The regional effects of Model I for Monohulled Sailboat

Figure 8(b): The regional effects of Model I for Catamarans

The graph shows that for monohulls, the trend of regional influence on different models is basically the same, and for Catamarans, there are some differences in the influence of regions on

different variants. For example, the degree of regional preference for variants is not consistent, and the prices of different variants of monohulls in the same region are roughly proportional. However, the price pattern of different variants of Catamarans in the same region is not obvious, generally, those with high prices are always higher than others.

5 Model II: Exploring Regional Effects

5.1 Modeling with the Regional Data Removed from Provided Data

Inspired by ablation experiments, to explore the impacts of regional effects,[8] we excluded regions from the training model. That is to say, we explore whether the inclusion of regional effects affects the model's pattern.

5.1.1 Establishment of Model II

The regional data are removed, which means a machine learning model (Decision Tree Regression with AdaBoost) is trained by python using the variable x (which contains Make, variant, Length, Year). The principle has been described in Section 4.

For the parameter selection, we choose the same parameters as Model I. Then, we use the existing data to solve the model. Then we can build a strong learner that can give an exact solution for the listed price with the region removed.

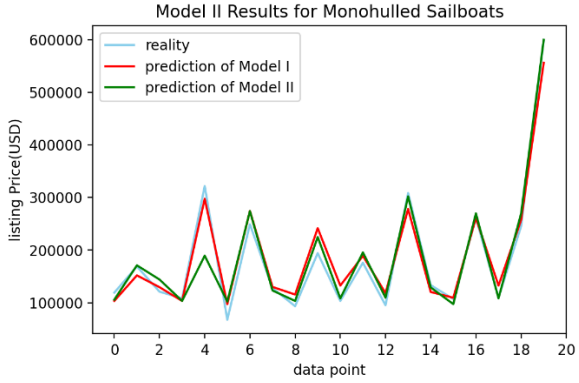
5.1.2 Model Evaluation

After solving the model, the R-Square scores are 0.829 and 0.776, respectively, which are both down by about 0.03 compared to Model I. That is, Model II is not as accurate compared to Model I, although it can also achieve a relatively high score.

5.2 Result for Model II

5.2.1 Use Model II to Explain the Effect of Region on Listing Prices

We make the same data output as the previously selected 20 data points.



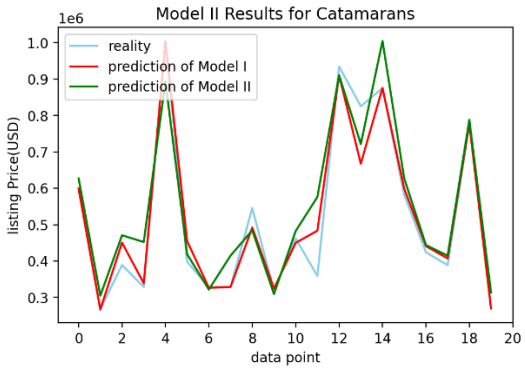


Figure 9(a): Model II Results for MS **Figure 9(b): Model II Results for CM**

The graph shows that when the model is trained with the region feature removed, the predictions are more biased than when the region feature is added, that is, region is important to the

model.

5.2.2 Address the Practical and Statistical Significance of any Regional Effects Noted

By calculating the Shapley value of model 1, the "significance" of the region in model 1 to MS is about 39097, and the "significance" of C is about 26215, and the significance of the region is extremely high compared with the other variables, which has a great impact on the model.

And when the model training rejects the features, it will be found that the prediction accuracy of the model is slightly reduced, reflecting the importance of considering the information region when pricing.

In the process of pricing second-hand sailboats, regions will have an impact on many aspects, such as the preference of people in different regions for a certain type of sailboat, the level of freight, the degree of market development in different regions, and the degree of economic development, which will lead to the difference in the listing price of the same second-hand sailboat.

