Prediction of The Price of Second-hand Sailboat Based on XGboost Regression Model

Jiuran Nie*

*School of Water Conservancy and Civil Engineering, Zhengzhou University, Zhengzhou, Henan Province, China

*13500872575@139.com

Abstract. With the development of economy and society, the flow of second-hand sailboats is increasing day by day. The sales market of second-hand sailboats has great potential. Studying the factors that affect the pricing of second-hand sailboats will help merchants to formulate accurate marketing strategies and obtain rich commercial profits. A comprehensive comparison of used sailboat price models developed by the random forest regression model and the XGBoost regression model using MSE, RMSE, MAE, and MAPE based on pre-processing and correlation analysis of advertising data and relevant supplemental data for approximately 3,500 sailboats of 36 to 56 feet in length sold in Europe, the Caribbean, and the United States in December 2020, concluded that The XGBoost regression model performs best in this problem. Based on the XGBoost regression model, the analysis focuses on the influence of region on used sailboat prices and makes a specific analysis of used sailboat prices in Hong Kong (SAR), with a view to providing a reference for pricing issues in the used sailboat market in Hong Kong.

Keywords: Second- hand sailboat pricing, XGboost regression, Random forest

1 Introduction

Sailboats are being used in a wide range of applications, from the most basic transportation and fishing operations to sports and recreation. The sailing industry will become the main body of the development of marine economy and develop rapidly[1]. With the development of social economy, the huge market of sailing boats is immeasurable. Pricing is generally regarded as one of the key factors to determine whether an industry can develop in the long run. Sailboats are often sold through brokers, but the complexity and opacity of the pricing system makes it inevitably confusing to set a price for a used sailboat. Therefore, studying the pricing of used sailboats is beneficial in enhancing brokers' understanding of the market value of used sailboats and helping them to make better pricing decisions. However, there is very little literature on the study of used sailboat pricing issues, making it even more important to study the pricing of used sailboats.

A. Rauf et al. (eds.), *Proceedings of the 3rd International Conference on Management Science and Software Engineering (ICMSSE 2023)*, Atlantis Highlights in Engineering 20,
https://doi.org/10.2991/978-94-6463-262-0_25

The price of a used sailboat is influenced by various factors, such as the characteristics of the sailboat itself, regional factors, and so on. The problem of setting prices for used sailboats can be transformed into a prediction problem. With the wide application of machine learning, more and more scholars have applied modern prediction methods to price prediction and their good performance in prediction is generally recognized, such as decision tree prediction[2], neural network forecasting[3], support vector machine forecasting[4], and deep learning forecasting[5], etc. XGBoost was proposed by Chen in 2016 and demonstrated the low computational complexity, fast running speed, and high accuracy of its model[6]. Therefore, using XGBoost model for used sailboat price prediction can not only improve the prediction accuracy but also increase the prediction rate, which can not only fill the gaps in the existing literature but also help brokers to set better prices for used sailboats.

2 Data and methodology

2.1 Data source

Data on monohulled sailboats and catamarans are from the website SailboatData.Com and economic data by year and by region are from The World Bank. Comparable listing price data for the Hong Kong (SAR) market are from https://sailing.org.hk/zh-hant, and cargo throughput data from Global Economic Data.

2.2 Data preprocessing

Data preprocessing, as an indispensable part of the machine learning application process, is linked to the good or bad conclusion. Among the known data and the collected and supplemented data, there are issues of missing values and varying numbers of each feature domain. Outlier cleaning, missing value filling, centering and normalization of the data are of great importance.

2.2.1.Potential outlier detection based on box line diagram.

When there are outliers in the data, especially when there are outliers with large deviations, it will bring errors to the data analysis and modeling. In this paper, the distribution of the data is not uniform, and the distribution of a part of numerical features does not conform to normal distribution, so the box line plot is chosen for the outlier detection of the numerical features,which has no requirement on the data distribution.

2.2.2.Missing and duplicate values processing.

Missing and duplicate values in the data are found by filtering the data. For missing values, the most frequent value of the attribute where the missing object is located is used to fill in the missing values according to the principle of plurality in statistics. For a small number of duplicate values, the deletion process is done.

2.2.3.Data Centricity and Normalization.

