A Study on Price Prediction of Used Sailboats Based on XG-Boost Algorithm

Linkang Song [1, *, &], Bangxu Li [2, &], Xinyao Sun [3, &] and Zhichen Dong [4, &]

[1] School of Life Sciences, Sichuan University, Chengdu, China

[2] School of Mechanical Engineering, Sichuan University, Chengdu, China

[3] Faculty of Architecture and Environment, Sichuan University, Chengdu, China

[4] Faculty of Economics, Sichuan University, Chengdu, China

* Corresponding Author Email: 2948565469@qq.com

[&] These authors contributed equally to this work

Abstract. The sailing industry is currently receiving more and more attention and there is a growing demand for sailboats. Their prices are constantly changing and inevitably contain useful trends and intrinsic correlations. Therefore, we developed a used sailboat price prediction model to allow relevant practitioners to better understand the sailboat market. First, we obtained some key data from the official websites of various sailboat manufacturers that may affect the price, including width, draft, displacement, sail area, headroom, and so on. We pre-processed the data. Next, based on the XG-Boost algorithm, we built a used sailboat pricing prediction model and estimated its accuracy, which reached about 91.5% in the test set. Finally, we did a one-way ANOVA on region and sailboat prices and found that there are significant differences in sailboat prices across regions. Therefore, we conducted a consistency test for region effects and found that the region effects for displacement and sail area were consistent.

Keywords: XG-Boost; one-way ANOVA; used sailboat prices.

1. Introduction

As people's standard of living continues to rise, the luxury of sailing is slowly coming to the forefront of people's minds, and although the numbers are still small, the demand is slowly increasing. However, it is worth noting that people are slowly starting to pay attention to sailing. This is not only conducive to improving the quality of life of a part of the population (e.g., watching races or experiencing them in person), but also drives the development of a series of supporting industries such as transportation, maintenance, refitting, management, etc. The sailing industry is becoming more and more complex. Then, it is essential to understand and analyze the price changes of sailboats in a timely and accurate manner, because it is related to the study of the development prospect of the industry [1].

In this paper, a used sailboat price prediction model is developed to account for selected predictors and used to train a more accurate model for estimating sailboat prices. The model explains the relationship between regional prices and sailboat prices and discusses the consistency of regional effects for sailboat variants.

2. Data preprocessing

We searched and summarized information on all types of sailboats by visiting Saiboatdata.com and the official websites of major sailboat dealers. Some factors that may affect the price were obtained, including width, draft, displacement, sail area, etc. As a result, we modeled the influencing factors of sailboats, as shown in Figure 1.

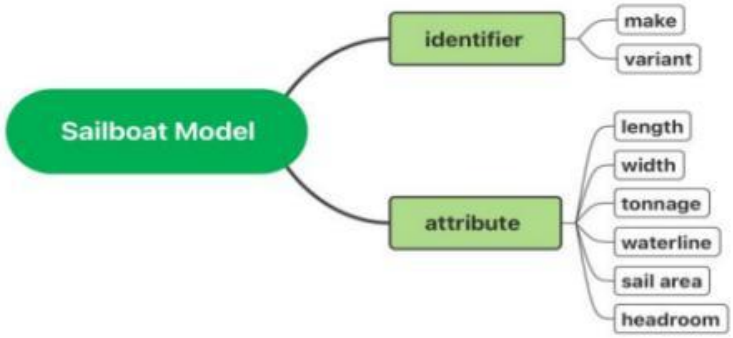

Fig. 1 Sailboat model identification and attributes

In addition, we collected the GDP of each geographic region from the official website of the World Bank (www.worldbank.org) as regional economic data.

3. Predictive model based on XG-Boost algorithm

3.1. Introduction to the XG-Boost algorithm

XGBoost [2] (Extreme Gradient Boosting) is an integrated decision tree-based learning al- gorithm that allows prediction in classification and regression problems. Unlike traditional decision trees, XGBoost trains the model with a gradient-boosting algorithm. In this competi- tion, we will use the XGBoost algorithm to predict the price of a used sailboat based on given features such as length, width, displacement, draft, and sail area.

The core idea of XGBoost is to use multiple decision trees to improve the model's predictive power incrementally. Each decision tree tries to learn some patterns and regularities in the data and makes predictions based on them. Eventually, the predictions from multiple decision trees will be combined into a final prediction [3].

Specifically, XGBoost uses a gradient-boosting algorithm to improve the model's predictive power gradually. The algorithm calculates the gradient of each sample in each training round based on the difference between the current model's prediction and the actual value and uses these gradients as the training target for the next decision tree. During the training of each decision tree, XGBoost also uses regularization techniques to avoid overfitting.

