Application Of Multiple Regression Algorithm to Price Prediction Model: Taking Sailboat Price as An Example

Jingyi Yang[1, *, #], Haitong Liu[2, #]

[1]International business college, South China Normal University, Foshan, Guangdong, 528225

[2]College of physics, Jilin University, Changchun, Jilin, 130012

*Corresponding author: jingyiyang23@outlook.com

[#]These authors contributed equally

Abstract. The sailboat market is becoming more and more mature, but market price prediction is rarely studied. Therefore, on the basis of considering the characteristics of sailboats themselves, this paper also considers the influence of regional effects on sailboat prices from three aspects: taxation, interest rate and economic conditions. Multiple linear regression model was used to fit the data, and the weight size of each index was analyzed to study the impact of each index on sailboat price. The results show that: (1) the MAE of the monolithic sailboats in the test set is 8478.0. However, the MAE of catamaran is 10,868.4. The fitting results are good, indicating that the multiple linear regression model is suitable for predicting sailboat prices. (2) The weight of the two indicators, tax change rate and GDP per capita, is higher in monolithic sailboats. The weight of the expected growth rate is high but negative, indicating that monolithic sailboats are greatly affected by regional economic factors. (3) Among catamarans, only the GDP per capita is larger than that of monolithic sailboats. The effect of the rate of tax change on catamaran prices is almost negligible. The expected growth rate, although negative, is much smaller than monolithic sailboats. (4) The monolithic sailboats is much more affected by region than the catamaran. The results can provide guidance for the simulation and prediction of sailboat market prices.

Keywords: Sailboats; Price forecasting; Multiple regression algorithm; Prediction model; Regional effect.

1. Introduction

Sailboat, as a kind of ship that uses wind power as forward power, is different from yacht, which mainly relies on engine gears for power. Sailing boats are mainly used for sports and entertainment. In coastal regions such as the United States, Europe, and the Caribbean, used sailboats are more popular in the market. However, the market value of a sailboat is not only affected by its own value, but may also be affected by various external factors such as economy, environment and policy.

Although there are few studies on the valuation methods of used sailboats, the pricing of goods and services has something in common. We believe that the pricing of used sailboats is highly relevant to the pricing of used cars and use it as a reference for our sailboat pricing model. The reasons are as follows: (1) Their attributes are all transportation vehicles; (2) When consumers consider the purchase of cars and sailboats, they will fully consider the performance of products, such as size, material and engine performance. (3) The academic research on the pricing of used car market is relatively mature.

In recent years, the research on used car price evaluation methods has been innovated to use machine learning to evaluate products, and the methods are relatively diverse [1-2]. Pudaruth [3] predicts the market price of used cars in Mauritius using four methods: multiple linear regression, k-nearest neighbor value, naive Bayes and decision tree, respectively. Other scholars also uses neural network and deep learning models to predict car prices [4-6]. The scholars use the collected data to let machines learn and evaluate prices, and the results are impressive and accurate. It can be said that the introduction of machine learning has refined the price evaluation model and enabled people to price used cars more accurately.

One of the most frequently used methods in forecasting and estimation is linear regression, which is a traditional method in predicting used car prices. It can show the relationship between dependent

factors and random factors [7]. Alhakamy [8] uses big data resources to collect the brand, model, mileage and other characteristics of used cars to predict their prices. The results show that linear regression model is good at predicting used prices and supports sustainability. Brahimi et al. [9] use text mining technology to extract and compile the description information of used cars in the Arab market. They use different models to predict the price of cars, and the results show that linear regression is the most effective prediction method.

The prediction of used sailboat prices is still a blank in the academic field. We will start from a traditional forecasting model, the linear regression model to simulate and forecast the price of sailboats. We wish more improvements will be made on this traditional model, if it is necessary.

2. Model Preparation

2.1 Data Description And Processing

In this paper, data on six characteristics, which are length, manufacturer, year, beam, draft and headroom of monolithic sailboats and catamaran are collected from used sailboat websites. Then the data collected are processed for missing values and cleaned. The characteristic variables of sailboats and their explanations are shown in Table 1.