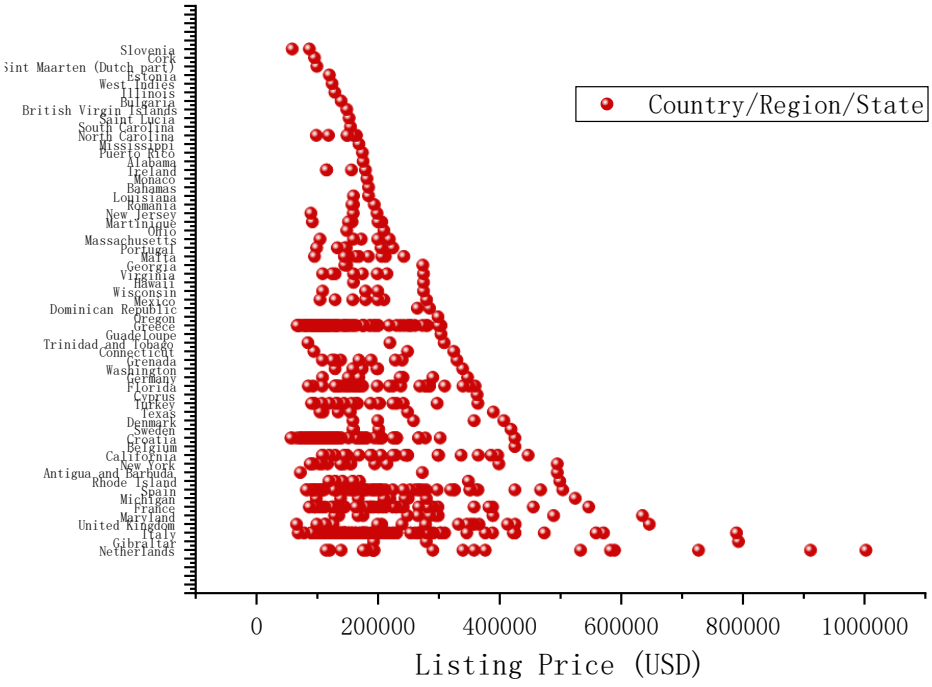


Figure 10 Distribution of Listing Prices by Region in 2006-2008

Figure 10 shows the distribution of transaction prices in different regions, and by comparing the maximum values of various regions, it is obvious that the most economically developed regions have higher maximum listing prices, such as the Netherlands, the United Kingdom, France, New York and other economically developed areas, the maximum value of second-hand sailing prices is significantly greater than Slovenia, West Indies, Estonia and other economically underdeveloped regions, and the statistical results are very significant and intuitive.

6 Model III: Application of the Model I to the Hong Kong Region and Model Differences

6.1 Prediction of Hong Kong Data Using Model I

We found about 20 sets of overlapping data for each type of ship by searching various websites(https://m.jyacht.com, https://www.simpsonmarine.com, https://hongkongboats.hk, https://www.luxboating.com).

Since the Hong Kong region is not included in the values available for the regional features in our model, we consider substituting the remaining features of the found data into the various regional features that are desirable for the model, and then take the average of the predicted values as the predicted values of the found data. The results are as follows.

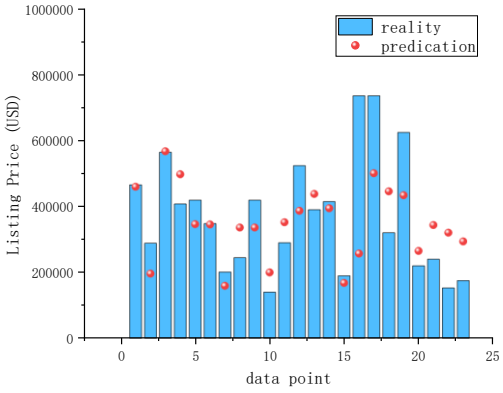


Figure 11(a): Model I Results for Monohulled Sailboat in HK

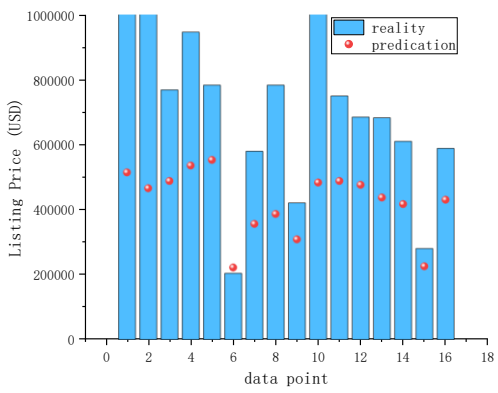


Figure 11(b): Model I Results for Catamarans in HK

Figure 11: Plot of predicted vs. actual listing prices for Hong Kong data using Origin Pro 2023b, data come from various websites (cited before).

