

Price Prediction Study of Used Sailboat Based on Random Forest Regression Model

Miao Liu¹, Yue Kong¹, Jun Liu^{2,*}

¹ Department of Mechanical Electronics, North China Electric Power University (Baoding), Baoding, Hebei, 071000

² Department of Electronics and Communication Engineering, North China Electric Power University (Baoding), Baoding, Hebei, 071000

* Corresponding Author Email: liuj31762@gmail.com

Abstract. As an ancient means of transportation, second-hand sailing boats are very frequently traded in the market. To accurately price used sailboats, this paper uses multiple linear regression to obtain variables that significantly affect prices based on preprocessed data, and removes variables with lower correlation. Subsequently, the pricing of used sailboats was predicted using random forest regression. To visualize the results, two methods are used: feature importance, and comparing the predicted results with the real results. Finally, the influence of different geographical areas on the price of sailing use is studied to help and guide governments and enterprises to develop development strategies and find effective ways to develop cities and regions more effectively.

Keywords: Random Forest Regression; Multiple linear regression; Used Sailboat.

1. Introduction

With the development of the market economy, the shipping industry is one of the few market sectors with significant assets that can be traded. The second-hand ship market has an important economic position in the shipping industry. Considering the diversity of factors affecting the second-hand ship market, further research on the changing patterns of second-hand ship prices is urgently needed. Therefore, it is of profound significance for us to study the price change patterns of second-hand ships.

In view of the importance of second-hand ship prices in the shipping market, domestic and foreign experts and scholars have never stopped to study the price of second-hand ship prices. Since the 1990s, many experts and scholars have carried out systematic research and discussion on the price fluctuation law of second-hand ships, evaluation and prediction, purchase decision, investment timing selection, investment risk and financing trend. Many novel ideas and scientific research methods are worth learning from and learning and have a positive effect on analyzing the relevant factors affecting the price of second-hand ships.

CAI Liming, Lu Chunxia (2008) ^[1] analyzed the time series data of Panama type (60-100 kdw) bulk used ship price provided by Clarkson Company from 1984-2004, established the corresponding ARIMA (p,d,q) model, and predicted the price from 2005-2008. The predictive performance of the model was measured against the mean absolute error (MAE), mean relative error (MPE), and root mean squared error (RMSE). After testing, the application of the model has achieved good results. Luo Fucai (2011) ^[2] established five BP neural network models for the study of second-hand ships and ships, but because there is no detailed classification of ships, it is difficult to accurately explain the prices of second-hand ships and ships. Liu Yang and Gao Jin (2012) ^[3] studied the value evaluation of second-hand ships from the perspective of the influencing factors of second-hand ship finance lease. Foreign scholars started early in the research of second-hand ship market and other segments, through time series model, econometrics theory, physical option theory, and mainly focused on the study of ship price volatility and ship investment transaction decision. In terms of ship price and investment decisions, Beenstock and Vergotis (1993) ^[4] for the first time used the model of asset pricing. Firstly, the whole ship trading market is assumed to be effective, and then according to the theory of supply and demand balance, is established. Stopford M Put forward the factor

classification theory, and classified the shipping market according to the predictability, namely, tangible, irregular, technical and behavioral.

2. Listing Price Prediction Model of Used Sailboats

On the basis of preprocessed data, multiple linear regression was used to identify significantly significant variables and remove those with less **associated**^[5]. Subsequently, the pricing of second-hand vessels was predicted using random forest regression. To visualize the results, two methods are used here: feature importance, and comparing the predictions to the real results.

2.1. Multiple linear regression model

When faced with multiple variables, in order to avoid overfitting in the subsequent forecasting process, this paper filters out the variables that have a significant effect on prices^[6]. Multiple regression analysis is a common method to screen out variables that have a significant effect on price.

The value of the dependent variable is predicted using a regression model and the difference between it and the actual observed value is called the residual. Calculating the distribution of the residuals can help us understand the nature of the residuals. Significance tests are performed using statistical methods, by which to determine whether the effect of each independent variable on the dependent variable is significant. If the p-value of an independent variable is less than the significance level, it indicates that it has a significant effect on the dependent variable, otherwise it indicates that it does not have a significant effect on the dependent variable. This paper selects the following variables for multiple linear regression: GDP. watercourse **rates** (WCR). displacement of each **sailboat** (DWT). mainsail **area** (MA). sailboat engine **horsepower** (SP). Width. Height.

$$p_i = \beta_{i0} + \beta_{i1}x_{GDP} + \cdots + \beta_{i7}x_{Height} + \varepsilon \quad (1)$$

Where ε is the random error follows a normal distribution $N(0, \sigma^2)$, β_{i0} , β_{i1} , $\cdots$, β_{i7} are regression coefficients.