It is necessary to centralize and standardize the data before modeling and analysis. Centralization of data can increase the orthogonality of basis vectors, and different features can have the same scale through standardization.

$$x' = x - \bar{x} \tag{1}$$

$$x'' = \frac{x-\bar{x}}{\sigma} \tag{2}$$

2.3 Correlation Analysis

In this paper, a multivariate descriptive statistical analysis is conducted to draw heat maps. The correlations between the price of sailboats and the different variables affecting the price of sailboats were explored, and then the multicollinearity between the independent variables was explored. Heat map, also known as correlation coefficient map, can determine the magnitude of correlation between variables based on the magnitude of correlation coefficients corresponding to the colors of different squares in the heat map.The formula for calculating the correlation coefficient between two variables is:

$$\rho_{x_1x_2} = \frac{Cov(X_1, X_2)}{\sqrt{DX_1 \cdot DX_2}} = \frac{E(X_1X_2) - EX_1 \cdot EX_2}{\sqrt{DX_1 \cdot DX_2}} \tag{3}$$

2.4 XGboost regression model building

XGboost is an integrated learning algorithm based on Boosting. The model has good fault tolerance for data sets with a small number of missing values, and can automatically learn the splitting direction of the decision tree by sparse perception algorithm. In this paper, decision tree is taken as the base learner.

Constructing the loss function and the objective function. By constructing the loss function, the deviation and variance of the model are reduced. Reducing the deviation is to reduce the error between the predicted result of the model and the real value. Reducing variance can be regarded as preventing over-fitting, which is generally achieved by introducing regularization terms into the model and reducing the complexity of the model. The objective function consists of loss function and regularization term, and the calculation formula is as follows:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) \tag{4}$$

Because Boosting follows the forward distribution addition, the predicted value of each step is determined by the predicted value of the previous step. So by superposition, the final expression of the objective function can be expressed as follows:

$$Obj = \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_i) \right] + const \tag{5}$$

Taylor expansion approximation of the objective function. Because the objective function of XGboost is too complex to be solved directly, Taylor polynomial is used to approximate the objective function here.

$$Obj^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{6}$$

Decision tree generation. The complexity of decision tree is determined by the number of leaves $T$and the weight of leaf nodes ω. Next, by scoring the structure of the tree, the greedy algorithm is used to find the splitting income of the tree, and the objective function after tree splitting is obtained as follows:

$$Obj = -\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda}\right] + 2\gamma \tag{7}$$

2.5 Random forest regression model

Random forest is an integrated learning method using decision trees as the base learner to improve the accuracy of prediction by integrating the prediction results of multiple trees7. Its generalization ability is high and it is able to measure the importance of explanatory variables.

Creating a random tree. In the process of constructing decision tree, by setting an expected error reduction value as the splitting attribute of tree nodes, it is set that when the standard deviation is less than 5%, the tree will stop splitting. The standard deviation reduction value is calculated as follows:

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} sd(T_i) \tag{8}$$

Select split property. In the binary classification problem, the random forest integrates multiple CART binary trees. Using the Gini index as the partitioning criterion of the CART algorithm, attributes are selected at each internal node of the decision tree, which is calculated as:

$$\begin{cases} Gini(S) = 1 - \sum_{i=0}^{c-1} P_i^2 \\ P_i = \frac{s_i}{S} \end{cases} \tag{9}$$

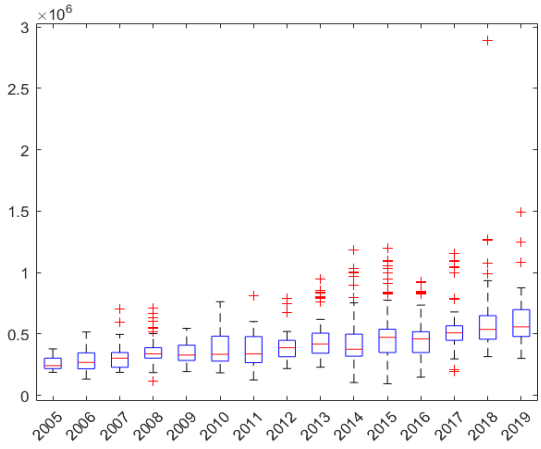
Pruning and integration tree. After the decision tree is built, some branches of the tree can only capture few samples, so they need to be subtracted. Integrating multiple decision trees can effectively avoid the instability of a single decision tree.

3 Results and discussion

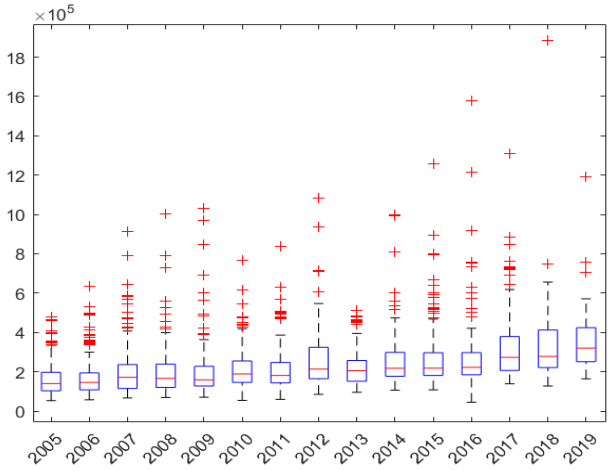
3.1 Box plot and analysis.