It can be seen that the price of used Monohulled Sailboat in Hong Kong is basically the same as the predicted value, while the predicted value of the model for catamaran is slightly smaller than the actual value, but it can also be evaluated by the model for its high or low price, so the model can be applied to Hong Kong in assessing the price ratio.

6.2 Model III: Decision Tree Regression with AdaBoost with Hong Kong Data Included

6.2.1 Establishment of the Model

We add the data of Hong Kong, which belongs to region. Since there are too many small regions, we consider modeling with data other than small regions, that is we use the variable x (which contains "Make", "variant", "Length", "Geographic Region", "Year") to train a machine

learning model using Python (Decision Tree Regression with AdaBoost). The schematic model has been described in Section 4.

For parameter selection, the base learner and its parameters are the same as for Model I.

For the parameters of AdaBoost, we still use the grid search and can derive the maximum number of iterations of the base learner as 150 for Monohulled Sailboat, the learning rate as 0.8 and the error function as 'square'.

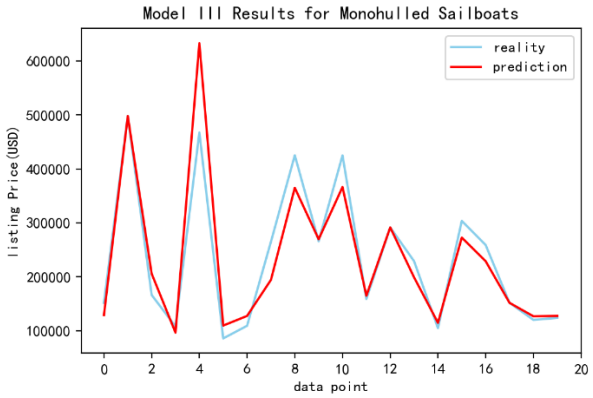
For Catamarans, the maximum number of iterations of the base learner is 150, the learning rate is 0.6, and the error function is 'square'.

6.2.2 Model Evaluation

For the models built by machine learning, we use model scores as a way to evaluate their accuracy.

The test set R-square score for the Monohulled Sailboat model is about 0.853, and the test set R-Square score for the Catamarans model is about 0.866.

We output 20 sets of comparison plots of predicted data and real data respectively, which are shown in Figure 12.



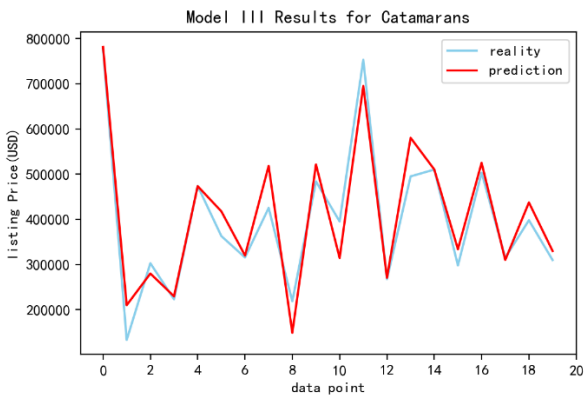


Figure 12(a): Model III Results for MS **Figure 12(b): Model III Results for CM**

Figure 12: The prediction results of Decision Tree Regression with AdaBoost trained by Python for the test set.

The graph shows that the solution model is relatively accurate and it has the ability to give an accurate price prediction for a certain characteristic sailboat.

6.3 Result for Model III: Discuss the Impact on a Subset of Data After Adding Hong Kong Data to Model

Model after adding Hong Kong data, then predict the values of the previous subset. Next, differentiate the predicted value from the forecast value of Module I and calculate the relative error. Finally, calculate the average of all relative errors, so we can obtain the model difference after adding Hong Kong data.