This paper estimates all the coefficients by OLS and test the significance of the regression coefficients by t-statistic, which can be significant when $p < 0.1$, and the results of the regression are as Table 1:

Table 1: Multiple regression coefficients and significance tests

β	x_{GDP}	x_{WCR}	x_{DWT}	x_{MA}	x_{SP}	x_{Height}	x_{width}
β_1	0.285	0.064	0.6	0.307	1.294	-0.06	-0.051
p_1	1.3214E-29	0.001	1.2279E-8	1.9208E-16	5.0779E-31	0.102	0.953
β_2	0.652	0.587	0.6986	1.395	0.532	-0.822	-0.038
p_2	1.0556E-25	1.5324E-9	0.003	1.7456E-17	1.7564E-7	0.356	0.248

By analyzing the calculation results, it is found that the ship height and ship width have little influence on the price. This means that in the future, these factors on the model. Instead, all other important variables will be focused on.

2.2. Random Forest Regression Prediction Model

To predict the pricing of used boats, this paper choses a random forest regression algorithm^[7]. This algorithm is a decision tree based regression model that can effectively identify the relationship between the independent and dependent variables^[8]. During the prediction process, the regression prediction will be performed using variables significantly associated with old boat pricing.

This paper divides the data into a training set and a prediction set, where 70% of the data is used for training the model and the remaining 30% is used for prediction. To perform hyperparameter tuning, this paper uses a grid search method to set different combinations of parameters and use methods such as cross-validation to determine the best hyperparameters.

This will help ensure the accuracy and generalization capability of the model, which is implemented in the Figure1 process:

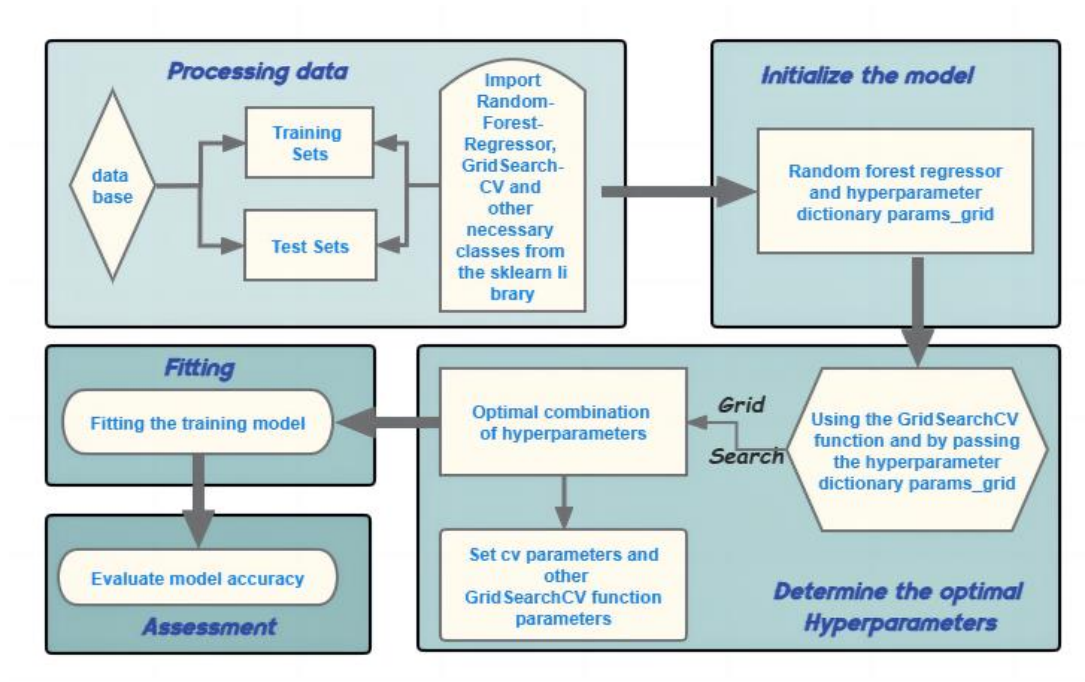


Figure 1: Algorithm flow chart

Using python to train the model with the best hyperparameters obtained according to the grid search method. The prediction results are as Figure 2. Figure 3. Figure 4:

