ISSN 2616-7433 Vol. 5, Issue 7: 102-108, DOI: 10.25236/FSST.2023.050718

Data analysis based on the second-hand sailing market

Xinyao Wu[1], Qihao Li[2], Zhenglong Liu[1,*]

[1]Department of Information Management and Information Systems, North Sichuan Medical College, Nanchong, China
[2]School of Artificial Intelligence, China University of Petroleum-Beijing, Beijing, China
[*]Corresponding author: nsmclzl@163.com

Abstract: *As the second-hand sailboat market continues to develop, the value of sailboats changes with their aging and changes in market conditions. This article mainly analyzed and predicted the second-hand sailing market. To predict and evaluate the changes in the second-hand sailboat market, we exclude human factors such as policies, technology, and environment. To build a model for the second-hand sailboat market, we first clean the data by removing duplicates and irrelevant information. We then extract feature variables and use Pearson correlation analysis to filter out irrelevant or low-correlation feature variables. Next, we construct an XGBoost+SHAP model to predict the impact of regions on the second-hand sailboat market, as well as the impact of regions on single and double sailboats, and perform comparative analysis. To highlight the effectiveness of the training and validate the model, we input the data into various models such as Random Forest, Decision Tree, and Logistic Regression. We find that the XGBoost+SHAP algorithm model has the highest accuracy. Finally, we evaluate the contribution of each feature variable to the predictive power by measuring the positive and negative correlations between individual sailboat variables and regions, revealing the impact of feature variables on second-hand sailboat prices under different values.*

Keywords: *XGBoost, SHAP, region, length*

1. Introduction

The international shipping market is a free and competitive market, and no country's regulations or measures can control the competitive situation of the international shipping market. Since 1994, an overcapacity situation has begun to emerge in the international shipping market, which has further intensified competition and reached a white-hot stage. The form of the international shipping market is not very clear, and the prospects are not optimistic. In order to get rid of the predicament, international shipping companies have been actively exploring ways to solve problems, such as making timely decisions on buying and selling ships, shipbuilding, and ship leasing. However, currently, for sailboats in particular, many ships have a relatively long service life, and some have already exceeded their economic life, resulting in severe corrosion of the hull, outdated equipment, aging electrical systems, and low technical conditions, which affect the overall economic value of the ship. At the same time, the prices vary in different marketing areas and for different uses.

2. Data collection and cleaning

Data are mainly collected from the Sailboat listings.com, SAILING THE WEB, and Hong Kong Sailing Federation. The data sources are summarized in Table 1. Based on the original data" 2023_MCM_Problem_Y_Boats", we searched for additional features of this data on the website and extracted a more complete and feature-rich dataset of sailboats[1].

Table 1: Data source

Database	Websites
function	https://www.sailboatlistings.com/sailboats_for_sale/
function	https://www.sailingtheweb.com/en/shipyard/fountaine+pajot
economic	https://sailing.org.hk/zh-hant/

ISSN 2616-7433 Vol. 5, Issue 7: 102-108, DOI: 10.25236/FSST.2023.050718

3. Sailboat price prediction model

The sailboat price prediction model is designed for the second-hand sailboat sales issue. It utilizes the methods of XGBoost and SHAP to score and select different features, aiming to establish a good relationship between features and the target variable (price) to complete the regression task. Combining the main factors affecting sailboat valuation in different regions, the model is applied to compare and analyze single and double sailboats, as shown in Figure 1.

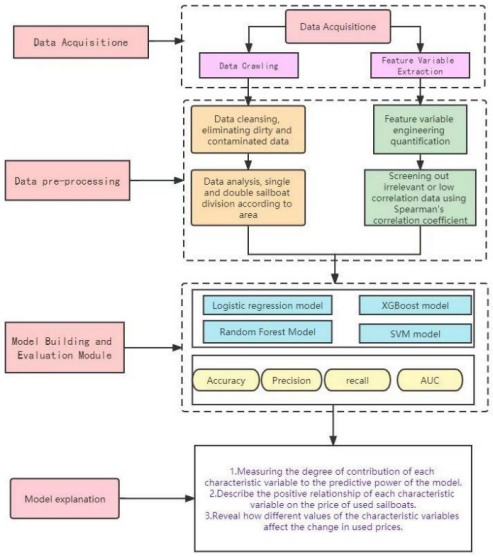

Figure 1: Model process diagram

3.1 Data presentation or Data visualization

Based on the data information provided in the problem, we can attempt to visualize the data and extract some relevant information from the dataset , as shown in Figure 2 and Figure 3.