The box line plot visualizes the median, upper quartile, lower quartile, upper and lower edges, and potential outliers of the variable data. In figure 1, the line in the middle part indicates the median, the upper and lower edges of the box indicate the upper quartile

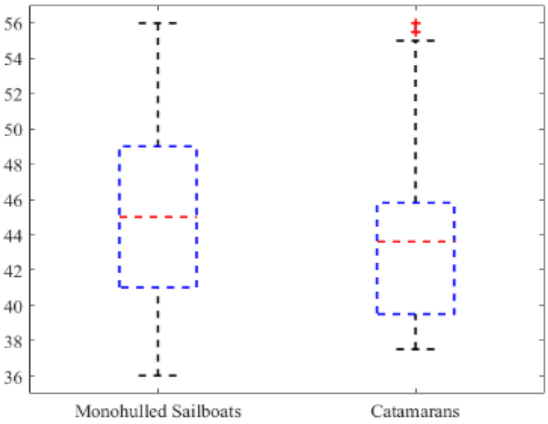
and lower quartile, respectively, and the upper and lower two black lines indicate the upper and lower edges; data points outside this range are outliers and are removed.



(a) Price data of monohulled sailboats.



(b) Price data of catamaran sailboats.

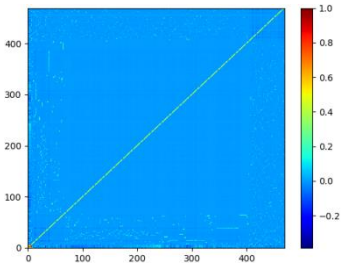


(c) Length data for the two sailboats.

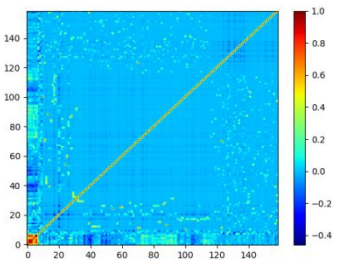
Fig. 1. Box diagram of price and length data for two models of sailboats.

3.2 Results of correlation analysis

Displacement, Beam, LWL (Low water level), Sail Area, Draft, Length , Year, GDP, Average cargo throughput, Average ratio of total logistics costs to GDP, and GDP per capita data collected were used as factors affecting prices. The two heat maps in figure 2 show the results of separate analyses for monohull and catamaran sailboats. Different correlation coefficients correspond to different colors. It can be seen that the linear correlation between the factors influencing the price of both monohull sailboats and catamarans is not high.



(a) Factors influencing the price of monohulled sailboats.



(b) Factors influencing the price of catamaran sailboats.

Fig. 2. Plot of correlation coefficients between factors influencing the price of sailboats.

3.3 Comparison and evaluation of the model

Price data for sailboats and data on sailboat characteristics including boat length, beam, draft, displacement, and sail area were used to analyze the effect of sailboat characteristics on price. MSE, RMSE, MAE, MAPE, and R^2 indicators were used to evaluate the regression results of the random forest model and the XGBoost regression model. The smaller the MSE, RMSE, MAE and MAPE values and the closer the R^2 value is to 1, the higher the accuracy of the model. Combining table 1 and table 2, the comparative analysis reveals that the XGBoost regression model performs best in this problem. Therefore, the XGBoost regression model is chosen to discuss the pricing issue of used sailboats.

Table 1. Evaluation of random forest model for the effect of sailboat characteristics on prices.

Categories	Sets	MSE	RMSE	MAE	MAPE	R^2
Monohulled Sailboats	Training set	0.144	0.379	0.296	226.641	0.846
	Test set	0.415	0.644	0.431	392.9	0.632
Catamarans	Training set	0.062	0.250	0.197	523.014	0.943
	Test set	0.224	0.474	0.351	1406.932	0.613

Table 2. Evaluation of XGBoost model for the effect of sailboat characteristics on price.

