ISSN:2790-1661 Volume-6-(2023)

Predicted the Used Sailboat Price by Decision Tree and Multiple Regression Model

Jia Wang [1], Yihang Zang [1], Kaihuang Wang [1], Yutong Li [2], Tao Liu [3, 4, a], Ruofeng Qiu [5], Yunfei Qi [5]

[1] School of Control Engineering, Northeastern University at Qinhuangdao, China;

[2] Sydney Smart Technology College, Northeastern University at Qinhuangdao, China;

[3] School of Mathematics and Statistics, Northeastern University at Qinhuangdao, China;

[4] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore;

[5] Eighth Geological Brigade of Hebei Bureau of Geology and Mineral Resources Exploration, China.

[a] liutao@neuq.edu.cn; tao.liu@ntu.edu.sg; math.taoliu@gmail.com

Abstract. In this paper, three mathematical models are established to explore the factors that influence the price of used sailboat, and predicts the price of used sailboat in different regions. First, clean and process the dataset, and the node centrality analysis is carried out. Next, a decision tree model is used to explain the price of the sailboat. In order to predict the price of sailing ships in different regions, this paper transforms geographical region variables into dummy variables, and uses multiple linear regression method to evaluate the influence of geographical region on the price. Then, an ANOVA model is established to analyze the price and regional impact of different types of sailboats. Finally, a cluster analysis algorithm is established to classify used sailboat by various classification factors.

Keywords: Used sailboats; Decision tree algorithm; Regression models; Analysis of variance.

1. Introduction

Due to the strong complexity and volatility of the sailing market [1-5], the characteristics of huge amount, high risk and long payback period of sailing investment, enterprises investing in sailing ships have strong uncertainty. Assuming that the company has stable capacity demand, it currently uses chartering to meet its own capacity demand, and is bullish on the future charter market. From the perspective of controlling costs, companies began to consider whether to buy sailboats. Since newbuilding requires a certain cycle time, consider a second-hand vessel that is immediately available, such as this one that is currently chartered, or a similar boat in the market that is interested in selling. The value of a sailboat changes with age and market conditions, and the price of a sailboat is also affected by many factors such as manufacturer, variant, length, and so on [6-10]. Based on this, the purpose of this article modeling is to study the pricing of used sailboats so that brokers can sell them better.

2. General Assumptions

The following basic assumptions are made to simplify problems.

(1) Suppose that all the data given in the question are reasonable.

(2) Suppose that sudden second-hand sailing price intervention does not exist.

(3) It is assumed that there will be no explosive changes in future used sailing price predictions.

(4) It is assumed that the encoding of the data in the model does not affect the reflection of the original information.

(5) Suppose that the data published by different sailing factories and different regions have the same statistical principles.

(6) Exclude small probability events in life (e.g., black swan events, abnormal situations).

3. ModelI: C5.0 Decision Tree Algorithm

This paper uses the C5.0 decision tree algorithm [11-17] to establish the box office prediction model, the main reasons are: on the one hand, in the research process of box office prediction, the decision tree method is relatively advanced; on the other hand, the target of the C5.0 algorithm needs to be categorical variables, and the box office index in this paper just meets the requirements. C5.0 is a very classic decision tree model algorithm, which is an improvement on the basis of C4.5. Compared with C4.5, the C5.0 decision tree algorithm introduces many new technologies, such as the use of Boosting technology to improve the accuracy of the algorithm; the use of cost matrix construction Cost-sensitive tree to reduce the probability of high-cost false positives.

C5.0 takes the given sample set as the root node, and then calculates the information gain ratio of each feature attribute in the current sample set separately, and then finds the feature attribute with the highest information gain ratio as a sub node of the current node, repeats the above steps as the root node, and continues to split downwards until all attributes in the subset belong to the same category, and the tree stops splitting.

Given that this dataset has more than 1000 dimensions, here, we select two of them, then select six sets of sample data, and then plot a decision surface on these two dimensions. This is shown in Figure 1.

As can be roughly seen from the figure above, most of its classification is for the red area part, which can be analyzed and predicted by subsequent analysis.

The predictive training learning curve of the second-hand sailing dataset using the C5.0 decision tree algorithm is shown in Figure 2.