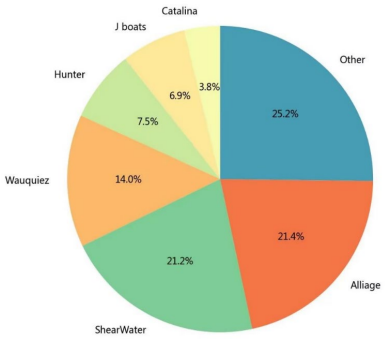

Figure 2: Pie chart showing the proportion of each manufacturer

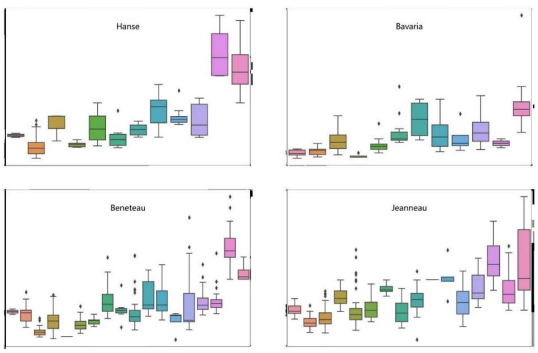

Figure 3: Box plot showing the product prices of each manufacturer

ISSN 2616-7433 Vol. 5, Issue 7: 102-108, DOI: 10.25236/FSST.2023.050718

3.2 Sailboat price model based on XGBoost and SHAP

3.2.1 XGBoost algorithm

The XGBoost algorithm efficiently implements the gradient boosting decision tree algorithm and has made many improvements, achieving good results in multiple fields. The XGBoost algorithm is essentially an iterative method of Boosting, involving additive models and forward distribution algorithms.

$$Obj = \sum_{i=1}^{n}(g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)) + \Omega(f_t) \tag{1}$$

Here, gi and hi are the first-order and second-order derivatives of the loss function with respect to the predicted boat market price $\hat{y}_i^{t-1}$ obtained from the first t-1 iterations.

The XGBoost algorithm introduces leaf score ω_m to represent the prediction value on each leaf node, and $q(x)$ represents the specific leaf node that the sample falls into, which is $\omega_q(x)$. The model complexity of the tree is defined as $\Omega(f_t)$ with the leaf node set I_j, and its expression is:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{2}$$

$$I_j = \{i|q(x_i) = j\} \tag{3}$$

By rewriting the sample set as a set of leaf nodes and letting $G_i = \sum_{i\in I_j} g_t$ and $H_i = \sum_{i\in I_j} h_i$, a new objective function is obtained.

Based on the XGBoost algorithm, we analyzed and constructed a sailboat price prediction model. The dataset for sailboats, denoted as X, includes features such as boat width, draft, displacement, rigging, sail area, hull material, engine hours, sleeping capacity, clearance, etc., while Y represents the sailboat price. The training dataset consists of n samples and m features, denoted as

$$D = \{(x_1, y_1), \cdots, (x_n, y_n)\} \tag{4}$$

Where $x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(m)})$. The training dataset is fed into the XGBoost algorithm for model training.

3.2.2 Interpretability with SHAP fusion

SHAP is mainly used to calculate the Shapley values for each feature, in order to reflect the contribution of each feature to the predictive ability of the entire model. The SHAP model's predicted value is interpreted as the sum of the attribution values for each input feature, as shown by the formula.

$$\hat{y} = f_0 + \sum_{i=1}^{M} f_i \tag{5}$$

Traditional feature importance can intuitively reflect the importance of features, but cannot reveal the relationship between features and the final prediction. SHAP calculates feature attribution values using feature attribution methods, which can clearly reflect the influence of each feature value on the final prediction, and can also reflect the positive and negative effects of the influence, thus increasing the interpretability of the model.

