Second-hand sailboat price assessment based on ANOVA

Chengxuan Zhu[1, *], Zihan Fang[2], Jiaying Zheng[1]

[1]College of International Education, Fujian Normal University University, Fujian, China, 350007

[2]College of Computer and Cyber Security, Fujian Normal University University, Fujian, China, 350007

*Corresponding author: jike68114@126.com

Abstract. As a luxury, the price of second-hand sailing will vary with the changes of years and markets, and its price may be affected by many factors, such as regional differences, various characteristics of sailing and the year of production of sailing. In order to determine the dominant factors affecting sailing prices and explain the changes of sailing prices, Spearman correlation analysis is used. Multivariate linear regression model is established to analyze and determine the weight of the indicators affecting the sailing, and finally get a formula to explain the sailing price. Based on the price model of sailing using ridge regression, ANOVA was used to further determine the effect of regions on different variants of sailing.Considering the consistent effects of different regions on the same variable, a two-way ANOVA was used to identify the inconsistent effects of different regions on different navigation variables.Finally, according to the results of variance analysis, it was found that the price of second-hand sailing varied significantly in different regions, and through two-way variance analysis, there was no consistency between different sailing variants.

Keywords: Ridge regression, Spearman correlation analysis, ANOVA,Second-hand sailboat price.

1. Introduction

1.1. background information

Sailboats originated in Europe, relying on the natural wind force for use on sails or engines, and being operated by people on water.The sport combines creativity, adventure and competition, and is popular in Europe and the United States. Boats are expensive, and many people choose to use second-hand sailboats.

In the second-hand market, the value of a sailboat is not only affected by the internal components, but also by age and market conditions.Therefore, it is important to value a second-hand sailboat before it is traded through a broker.

1.2. Introduction to data processing

Missing value and outlier processing.

1.Delete data that with missing items in the data table.

2.Delete data about manufacturers and variants that do not exist in reality.

3.Delete data with a build time earlier than the variant's first build time.

Data expansion and collection.

In order to more specifically represent the differences between different sailboats variants, detailed data needs to be collected for each variant. Hence data items expansion is necessary.As shown in Table 1:

Table 1: Added list of data items

Data Source Names	Data Source Websites	Type
SailboatData	https://sailboatdata.com	Text
Yacht World	https://www.yatchworld.com	Text
SailboatsForSale	https://www.sailboatsforsale.com	Text
Sailboat listings	https://www.sailboatlistings.com	Text
Boats	https://www.boats.com	Text
Lagoon	https://www.lagoon.com	Text
Fountaine Pajot	https://www.fountaine-pajot.com/	Text
Tofinou	https://tofinou.com	Text

Data Standardization.

To eliminate the effects of different dimensions and improve the accuracy of the model, the data were transformed to a standard normal distribution (mean 0, standard deviation 1).

Data encoding.

When dealing with classification problems, categorical variables usually need to be converted into numerical variables to be processed using algorithms. This process is often known as encoding or conversion.

One-Hot encoding and Target Encoding were used to achieve the quantification of categorical data.

One-Hot encoding :

The results of one-hot enconding are shown in Figure 1:

Figure 1: One-Hot encoding

Target Encoding :

Target encoding is a categorical feature encoding technique that replaces the values of each category variable with the mean target variable value for that category.

2. Established model

2.1. Sailboat price interpretation model

2.1.1 Data category selection

To realize the universality of the model while preventing overfitting caused by too many variables, using Spearman correlation analysis to identify as few relevant independent variables is an optional way.

In order to take the region and manufacturer of the sailboat into account, the region of the sailboat is encoded using One-Hot coding, labeled V0-USA, V1-Europe, V2-Caribbean, and the sailboat manufacturer is encoded using the target coding, labeled Make(encoded).