Table 1: characteristics of sailboats

Variable	Description	Unit
Length	The length of the boat in feet.	foot
Manufacturer	The name of the manufacturer of the boat.	/
Year	The year the boat was manufactured.	/
Beam	The height available to stand up in the cabin.	foot
Draft	The minimum depth of water required to float a boat without touching the bottom.	foot
Headroom	The height available to stand up in the cabin.	meter

Table 2: Data source collection

Data Name	Concrete Index	Data Sources	The amount of data collected
characteristics of sailboats	Length	https://www.yachtworld.com/	2585
	Manufacturer		2585
	Year		2585
	Beam		2585
	Draft		2585
	Headroom		2585
Economic data	per capita GDP in 2020	https://data.oecd.org/	28
	Tax on goods and services(%GDP) from 2019 to 2020		44
	Short-term interest rate forecast from 2019 to 2020		30

In addition, the impact of tax, interest rate and economic situation on the sailing price is considered in order to measure the regional effect. The sources of all concrete indexes are shown in Table 2. Specifically, the following three indicators are rate of change in taxes, expected rate of interest growth and GDP per capita. The calculation formula is as follows:

$$RCT = \frac{\tau_t/GDP_t - \tau_{t-1}/GDP_{t-1}}{\tau_{t-1}/GDP_{t-1}} \quad (1)$$

In this formula, RCT denotes the Rate of change in taxes, that is, the change in the share of taxes on goods and services in GNP, which is the gross national product, and *t* denotes the year. It is able to measure the tax changes in a country or region. We argue that when taxes change, the demand for used sailboats also changes, which affects its price. Moreover, the tax changes of each country in the same period are different, with regional heterogeneity, so the Rate of change in taxes can be used to measure the regional effect.

$$ERI = \frac{r_{E_t} - r_{E_{t-1}}}{r_{E_{t-1}}} \quad (2)$$

In this formula, ERI denotes the expected rate of interest growth, which measures the change in the expected rate of return on bonds in the financial market. Let r_{E_t} denotes the expected short rate, where short means three months and t is the year. When changed, it means that the expected interest rate has changed. According to modern portfolio theory [10], with other conditions unchanged, the increase in the expected return rate of an asset relative to other assets increases the demand for the asset. As sailboats are real assets, when the expected interest rate in the bond market changes, the demand for sailboats will also change, and so will the price of sailboats. Each country has different expected interest rate changes over the same period and has regional heterogeneity, so it can be used to measure regional effects.

$$PGP = \frac{real\ GDP}{Poputation} \quad (3)$$

In this formula, PGP denotes the GDP per capita, which is calculated in this paper by dividing the real GDP of a country or region in a year by the permanent population of the country. It can be used to understand the macroeconomic operation of a country or region. Different countries or regions have different macroeconomic conditions, which can be used to measure their regional effects.

2.2 Model Establishment

One of the most commonly used methods to analyze the relationship between multiple variables is regression analysis. It attempts to fit a straight line hyperplane to the data set. This procedure can be performed under a variety of conditions and can reveal the effect of the variable on the dependent variable. Regression analysis can also be used to predict an unknown dependent variable based on the independent variables.

The price of a sailboat is composed of many factors, each of which has a different degree of influence on the price. We use the multiple linear regression model to express the price of sailboats as:

$$Y = (\beta_1\alpha_1 + \beta_2\alpha_2 + \ldots + \beta_i\alpha_i) + c \quad (4)$$

Where α_i is the various factors, β_i is the weight assigned to each factor according to the different degree of impact on sailboat prices, and *c* is a constant.

To express this relationship more intuitively, we define two matrices *X*, *m*.

$$X = \begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1i} \\ \vdots & \ddots & \vdots \\ \alpha_{n1} & \cdots & \delta_{ni} \end{bmatrix} \quad (5)$$

Where X is the matrix composed of various factors affecting sailboat prices. i represents the number of factors, and n is the number of data for each factor.

m is a coefficient vector composed of weights β_i:

$$m = [\beta_1 \quad \beta_2 \quad \dots \quad \beta_i] \tag{6}$$

So, the price of a sailboat can be expressed as:

$$Y = mX + c \tag{7}$$

2.3 Precision of Our Estimate

To derive the best value for the corresponding weight of each factor, we introduce the root mean square error RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}} \tag{8}$$

When the root mean square error is minimum, we believe that the weight of each factor reaches the optimal value.