For a dataset with an m-dimensional entries, the XGBoost model can be expressed as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F(i = 1,2,\ldots,n) \tag{1}$$

$$F = \{f(x) = w_{q(x)}\}(q: R^m \rightarrow \{1,2,\ldots,T\}, w \in R^T) \tag{2}$$

The above equation is the set of CART decision tree structures, q is the tree structure of samples mapped to leaf nodes, T is the number of leaf nodes, and w is the real fraction of leaf nodes. When constructing the XGBoost model, it is necessary to find the optimal parameters according to the principle of minimizing the objective function, so as to build the optimal model.

The objective function of the XGBoost model can be divided into an error function term L and a model complexity function term Ω. The objective function can be written as

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{3}$$

Where L(θ) is the loss function, which represents the error between the predicted and actual values of the model; Ω(θ) is the regularization term, which is used to control the complexity of the model. The loss function of XGBoost can be chosen in different forms according to the specific problem. For regression problems, we usually choose the mean squared error (MSE) as the loss function:

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (4)$$

Where yi is the true value of the ith sample and i is the predicted value of the model for the ith sample.

The regularization term can be chosen from L1 or L2 regularization or a combination of both. For this contest, we use the L2 regularization, which takes the form of

$$\Omega(\theta) = \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \quad (5)$$

Where T is the number of leaf nodes in the decision tree, ωj is the score of the jth leaf node, and λ is the regularization hyperparameter.

The training process of XGBoost can be divided into two stages: the first stage is to train a new decision tree by calculating the gradient and weight of each sample based on the prediction results of the current model; the second stage is to add the new decision tree to the model and update the weights of each leaf node. Specifically, the steps of the XGBoost training process are as follows:

Step 1: The predicted value of the initialized model is the average of all samples.

Step 2: For each decision tree:

Calculate each sample's gradient and weight as the decision tree's training target.

(1) Use the greedy algorithm to construct the decision tree and calculate the score of each leaf node.

Use L2 regularization to update the score of each leaf node.

(2) Add the new decision tree to the model and update the predicted values for each sample.

Step 3: Step 2 is repeated until the specified number of training rounds is reached, or the model's predictive power has been optimized.

The advantages of XGBoost are that it can automatically handle missing values and outliers and handle high-dimensional features. In addition, XGBoost provides some unique features such as feature importance analysis and cross-validation.

3.2. Application of the XG-Boost algorithm

Based on the data we collected, we added width, headroom, sail area, displacement, and draft as additional feature dimensions for price prediction. And dummy variables for variables such as Make, Variant, etc. are transformed from fixed class variables to quantitative variables. We used XGBoost to train the model and performed multiple rounds of conditioning to obtain a tree depth of 6, and a gradient-boosting tree with dropout as the base learner.

Our model is trained to obtain the following feature importance ranking, refering to Table 1. The length of the boat and the manufacturer are the two most important indicators, accounting for more than 60%. This is also very consistent with our perception that the bigger the boat is, generally speaking, the more expensive it is different brands of boats will have different brand effects, and boats made by more high-end brands will have higher brand premiums.

Table 1. Importance of features

Feature	Importance
Length (ft)	33.40%
Make	28.20%
GDP (in USD billion)	9.60%
Year	6.00%
Width (m)	4.90%
Variant	4.00%
headroom(m)	3.60%
Sail area (m^2)	3.20%
displacement (kg)	3.10%
draft (m)	1.90%
Country/Region/State	1.50%
Geographic Region	0.90%

As shown in Table 2, the accuracy on the test set was over 90%, reaching 0.915. It shows that our model explains well the relationship between the price of used sailboats and different factors and that the data we collected is informative, accurate and reliable.

Table 2. Model evaluation results

	MSE	RMSE	MAE	MAPE	R^2
Training set	981672354.9	31331.651	22391.372	8.744	0.975
Cross-validation sets	6102524785	76527.691	42001.611	14.6	0.85
Test set	3196996571	56541.989	37504.324	13.382	0.915

3.3. Discussion on the prediction precision of different variants

Our model is based on XGBoost, as shown in Figure 2, and its prediction results are very close to the actual prices, which indicates that our model has very high accuracy and reliability. During the model development process, we have gone through repeated experiments and optimizations to ensure that our model can accurately predict prices.

