DOI: 10.54254/2755-2721/68/20241449

A used sailboat pricing model based on multiple linear regression

Qiman Li[1,2,*], Zhiyong Zhang[1,3]

[1]School of Information Engineering, Northwest A&F University, Shanxi, China

[2]Maztina@nwsuaf.edu.cn
[3]zzy@nwsuaf.edu.cn
*corresponding author

Abstract. The value of sailboats varies depending on the region and market conditions. Brokers need a better understanding of the sailboat market and make informed decisions to mitigate potential risks. To address these issues, this paper employs methods of multiple linear regression analysis and the grey correlation forecasting method. By utilizing data analysis techniques, a commonly used sailboat pricing model is established. Specifically, we first preprocess the given dataset and encode the processed data for subsequent model use. Additionally, the paper selects the three most influential factors on the pricing model: length, make, and year, and employs the method of independent weighting coefficients for multiple linear regression analysis calculation. Furthermore, we analyze the regional impact on listing prices of used sailboats. Based on these analyses, the proposed model can provide data analysis and sales recommendation reports for brokers. Lastly, the model proposed in this paper demonstrates strong accuracy and robustness, making it easy to implement with other datasets.

Keywords: Multiple linear regression, Gray correlation prediction, Independence power coefficient method, Paired sample t-test, ANOVA.

1. Introduction

Sailing, as a fascinating water sport, boasts a wide range of enthusiasts worldwide. However, the value of sailboats is not fixed but varies depending on regional and market conditions. Significant fluctuations in sailboat prices across different regions and market environments present challenges for brokers and consumers, necessitating a better understanding of the sailboat market and prudent risk mitigation strategies for potential future market scenarios.Given the complexity and variability of the sailboat market, in today's highly developed information age, employing data analysis and modeling methods to reveal patterns and trends becomes particularly crucial.

In various related research work, the utilization of the BP Neural Network model has been notable[1]. Moreover, an innovative approach to modeling price fluctuations has emerged[2].Ming [3] et al. crafted an econometric framework for assessing second-hand tanker prices, focusing on the analysis of various independent factors. This approach integrates identified mean and variance structural break points into the GARCH model for enhanced analysis.Fan [4] et al. utilized a multiple linear regression approach to construct a predictive model for typhoon frequency spanning the years 1965-1999. This model holds potential for operational seasonal forecasting of typhoon frequency in the Western North Pacific (WNP)

DOI: 10.54254/2755-2721/68/20241449

region.Kavussanos [5] and A.H. Alizadeh substantiated the pricing dynamics of second-hand bulk carriers under the rational expectations hypothesis. Their employed the GARCH-M model for corresponding interpretation.[6]developed a random forest-based predictive model for forecasting second-hand sailboat prices. It employed hypothesis testing to assess the correlation between various features and the predictive model but did not investigate the influence of regional factors on prices.

This study aims to develop a comprehensive sailboat pricing model using multiple linear regression analysis and grey relational prediction methods. The model is intended to provide brokers and consumers with more accurate and comprehensive market information and decision support. Additionally, it will account for regional variations in sailboat pricing. Ultimately, this research contributes by offering a novel, data-driven decision-making approach for participants in the sailboat market. By establishing accurate and reliable pricing models and market forecasting methods, we can assist brokers and consumers in better understanding and responding to market dynamics, thereby enhancing their decision-making efficiency and market competitiveness.

2. Data preprocessing

2.1. Missing data processing

Like most real-world datasets, it may have missing data. We use the ismissing function that comes with matlab to find out where the null values are. We find that the missing values are all regions, and for the missing regions in Europe, we fill them using the plural of the European countries; for the missing regions in the US, we fill them using the plural of the US regions. The problematic areas are shown below:

Table 1. Tables with missing values

Make	Variant	Length(ft)	Geographic Region	Country/Region/State	Listing Price(US)	Year
Beneteau	Oceanis 54	54	USA		$479,805	2013
Delphia	46 cc	46	Europe		$314,606	2013
Bavaria	Cruiser 46	46	Europe		$201,460	2014

2.2. Abnormal value processing

While coding the data, we found that some of the categorized data which should have the same coding (identical categorized names) were coded differently. After examining the original data set, we found that some of the variant data items had spaces at the end of the data, which caused the program to recognize them as different names, for example, when the variant was "38 Cruiser" the number was 2, and when the variant was For example, when the variant is "38 Cruiser" the number is 3. The spaces cause the same type of variant to have different codes. In response, we wrote a program to remove these null characters uniformly. The comparison between before processing and after processing is as follows:

Table 2. Abnormal data

Make	Variant	Length(ft)	Geographic Region	Country/Region/State	Listing Price(US)	Year
2	2	38	1	2	$75,178	2005
2	2	38	1	2	$66,825	2005
2	2	38	1	2	$54,661	2005
2	$\boxed{3} \rightarrow 2$	38	1	2	$53,447	2005
2	2	38	1	3	$91,101	2005

DOI: 10.54254/2755-2721/68/20241449

2.3. Data de-duplication

We found that some of the data items in the given dataset were identical except for the Listing Price data item, which we thought might be sold at different times of the year. Since our evaluation of time is only accurate to the year, we combined these data into one data item by taking the average value of Listing Price. The comparison before and after the merge is as follows:

Table 3. Comparison before and after data consolidation

Make	Variant	Length(ft)	Geographic Region	Country/Region/State	Listing Price(US)	Year
Bavaria	38 Cruiser	38	Europe	Croatia	$75,178	2005
Bavaria	38 Cruiser	38	Europe	Croatia	$66,825	2005
Bavaria	38 Cruiser	38	Europe	Croatia	$54,661	2005
Bavaria	38 Cruiser	38	Europe	Croatia	$53,447	2005
Average value					$62,528	

3. Model Construction and Analysis

3.1. Pricing model for a sailboat

3.1.1. Selection of predictive factors

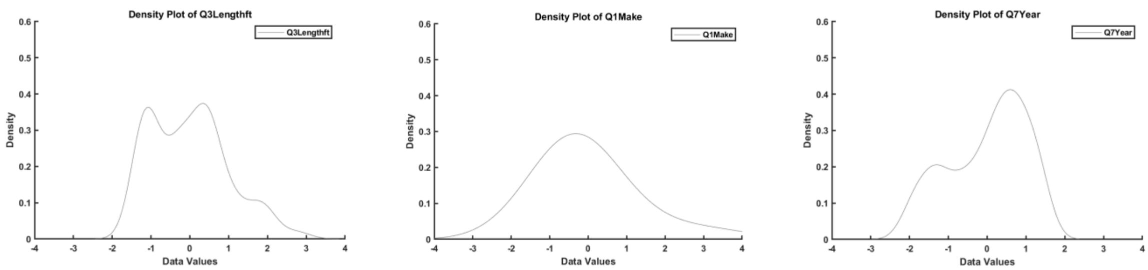
Since not all the data items given in the question have a very strong relationship with the final price, some of the indicators have a strong covariance, and this can have a large impact on our modeling results. In order to better remove the correlation between the factors, we use the independence weight coefficient method to select the top three factors that most affect the pricing results for multiple regression analysis.After performing the analysis of the related experiments, we selected three factors with higher importance of Length, Make, and Year as predictors.

3.1.2. Multiple linear regression modeling

In order to build a multiple linear regression model, we need to make the three selected discrete indicators continuous.First, the Z-score method was used to eliminate the magnitudes of the three variables and stan- dardize the data with the following formula:

$$x = (x - \mu)/\sigma \tag{1}$$

where x is the original data, μ is the mean of each variable, and σ is the standard variance of each variable. The obtained z-scores have the same magnitude, and here we obtain three sets of data showing a skewed distribution, as follows:



(a) Length (b) Make (c) Year

Figure 1. Skewed distribution

Since the data did not present a standard normal distribution with standard deviation of 1, we normalized the obtained Z-scores and mapped the data to the range [0,1] using the Min-Max normalization method (Min-Max Normalization).The three sets of variables after normalization are then used as independent variables, and the data are randomly assigned, of which 80% is used as the training set and 20% as the test set, and multiple linear regression is performed using the training set to obtain the model:

$$AvgPrice = 241772.5554 + 64737.6655 * Q1Make + 74326.2587 * Q3Length + 51629.5699 * Q7Year + \varepsilon \tag{2}$$

The paired sample t-test determines whether there is a significant difference between the means of the two sample data by calculating the p-value, kurtosis, and skewness of the two samples.We name the listed price obtained from the test set as Pred value and the Listing Price obtained from the training set as True value, and use the paired-samples t-test method to determine whether there is a significant difference between Pred value and True value.First, we used the Shapiro-Wilk test to test whether the data satisfied the normal distribution and obtained the p-value. If the p-value was less than 0.05, the data did not satisfy the normal distribution and could not be compared between groups. The results of the normality test for paired differences are shown in the following table:

Table 4. Normality test results

Variable Name	Sample size	Average value	Standard deviation	Bias value	Peak	S-W test	K-S test
True_Value	361	252430.5	182143.534	3.678	22.45	......	......
Pred_Value	361	245010.807	117889.321	0.424	0.289	......	......
True_value pairs Pred_Value	361	7419.693	124936.568	3.479	24.158	0.764(0.000***)	0.137(0.000***)

Note: ***, **, * represent 1%, 5%, 10% significance level respectively

As can be seen from the above table, P is 0.000***, the peak value is 24.158, and the skewness is 3.479. Here the P value is less than 0.05, the data do not satisfy the normal distribution, but usually the realistic research situation is difficult to satisfy the test, the absolute value of its peak is greater than 10 and the absolute value of skewness is greater than 3, then combined with the normal distribution histogram can be described as basically in line with the normal distribution. The output of the True value-Pred value data normality test is shown in the following histogram:

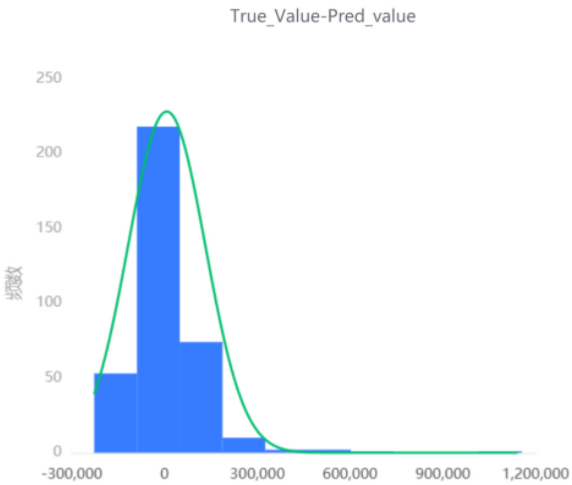


Figure 2. Histogram of normality test

This normal plot shows a bell shape, which is not absolutely normal, but can be accepted as a normal distribution. Further, we were able to perform a t-test based on True value paired with Pred value to obtain the significance p-value. The results of the paired samples t-test are presented in the following table:

Table 5. Normality test results

Pairing variables	Average value ± Standard deviation			t	df	p	Cohen's d
	Pairing 1	Pairing 2	Pairing difference (pairing 1 - pairing 2)				
True_value pairs Pred_Value	252430.5±182143.534	245010.807±117889.321	7419.693±64254.213	1.13	360	0.03	0.059

Note: ***, **, * represent 1%, 5%, 10% significance level respectively

The above table shows a significant p-value of 0.260, which does not present significance, no significant difference between True value and Pred value, and a Cohen's d-value of 0.059, which is a small difference in magnitude.The comparison chart between True value and Pred value is as follows:

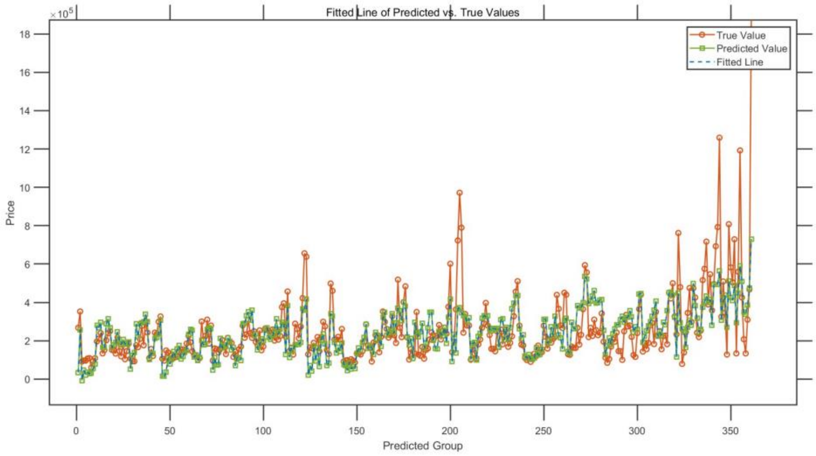


Figure 3. The degree of the model fit

It can be concluded from **Figure 3** that the model fits with a high accuracy, which can reach 94.1%.