The heat maps drawn from the above data are shown in Figure 2 :

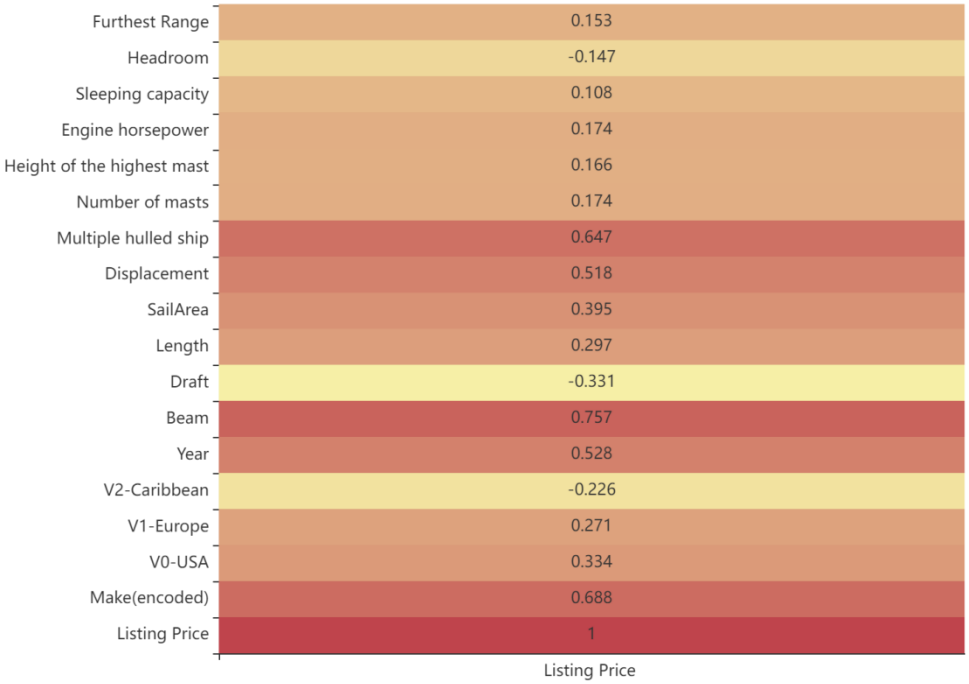

Figure 2 : Spearman correlation analysis result heat map

According to Spearman correlation analysis:

Beam, Make(encoded), Multiple hulled ship, Year, Displacement, SailArea, Draft, Length, V0-USA,V1-Europe, V2-Caribbean.There is a strong correlation between the above parameters and the listing prices.

Therefore, the independent variables of model construction are selected as the above categories.

Identification data are presented in Table 2:

Table 2: Notations used in this paper

Symbol	Description	Dimension
D_i, D_i'	Displacement	lb
B_i, B_i'	Beam	ft
F_i, F_i'	Draft	ft
Y_i, Y_i'	Year	$year$
L_i, L_i'	Length	ft
S_i, S_i'	Sail Area	ft^2
H_i, H_i'	Horsepower	hp
$V_0 - USA, V_1 - Europe, V_2 - Caribbean$	Region	/
G_i	GDP	/
S_x	Water Area	ft^2
M_i	Make encoding	/
S, S'	Multiple hulled ship	/

2.1.2 Model selection

Since the realistic meaning of Displacement in the above category is related to Draft, Beam, and Length, Spearman correlation test is carried out for the four categories.

The heat maps drawn from the above data are shown in Figure 3 :

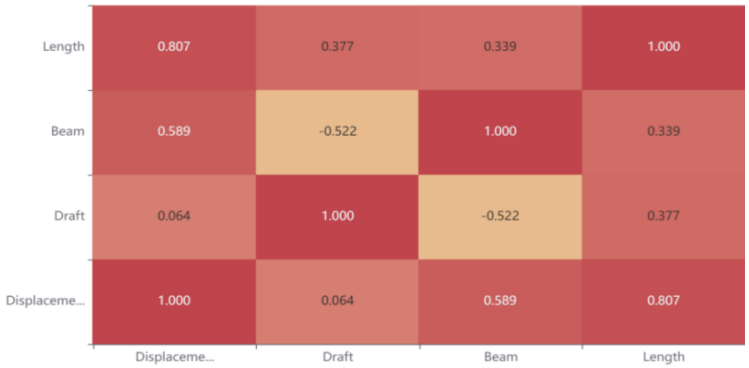

Figure 3 : Four factor Spearman correlation analysis results heat map

Due to the strong correlation between the four, it can be considered that some data categories in the original data have collinearity. Therefore, ridge regression method is used to fit the data, so as to reduce the error caused by collinearity.