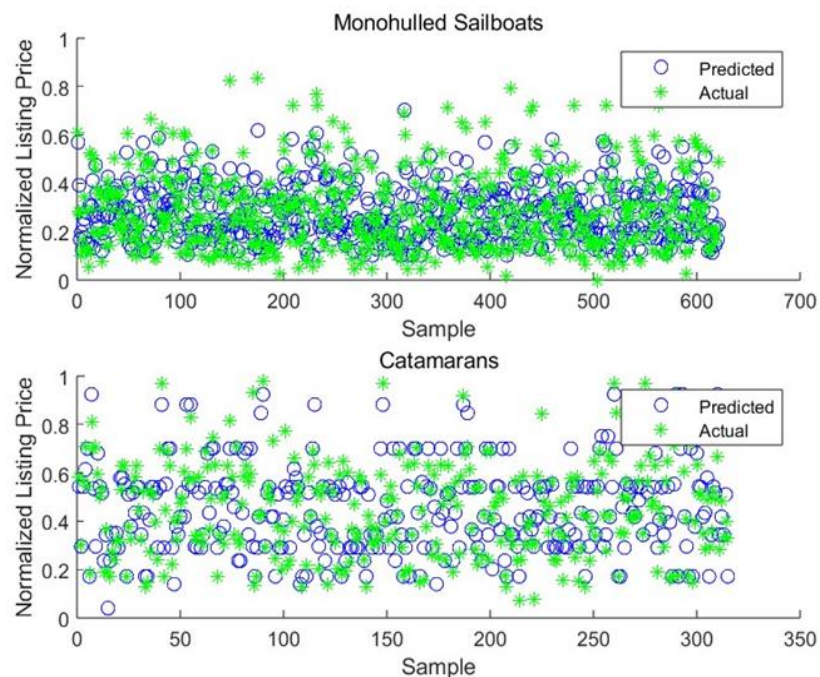


Figure 2: Comparison of predicted and actual values

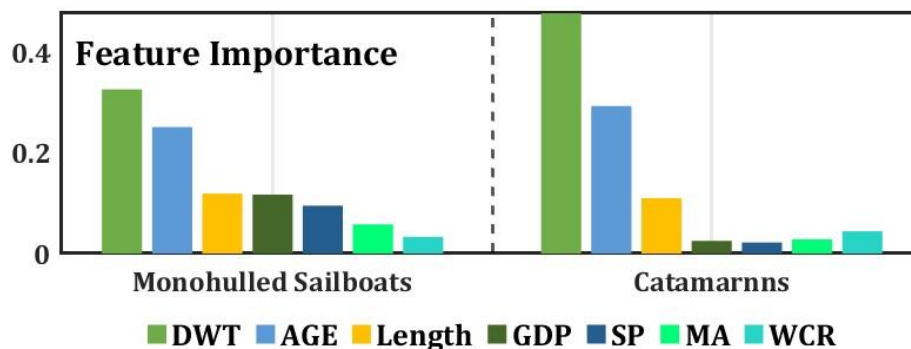


Figure 3: Feature Importance

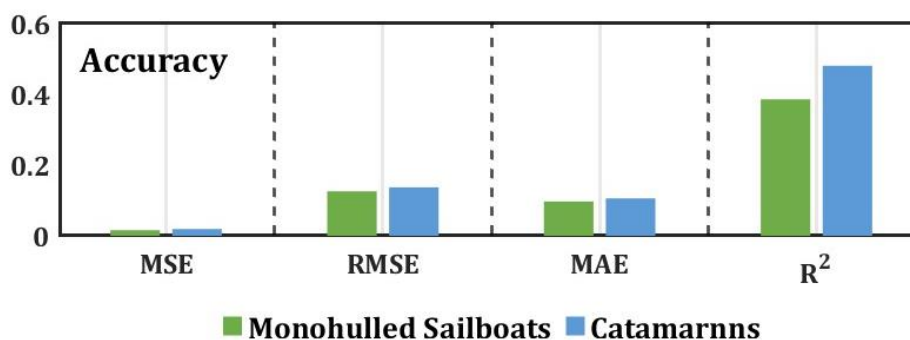


Figure 4: Accuracy

At first, this paper uses the method of comparing predicted results with true results to understand the gap between predicted and true values. By this method, this paper finds that the gap between the predicted and true results is relatively small, indicating that the model has good predictive power. This paper then compares feature importance to understand which variables have the most impact on the prediction results. Among them, the main differences between monohull and catamaran are shown in GDP, SP, MA and WCR, and these variables have a greater impact on the listed price of monohull. With this approach, this paper finds that the three variables of displacement, age, and length of the sailboat have the greatest and consistent order of influence on the pricing of both monohulls and catamarans.

In summary, the pricing of used vessels can be better predicted by using multiple linear regression and random forest regression and outcome visualization.

3. Regional Effect Model