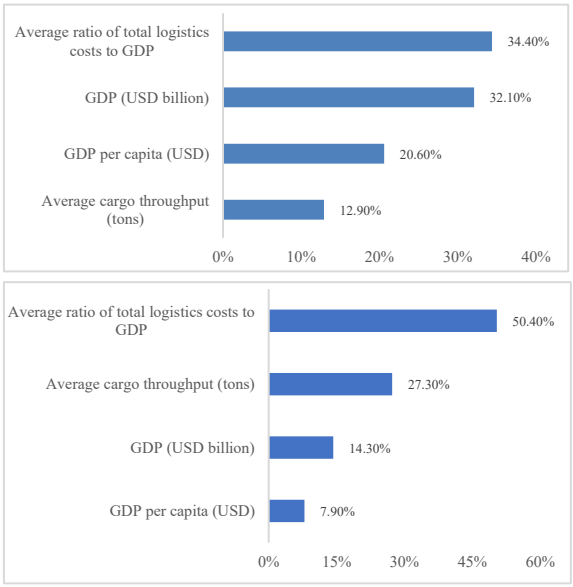
Categories	Sets	MSE	RMSE	MAE	MAPE	R^2
Monohulled Sailboats	Training set	0.005	0.072	0.054	47.310	0.994
	Test set	0.386	0.622	0.402	281.520	0.658
Catamarans	Training set	0.001	0.027	0.015	9.389	0.999
	Test set	0.273	0.523	0.388	111.659	0.651

3.4 Analysis of the impact of region on the price of used sailboats

To analyze the influence of regional factors on the price of used sailboats, based on the XGboost regression model, we focused on the influence of regional GDP, GDP per capita, Average ratio of total logistics costs to GDP, and Average cargo throughput on the price of second-hand sailboats. Separate analyses were conducted for monohulled sailboats and catamarans. Data of 70% of the prices for monohulled sailboats and catamarans and their influencing factors were used as training samples, respectively, and calculated using the XGboost model.

The results in figure 3 indicate that region has an effect on the listing price of sailboats, while the effect on the listing price of monohulled sailboats and catamarans is not consistent. The similarity is that the most significant factor affecting the price of both monohull and catamaran sailboats is Average ratio of total logistics costs to GDP;

the other more significant factors affecting the price of monohull sailboats are GDP, GDP per capita and Average cargo throughput, in that order; and for catamaran sailboats are Average cargo throughput, GDP, GDP per capita, in that order.



(a) Monohulled Sailboats

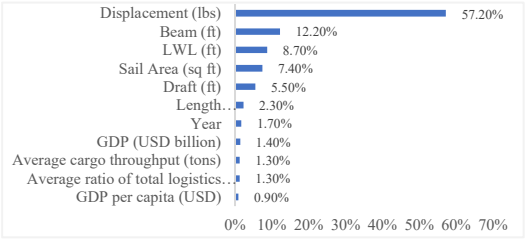
(b) Catamaran boats

Fig. 3. Order of the importance of regional factors affecting the price of sailboats.

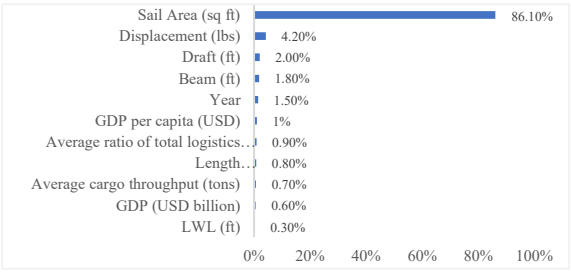
3.5 Analysis of used sailboat prices in Hong Kong (SAR)

The used sailboat market in Hong Kong (SAR) is analyzed by adding comparable listing price data from the Hong Kong (SAR) market. The data of 70% of monohull sailboats and 70% of catamarans were entered into the XGboost model as training samples for calculation, respectively, and then the factors influencing the prices of both sailboats in Hong Kong (SAR) were analyzed and predicted.

The results (Figure 4) illustrate that in Hong Kong (SAR), the most important factor influencing the price of monohulled sailboats is Displacement, followed by data on hull characteristics, such as Beam, Low water level, Sail area, Draft, and Length, while regional factors such as GDP, Average cargo throughput, Average ratio of total logistics costs to GDP, and GDP per capita have less influence on monohulled sailboats. The most important factor affecting the price of catamarans is Sail area. And regional factors have less influence on the listed price of catamarans than monohulled sailboats.