2.1.3 Introduction to Ridge Regression

Ridge regression is a biased estimation regression method to deal with multicollinearity of independent variables (VIF value is generally greater than 10) for small sample data. Its mathematical model is as follows:

$$\hat{\beta}(k) = (X'X + kI)^{-1}X'y \quad (1)$$

Where, the regression parameter β is estimated by ridge regression, k is the ridge parameter. Based on the sample data, the parameter K=0.135 is set by using the variance inflation factor.

2.2. Analysis of variance for regional factors

2.2.1 Consider whether different regions are consistent for different sailing variants

Statistical analysis of the data is required.Using statistical methods, such as analysis of variance (ANOVA), to compare the performance of the sailing variants across different regions. ANOVA can help us determine whether there are significant differences in performance between the sailing variants in each region, as well as whether there are significant differences in performance between the regions for each sailing variant.

The price of sailing boats in different regions is shown in Figure 4:

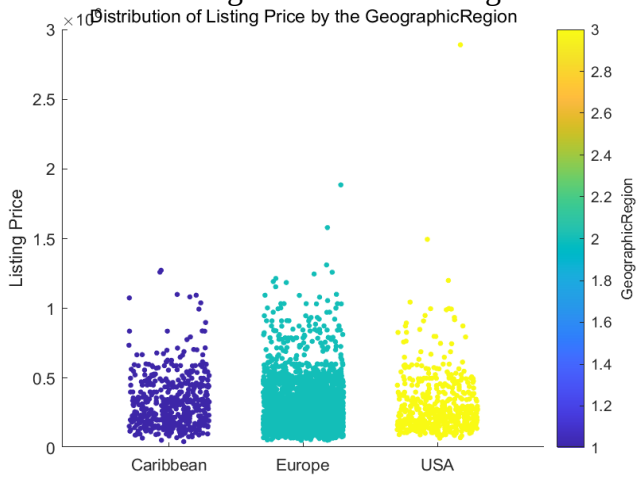

Figure 4: Scatter plot of individual sailing prices classified by region

2.2.2 Two-factor ANOVA

ANOVA is a commonly used statistical method to compare differences between two or more samples or treatments.

Considering that the regional data is classified data, the analysis of variance can be used to analyze the difference of the data of multiple classified fields. Using analysis of variance to study the difference between a defined field (X) and one or more quantitative fields (Y).

ANOVA can be used to determine whether regional factors have some influence on price, but whether there is consistency in the influence of region on the price of different sailing variants, and whether there is a greater or smaller influence on some variants, in order to verify this problem, introducing a second variable is necessary.

The sailing variable was introduced, which is also a categorical variable, and therefore a two-way ANOVA was used to verify consistency.

Two-factor ANOVA is used to analyze the influence of two independent variables on a dependent variable. The basic principle is to classify the observed data according to the combination of two factors, then calculate the mean and variance of the dependent variable under each combination, and test whether the influence of the two factors and their interaction on the dependent variable is significant through ANOVA.

3. Results

3.1. Establish a ridge regression model

3.1.1 Establish a model to describe the influence of characteristics

Model-building coefficients are shown in Table 3:

Table 3: Building ridge regression model diagram

K=0.135	Non-normalized coefficients		Normalized coefficients	Symbol	t	R^2	F
	B	Standard error	Beta				
Constant	0	0.01	-	-	0	0.742	597.384(0.000***)
V'_1	0.02	0.005	0.02	β_1	3.718		
V'_0	0.032	0.007	0.032	β_2	4.341		
V'_2	-0.058	0.007	-0.058	β_3	-7.75		
H'_i	-0.012	0.011	-0.012	β_4	-1.17		
Y'_i	0.281	0.01	0.281	β_5	27.182		
B'_i	-0.075	0.036	-0.075	β_6	-2.071		
F'_i	0.046	0.02	0.046	β_7	2.34		
D'_i	0.367	0.019	0.367	β_8	19.209		
S'_i	0.041	0.013	0.041	β_9	3.229		
L'_i	0.092	0.019	0.092	β_{10}	4.874		
S'	0.367	0.036	0.367	β_{11}	10.183		
M'_i	0.218	0.018	0.218	β_{12}	12.426		