We take the listed price of sailboats as the independent variable and brand, variant, length, country/region/continent as the dependent variables. We use the positive/negative signs of the regression coefficients to reflect the correlation between the independent variable and the listed price of sailboats. Then, we use SHAP to calculate the feature attribution values. From this, we can know that the independent variables of brand, variant, length, and country/region/continent are positively correlated with the price of used sailboats[2].

We imported the data into Python and conducted descriptive statistics on the quantitative variables,

ISSN 2616-7433 Vol. 5, Issue 7: 102-108, DOI: 10.25236/FSST.2023.050718

obtaining 8 indicators including brand level, length, list price, year, used price, variant level, geographic region, and country/region/continent.

Next, we used Stata to automatically remove variables with multicollinearity, obtained standardized regression coefficients, and performed statistical analysis on them while also plotting the residuals against the fitted values. As the fitted values for the evaluation variable can be negative, we conducted heteroskedasticity BP and heteroskedasticity White tests, and then used OLS with robust standard errors, followed by stepwise regression.

3.2.3 Algorithm evaluation metrics and Comparison between algorithms

In this article, the determination coefficient (R^2), mean absolute error (MAE), and root mean square error (RMSE) were selected as evaluation metrics for measuring the effectiveness of the models. The formulas for calculating these metrics are shown in the following equations:

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})}{\sum_{i=1}^{n}(y_i - \hat{y})} \tag{6}$$

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}| \tag{7}$$

$$RESE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} \tag{8}$$

y_i represents the true value of the i-th sample, $\hat{y_i}$ represents the predicted value, and n is the number of samples.

The coefficient of determination (R^2) reflects the proportion of the total variability of the dependent variable that can be explained by the regression relationship with the independent variables. R^2 takes values between 0 and 1, and a higher value indicates a better model performance. MAE and RMSE measure the difference between the predicted values and the true values of the model, with both having the same unit of measure. However, RMSE calculates the square root of the sum of squared errors, magnifying the difference between larger errors. Compared to MAE, the values of RMSE are generally larger and provide a more intuitive measure of the actual errors.

After preprocessing and feature engineering on the raw data, the features and labels of the dataset are input into the XGBoost ensemble algorithm, which is then divided into training and test sets. The model is trained on the training set and used to predict the test set data. Based on the evaluation metrics of the algorithm, the model's performance is assessed and parameter tuning is performed.

The XGBoost algorithm has parameters such as n_estimators, Learn_rate, Max_depth, Objective, Subsample, Reg_alpha, and Reg_lambda. The optimal hyperparameters we set up are shown in the following table 2.

Table 2: Table displaying various hyperparameters of the XGBoost model

Parameters	value
n_estimators	500
Learn_rate	0.1
Max_depth	5
Objective	5
Subsample	1
Reg_alpha	0
Reg_lambda	1

After utilizing feature engineering algorithms and machine learning techniques, we were able to achieve more accurate comparison results among the six machine learning algorithms. Following the training process, we obtained the following results:

According to the table 3, XGBoost outperforms mainstream machine learning algorithms in all three evaluation metrics. Logistic regression and random forest also have good fitting performance. Decision tree, random forest, and XGBoost have the highest training set accuracy.

ISSN 2616-7433 Vol. 5, Issue 7: 102-108, DOI: 10.25236/FSST.2023.050718

Table 3: Table displaying the results of 6 machine learning methods

algorithms	Training accuracy	test set accuracy	Mean Absolute Error(MAE)
neural network	0.892	0.694	0.3055555
Logistic Regression	0.712	0.826	0.1944444
Decision tree.	1.000	0.583	0.4166666
Random Forest	1.000	0.722	0.2777777
LGBM	0.901	0.611	0.3888888
XGBoost	0.997	0.856	0.1678222

In summary, XGBoost outperformed other machine learning algorithms in all three evaluation metrics, achieving the best training set performance and the most accurate predictions, which can accurately reflect the relationship between sailboat performance and market prices. Based on feature selection and preprocessing, XGBoost uses second-order Taylor expansion to optimize the objective function, which preserves more information and adds regularization to control complexity, resulting in good prediction performance. The XGBoost model can better predict the price of sailboats.