3.1.3. Optimize the model

Because the multiple linear regression model contains only three data items, Length, Make, and Year, some of the predicted values differ greatly from the actual sales price, we added the GDP data of the region we are in to explore the impact of regional factors on our model. By adding it to the multiple linear regression model, the trained model is as follows:

$$\boldsymbol{AvgPrice} = -\mathbf{87945.0811} + \mathbf{196655.3026} * \boldsymbol{Q1Make} + \mathbf{318423.7158} * \boldsymbol{Q3Lengthft} + \mathbf{159444.2697} * \mathbf{Q7Year} + \mathbf{84994.529} * \mathbf{Q8GDP} + \boldsymbol{\varepsilon} \tag{3}$$

Same as the paired-sample t-test method in step 3, we also divided the data into training set and test set, and passed the paired-difference normality test, True value-Pred value data normality test and paired-sample t-test. The obtained t-test results are as follows:

Table 6. Normality test results

Pairing variables	Average value ± Standard deviation			t	df	p	Cohen's d
	Pairing 1	Pairing 2	Pairing difference (pairing 1 - pairing 2)				
True_value pairs Pred_Value	243536.443 ±107721.024	250692.348± 178262.468	(-7155.904)±(-70541.444)	-1.01	360	0.315	0.059

Note: ***, **, * represent 1%, 5%, 10% significance level respectively

A comparison chart between True value and Pred value can be obtained as follows:

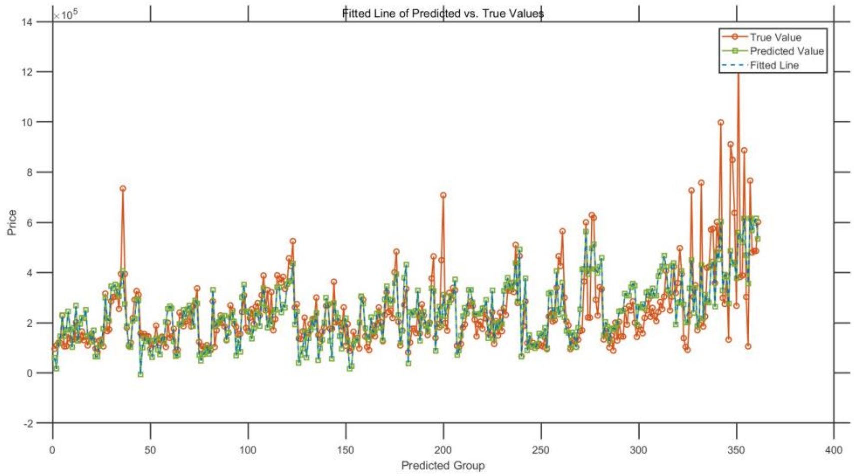


Figure 4. The degree of the new model fit

From the above figure, we can obtain that the accuracy of the optimized model has improved, with a fit accuracy of 94.7%, and incorporates GDP, a data item that reflects regional factors.

3.2. The impact model of region on sailboat listing prices

3.2.1. Relevant data research

First,we conduct the statistics of the total number of regional listed boats on the dataset to observe the best-selling degree of monohull and catamaran boats in each region.Their quantity statistics are shown in the following figure:

Table 7. Quantity statistics

	Europe	USA	Caribbean
Monohulled Sailbs	1783	385	178
Catamarans	735	108	302

From this we can see that: the United States has more monohull listings, while the Caribbean has the most catamaran listings. We can also see that the distribution of listings of monohulls and catamarans is different in each region, for example, Europe has more monohulls than catamarans, while

DOI: 10.54254/2755-2721/68/20241449

the Caribbean has more catamaran listings than monohulls.This is related to the economy of each region and the daily habits of the local population.
Next, we would like to analyze the impact of region on the listing price.Here, we test whether Region has an impact on the final listing price by Spearman correlation analysis.The correlation coefficient table is as follows:

Table 8. Correlation coefficient table

	AVG Listing Price (USD)	Country/Region/State	Geographic Region
AVG Listing Price (USD)	1(0.000*)	0.098(0.000***)	0.007(0.777)
Country/Region/state	0.098(0.000***)	1(0.000***)	0.644(0.000***)
Geographic Region	0.007(0.777)	0.644(0.000**)	1(0.000***)