The price of used boats varies from region to region, which is influenced by a series of factors, such as the level of economic development, population density, water coverage, etc. in different regions. In addition, the culture of sailing consumption in different regions can also have an impact on the price of used boats. For some cities and regions where sailing consumption is more prosperous, the size of the used boat market and the number of people exposed to it are also higher. Therefore, this paper builds this model in order to study the influence of different geographical areas on used sailboat prices, which can be used to assist and guide the government and enterprises in their development strategy formulation, and to find effective ways to develop cities and regions more effectively.

3.1. Regional Clustering

Based on the available data, this paper uses the GDP and watercourse rate of each region to divide into several different regions. The optimal number of clusters was determined from the elbow plot as 3, The elbow plot is as Figure 5:

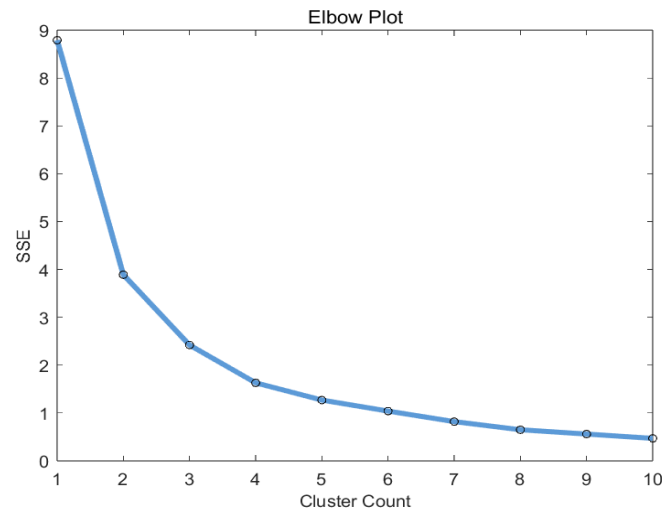


Figure 5: Elbow Plot

As can be seen from the figure, the degree of distortion varies the most for K values from 1 to 3. After exceeding 3, the change in the degree of distortion decreases significantly. Therefore, the elbow is K=3, so the number of categories can be set to 3. The results of clustering using K-means are as Figure 6.