By normalizing the confusion matrix, we conclude that after applying graph learning, decision tree classification prediction is better.

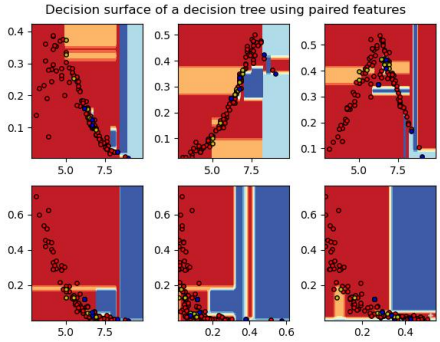

Figure 1 Decision tree decision plane based on paired features

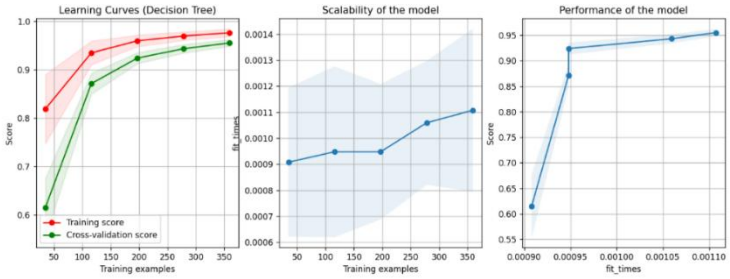

Figure 2 Decision tree learning training curve

4. Model II: Multiple Regression Model

General |r|>0.95, with significant correlation; |r|<0.3 The relationship is very weak and considered irrelevant. 0.5≤|r|≤0.8 is moderately correlated, and 0.3≤|r|≤0.5 is considered low.

Pearson Coefficient method: Calculation of data from a fixed distance variable. The formula is

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{1}$$

where r is the correlation coefficient, $\bar{x}$ $\bar{y}$ are the means of variables x and y, respectively, x y x_i y_i i are the first observations of the variables, respectively.

Using SPSS19.0, distance correlation analysis was performed to investigate the degree of proximity between similar or dissimilar measurements.

SPSS 19.0 was used to analyze the correlation between the two, and the indicators with large correlation ($n2\leq\leq n10$) were selected as the correlation indicators $f_k(x_i)$.

$$x_{i,}x_{j,}x_m.....$$

The regression equation is established as follows:

$$\begin{cases} f_1=f\left(x_i,x_j,x_m.....\right) \\ f_2=f\left(x_k,x_p,x_z.....\right) \\ f_3=f\left(x_i,x_p,x_l.....\right) \\ f_2=f\left(x_i,x_j,x_m.....\right) \end{cases} \tag{2}$$

Table 2 shows descriptive statistics.

Table 2 Descriptive statistics

	average value	standard deviation	Number of cases
Decoration	2.14	.945	58
type	1.69	.467	58
color	2.69	1.810	54
weathering	1.59	.497	58

Table 3 shows the correlations.

Table 3 Correlation analysis

	Decoration	type	color	Surface weathering
Decoration	1	.099	.194	.049
type	.099	1	.424**	.344
color	.194	.424**	1	-.115
Surface weathering	.049	.344**	-.115	1
**. At level 0.01 (double-tailed), the correlation is significant.				

The above table shows that the surface weathering of glass artifacts has the highest correlation with glass type.

The following is a regression analysis, which is as follows:

Table 4 Model summary

model	R	R-square	Adjusted R side	Error in standard estimates	Durbin Watson
1	. 678a	.104	. 072	. 386	0.860

Party R in the table is 0 678, greater than 50%, indicating that the model prediction is accurate and the study of the model is meaningful.

Table 5 ANOVA

model		Sum of squares	degree of freedom	mean square	F	Salience
1	regression	2.414	3	.805	3.685	.018b
	Residuals	10.919	50	.218		
	total	13.333	53			

This table considers whether the regression equation makes sense, significance 0.005<0.05, so the equation makes sense.