Ridge regression results show that: based on the significance of F test, P value is 0.000***, showing significance at the level, indicating that there is a regression relationship between the independent variable and the dependent variable. Meanwhile, the goodness of fit R^2 of the model is 0.74, indicating that the model performs well.

$$LP = \beta_1 \times V'_1 + \beta_2 \times V'_0 + \beta_3 \times V'_2 + \beta_4 \times H'_i + \beta_5 \times Y'_i + \beta_6 \times B'_i + \beta_7 \times F'_i + \beta_8 \times D'_i + \beta_9 \times S'_i + \beta_{10} \times L'_i + \beta_{11} \times S' + \beta_{12} \times M'_i \quad (2)$$

3.1.2 Model testing and analysis

The model-fitting results are shown in Figure 5:

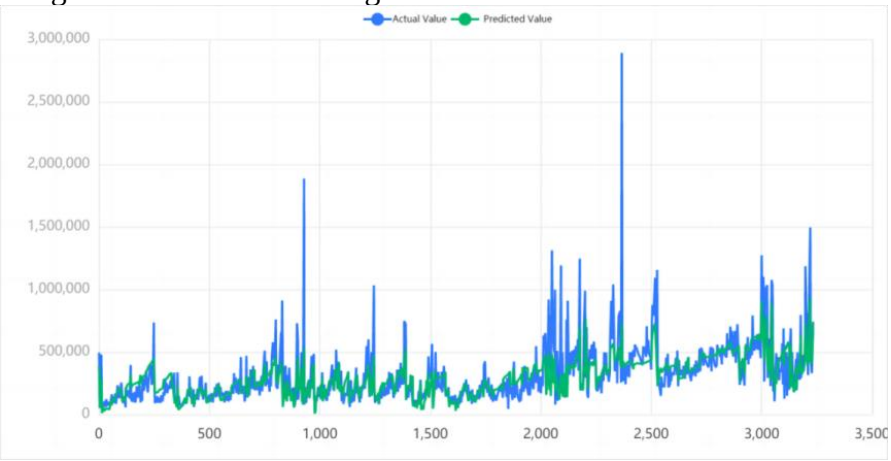

Figure 5: Ridge regression result

It can be seen that the fitting effect is excellent.

3.2. ANOVA to determine the regional influence

The results of ANOVA are shown in Table 4:

Table 4 : Analysis of variance results table

variable	Value	Sample	Average value	standard deviation	F	P
LP	Caribbean	461	339127.154	169163.14	11.571	0.000***
	USA	453	282842.285	180832.641		
	Europe	2319	295352.74	203671.11		
	Total	3233	299841.688	196652.25		

The above table shows the results of analysis of variance, including mean ± standard deviation results, F-test results and significance P-value.

Analysis: The average values of Caribbean, USA and Europe on LP are 339127.154*/282842.285*/295352.740*; The P value of variance analysis is 0.000***≤0.05, so the statistical result is significant, indicating that different Geographic regions have differences in price.

The results of the effect quantification analysis are shown in Table 5:

Table 5 : Effect quantization map

Analysis item	Difference between groups	Total deviation	(Partial η^2)	Cohen's f
LP	889120521439.667	124988251327837.72	0.007	0.085

The results of effect quantification analysis showed that, based on LP, Eta square (η^2 value) was 0.007, indicating that 0.7% of the data differences were due to differences among different groups. Cohen's f value is 0.085, indicating that the difference degree of effect quantification of data is minimal.

Result analysis shows that different regions have a high difference in price, indicating that different regions affect price to some extent. However, through quantitative effect analysis, it is known that different regions have little difference in price, indicating that regions have little influence on price.

3.3. Whether the regional factors are consistent for the different variants

The results of a two-way ANOVA analysis are shown in Table 6:

Table 6 : Analysis of variance results table

item	quadratic sum	degree of freedom	mean square	F	P
Variant	33349241193232.203	345	96664467226.76	3.928	0.000***
Geographic Region	4945159915.403	2	2472579957.701	0.1	0.904
Variant * Geographic Region	46568955802341.99	690	67491240293.249	2.742	0.000***

It can be seen from the figure that for the same variant, the influence of region is not high. Through the cross effect analysis of variant and region, it can be seen that there is cross effect between variant and region, that is, the influence of region may have particularity for different variants, so there is no consistency in the influence of region on price of different variants.