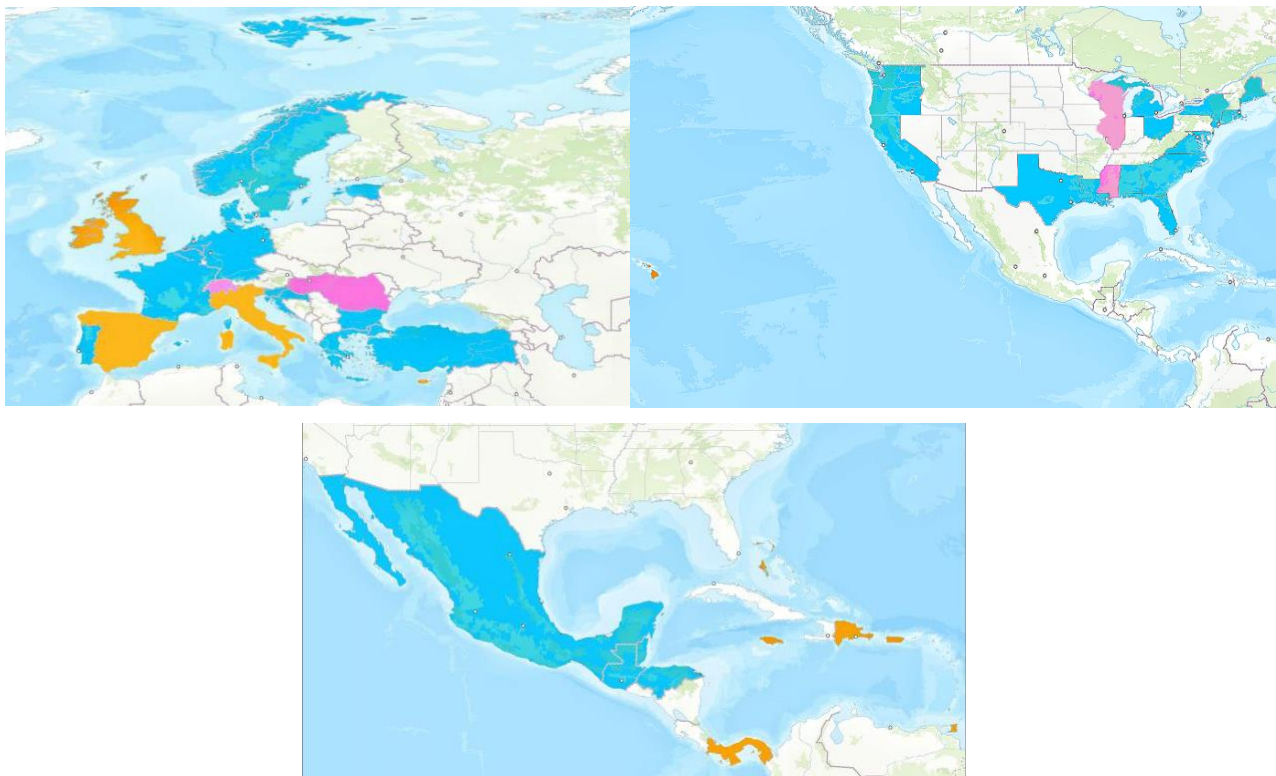


Figure 6: Clustering Result

Based on the clustering results, all regions are divided into three categories: inland regions, coastal regions and islands. This division is only an imprecise division based on clustering visualization and may differ from the actual situation of the regions. Inland areas have a low GDP and a low WCR, which means that there is less economic activity and resource use in the area and less demand for sailboats. Island regions have significantly lower GDP but the highest WCR, so they have the highest demand for sailboats, but their prices are affected by the low GDP. Coastal areas have the highest GDP and not a low WCR, so their prices are likely to be higher.

3.2. Feature Importance Analysis

This paper divided the data into three groups based on the clustering results^[9-11], and used a random forest regression prediction model^[12] to predict the prices of used boats in each of the three regions and compare the differences in feature importance, the results are shown in Figure 7:

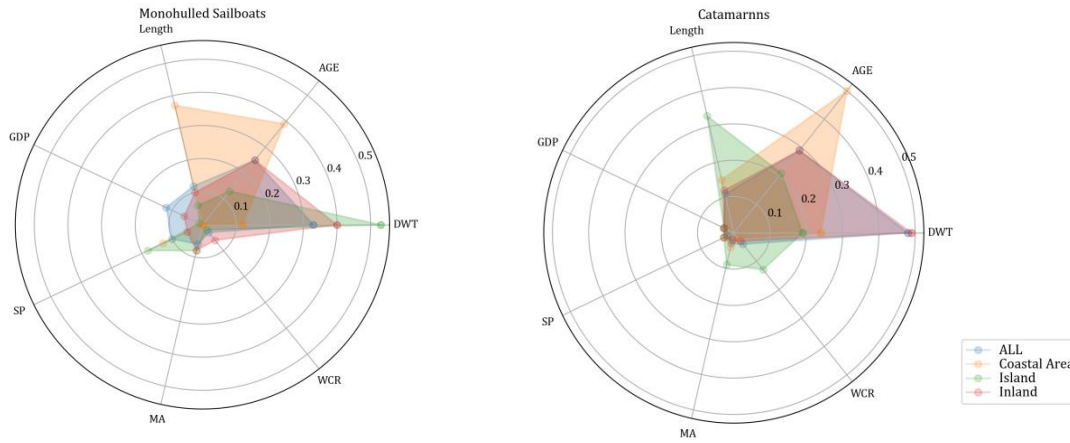


Figure 7: Regional Effect of Feature Importance

The forecast results allow us to conclude the following:

1. The feature importance of variables in the second-hand sailboat listing prices prediction model for three clusters after clustering shows significant differences.
2. For monohulls, the most important influencing factors in coastal areas are length and age of the boat; in island areas displacement has a strong influence on pricing; and inland areas displacement and age of the sailboat are the most important influencing factors.
3. For catamarans, the most important influencing factor in coastal areas is age, followed by sailboat length and displacement; in island areas length is the most important influencing factor, followed by age, displacement and water coverage of the sales area; inland areas sailboat displacement is the most important influencing factor, followed by age.