3.3 Explaining the impact of regions on the listing price

Regarding the region, the distribution of Monohull and Catamaran numbers is shown in the following table 4.

Table 4: Table showing the distribution of Monohull and Catamaran numbers

	Europe	Caribbean	USA
Monohulls	735	302	108
Catamarans	1783	385	178

We can see from this that the United States has the largest number of Monohull boats listed, while the Caribbean has the largest number of Catamarans listed. We can also see that the distribution of monohull and catamaran listings is different across regions. For example, Europe has more monohull listings than catamaran listings, while the Caribbean has more catamaran listings than monohull listings.

The explanation in this paper is mainly based on the Spearman rank correlation coefficient.

Spearman rank correlation coefficient is a method used to measure the correlation between two waveforms. When two waveforms are exactly the same, the value of Spearman rank correlation coefficient is 1. When two waveforms are completely opposite, the value is -1. In other cases, the value varies between -1 and 1. The specific idea is to rank the values in the two waveforms, and then calculate the differences between the ranks to determine the correlation.

Firstly, the waveforms in the sequence $x = \{x_1, x_2, \cdots, x_n\}$,are arranged in ascending or descending order to obtain the arranged sequence $a = \{a_1, a_2, \cdots, a_n\}$, and the position of each element x_i in the sequence a is recorded as the rank r_i of element x_i. Thus, the rank of sequence x is obtained as r_0. The other waveform sequence $y = \{y_1, y_2, \cdots, y_n\}$ is arranged in the same way to obtain sequence $b = \{b_1, b_2, \cdots, b_n\}$, and the rank sequence s_0 of sequence y is obtained accordingly. We subtract each element in sequence r from its corresponding element in sequence s to obtain the rank difference sequence $d = \{d_1, d_2, \cdots, d_n\}$, and then substitute it into the Spearman rank correlation coefficient formula:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{9}$$

In the formula, n represents the number of data points in the sequence, corresponding to the number of sampling points in a window; ρ represents the Spearman rank correlation coefficient.

When the system is normal, the two waveforms on both sides are completely negatively correlated, and theoretically, $= -1$. We can obtain the regularity of their rank order in the completely negatively correlated state:

1) The sum of the rank of the two corresponding points at the same moment is equal to n+1;

2) The rank difference sequence d is an arithmetic sequence with a common difference of -2 for (n-1)~-(n-1);

We apply this ranking method to two perfectly negatively correlated sine waves, and their standard rank sequence is shown in the figure 4 and Figure 5 below:

ISSN 2616-7433 Vol. 5, Issue 7: 102-108, DOI: 10.25236/FSST.2023.050718

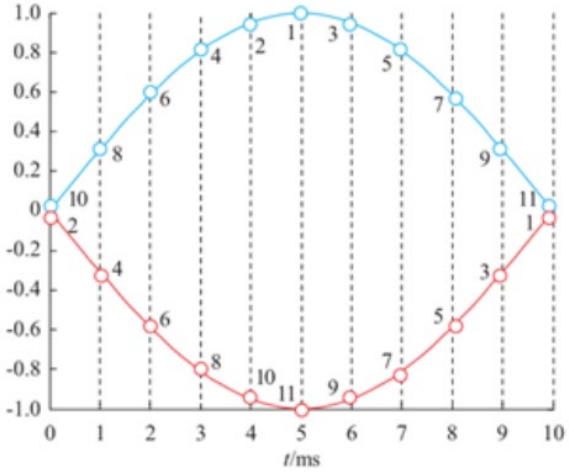

Figure 4: Ranking diagram

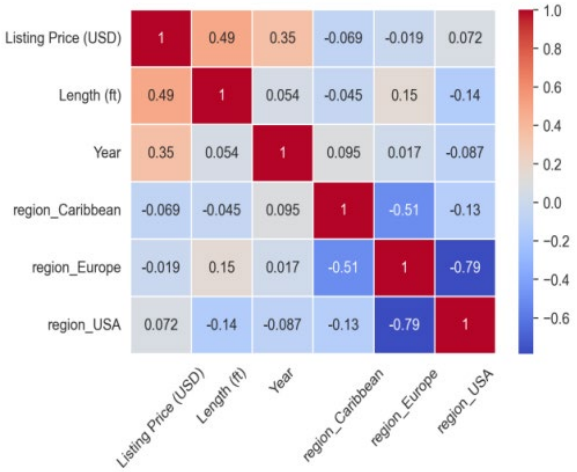

Figure 5: Spearman heatmap

In the end, we can obtain the result as follows:

From the above graph, it is evident that the geographic location has a significant impact on the listing prices, and the listing prices for the same sailboat can vary across different regions[3].