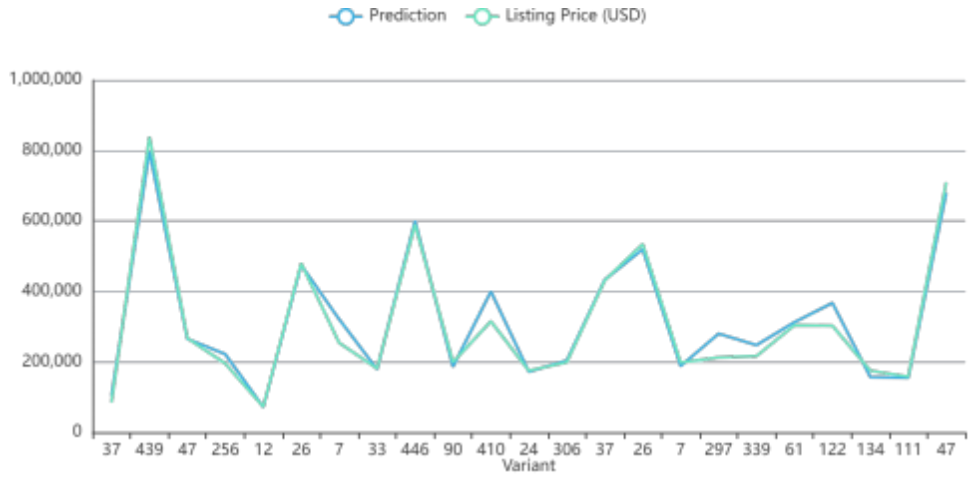

Fig. 2 Predicted and actual fitted curves

In addition to high-accuracy prediction results, our model has good generalization capability to cope with price prediction in various variant cases. We have developed our model to take into account the effects of various characteristics of sailboats, resulting in a comprehensive and accurate model.

4. Measuring the impact of price on price

4.1. One-way analysis of variance (ANOVA)

In this section, we first did a variance analysis of the effect of region on price to see if there is a significant difference between different regions and to determine if the region affects the price. We did a correlation analysis between regional and sailboat prices based on a one-way analysis of variance (ANOVA) [4].

Sailboat prices were grouped according to the region and tested for normality to see if they follow a normal distribution in general. The normality test and histogram results are shown in the Table 3 and Figure 3 below.

Table 3. Model evaluation results

Variate	Size	Median	Mean	Standard deviation	Skewness	Kurtosis	S-W	K-S
price	3427	241725	289366.01	164939.944	1.057	0.8	0.912	0.119

We applied the Shapiro-Wilk test to the data on sailboat prices. We can see that its significance level is highly significant, rejecting the original hypothesis (that the data meet the normal distribution), so the data do not meet the normal distribution. However, its kurtosis is less than ten, and skewness is less than 3. We can further analyze it to get that it basically meets the normality test.

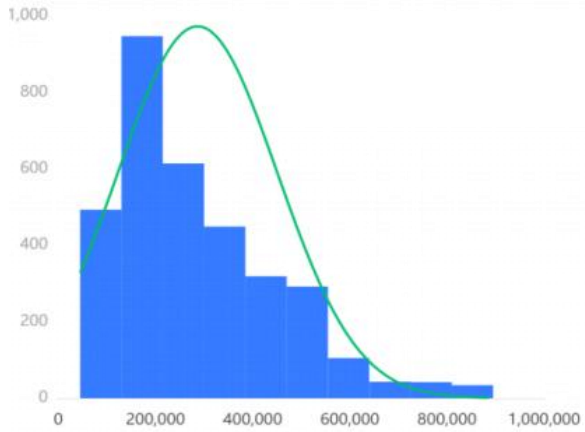

Fig. 3 Two or more references

The above Figure 3 shows the normality test results for sailboat prices (in US dollars). The histogram shows high middle and low sides, indicating that sailboat prices generally obey a normal distribution and satisfy the normality test.

Sailboat prices were grouped according to regions, and the chi-square test was conducted to determine if the p-value was less than 0.05 and if the chi-square was satisfied.

According to the results of the standard deviation, F-test results, and significance P-value obtained from our analysis, the significance P-value was 0.000***, a highly significant level, so the original hypothesis was rejected (i.e., variance chi-squareness was satisfied), and the sailboat price did not satisfy variance chi-squareness. However, in the actual analysis process, many results do not satisfy the chi-squareness and do not require rigorous use of one-way ANOVA. Here we can use one-way ANOVA to analyze regional and sailboat prices.

A one-way ANOVA was conducted to determine the significance and conclude whether the region affects the price of sailboats.

By analyzing the means of the prices, we compare the mean values, and we can get that their mean values keep changing with the region and have more significant differences.

We performed an ANOVA on regional and sailboat prices and obtained results including the mean ± standard deviation, the results of the F-test, and the results of the significant level P-value, where the F-value was 3.413; the P-value was 0.000***0.05, and the sta- tistical results were significantly different,i.e., the prices of sailboats differed significantly from one region to another.