Table 4 The Difference between Model I and Model III

sailboat variance	Monohulled Sailboat	Catamarans
average relative error	0.008928	0.018794

Analysis of the difference between Monohulled Sailboat and Catamarans:

For Monohulled Sailboat, the total numerical trend increases only a little after adding the model, and the graphical change is not obvious, while for Catamarans, we can find that the total trend of the model increases by about 2%. We make a relative residual plot based on the obtained data, which is shown in the figure below.

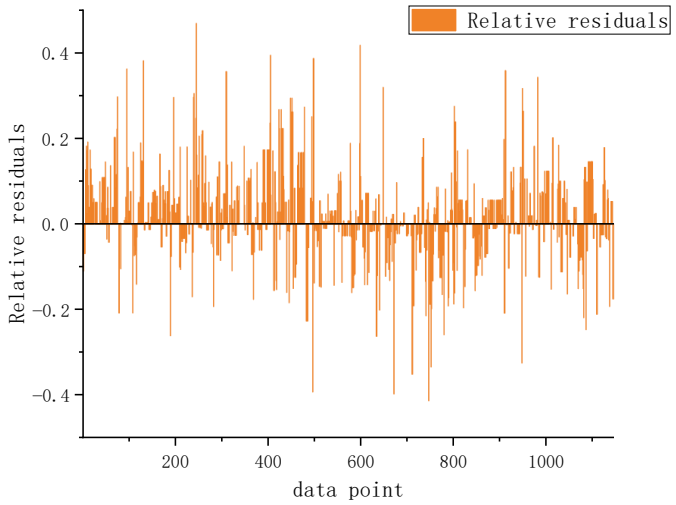


Figure 13 Relative Residuals Histogram

Figure 13: A residual plot of the predicted values of Model III against the predicted values of Model I using Origin Pro 2023b with Model I as the base.

The number of Hong Kong data we add is only about 20, which produces a significant overall improvement in the total model, so in fact, adding a large number of Hong Kong prices should result in an overall increase in the model predictions.

It can be seen that most of its values have improved by some amount, and a small number of values have produced a decrease.

7 Sensitivity Analysis of Model I

For our previous exact prediction model (Model I), we add a random noise between -1000 and 1000 to the listing price of the data set, and then by comparing the changes in the two outputs, we use Origin to plot.

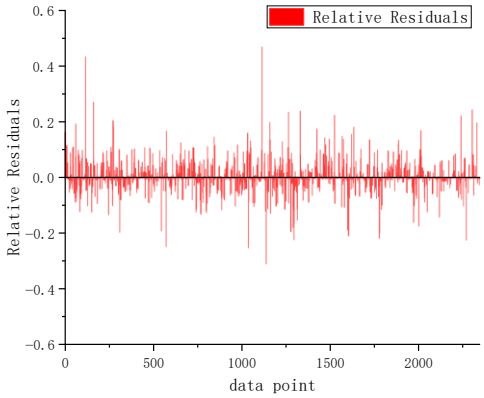


Figure 14(a): Model I Sensitivity Analysis for Monohulled Sailboat

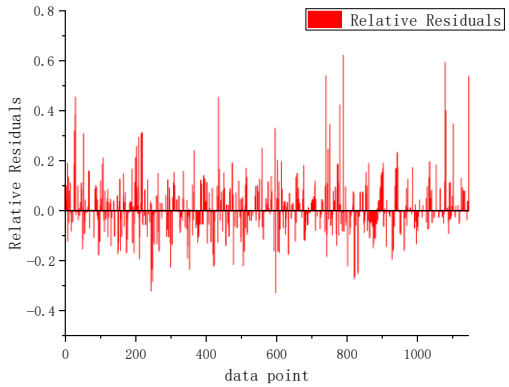


Figure 14(b): Model I Sensitivity Analysis for Catamarans

Figure 14: Graphs of relative residuals of changes to the dataset using Python-trained Model I.

