Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/41/20240628

The model of price of sailing ships based on Lasso regression

YueyingZhang[1,3,4], XinyiZhou[2,5], YingfeiWang[2,6], DongminWang[2,7]

[1]Information and Computering Science, Jinan University, Guangzhou city, China
[2]Mathematics and Applied Mathematics, Jinan University, Guangzhou city, China
[3]Corresponding author

[4]3545841780@qq.com
[5]3196558422@qq.com
[6]ywang2739@gmail.com
[7]1530846815@qq.com

Abstract. For the sample data of sailing ships and the listed price prediction of sailing ships based on the characteristics of sailing ships found on the website, we first conducted data cleaning on the original data obtained. In this stage, there were many missing values and outliers in the original data. After filling the missing values with mode, We transform the classified variables into dummy variables, and finally normalize them to convert the original data into the training data of the model. Then, we obtained the predicted value of Listing Price (USD) through multiple regression fitting. By calculating R^2 as 0.929, it was found that the model fitting effect was perfect, but there were too many variables due to the conversion of attribute variables to dummy variables, so it was necessary to compress model variables to select key variables. Since this topic is the explanation of Listing Price (USD), the coefficient of each variable needs to be known, so tree model is not adopted. In linear model, Lasso regression mainly screens model variables. In this case, Lasso is the main screening method. The mean square error of the listing price predicted by the multiple regression model based on Lasso regression adjustment parameters is 0.125, indicating that the model has high accuracy and the simulated listing price predicted is relatively high.

Keywords: Lasso regression, Dummy variables, Multiple regression.

1. Introduction

As with many luxury goods, the price of a sailboat in the sailing market changes as the boat ages and market conditions change. Since the COVID-19 epidemic, the consumption pattern of second-hand sailing boats has been gradually accepted by consumers, and the second-hand sailing trading market has gradually flourished, and the circulation demand of second-hand sailing boats is also increasing. In the process of second-hand sailboat trading, the most difficult and important problem is the valuation of second-hand sailboats, which is also the most relevant problem for traders. Second-hand sailboats are different from general second-hand products, and there is a complexity of ”one boat, one condition”. First, the price of second-hand sailboats is not only affected by their own configuration, such as model, boat width, sail area, displacement, and other factors, but also affected by the region, market price, year of manufacture. As a result, the price of second-hand sailboats cannot be evaluated in batches, which reduces the valuation efficiency of the second- hand sailboat market. However, there is no complete and

DOI: 10.54254/2753-8818/41/20240628

reasonable pricing system in the second-hand sailing market at present. Therefore, it is urgent to find a more accurate and reasonable valuation method and establish a sound evaluation system for second-hand sailing market. In the age of the Internet, boaters provided COMAP with valuable economic and research data on used sailboats sold in Europe, the Caribbean, and the United States in December 2020. With the help of ever- advancing scientific algorithms and mathematical tools, how to efficiently analyze and process these data and then find a suitable valuation model to determine the transaction price of second- hand sailboats is the focus of current research.

2. Model building and solution

2.1. Data cleaning

2.1.1. Fulling the missing value

First, we took the data given in the title and the data[1, 2] we found about the characteristics of sailing boats as the original data. In the original data, there were many missing values, and we carried out a visual analysis of the missing values. The following figure shows the situation of missing values of each characteristic variable. Therefore, we choose to directly delete the feature variables with more missing values; For the characteristic variables with few missing values, we adopt the mode filling method to process the missing values.

2.1.2. Check and deal the abnormal

After we deal with the missing value of the original data, we will find that there are still some outliers in the data. First, we need to judge outliers, for which we use boxplot to visualize the data. Some sample points in the sample that deviate significantly from the residual values are called outliers.

As for the outliers caused by dimensional errors in the samples, we adopt the method of dimensional correction to deal with them. For outliers caused by other reasons, to reduce the errors in the model training process, we adopt the method of deleting outliers.