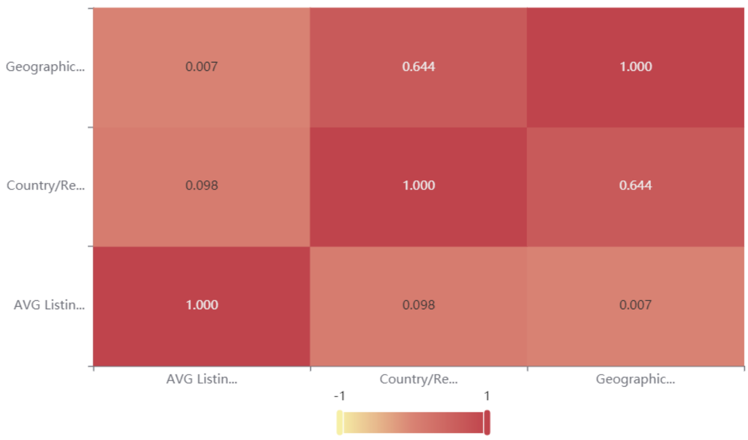


Figure 5. Correlation coefficient table

Based on this table we can see that the Country/Region/State data has a β-factor effect on the model compared to the data items in the other tables, but it is not a major influence. The Geographic Region data, on the other hand, has almost no effect on the results. It can be seen that factors such as the manufacturer of the sailboat, the skipper and the year of production of the boat have more influence on the modeling results than the region.

3.2.2. Multiple linear regression modeling

From an overall perspective the effect of region on listing price is not significant. However, in order to analyze whether the effect of region on listing prices is common to Europe, the United States, and the Caribbean, we will use multiple linear regression to discuss whether there are regional factors (latitude and longitude, land and sea location, regional area e.g.) that are significant factors in sailing prices.Here we use a multiple linear regression model to analyze the effect of region on used sailboat prices. In this model, we transform geographic regions and countries/regions/states as categorical variables into dummy variables (also called dummy variables) and then add them to the multiple linear regression model. The equation of the model is as follows:

$$\boldsymbol{Price} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} * \boldsymbol{Q1Make} + \boldsymbol{\beta_2} * \boldsymbol{Q2variant} + \boldsymbol{\beta_3} * \boldsymbol{Q3Lengthft} + \boldsymbol{\beta_4} * \boldsymbol{Q7Year} + \boldsymbol{\beta_5} * \boldsymbol{Q8GDP} + * \boldsymbol{dummy1} + \aleph_2 * \boldsymbol{dummy2} + \aleph_3 * \boldsymbol{dummy3} + \ldots + \aleph_{72} * \boldsymbol{dummy72} + \aleph_{73} * \boldsymbol{dummy}73 \quad (4)$$

DOI: 10.54254/2755-2721/68/20241449

The country/region/state dummy variables are denoted as dummy1,dumm2,etc.The fitted coefficients before each variable are shown in the following table (due to space limitations, only some are shown)

Table 9. Variable coefficient

Variable			Coefficient
dummy5	Europe	Greece	-0.043316
dummy6	Europe	Turkey	-0.029346
dummy7	Europe	Croatia	-0.057557
dummy8	Europe	Netherlands	0.056093
dummy9	Europe	Germany	-0.02866

The model allows us to derive price coefficients for each geographic region and country/region/state that represent the offset of the sailing price in that region relative to the benchmark price. We performed an ANOVA on the resulting data to test for significant price differences across countries/regions/states. If the p-value of the ANOVA test is below 0.05, it means that prices in this country/region/state are significantly different from those in other regions.The results of the ANOVA made are as follows:

Table 10. Analysis of variance results

Variable			Coefficient	pValue
dummy5	Europe	Greece	-0.043316	0.44997
dummy6	Europe	Turkey	-0.029346	0.61041
dummy7	Europe	Croatia	-0.057557	0.3152
dummy8	Europe	Netherlands	0.056093	0.33212
dummy9	Europe	Germany	-0.02866	0.62046

The dummy variable for each region is less than 0.05. indicating that none of the region effects are particularly significant.In order to find intra-regional differences, price heat maps were created to visualize the differences in listing prices between regions.