The R-Square score of Monohulled Sailboat becomes 0.825, and the R-Square score for the Catamarans becomes 0.861.

It can be seen that the relative residuals are small, only a few data points have relative residuals greater than 0.1, we select data points greater than 0.1 as deviation points, and points less than 0.1 as stable points. Define the model stability rate as follows.

$$R = \frac{num_{stable}}{N} \tag{13}$$

If the number of deviation points for **Monohulled Sailboat** is 80 and the number of deviation points for **Catamarans** is 130, the stability rate R can be calculated as shown in the following table**.**

Table 5 Stability Rate of Model I

Sailboat variant	Monohulled Sailboat	Catamarans
Stability rate	0.966	0.887

It can be seen that the model is relatively stable and the model score does not change much.

8 Strengths and Weaknesses

8.1 Strengths

8.1.1 Advantages of Decision Tree Regression for Base Learners

• Simple and intuitive, the model is easy to train, and the training time for large samples is short.

• There is basically no need for preprocessing, no need for early normalization, and missing values can be handled.

• Ability to work with both data-type and typed attributes, i.e. continuous and separated scattered variables.

• Compared to black-box classification models such as neural networks, decision trees can be interpreted logically well.

• Trees are inherently tolerant to outliers because the choice of segmentation depends on the ordering of the values, not the absolute size of those values.

8.1.2 Strengths of AdaBoost

• Good use of weak classifiers for cascade.

• The AdaBoost algorithm has high accuracy, and the modeling in the paper gives more accurate results.

• Compared with the bagging algorithm and the Random Forest algorithm, each AdaBoost algorithm fully considered by AdaBoost has high accuracy, and the modeling in this paper gives the weight of the more accurate result classifier.

• Since AdaBoost evaluates weights based on the accuracy of weak learners, overfitting is not prone to occur.

8.2 Weaknesses

• Decision tree regression models are sensitive to changes in the data, and even minor changes can result in very different splits, but leveraging ensemble learners to solve this problem well by measuring weights.

• AdaBoost is sensitive to samples, and abnormal samples may receive higher weights in iterations, affecting the prediction accuracy of the final strong learner.

• The number of AdaBoost iterations, that is, the number of weak classifiers, is not easy to set, but we use the grid search method to obtain the optimal model parameters, which solves this problem well.

9 A Report to Hong Kong Broker

We are the team from MCM that researches the pricing of used sailboats. We trained the Decision Tree Regression with AdaBoost model, which can accurately predict the price of certain used sailboat, based on data provided by a sailing enthusiast and Hong Kong data we searched. Our research conclusions are as follows.

I. Our Model can price the used sailboat accurately.

By establishing Decision Tree Regression with AdaBoost to achieve accurate prediction of the pricing of a certain sailboat, it can be seen from Figure 14 that the predicted value corresponding to the red line is strongly fitted to the true value corresponding to the blue line, and the test set R-square is 0.852 for the Monohulled Sailboat and 0.828 for the Catamarans. The relative error of the variant prediction is basically within 0.1, which means the model has high accuracy and strong usability. We also do sensitivity analysis and the stability rate is about 0.9, which proves that the model's stability is strong.