80% of the data set is used as the training set to fit the model, and the rest of the data is used for testing. For the accuracy verification of the model, root mean square error (RMSE) and mean absolute error (MAE) are used to reflect the accuracy of the model

The mean absolute error can be expressed as:

$$MAE = \frac{\sum_1^i |y - y|}{i} \tag{9}$$

3. Result Analysis

3.1 Price Simulation And Prediction

For monolithic sailboats, the effect diagram of model training and testing is shown in Figure 1 and Figure 2:

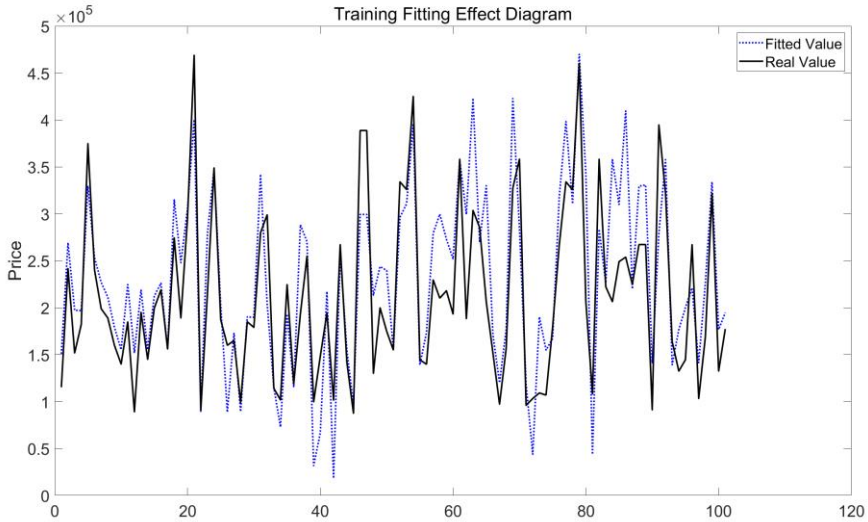

Figure 1 Model training result of monolithic sailboats

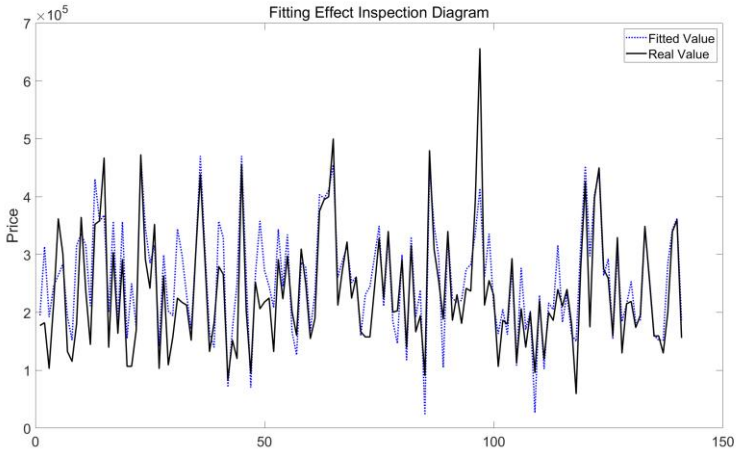

Figure 2 Model testing result of monolithic sailboats

According to the fitting results, the RMSE is 24674.17 in the training set and 44715.78 in the test set, the MAE is 8478.0, that is, the average error between the fitted value and the actual value is $8478, indicating a good model fitting effect.

For the weight of each factor, the fitting results of the model are shown in Table 3:

Table 3: Fitting results of the model for monolithic sailboats

	Length	Year	Beam	Draft	headroom
Weight	0.41	0.31	0.26	0.19	-0.13
	RCT	ERI	PGP	ATV	
Weight	0.16	-0.15	0.25	0.18	

For catamarans, the effect diagram of model training and testing is shown in Figure 3 and Figure 4:

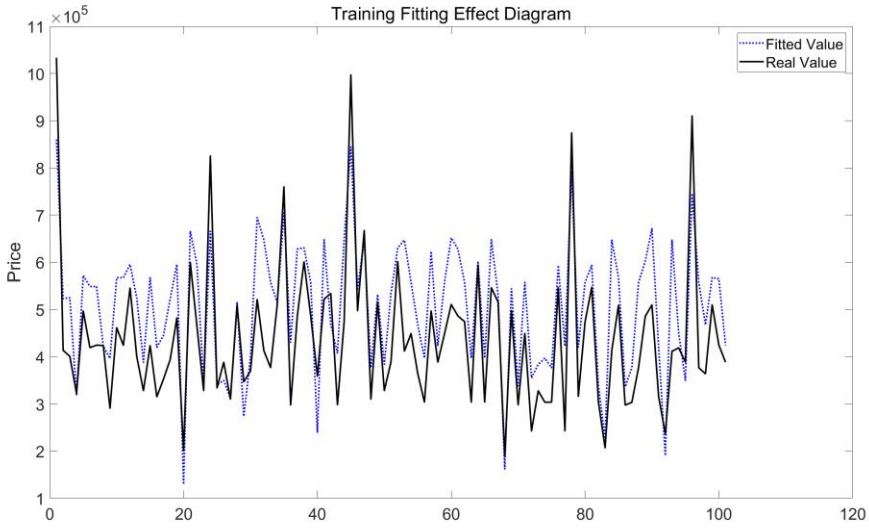

Figure 3 Model training result of monolithic catamarans

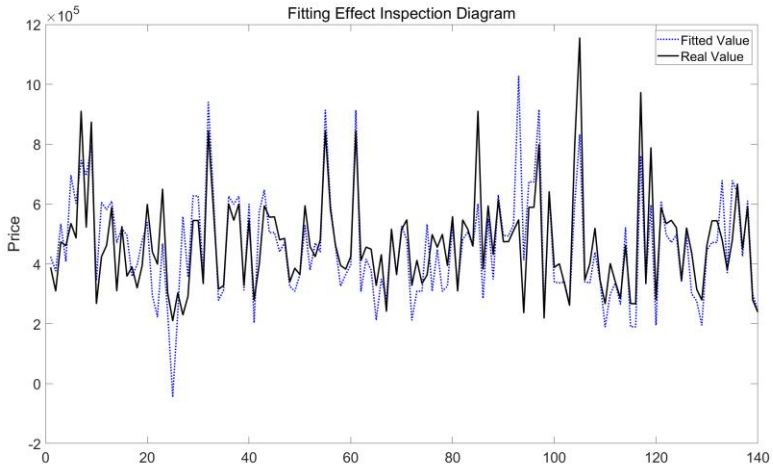

Figure 4 Model testing result of monolithic catamarans

According to the fitting results, the RMSE is 23179.23 in the training set and 46753.19 in the test set, the MAE is 10868.4, that is, the average error between the fitted value and the actual value is $10868, indicating a good model fitting effect.

For the weight of each factor, the fitting results of the model are shown in Table 4:

Table 4: Fitting results of the model for catamarans

	Length	Year	Beam	Draft	headroom
Weight	0.45	0.25	-0.04	0.20	-0.09
	RCT	ERI	PGP	ATV	
Weight	-0.01	-0.08	0.14	0.12	

According to the model fitting results of the two types of sailboats, the weight of length and draft is relatively large, and weight of length in the two types of sailboats is the largest. These two values largely represent the model of sailboats. Therefore, the model of sailboats is the most important factor affecting the price of sailboat, and the larger the model is, the higher the price of sailboats is. The weight of year is the second in both monolithic sailboats and catamarans, which is positive, indicating that the newer the sailboats are, the higher the price is. The weight of headroom is negative in both types of boats, which is counter-intuitive. But observing the distribution of the data, it can be found that as the headroom increases. The price of the sailboat first rises and then falls, indicating that when the boats are small, more space will increase the price of the sailboats. But when the boats are too large, the price of the sailboats will decrease. It shows that space utilization is very likely to be an important factor in determining the price of sailboats. Average transaction volume also accounts for a considerable weight in both types of sailboats, indicating that the manufacturer of sailboats also has a great influence on the price of sailboats. At the same time, the weight of the three selected indicators reflecting regional factors has reached a certain level, indicating that the region where sailboats are sold will also affect the price of sailboats.