2.1.3. Dummy variable transformation

Before screening characteristic variables, we need to convert the types of characteristic variables. Among all variables that affect the listing price of second-hand ships, characteristic variables such as Make, Variant and Geographic Region are disordered multi-classification variables. To quantify the data, we usually assign values of 1,2,3,4. However, 1,2,3 and 4 have the order relation from small to large, but in fact, there is no such size relation among classification variables, and they are equal and independent. If 1,2,3 and 4 are substituted into the model, the result obtained is also unreasonable, so we need to convert them into dummy variables. The value 0 or 1 reflects the different properties of the variables.

2.1.4. Normalization

Before putting data into the training model, different characteristic variables often have different dimensions and dimensional units, so direct input into the model will affect the final training results. To eliminate the dimensional influence between different characteristic variables, it is necessary to conduct standardized data processing to solve the comparability between data. The most typical method is to conduct normalized data processing.

The normalization method adopted here is maximum and minimum normalization, that is, the original data is linearly transformed into the range of [0,1] through linear function, and the calculated results are normalized data. The dimensionless expression is transformed into a dimensionless expression through transformation. The specific formula is as follow:

$$x' = \frac{x - x_{min}}{x_{max} - x}$$

DOI: 10.54254/2753-8818/41/20240628

2.2. Model Preparation

2.2.1. Model evaluation coefficient: mean square error

Mean-square error (MSE) is a measure that reflects the difference between the estimator and the estimator. MSE is a statistical measure and loss function commonly used in ML regression models, such as linear regression. Its formula is shown in the figure:

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$$

Where y_i is the true value and $\hat{y}_i$ is the predicted value.

In this paper, the estimator and the estimator are the listing price.

2.2.2. Adjust the compression penalty parameter λ

$$Lasso : \min_{\beta} \frac{1}{2n}\|y - X\beta\|_2{}^2 + \lambda\|\beta\|_1$$

where

$$\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$$

Where λ is the regulating parameter, sometimes called a hyper-parameter. $\lambda // \beta //_1$ is the compression penalty, and P is the number of arguments. Different λ will result in different mean square errors of regression models with variables selected through the L_1 regularization process. We calculate the coefficients of λ and variables corresponding to the minimum mean square errors to determine the optimal degree of the model.

2.3. Select characteristic variable

After the data cleaning of the original data, we get the processed data. Next, we adopt the optimization stepwise regression and neural network to screen the characteristic variables that have a great impact on the listing price and take the intersection of the variables screened by the two methods as the final characteristic variable.

2.3.1. Model overview

Firstly, a multiple regression analysis model was established for n regression independent variables $x_1, x_2, \cdots, x_n$ and co-dependent variables $Y = \beta_0 + \beta_i x_i + \varepsilon$, $i = 1,2,3,\cdots,n$. Each feature, that is, the independent variable, has a corresponding slope coefficient βi . When we calculated the coefficient βi through Python multiple regression analysis, we obtained the correlation and significance level of the corresponding independent variable and dependent variable.

Then, we used Lasso regression and neural network to discard independent variables with poor correlation and significance level and selected independent variables with strong correlation xi for mathematical modeling again.

Meanwhile, in the process of obtaining the correlation table above, we will discuss the collinearity between independent variables xi. If the collinearity between independent variables is strong, we will screen out variables with relatively large characteristic parameters by adjusting α parameters in Lasso regression. Variable parameters with relatively small mean square error are selected to build a model as our final valuation model to explain Listing Price (USD).

2.3.2. Lasso regression model was established Lasso

Lasso regression[4, 5, 6] is a linear model, and this method is a compressed estimate. It obtains a more refined model by constructing a penalty function, making it compress some regression coefficients, that is, the sum of the absolute values of the force coefficients is less than a certain fixed value. It is also a biased estimation for complex collinear data.

DOI: 10.54254/2753-8818/41/20240628

For the linear regression problem with multiple variables, the fitting model is relatively complicated due to excessive parameters. However, in order to prevent the overfitting phenomenon, the model should be simplified as much as possible, and the majority of variables should be replaced by a finite few variables to explain the estimated quantity. The commonly used methods for parameter selection include sequence forward selection, sequence backward elimination, sequence forward selection and backward elimination combination, and Lasso compression variable.