These findings explain well the regional effect of the listing price of used sailboats, which is mainly reflected in the fact that the importance of the characteristics of their variables differ significantly in the models trained with data from different regions

3.3. Regional Effects of Different Brands

In the previous model, this paper ignored the variability of the regional effect among individual brands, so by comparing the average of the listing prices of each brand in different regions, this paper determines whether the region has a significant impact on the sailing prices of all different brands. Since there are many types of sailing boats, this paper has selected the top 20 selling monohulls and catamarans respectively for our study.

Average of listing prices for each brand in region i :

$$\bar{P}_{ij} = \frac{\sum_{j=1}^{m_i} P_{ij}}{m_i} \quad (2)$$

Where m_i denotes the sales of each brand in region i , P_{ij} denotes the pricing of the j th sailboat of the brand in region i , the statistical results of all the brands are shown in Figure 8 and Figure 9:

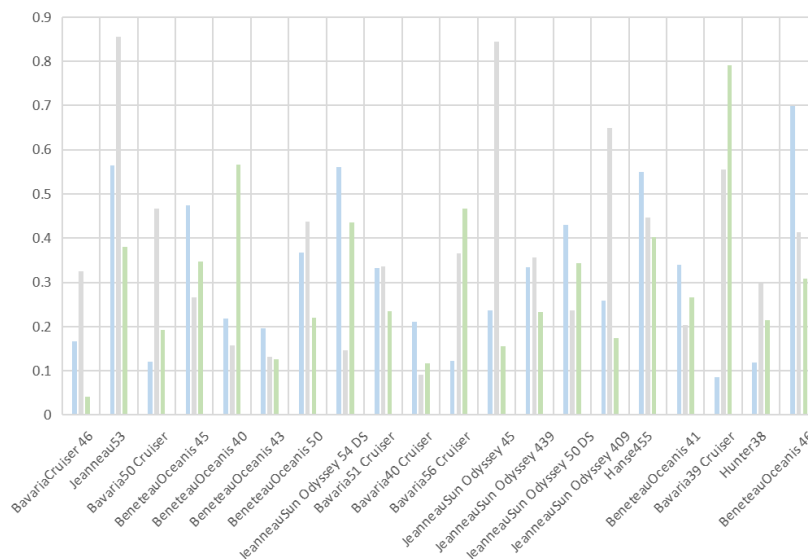


Figure 8: Regional effects of different brands in monohull ships

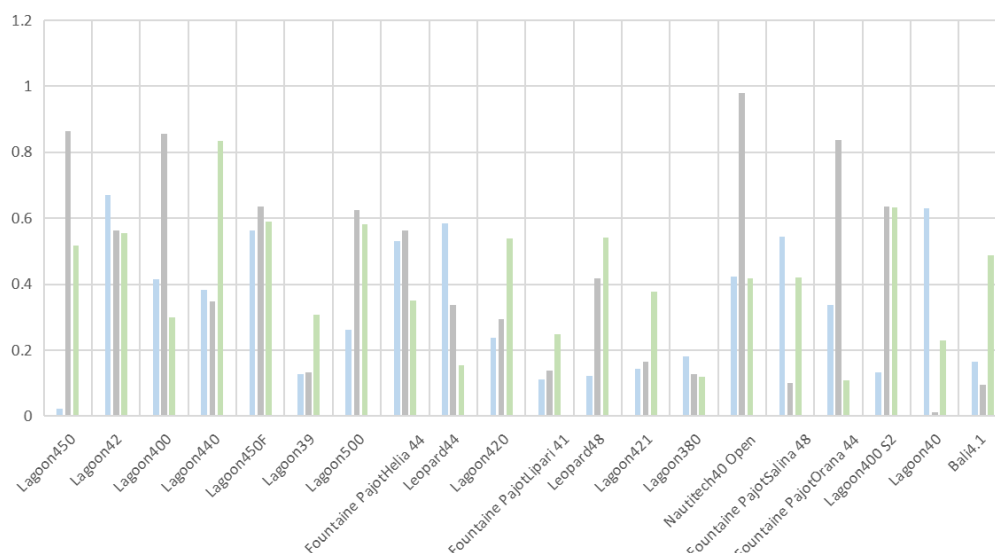


Figure 9: Regional effects of different brands in catamarans

It is found that in the monohull and catamaran market, most brands have regional effects, but some do not. Specifically, 85% of the monohull brands have a regional effect and 90% of the catamarans. This suggests that consumer demand and preference for brands varies across regions.

For brands that do not have a regional effect, it may be because they are not well known and credible enough in the region, or because there is not a high demand for the brand among consumers in the region. In this case, the brand may need to reevaluate its marketing strategy and product positioning in different regions to increase its awareness and sales in that region. In addition, the effects of some brands vary across regions. For example, the price impact of certain brands is higher in coastal regions, but relatively lower in inland regions. This suggests that consumer demand and preference for brands differ across regions.

In summary, although the regional effect does not apply to all brands, most brands in the monohull and catamaran markets are still affected by the regional effect. Therefore, brands need to adapt to local market conditions in order to increase their awareness and sales in the region.

4. Conclusions

The value of sailing boats is influenced by various factors, such as market conditions, age, etc. This paper studies the pricing model of sailing trading, which helps sailing dealers to determine the