Thus, we conclude that region has a significant effect on the price of sailboats, which varies from region to region, and that the price of sailboats in economically developed regions is higher relative to those in less developed regions, which we speculate may be due to the disparity in income and consumption levels among residents in different regions.

4.2. Consistency test for regional effects

The level of craftsmanship and production standards can vary significantly from region to region due to the different levels of development, so the characteristics we collected may differ significantly between the sailing variants. However, some of them may also be consistent. Therefore, the question asks whether there are regional effects that are consistent among our sailboat variants.

To discuss regional consistency, this requires us to conduct separate one-way analyses of each characteristic of the sailing variant between different regions to determine whether each characteristic is significantly different, and if there is a significant difference, the regional effect is not consistent; if there is no difference or the difference that exists is small, then it can be shown to be consistent [5].

Table 4. Results of the Quantitative Analysis of Effects

Element	Inter-group difference	Total deviation	Partial η2	Conhen's f-values
Price (USD)	6.42311E+12	8.98196E+13	0.072	0.278
Draught(m)	65.276	642.778	0.102	0.336
Width(m)	1510.698	9335.07	0.162	0.439
Displacement(kg)	1307104940	42347982031	0.031	0.178
Sail area(m^2)	185465.826	5684593.679	0.033	0.184
Clearance(m)	5538588.921	52028464.13	0.106	0.345

The above Table 4 shows the results of the quantitative analysis of effects, including between-group differences, total differences, Partial η^2, and Conhen's f values. We can analyze the differences between the databased on these results. Additionally, we can quantify the difference by combining the Partial η^2, and Conhen's f-values.

In which the Partial bias η^2: between 0 and 1, the larger the value, the larger the magnitude of the difference. For example, Partial η^2 is 0.1, which means that 10% of the difference in the data is from the difference between different groups. In general, Partial η^2 is very small. When

using Partial η^2 to indicate a large effect size, the threshold points for differentiating between small, medium and large effect sizes are: 0.01, 0.06 and 0.14, respectively.

Cohen's f-value indicates the effect size, and the threshold points for differentiation of small, medium and large effect sizes are: 0.1, 0.25 and 0.40, respectively.

Our analysis of effect sizes shows that based on price, Partial η^2 is 0.072, indicating that 7.2% of the variation in the data comes from differences between groups. Cohen's f value of 0.336 indicates a moderate degree of variation in the quantified effects of the data; based on width, Partial η^2 is 0.162, indicating that 16.2% of the variation in the data originates from differences between groups. Cohen's f value of 0.439 indicates that the degree of variation in the quantified effects of the data is large; based on displacement, Partial η^2 is 0.031, indicating that 3.1% of the variation in the data is from differences between groups. Cohen's f value of 0.178 indicates that the degree of variation in the quantification of the effect of the data is small; based on the Sail area, Partial η^2 is 0.033, indicating that 3.3% of the variation in the data originates from the variation between groups. Cohen's f value is 0.345, which indicates that the quantified effect of the data is a small degree of variation; based on Clearance, Partial η^2 is 0.106, which indicates that 10.6% of the variation of the data is from the difference between groups.

Therefore, from the above results, there are fewer differences in the two factors of discharge and sail area, i.e., these two characteristics can be considered consistent regional effects.

5. Summary

This paper discusses the impact of various attributes related to sailboats on their selling prices. We quantify the extent to which various attributes affect the selling price of a sailboat, with length, brand, and GDP having a significant effect on the selling price of a sailboat, with a characteristic importance of 33.4%, 28.2%, and 9.6%, respectively.

In the analysis of the area effect, we clarified that the area effect has a significant impact on the selling price of sailboats, and the degree of impact is very significant. Based on the results of the previous question, we initially concluded that the area effect acts on the selling price of sailboats through GDP. We then tested the consistency of the area effect on sailboat attributes. We conclude that the area effect is consistent for displacement and sail area, while it differs for the other attributes.

References

[1] The ocean. american sailing market [J].Boat, 2008, No.291 (05) : 46-51.

[2] Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System [J]. ACM, 2016.

[3] Jingjuan Guo , Yaowei Liu . Research on highway project estimate prediction method based on XGBoost [J]. Highway and Transportation Science and Technology,2023,40(3):58-68.

[4] Dai Jinhui, Yuan Jing. Comparison of one-way ANOVA and multiple linear regression analysis [J]. Statistics and Decision Making, 2016, No. 453 (09) : 23-26.

[5] Song Jabang, Yu Haiyang, Wang Songchen, et al. Consistency test method for stratigraphic oil and gas high temperature phase experiment [J]. Special oil and gas reservoirs,2023,30(1):93-99.