However, in this case, due to the variants of sailboats, there are too many dummy variables, and the efficiency is too low whether the series is forward selection or backward elimination. Therefore, Lasso compression variable model is adopted, and the coefficient of irrelevant variables is reduced to zero by adding penalty term.

2.3.3. Concrete mathematical expression

Linear regression optimization objective:

$$\beta^* = \text{argmin}_\beta \frac{1}{n}\sum_{i=0}^{n}((y_i - \hat{y}_i) - \beta^\top(x_i - \hat{x}_i))^2$$

Optimization objectives after regularization:

$$\beta^* = \text{argmin}_\beta \frac{1}{n}\|y - X\beta\|_2{}^2 + \lambda\|\beta\|_1$$

Where $\|\cdot\|_2$ is the binary norm, that is, R^n in the vector space, let $x = (x_1, x_2, \cdots, x_n)^\top$.

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

$$\|x\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2}$$

2.3.4. Concrete modeling process

In the first question, the data table given in the question includes the manufacturer, sailboat model, region, year, and price. To meet the requirements in the question, we need to analyze the sailboat characteristics and regional economic conditions related to the price. We collected the relevant data of the types of sailboats given in the title on the website of second-hand sailboats and collected the economic conditions of cities in relevant regions. We decided to use a variety of GDP-related data to express the regional economy and differentiated the sailboats in different regions by giving different characteristic values through dummy variables. Finally, because there is no clear requirement in the title, the data of single and double sails are combined in this question to facilitate the larger data set to have better fitting effect in the subsequent multiple regression analysis. For sailing-related data, we collected the waterline length LML, boat width, draft, displacement, sail area and average cargo throughput. Shown in the following Table 1.

Table 1. HuTll data chart

LWL	Beam	Draft	Displacement	Sail Area	Average cargo throughout	GDP	GDP per capital
37.24	12.63	3.94	22046.0	824.0	45350000.0	2939.0	44494.0
36.06	12.99	6.07	15432.0	721.0	595000.0	57.8	13647.0
36.06	12.99	6.07	15432.0	721.0	595000.0	57.8	13647.0
36.06	12.99	6.07	15432.0	721.0	595000.0	57.8	13647.0
36.06	12.99	6.07	15432.0	721.0	595000.0	57.8	13647.0
36.06	12.99	6.07	15432.0	721.0	3150000.0	204.0	19147.0
37.07	13.02	6.23	19621.0	776.0	3150000.0	204.0	19147.0

DOI: 10.54254/2753-8818/41/20240628

Then for the newly formed number table, we use Python to carry out multiple regression analysis. First, we standardized all the data to avoid the abnormal impact of measurement on the price of sailboats. After standardized processing, we tabulated the data in Excel and understood the correlation between various independent variables through the preliminary observation of the heat map followed by Figure 1:

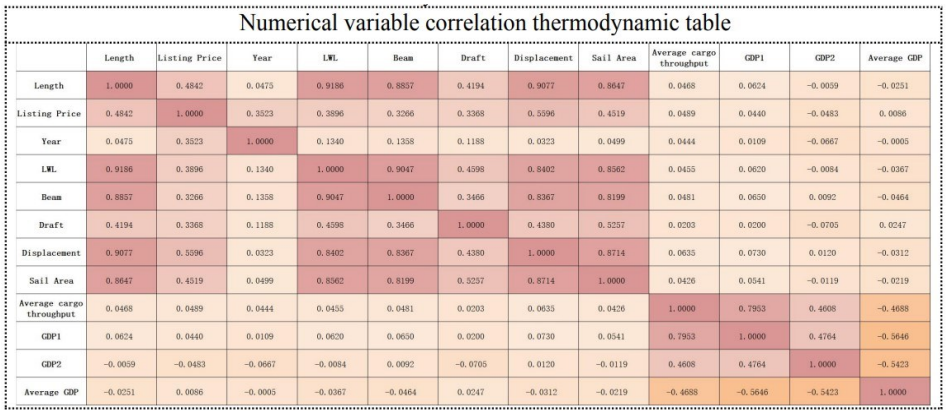

Figure 1. Numerical variable correlation thermodynamic

For the standardized data, we conducted preliminary multiple linear regression analysis, and we could get the parameter table as shown in the following figure to represent the correlation level between each independent variable and dependent variable and the fitting of the multiple linear regression model. At the same time, according to the results followed by Table 2, we know that each independent variable has strong collinearity.