Table 6 VIF diagnosis

model		Unstandardized coefficients		t	Salience	Collinearity statistics	
		B	standard error			Tolerance	VIF
1	(constant).	.930	.264		3.515	.001	
	Decoration	.040	.070	.074	.568	.573	.062
	type	.466	.149	.442	3.128	.003	.316
	color	-.088	.040	-.317	-2.215	.031	-.115

The VIF values in the table are not greater than 5, indicating that there is no multicollinearity between the independent variables, which reflects the accuracy and reliability of the regression model.

Based on all the above analysis, the regression equation for the independent and dependent variables is:

$$y = 0.04x1 + 0.466x2 - 0.088x3 + 0.930$$

5. Model III: Cluster Analysis

Fuzzy C-means clustering algorithm is a clustering method that blurs the definition of classical division and uses the degree of membership to determine the degree of belonging to a certain cluster. Among them, there are two important parameters, namely the number of clusters c; Fuzzy weighted index. The algorithm divides vectors into groups m n $x_k \subset R^5$ c(); $k = 1,2,\ldots,n$ s is the dimension of the vector and finds the clustering center for each group x_k.

The basic steps are:

(1) Meet the constraints of equation (3).

$$\sum_{i=1}^{c} u_{ik} = 1 u_{ik} \in (0,1) \tag{3}$$

(2) Calculate the cluster center V according to equation (4).

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m}, i = 1,2,\cdots,c \tag{4}$$

(3) Calculate the objective function according to equation (5).

$$J(U, v_1, \cdots, v_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m d_{ik}^2 \tag{5}$$

$$u_{ik}=1/\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}},i=1,2,\cdots,c$$
$$k=1,2,\cdots,n \quad (6)$$

In equations (3) and (4), is the Euclidean distance between the first fuzzy group and the first cluster center; is the fuzzy weighted index, $d_{ik}=\sqrt{\sum_{q=1}^{s}(x_{kq}-v_{iq})^2}$ k i m , m∈(1,+∞), 2≤c≤n. If the variable is less than some determined threshold relative to the last result, the algorithm stops and outputs the final membership matrix ε U and cluster center V, otherwise calculates a new membership matrix and returns to the step (3).

It is calculated from the final membership matrix $U=\{u_{ik}\}_{cxn}$

$$i=arg\max_{1\le i\le c}u_{ik},k=1,2,\ldots,n \quad (7)$$

It can be seen that after learning, the sample is gathered into the class x_k i $i=1,2,\ldots,c$.

According to the above principal analysis, we will collect data samples and feature dimension collections. The characteristic dimensions are beam, draft, displacement, rigging, sail area and other information.

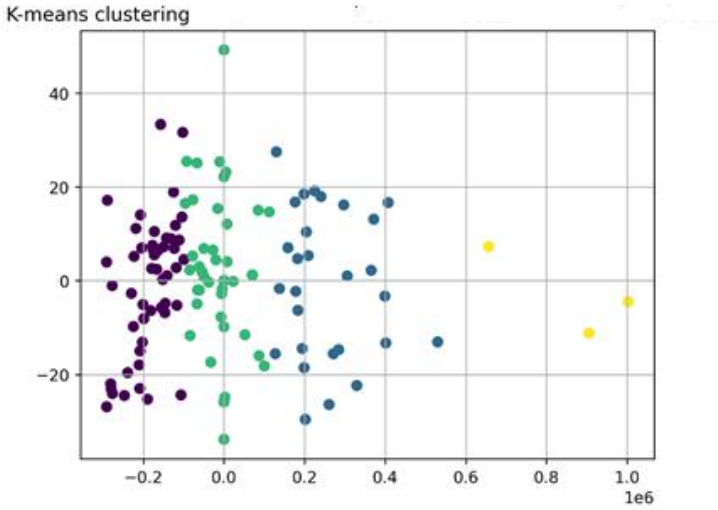

Figure 3 K-means clustering results

The results obtained by the training of the K-means clustering algorithm are shown in Figure 3, and it can be seen that the dataset is divided into four categories, and the classification effect is better.

6. Sensitive Analysis

In order to verify the stability of the model, we adjusted the data set three times and then predicted, and found that the prediction accuracy was above 90%, indicating that the prediction effect was very good and the stability of the model was high.