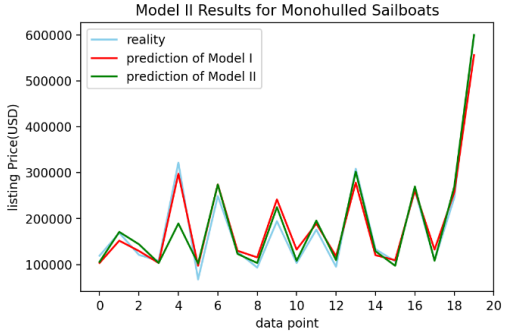


Figure 15(a): Model II Results for MS

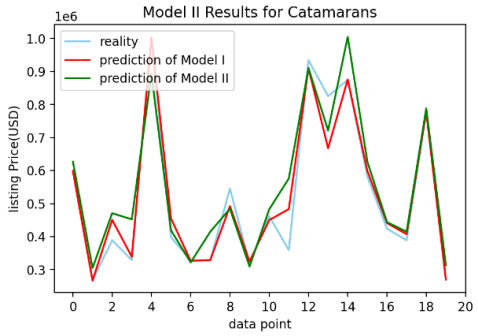


Figure 15(b): Model II Results for CM

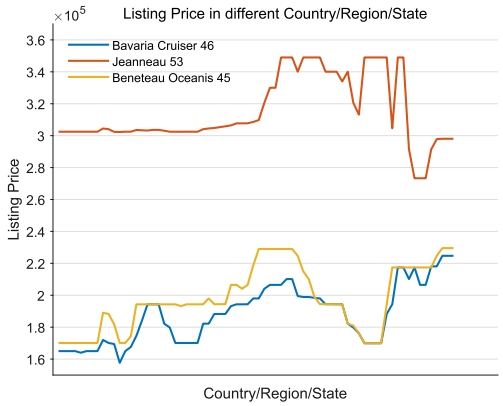
II. Regional factors matter a lot in pricing.

The region factor is critical to sailing pricing. When the variable feature of region is removed from the model, the trained model had a reduced fit effect on both Monohulled Sailboat and Catamarans. The comparison of R-Square is as follows.

Table 6 R-Square for Model I and II

Type / Model	Monohulled Sailboat	Catamarans
Model I	0.852	0.828
Model II	0.829	0.776
Δ	-0.23	-0.52

The figure 16 shows the pricing forecasts for different regions in 2012 for the three models with the largest sample size forecasted by Model I. Judging by the curve trend, the listing price of the same model in different regions is not the same, but its overall change trend is roughly the same. This is mainly affected by the local economy, but in some regions, there is an inconsistency in the listing price, which may be determined by factors such as the preference of local people for some specific models.



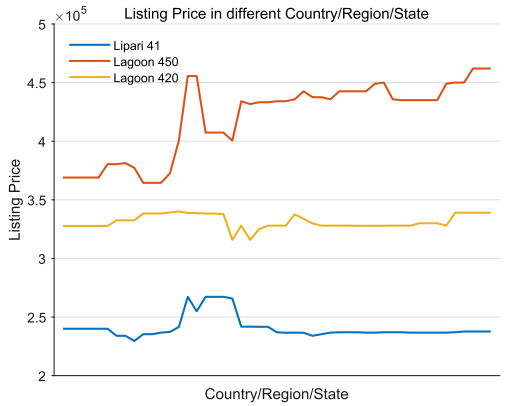
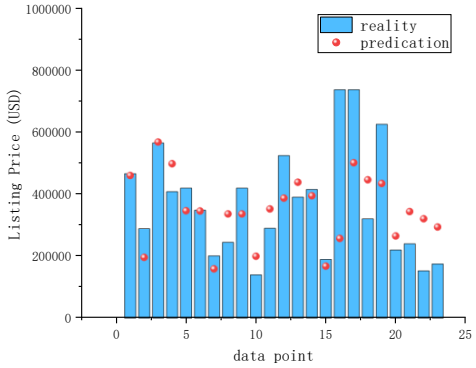


Figure 16(a): Model I Results for MS **Figure 16(b): Model I Results for CM**

III. Model I can be used to predict price of used sailboats in Hong Kong.

The results of directly using Model I to predict the searched Hong Kong data are shown in Figure 16, which shows that the Monohulled Sailboat can basically directly apply Model I for prediction, while the actual data of the catamaran is too large, and it should be retrained with Hong Kong data to obtain a more accurate model.



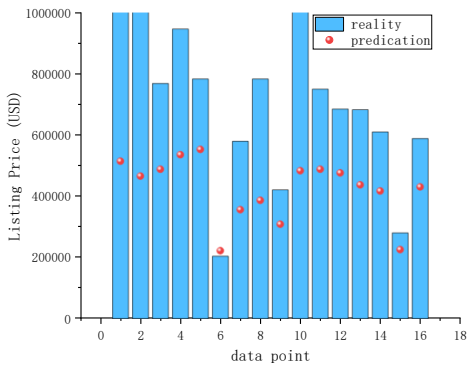


Figure 17(a): Model I Results for MS **Figure 17(b): Model I Results for CM**

IV. Adding data in Hong Kong will cause the predicted price to increase.

Model III is trained after adding about 20 Hong Kong data searched from websites to the training set of Model I. The average relative error of the prediction of the original subset is shown in Table 7.