3.2 Regional Effect

In order to study the regional effect of sailboat prices, we selected three indicators RCT, ERI, and PGP. According to the weight of these three indicators in the multiple linear regression, the influence of regional effect on the sailboat price is studied. The weights of these three indicators in the two types of sailboats are shown in Table 5:

Table 5: The weights of these three indicators in the two types of sailboats

	RCT	ERI	PGP
Catamarans	-0.01	-0.08	0.14
Monolithic sailboats	0.16	-0.15	0.25

We find that the weights of the two indicators, RCT and PGP, are both high and positive, and the absolute weights of ERI are also high but negative in monolithic sailboats. It shows that the monolithic sailboats is greatly affected by regional economic factors. Moreover, the larger absolute weight of ERI indicates that people may use sailboats as a store of value.

Among catamarans, however, only PGP has a large weight. The effect of RCT on catamaran prices is almost negligible. While the weight of ERI is negative as in the case of monolithic sailboats but the value is much smaller than that of monolithic sailboats. This indicates that catamarans are less affected by regional economic factors.

Monolithic sailboats differ considerably from catamaran sailboats in the extent to which they are affected by territorial factors, so we do not expect regional effects to be consistent across any sailboat variant.

4. Conclusion

On the basis of considering the characteristics of sailboats themselves, this paper also considers the influence of regional effect on sailboat price from three aspects: tax, interest rate and economic condition. Multiple linear regression model was used to fit the data, and the weight size of each index was analyzed to study the impact of each index on sailboat price. The results show that: rate of change in taxes, expected rate of interest growth and GDP per capita

(1) The MAE of monolithic sailboats in the test set is 8478.0. However, the MAE of catamarans is 10,868.4. The fitting results are excellent, indicating that the multiple linear regression model is suitable for predicting sailboat prices.

(2) The weight of the two indicators, rate of change in taxes and GDP per capita, is higher in monolithic sailboats. The weight of the expected rate of interest growth is high but negative, indicating that monolithic sailboats are greatly affected by regional economic factors.

(3) Among catamarans, only the GDP per capita is larger than that of monolithic sailboats. The effect of rate of change in taxes on catamaran prices is almost negligible. The expected rate of interest growth for catamarans, while negative as for monolithic sailboats, is much smaller than for monolithic sailboats. The monolithic sailboats are more affected by region than the catamarans.

References

[1] Harahap, F.; Harahap, A.Y.N.; Ekadiansyah, E.; Sari, R.N.; Adawiyah, R.; Harahap, C.B. Implementation of Naïve Bayes Classification Method for Predicting Purchase. In Proceedings of the 2018 6th International Conference on Cyber and IT Service.

[2] Asim, M.; Khan, Z. Mobile price class prediction using machine learning techniques. Int. J. Comput. Appl. 2018, 179, 6-11.

[3] S. Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques,” International Journal of Information & Computation Technology, vol. 4, no. 7, pp. 753-764, 2014.

[4] Hao, H.; Zhang, Q.; Wang, Z.; Zhang, J. Forecasting the number of end-of-life vehicles using a hybrid model based on grey model and artificial neural network. J. Clean. Prod. 2018, 202, 684–696.

[5] Zhang Yuansen. used Car Price Evaluation Model Based on Neural Network [D]. Tianjin University.2018(07)

[6] Al-Mubayyed, O.M.; Abu-Nasser, B.S.; Abu-Naser, S.S. Predicting Overall Car Performance Using Artificial Neural Network. 2019.

[7] Puteri, C.K.; Safitri, L.N. Analysis of linear regression on used car sales in Indonesia. J. Phys. Conf. Ser. 2020, 1469, 012143.

[8] Alhakamy A'aeshah,Alhowaity Areej,Alatawi Anwar Abdullah,Alsaadi Hadeel. Are Used Cars More Sustainable? Price Prediction Based on Linear Regression[J]. Sustainability,2023,15(2).

[9] Brahimi, B. Arabic Text Mining for Used Cars and Equipments Price Prediction. Comput. y Sist. 2022, 26, 1015–1025.

[10] Markowitz, Harry M. 1952, "Portfolio Selection", Journal of Finance, vol.7.NO1 (March) :77-91.