(a) **(b)** **(c)**

Figure 6. Thermodynamic diagram

This leads us to the following consistent regional pricing conclusion: in general, region is not a major factor in listing price compared to factors such as vessel length and vessel type. However, when it comes to a specific country or continent, boats are priced higher in regions close to the oceans and more economically developed, while boats are priced lower in regions close to the inner sea or near smaller waters or less economically developed. This is similar to the sales situation of other products.

DOI: 10.54254/2755-2721/68/20241449

4. Sensitivity analysis

The purpose of sensitivity analysis is to check the robustness of the model,i.e., whether the model can maintain the prediction accuracy in the case of large changes in the input parameters or outliers. Based on the results of sensitivity analysis, we can adjust the range of values of input parameters, optimize the experimental design, and improve the prediction ability and accuracy of the model.Based on the significance levels of each variable obtained from 3.1 section, the input variables that require local sensitivity analysis are identified as Make and Lengthft, respectively.Next, the sensitivity of the AvgPrice pricing model to Make and Lengthft in this price model is analyzed separately, keeping the other parameters unchanged from those used in 3.1 section.The coefficients of Make and Lengthft were reduced by 10% respectively, and the model was retrained, and the predicted values were t-tested against the true values. The results of the paired-sample t-test showed that the magnitude of the difference between Lengthft was 0.15 and the magnitude of the difference between Make was 0.09, and the magnitude of the difference between the two main factors did not change much compared to the original model; the fitted curves and the magnitude of the difference between the two are as follows:

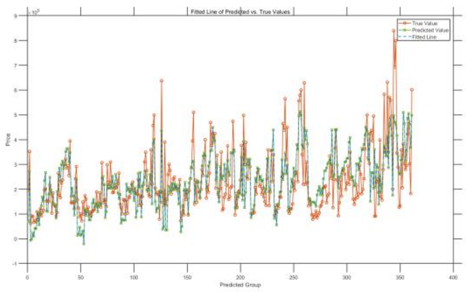


Figure 7. Make sensitivity analysis

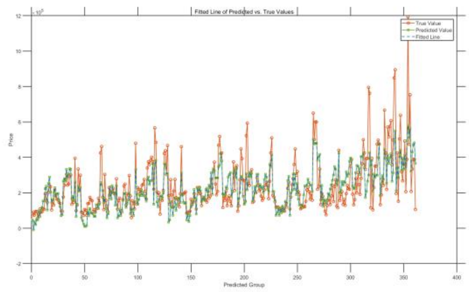


Figure 8. Length sensitivity analysis

This can show that our model has strong stability and is a very good model.

5. Conclusion

In this paper, we delve into the variations in the value of sailboats across different regions and market conditions. We recognize the need for brokers and consumers to gain deeper insights into the sailboat market to make informed decisions in potential future market scenarios. To address these issues, we employ multiple linear regression analysis and grey correlation forecasting methods to construct a data-driven used sailboat pricing model and conduct research and optimization on it.We select the three most influential factors on the pricing model: length, make, and year, and employ the method of independent weighting coefficients for multiple linear regression analysis calculation. Subsequently, we investigate the impact of regional variations on sailboat value and draw corresponding conclusions.Furthermore, the model proposed in this paper demonstrates strong accuracy and robustness, thereby providing valuable insights for brokers to better understand the sailboat market and make informed investment and trading decisions.

References

[1] Li Jianhua, "Research on Prediction of Second-Hand Ship Price Index Based on BP Neural Network," Master's thesis, Dalian Maritime University, 2008.
[2] Dai Lei, "An Empirical Study on the Volatility of the International Dry Bulk Shipbuilding and Second-Hand Ship Market," Doctoral dissertation,Shanghai Jiao Tong University, 2015.
[3] M. Zhong and Q. Shi, "Econometric Modeling of Second-hand Tanker Prices," in Automation and Logistics, 2007 IEEE International Conference on, 2007.
[4] F. Ke, "New predictors and a new prediction model for the typhoon frequency over western North Pacific," Science China: Earth Science English Edition, no. 9, p. 7, 2007.

DOI: 10.54254/2755-2721/68/20241449

[5] M. G. Kavussanos and A. H. Alizadeh, “Efficient pricing of ships in the dry bulk sector of the shipping industry,” Maritime Policy &Management, vol. 29, no. 3, pp. 303–330, 2002.

[6] J. Liu and Z. Hou, "Establishment of second-hand sailboats price prediction model based on random forest and exploration of influencing factors," 2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA), Dalian, China, 2023, pp. 1337-1342.