3.4. Practical and statistical significance

The model based on the ridge regression can predict the price of second-hand sailboats by understanding the general data of sailboats.To better understand market price trends and thus make purchase decisions more wisely.

The practical implication is that if there is a regional correlation effect on the price of sailboats, then this means that it may need to be taken into account when buying or selling sailboats in certain areas.This conclusion may also provide market participants with more comprehensive market information and help them make more informed decisions.

In terms of statistical significance, the results of effect quantification analysis show that there is little price difference in different regions, indicating that different regions have little influence on price. This conclusion is very important for statistical significance because it indicates that price differences between regions are not significant, meaning that these differences are likely due to random factors rather than real differences.

4. Conclusion

4.1. interpretation of result

Through the ridge regression equation, a model for predicting the price of second-hand sailboats is determined, with Beam, Make, Multiple hulled ship, Year, Displacement, SailArea, Draft, Length and regional parameters roughly available.

Prices vary significantly across geographic regions according to ANOVA results table, that different Geographic regions have differences in price.However, through quantitative effect analysis, it was found that prices in different regions were not differ, indicating that different regions had little influence on prices. Through the cross effect analysis of variation and regions, we can see that there is a cross effect between variation and regions, that is, the influence of regions on different variants may be special, so there is no consistency in the influence of regions on different variant prices.

This conclusion is very important for statistical significance because it indicates that price differences between regions are not significant, meaning that these differences are likely due to random factors rather than real differences.

4.2. Future feasibility and application prospects

In the future, with the continuous development of artificial intelligence technology, the second-hand sailing price evaluation method based on variance analysis can also be combined with artificial

intelligence technology to form a more intelligent evaluation model. By analyzing a large amount of historical data, machine learning algorithms can gradually improve their ability to predict the price of second-hand sailboats. The factors identified in this paper can predict the price of used sailing boat more accurately and scientifically.

The second-hand sailing market is large and in great demand. With the increasing demand for recreation and Marine sports, the market demand for second-hand sailing boats is also expanding. There is also an increasing demand to evaluate and forecast the price of used sailing by establishing the price model.

References

[1] Bernhard S,Norbert K,Peter R, et al. Minimal sample size in balanced ANOVA models of crossed, nested, and mixed classifications[J]. Communications in Statistics - Theory and Methods,2023,52(6): 1728-1743.

[2] Shang G,Guohun Z,Alina B, et al. Stroke Localization Using Multiple Ridge Regression Predictors Based on Electromagnetic Signals[J]. Mathematics,2023,11(2): 464-464.

[3] Xiaoyu W,Xingyuan W,Bin M, et al. High-performance reversible data hiding based on ridge regression prediction algorithm[J]. Signal Processing,2023,204.

[4] Braga M M D,Dimitri D,Gabriela M O M, et al. Beyond ANOVA and MANOVA for repeated measures: advantages of GEE and GLMM and its use in neuroscience research.[J]. The European journal of neuroscience,2022,56(12): 6089-6098.

[5] Borislava T. ANOVA bootstrapped principal components analysis for logistic regression[J]. Croatian Review of Economic, Business and Social Statistics,2022,8(1): 18-31.

[6] Liu S X. Bias Correction for Eta Squared in One-Way ANOVA[J]. Methodology,2022,18(1): 44-57.

[7] Wilcox R. One-Way and Two-Way ANOVA: Inferences About a Robust, Heteroscedastic Measure of Effect Size[J]. Methodology,2022,18(1): 58-73.

[8] Someswara C R,G.N.V.G. S,Butchi K R, et al. Method for identification of 10 SSR markers from monkey genomes and its statistical inference with One & Two-way ANOVA[J]. MethodsX,2022,9: 101833-101833.

[9] R R W. Two-way ANOVA: Inferences about interactions based on robust measures of effect size.[J]. The British journal of mathematical and statistical psychology,2021,75(1): 46-58.

[10] A B,H S,H R N, et al. Relationship of quality management system standards to industrial property rights in Indonesia using Spearman Correlation Analysis Method[J]. IOP Conference Series: Earth and Environmental Science,2021,623(1): 12092-12092.