Table 7 Model accuracy under different parameters

	Parameter 1	Parameter 2	Parameter 3
Accuracy	92.3%	91.9%	93.2%

7. Conclusion

According to the above analysis, the prices of second-hand sailing boats in different regions are predictable and fluctuate in the short term. By establishing the decision tree model, the second-hand sailboats can be classified and predicted, so as to improve the accuracy of prediction. At the same time, multiple regression model is established to virtualize geographical variables, so as to evaluate the influence of different regions on sailing prices. Finally, this paper also uses cluster analysis to classify sailing boats, so as to better classify second-hand sailing boats. In a word, the accuracy rate of predicting second-hand sailboat prices through modeling is as high as 90%, with good prediction effect and high model stability.

Acknowledgements

This work was supported by the Natural Science Foundation of Hebei Province of China (A2020501007), the Technical Service Project of Eighth Geological Brigade of Hebei Bureau of Geology and Mineral Resources Exploration (KJ2022-021).

References

[1] Badia-Miró M, Carreras-MarÍn A, Huberman M. Smooth sailing: Market integration, agglomeration, and productivity growth in interwar Brazil. European Review of Economic History, 2023, 27(1): 45-69.

[2] Liu T. Porosity reconstruction based on Biot elastic model of porous media by homotopy perturbation method. Chaos, Solitons & Fractals, 2022, 158: 112007.

[3] Schill M J. Sailing in rough water: Market volatility and corporate finance. Journal of Corporate Finance, 2004, 10(5): 659-681.

[4] Aspers P, Sandberg C. Sailing together from different shores: Labour markets and inequality on board merchant ships. Global Networks, 2020, 20(3): 454-471.

[5] Centobelli P, Cerchione R, Maglietta A, et al. Sailing through a digital and resilient shipbuilding supply chain: An empirical investigation. Journal of Business Research, 2023, 158: 113686.

[6] An Y, Yu J, Zhang J. Autonomous sailboat design: A review from the performance perspective. Ocean Engineering, 2021, 238: 109753.

[7] Liu T. Parameter estimation with the multigrid-homotopy method for a nonlinear diffusion equation. Journal of Computational and Applied Mathematics, 2022, 413: 114393.

[8] Deng Y, Zhang X, Zhang Q, et al. Event-triggered composite adaptive fuzzy control of sailboat with heeling constraint. Ocean Engineering, 2020, 211: 107627.

[9] Lam H F, Hu J, Zhang F L, et al. Markov chain Monte Carlo-based Bayesian model updating of a sailboat-shaped building using a parallel technique. Engineering Structures, 2019, 193: 12-27.

[10] Shen Z, Ding W, Liu Y, et al. Path planning optimization for unmanned sailboat in complex marine environment. Ocean Engineering, 2023, 269: 113475.

[11] Guo Z, Shi Y, Huang F, et al. Landslide susceptibility zonation method based on C5. 0 decision tree and K-means cluster algorithms to improve the efficiency of risk management. Geoscience Frontiers, 2021, 12(6): 101249.

[12] Kristóf T, Virág M. EU-27 bank failure prediction with C5. 0 decision trees and deep learning neural networks. Research in International Business and Finance, 2022, 61: 101644.

[13] Liu T, Xia K, Zheng Y, et al. A homotopy method for the constrained inverse problem in the multiphase porous media flow. Processes, 2022, 10(6): 1143.

[14] Liu T, Yu J, Zheng Y, et al. A nonlinear multigrid method for the parameter identification problem of partial differential equations with constraints. Mathematics, 2022, 10(16): 2938.

[15] Salman Saeed M, Mustafa M W, Sheikh U U, et al. An efficient boosted C5. 0 decision-tree-based classification approach for detecting non-technical losses in power utilities. Energies, 2020, 13(12): 3242.

[16] Zhu B, Hou X, Liu S, et al. Iot equipment monitoring system based on c5. 0 decision tree and time-series analysis. IEEE Access, 2021, 10: 36637-36648.

[17] Kadhm M S, Ayad H, Mohammed M J. Palmprint recognition system based on proposed features extraction and (c5. 0) decision tree, k-nearest neighbour (knn) classification approaches. Journal of Engineering Science and Technology, 2021, 16(1): 816-831.