At the same time, we pass $P > |t|$ on the income form, the significance level of sorting, through technical processing choose strong correlation between independent variables, to build a new mathematical model. In the specific case of this question, we chose Lasso to select the independent variables in the question, instead of the method of stepwise regression. The reason is that, given the unique background of the data in the question, there are many independent variables and many dummy variables. If stepwise regression is used, there will be many data cycles, which will occupy a large amount of storage space. Second, the existence of meaningless data loops, will slow down the efficiency of the code. Therefore, we used this method to analyze model fitting for VIF value and R^2 value, to find out several independent variables with great correlation influence and reserve them.

Table 2. Regression coefficient analysis table

Name	coefficient	Standard error	t	$P > \|t\|$
Sail Area	-3301.72	1.31E+04	-0.252	0.801
Length	8314.47	1.76E+04	0.471	0.638
GDP2	7859.72	5756.478	1.365	0.172
GDP1	-8533.72	5757.049	-1.482	0.138
Average Cargo Throughput	-19640.00	1.00E+04	-1.962	0.050
LWL	32720.00	1.46E+04	2.24	0.025
Beam	109400.00	3.64E+04	3.009	0.003
Average GDP	-9556.25	3245.174	-2.945	0.003
Constant	377200.00	1.79E+04	21.059	0.000
Year	68790.00	2235.4	30.774	0.000
Draft	-103200.00	1.20E+04	-8.569	0.000
Displacement	67930.00	1.12E+04	6.081	0.000

DOI: 10.54254/2753-8818/41/20240628

For the selected independent variable, we can find that the multicollinearity problem in the multiple regression analysis problem is obvious to the retained independent variable through variance expansion coefficient. We analyze the possible collinearity problem through Lasso regression through parameter adjustment and seek the optimal situation. Through the improved least square method and $L1$ regularization, we analyzed the collinearity of the data. On the one hand, we carried out "feature screening" for the dependent variables. On the other hand, we find a more meaningful independent variable X with this method, which minimizes the mean square error of the model. The result is followed by Figure 2:

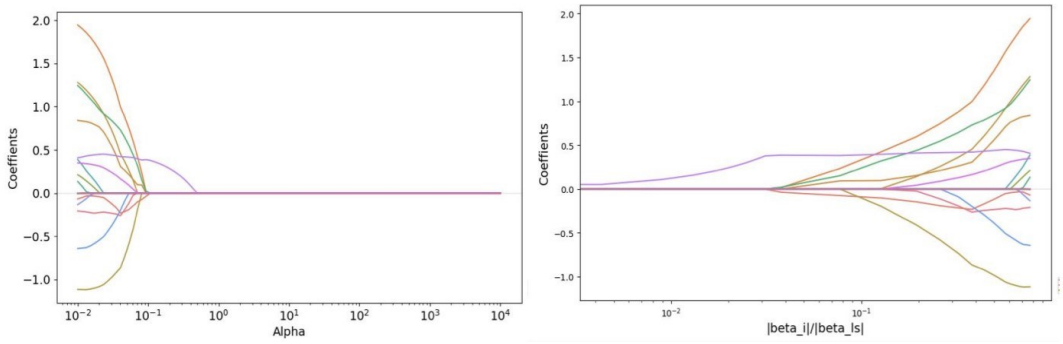

Figure 2. Regularized path diagram

The specific algorithm process is shown in the figure above. Finally, the coefficient of each variable is returned, and then the predicted value is calculated to obtain the mean square error to get the optimization model. The predicted value is explained according to the coefficient of each variable, namely the Listing Price.

