Sailboat price prediction based on statistical and machine learning methods

Shengjie Xia*
School of Information Science & Engineering, Lanzhou University, Lanzhou, 730000, China
xiashj21@lzu.edu.cn

ABSTRACT

Pricing of sailboats is a topic that many extreme sports enthusiasts like to pay attention to. For the sailboat market, price prediction is crucial and is a key factor in making decisions such as marketing strategies, cost calculation, and profit calculation. By using regression models and machine learning, companies can better grasp market risks, improve product pricing accuracy, and increase their profitability. In addition, this method is very valuable in various industries such as e-commerce, retail, finance, and real estate. This paper aims to establish the relationship between sailboat prices and various factors using typical models based on statistical methods and machine learning, and explore the regional and sailboat type effects on prices, particularly the regional effects of specific areas such as Hong Kong (SAR). By comparing the accuracy of each model, we can obtain some insights into the above issues.

CCS CONCEPTS

• **Computing methodologies** → Modeling and simulation; Simulation evaluation.

KEYWORDS

Sailboat Price Prediction, Ridge Regression, Two-way ANOVA, Machine Learning, Paired T-test

ACM Reference Format:
Shengjie Xia*. 2024. Sailboat price prediction based on statistical and machine learning methods. In *2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence (DEAI 2024), January 19–21, 2024, Hongkong, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3675417.3675475

1 INTRODUCTION

In the current business and financial field, price prediction is a very important task. It can provide decision support and strategic planning for businesses and investors, enabling them to better adapt to market changes and improve competitiveness. Traditional price prediction methods usually rely on statistical models and economic indicators. In recent years, machine learning and deep learning methods have been widely applied in price prediction. These methods make full use of historical price data and other relevant data (such as supply and demand, market sentiment, macroeconomic indicators, etc.) for training, and can improve the accuracy and predictive capabilities of price prediction. This article constructs a comprehensive model for sailboat price prediction and regional effect analysis to explore the factors influencing sailboat prices and the specific content of regional effects.

2 MATERIALS & METHODS

2.1 Data

The current dataset contains 2346 entries for monohulled sailboats and 1145 entries for catamaran sailboats. Each entry includes the brand, model, length, geographic region, country/region/state, listing price, and year of the sailboat.

2.1.1 Collection. For the required data, additional data needs to be collected to supplement the existing data. The make, type, length, beam, draft, displacement, rigging, sail area, hull material, engine hours, berth capacity, headroom, electronics, etc. of the sailboat need to be known. As data on the price of sailing vessels in Hong Kong (SAR) is not provided in the original dataset and it is generally difficult to find online, a regional simulation model based on indicators related to the economy and freight levels needs to be constructed. To achieve this, corresponding port container throughput data needs to be collected. Through the above discussion, the following metrics were added to the dataset, GDP per capita (in constant local currency) for countries around the world from 1960 to 2021, the real GDP per capita of each US state, the parameters of the boat like make, type, length, beam, draft, displacement, rigging, sail area, hull material, engine hours, berth capacity, headroom, electronics and corresponding port container throughput data.

2.1.2 Preprocessing. For real-world datasets, possible missing data in data items is inevitable, and we found three items with missing data in the table for monohulled sailboats. To avoid affecting the accuracy of our model, we excluded these items. At the same time, we need to make the values of the listing price, geographical area and variants of the sailboat approximately satisfy a normal distribution, only then can we proceed to the next step of the analysis, so we need to eliminate the corresponding outliers (i.e. values that do not satisfy the 3σ principle). A normal distribution transformation where the original data is cleaned and transformed into a lognormal distribution. In this way, it still retains the relevant properties of the normal distribution.

The data classes with more categories are preprocessed, and only the categories with more samples are kept, treating the rest as ”others”. All brands with less than 60 sailboats under the brand are classified as other brands. Table 1 displays the distribution of brands for Monohulled Sailboats.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DEAI 2024, January 19–21, 2024, Hongkong, China

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1714-7/24/01
https://doi.org/10.1145/3675417.3675475

Table 1: Makes of Monohulled Sailboats.

Makes	Value
Jeanneau	21%
Hanse	8%
Dufor	7%
Beneteau	21%
Bavaria	14%
Others	29%

Due to the presence of many categorical variables in the dataset, such as brand, model, geographical region, etc., it is necessary to perform ordinal encoding on them.

2.2 Methodology

The factors that may influence the listing price, such as make, species, length, beam, draft, displacement, rigging, sail area, hull material, engine hours, berths, headroom, electronic equipment, geographical area, and economic situation, were first identified to explore the relationship between listing price and these influencing factors.

Ridge regression and other machine learning methods were chosen for modeling. The accuracy of the developed model was discussed using appropriate model evaluation metrics (e.g., Root Mean Square Error (RMSE), Goodness of Fit (R-square), etc.).

Then, a two-factor ANOVA with hypothesis testing was performed to determine the significant relationship between geographic area and sailboat type on listing price. This aimed to ascertain whether geographic area and sailboat type have a significant effect on listing price.