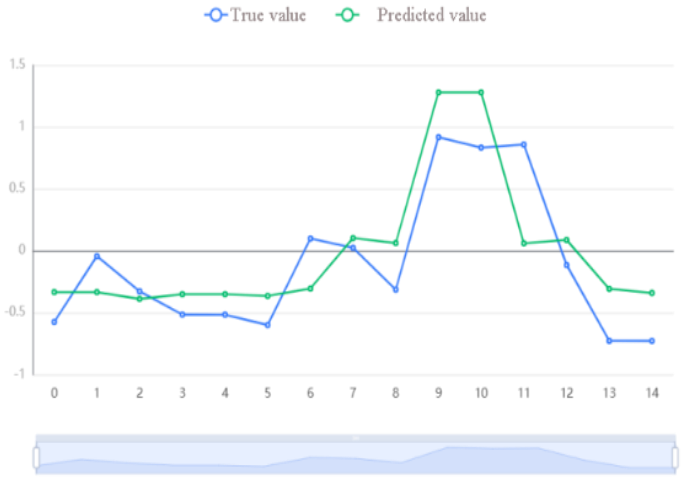
(a) Monohulled sailboats



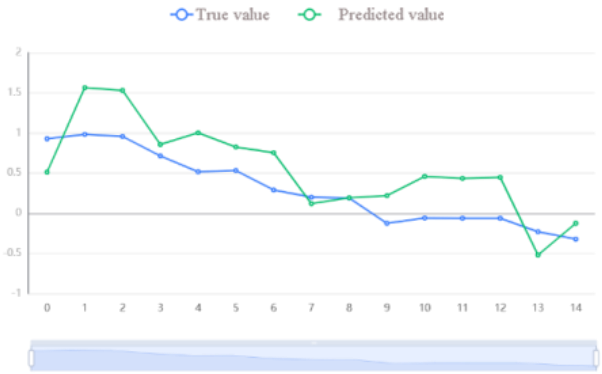
(b) Catamaran boats

Fig. 4. Order of the importance of factors affecting the price of two types sailboats.

The fitting of the predicted and true values in the test set is shown in figure 5. The predicted values are closer to the true values, indicating that the errors are small and the model fits well.



(a) Monohulled sailboats



(b) Catamaran boats

Fig. 5. Real values and predicted trends for the price of two types sailboats.

4 Conclusion

The descriptive statistics of the data show that the price of monohull is higher than the price of catamaran. As shown by the XGboost regression results, there is a significant difference in the importance of the effect of Average cargo throughput and GDP per capita on the price of the two types of sailboats. The influence of Average cargo throughput on the price of catamaran is significantly higher than that of Average cargo throughput on the price of catamarans is significantly higher than that of monohulled sailboats, while the influence of GDP per capita on the price of monohulled sailboats is significantly higher than that of catamarans. Since catamarans are much wider than monohulls, their stability is significantly better than that of monohulls and they have the ability to withstand larger wind and waves, catamarans may be used more for cargo transportation and are therefore more influenced by the factor of Average cargo throughput. In Hong Kong (SAR), the most important factor affecting the price of monohulled sailboats is Displacement, followed by vessel characteristics such as Beam, Low water level, Sail area, Draft and Length, while regional factors have less influence on monohulled sailboats. The most important factor affecting the price of catamarans is Sail area. And regional factors have less influence on the listed price of catamarans than monohulled sailboats.

References

1. Zhang Y 2019 Research on Design of 110-Feet Catamaran Sailboat *D. Dalian University of Technology*. https//doi.org/10.26991/d.cnki.gdllu.2019.003269.
2. Panigrahi S S and Mantri J K 2015 A text based decision tree model for stock market forecastin *Proc. Int. Conf. on Green Computing and Internet of Things (ICGCIoT)* (Galgotias Educ Inst, Greater Noida, India OCT, 08-10, 2015) p 405-11

3. Ticknor J L 2013. A Bayesian regularized artificial neural network for stock market forecasting *Expert Systems with Applications* J **40** 5501-06.
4. Kong F, SONG G P 2010 Stock price combination forecast model based on regression analysis and SVM *Applied Mechanics & Materials* J **39** 14-18.
5. Gao T W, Li X, CHAI Y T and Tang Y H 2017 Deep learning with stock indicators and two-dimensional principal component analysis for closing price prediction system *Proc. Int. Conf. on Software Engineering and Service Science* (China Hall Sci & Technol, Beijing, PEOPLES R CHINA AUG, 26-28, 2016) p 166-9
6. Chen T and GUESTRIN C 2016 XGBoost：a scalable tree boosting system *ACM SIGKDD Int. Conf.on Knowledge Discovery & Data Mining* (San Francisco, CA AUG, 13-17, 2016) p 785-94
7. Friedman J, Hastie T, Tibshirani R 2001 *The Elements of Statistical Learning* (New York: Springer series in statistics)