In the above heat map, the lighter the color, the higher the correlation between the two features, while the darker the color, the lower the correlation between the two features. The correlation heatmap indicates that the features region_Europe andregion_USA have a strong correlation (negative correlation).

3.4 Model transfer to Hong Kong

In order to simulate the impact of Hong Kong on the prices of some sailboats, based on the models, we preprocessed the collected data of sailboats sold in Hong Kong using Python, removing corresponding dirty and polluted data. Then, models were constructed to explain the effect of the Hong Kong market on the price of each sailboat , as shown in Table 5, Table 6 and Table 7.

Firstly, the sailboat dataset for sales in Hong Kong was obtained through web scraping. The following are the relevant information of the sailboat dataset in Hong Kong:

Table 5: Distribution table of monohull and catamaran numbers in Hong Kong

	Europe	Caribbean	USA
Monohulls	1277	295	103
Catamarans	351	161	60

With the use of Xgboost model, we processed the dataset of Hong Kong sailboat sales and trained the model, resulting in the following outcomes.

ISSN 2616-7433 Vol. 5, Issue 7: 102-108, DOI: 10.25236/FSST.2023.050718

Table 6: Monohull regression results

Index	Coef	Std_Error	t-Statistic
Const	-1.61e+07	1.96e+06	-8.214
Year	7841.1812	976.583	8.029
Region_Caribbean	-7.154e+04	1.12e+04	-6.367
Region_Europe	-4.915e+04	6400.938	-7.678
Length	1.307e+04	509.503	25.643
Region_HongKong	-2.324e+04	896	4.3567

Table 7: Catamaran Regression results

Index	Coef	Std_Error	t-Statistic
Const	-2.368e+07	1.65e+06	-14.344
Length	2.605e+04	774.214	33.651
Year	1.533e+04	1093.836	14.011
Region_Caribbean	-7.925e+06	5.51e+05	-14.389
Region_Europe	-7.891e+06	5.51e+05	-14.320
Region_USA	-7.86e+06	5.49e+05	-14.320
Region_HongKong	1.86e+06	5.89e+05	-6.382

4. Evaluation of the models

4.1 Strengths

In data preprocessing, we use methods such as filling missing values, data type conversion, and feature engineering to clone and process data, further improving the rationality of the data and the feasibility of subsequent analysis.

When establishing a second-hand ship prediction model, multiple models were compared. XGBoost+SHAPSVM prediction algorithm and exponential simulation algorithm were trained, with RMSE and mean absolute error (MAE) as the determining factors. This article verifies the accuracy and rationality of the model.

GBDT is based on the traditional CART as the base classifier, while xgBoosting supports linear classifiers, which is equivalent to introducing L1 and L2 regularization terms of logistic regression (for classification problems) and linear regression (for regression problems).

4.2 Weaknesses

While solving the regional effect model based on the XGBoost algorithm, although our model incorporated the optimal machine learning algorithm, we can further explore its performance by comparing it with linear regression in subsequent research, in order to identify the most suitable model.

For important parameters such as the ship's place of production, sensitivity analysis can be considered in the future to further observe the relative importance of ship performance and the sensitivity of other prices.

References

[1] Cruz M, Miranda B P, Carvalho M, et al. Visual Data Analysis for Hydrological Cycle Classification Based on Physico-Chemical Parameters[C]// Information Visualisation. IEEE, 2016.
[2] Zhang Renyi. Technical and economic demonstration method of ships. Shanghai: Shanghai Jiao Tong University Press, 1989.
[3] Yang Z, Daamen W, Vellinga T, et al. Impacts of wind and current on ship behavior in ports and waterways: A quantitative analysis based on AIS data[J]. Ocean Engineering, 2020, 213:107774.