Table 7 The Difference between Model I and Model III

sailboat variance	Monohulled Sailboat	Catamarans
average relative error	0.008928	0.018794

It can be seen that the trained model does not change much for the market price prediction value of Monohulled Sailboat, while the market price prediction value of Catamarans has been improved. If the amount of data in Hong Kong increases, the overall upward trend will be more pronounced. As analyzed by regional differences, Hong Kong's economy is relatively developed, so the price of second-hand sailing boats is generally on the high side in Hong Kong.

References

[1] Xia Q, Chen F. Shipping Economics Development: A Review from the Perspective of the Shipping Industry Chain for the Past Four Decades[J]. Journal of Shanghai Jiaotong University (Science), 2022, 27(3): 424-436.

[2] Kim Y S. Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size[J]. Expert Systems with Applications, 2008, 34(2): 1227-1234.

[3] Xu M, Watanachaturaporn P, Varshney P K, et al. Decision tree regression for soft classification of remote sensing data[J]. Remote Sensing of Environment, 2005, 97(3): 322-336.

[4] Vezhnevets A, Vezhnevets V. Modest AdaBoost-teaching AdaBoost to generalize better[C]//Graphicon. 2005, 12(5): 987-997.

[5] Schapire R E. Explaining adaboost[J]. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, 2013: 37-52.

[6] Chicco D, Warrens M J, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation[J]. PeerJ Computer Science, 2021, 7: e623.

[7] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. “Learning Important Features Through Propagating Activation Differences”. In: arXiv preprint arXiv:1704.02685 (2017).

[8] Milos F S. Galileo probe heat shield ablation experiment[J]. Journal of Spacecraft and Rockets, 1997, 34(6): 705-713.

Appendices

Appendix 1

Introduce: Data Processing

```
def read(keywords, i):
    df = pd.read_excel(keywords, sheet_name = i)
    for i in range(0, df.shape[1], 1):
        dfNull = df[df.iloc[:, i].isnull()]
        df.drop(index = dfNull.index, inplace = True)
    df['Variant'] = df['Variant'].astype(str)
    for j in range(0, df.shape[1], 1):
        for i in range(0, df.shape[0], 1):
            if type(df.loc[:, :].values[i ,j]) == type("1"):
                if ' ' in df.loc[:, :].values[i ,j]:
                    df.iloc[i, j] = df.iloc[i, j].strip()
                number = re.findall("\d+", df.loc[:, :].values[i ,j])
                num_str = "".join(number)
                str_ = ''.join(re.findall(r'[A-Za-z]', df.loc[:, :].values[i ,j]))
                sss = num_str + '' + str_
                df.iloc[i, j] = sss
    le = preprocessing.LabelEncoder()
    for i in ['Make', 'Variant', 'Geographic Region', 'Country/Region/State ']:
        df[i] = le.fit_transform(df[i].values)
    return(df)
```

Appendix 2

Introduce: Model Building

```
param_grid = {'n_estimators':(100, 300), 'learning_rate':(0.6, 0.8)}
search = GridSearchCV(AdaBoostRegressor(DecisionTreeRegressor(max_depth = 20,
min_samples_split = 2, random_state = 16), loss = 'square', random_state = ran-
dom_seeds), param_grid, cv = 5)
search.fit(X, y.ravel())
best_parameters = search.best_params_
adb_best =\
AdaBoostRegressor(DecisionTreeRegressor(max_depth = 20, min_samples_split = 2),
                                        n_estimators = list(best_parameters.values())[1],
                                        learning_rate = list(best_parameters.values())[0],
                                        loss = 'square', random_state = random_seeds)
adb_best.fit(X_train, y_train)
pred_best = adb_best.predict(X_test)
```