most appropriate listing price, thus promoting the success of sailing trading. Before examining the pricing model for secondary sailing, this paper used multiple linear regression to screen for significantly associated variables. Using the prediction of random forest regression to study the characteristic importance of different variables. The results show that the displacement characteristic importance of monohull vessel and catamaran reached 35% and 47%, respectively, indicating the greatest impact on the market price. Furthermore, the characteristic importance of age reached 28% and 31%, respectively, and other variables were less important in the model.

To investigate the regional impact of prices, K-means clustering was used to classify all regions as coastal, inland and island according to GDP and channel rate (WCR). The random forest regression prediction model was trained and differences in feature importance were compared. Moreover, in order to investigate whether all sailing brands have regional effect, this paper selected some large capacity brands and found that 85% of single-shell brands had regional effect and 90% of catamarans had regional effect.

References

- [1] Cai Liming, Lu Chunxia. Application of ARIMA model in predicting second-hand ship prices [J]. China Shipping, 2008 (06)
- [2] Luo Fucui. Research on Ship Assets Appraisal [D]. Dalian University of Technology, 2011
- [3] Liu Yang, Gao Jin. Constraints and Countermeasures for Mortgage Financing of Secondhand Ships [J]. Foreign Economic and Trade Journal, 2012,19 (01): 114-115
- [4] M Beenstock, A Vergottis. Econometric modeling of world shipping [M]. Chapman & Hall, 1st edition, 1993.
- [5] Cai Gejing, Fu Haibin, Jiang Renbin, et al. Fishery Economy Prediction and Optimization Based on ARIMA Model [J]. Computer and Modernization, 2019 (4): 5.
- [6] AJia Pengxiang. LightGBM-based price prediction of used cars [D]. Jinan: Shandong Normal University, 2021.
- [7] Jia Pengxiang. LightGBM-based price prediction of used cars [D]. Jinan: Shandong Normal University, 2021.
- [8] Lu W X, Wu H C, Wan L Y. Prediction of precipitation by PLS based on fused random forest algorithm [J]. Statistics and Decision Making, 2020, 36(18): 27-31.
- [9] Research and application of online machine learning algorithm under big data [J]. Science and Technology Innovation Herald, 2020, 17 (23): 3.
- [10] Tao Y, Du J L. Long- and short-term memory network temperature prediction based on random forest [J]. Computer Engineering and Design, 2019, 40(3): 737-743.
- [11] Chen Zhenyu. Research and Discussion on Machine Learning Algorithms Based on Big data Technology [J]. Archives and Periodicals, 2020, 000 (021): 201.
- [12] Tao Yun Cao. A study of variable importance based on random forest [J]. Statistics and Decision Making, 2022, 38(4): 60-63.