Finally, the MSE of the optimal model is 0.125/2, so the multiple linear regression model based on lasso regression parameter optimization has good goodness of fit. The result is followed by Table 3:

Table 3. Final parameters selected table

serial number	Characteristic variable name	15	Variant Swan 54
1	CRSSVG	16	Make Hallberg Rassy
2	Make HH Catamaras	17	Make Southerly
3	Make Discovery	18	Make Boreal
4	Make Nautor	19	Variant Pilot Saloon 48
5	Variant Series 5	20	Make Oyster
6	Make Nautitech	21	Length
7	Make Bestevaer	22	Year
8	Variant 52 Sport	23	LWL
9	Variant Atlantic 49MF	24	Beam
10	Variant SABA 50 Maestro	25	Sail Area
11	Variant 52	26	GDP
12	Variant 52F	27	AGDP
13	Variant V50 Mills	28	Europe
14	Make Outremer	29	USA

For the selected parameters in the table, the first 20 are dummy variables that have a relatively high impact on the listing price, and their impact on the listing price is at the level of 1%−10%. The relevant dummy variables include the manufacturer of the sailboat, the model of the sailboat variant and the regional influence variable. The last nine are sailing-related characteristic data (such as ship width, ship

DOI: 10.54254/2753-8818/41/20240628

length, displacement, etc.) and relevant regional economic characteristics, which are all independent variables selected in lasso regression analysis with strong correlation to listing price. For dummy variables, the influence of region is mainly caused by regional characteristics, which will be discussed together with regional economic factors in the second follow-up question. As for the influence of the sailboat manufacturer and the sailboat variant on the price, we can understand the premium generated by the brand effect. Besides, the characteristic variables related to the sailing ship itself, the length and width of the ship, on the one hand, determines the size and habitability of the space, on the other hand, determines the number of materials used in the hull and the more scientific design structure needed for larger ships, so the captain and width of the ship have a significant impact on the listing price. Moreover, considering the year of production of the ship, it also reflects the usable time of the ship, and considering the survival characteristics of the sailing ship, the production quantity of each type of ship is limited. Such uniqueness, like luxury goods, also has a significant impact on the change of listing price caused by the year. Nautical miles can also indicate the range of the vessel in a refueling situation, which is relatively specific to the buyer, so nautical miles have a significant impact on the listing price.

3. Conclusion

Compared with multiple linear regression, Lasso regression analysis adds a penalty norm $L1$. The existence of the norm increases the stability of our model and makes the screening model more effective. In the process of variable screening, Lasso controls the screening process through the hyperparameter real lambda between $(0,1)$ to ensure that the screening is a continuous process, while making the screening more robust without losing the interpretability.

Lasso is suitable for the model with larger data volume and more missing values, and when the meaningful variables are relatively limited, this kind of analysis effect is better. Because $L1$ norm tends to produce sparse coefficient, Lasso regression has built-in feature selection. Meanwhile, the solution of $L1$ norm is sparse, so it is more efficient in calculation when used together with sparse algorithm.

References

[1] https://www.ayc-yachtbroker.com/alliage-44
[2] https://www.yachtworld.com/yacht/2005-alliage-alliage-44-8666783/
[3] https://itboat.com/search?text=alubat+cigale+16
[4] Reducing bias and mitigating the influence of excess of zeros in regression covariates with multioutcome adaptive LAD-lasso [J]M¨ott¨onen Jyrki;L¨ahderanta Tero;Salonen Janne;Sillanp¨a¨a Mikko J. Communications in Statistics - Theory and Methods. Volume 53 , Issue 13 . 2024. PP 4730-4744
[5] Lasso regression under stochastic restrictions in linear regression: An application to genomic data[J] Gen¸c Murat;Ozkale M. Revan Communications in Statistics - Theory and Methods. Volume 53 ,¨ Issue 8 . 2024. PP 2816-2839
[6] High-dimensional nonconvex LASSO-type [formula omitted]-estimators [J] Jad Beyhum;Fran¸cois Portier Journal of Multivariate Analysis. Volume 202 , Issue . 2024. PP 105303-