Furthermore, the random forest regression method was selected to adjust the listing price of sailboats in Hong Kong based on economic conditions and freight levels, in order to investigate the regional effect of Hong Kong (SAR) on the listing price of sailboats.

A t-test was utilized to analyze the regional effects in Hong Kong, aiming to minimize the impact of errors. Finally, some interesting features of the original dataset were revealed through the analysis.

3 RESULTS & DISCUSSION

3.1 Regression Model for the Listing Price of Sailing Boats

3.1.1 Ridge-based Regression Results. Ridge Regression is a widely used statistical method for analyzing data and predictive modelling that has many applications [1–4].

The data on n individuals, $i = 1, \ldots, n$, where each individual has p traits. Then, $\mathbf{X}$ is the $n \times p$ matrix of characteristics, standardized such that $\mathbf{X'X}$ is in correlation form. $\mathbf{Y} = (Y_i, \ldots, Y_n)$' is a n-dimensional vector of predictive variables (sailboat listing price). The linear regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ is commonly used to model the relationship between characteristics and sailboat listing price. $\beta = (\beta_1, \ldots, \beta_p)'$ is a column vector of p regression coefficients, one for each feature of yacht, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$' is a vector of independent and identically distributed normally distributed random errors, $(\in_i)$= 0 and $(\in_i^2) = \sigma^2$. Ridge estimates of the regression coefficients are given by

$$\hat{\beta}^{Ridge} = \left(\mathbf{X}^{'}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{'}\mathbf{Y} \tag{1}$$

λ is the ridge parameter, which controls the amount of shrinkage of the regression coefficients, and $\mathbf{I}$ is the identity matrix. The equation above can be converted into a general form:

$$\hat{\beta}^{Ridge} = arg\min_{\beta}\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right] \tag{2}$$

It can be seen from the equation above that ridge regression is a least squares regression with a two-parametric penalty.

Based on the previous discussion, ridge regression was used to identify the relationship between sailboat prices and various factors.

Figure 1 displays the ridge trace distribution plots for each variable obtained from ridge regression. K = 0.185 determined by variance expansion factor method.

Figure 2 shows the fitting results of ridge regression on the dataset.

The goodness of fit R^2 of the model was 0.708 and the model performs relatively well.

3.1.2 Machine learning-based Regression Results. The following are several machine learning algorithms used in this study.

Decision Tree is a classification algorithm based on a tree-like structure. It recursively partitions the dataset into different nodes, with each node splitting based on the threshold of feature values, forming a tree-shaped structure. Decision trees have good interpretability and visualization capabilities, and can handle non-linear relationships.

Random Forest is an ensemble learning model composed of multiple decision trees. Each tree is trained on a randomly selected subset of features and samples. The final prediction result is obtained by averaging or voting the predictions of all decision trees. Random Forest has high accuracy, generalization ability, and robustness in feature selection.

MLP is a neural network-based model commonly used for classification and regression problems. MLP consists of multiple dense layers stacked together, with each layer containing multiple nodes that are connected to all nodes in the previous layer. Backpropagation algorithm is used to adjust the weights to minimize the loss function and improve prediction accuracy. MLP has strong expressive power in solving complex non-linear problems.

LGBM is a learning model based on Gradient Boosting Decision Trees. LGBM adopts new techniques such as histogram-based splitting and leaf-wise tree growth acceleration based on leaf values

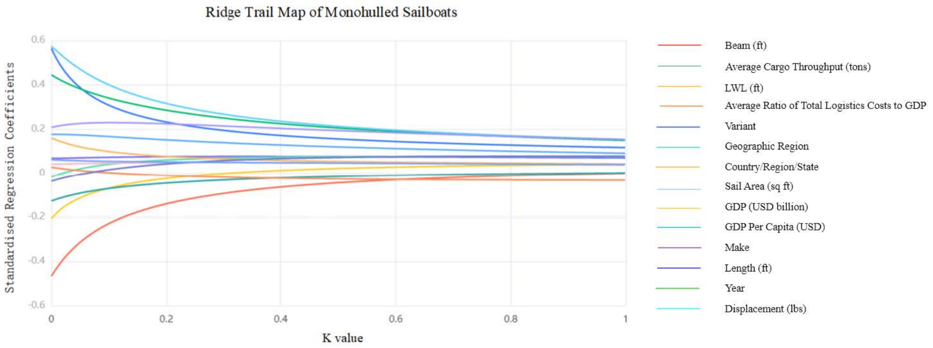

Figure 1: Ridge Trail Map of Monohulled Sailboats.

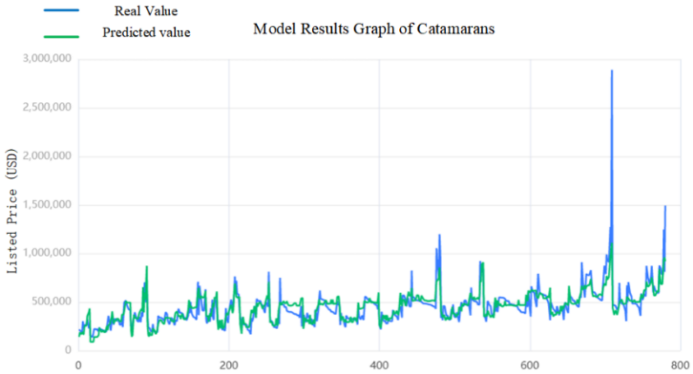

Figure 2: Model results graph.

Table 2: Prediction results for each model of Monohulled Sailboats.

Indicator	MLP	Linear Regression	Decision Tree	Random Forest	LGBM	XGBoost
Testing Set MAPE	0.370	0.351	0.201	0.153	0.170	0.167
Training set MAPE	0.350	0.438	0.032	0.082	0.129	0.077
R^2 Score	0.19	0.48	0.77	0.78	0.75	0.73

ordering to improve training speed and accuracy. This model is suitable for large-scale datasets and high-dimensional feature spaces, and has good generalization ability.

XGBoost is a machine learning model based on gradient boosting decision trees. It combines regularization techniques, shrinkage, and random subsampling strategies to build interpretable and accurate tree models. XGBoost has advantages in terms of training speed and accuracy, and is widely used in various machine learning tasks.

Those models were used for training, and the performance of each model in prediction was shown in Table 2. The ratio of training set to test set in this study is 8:2. The general form of MAPE is: $J = \frac{100\%}{n} \sum_{i=1}^{n} |\frac{\widehat{y_i} - y_i}{y_i}|$, where n is the number of samples in the training set, $\widehat{y_i}$ is predicted values and y_i is real value. MAPE of 0% indicates a perfect model, while MAPE greater than 100% indicates a poor model.

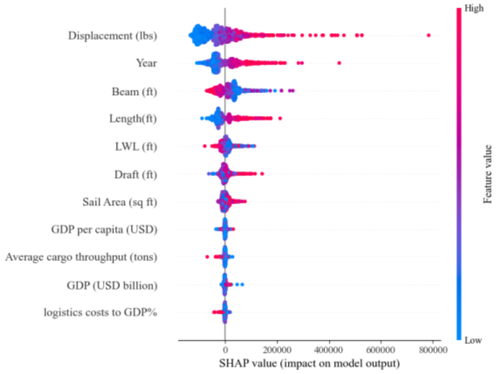

Figure 3: SHAP Value Interpretation.

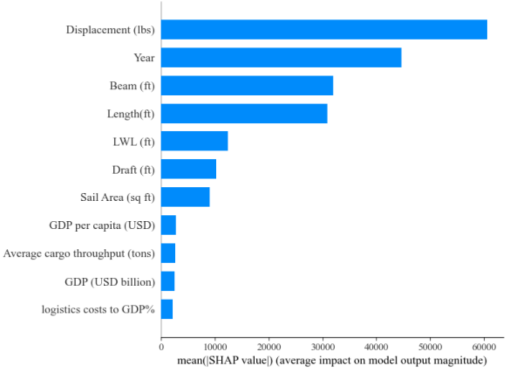

Figure 4: Mean SHAP Value Model Interpretation.

As is shown in Figure 3 and Figure 4, the results for the catamarans were similar to above, with a better random forest fitted.

Data-driven random forest modelling methods based on high-dimensional data adaptation and classification capabilities have been widely used in industry, medicine and other fields [5–7].

Therefore, this article will use the random forest method to predict the listing prices of sailboats in the Hong Kong (SAR) area.

3.2 Test Model for the Correlation between Price of Vessels and Region

Analysis of variance (ANOVA) is simply an example of the general linear model (GLM) that is commonly used for factorial designs. A factorial design is one in which the experimental conditions can be categorized according to one or more factors, each with two or more levels [8]. For example, an experiment might present two types of visual stimuli (e.g., faces and houses), each at three different levels of eccentricity. This would correspond to a 2 × 3 ANOVA, in which the six conditions correspond to unique combinations of each level of the 'stimulus-type' and 'eccentricity' factors [9].

Table 3 presents the results after conducting analysis of variance (ANOVA).

Figure 5 and Figure 6 display the distribution of price in specific geographic regions.

As can be seen from the chart above, prices vary across different countries/regions/states. Prices for the same model of yacht differ from area to area, for different models of yacht in the same area and for different models of sailboat in different areas.

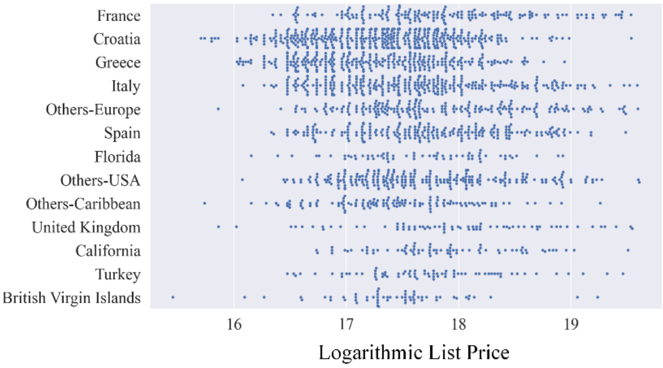

Figure 5: Price According to Specific Geographical Areas (Hexbin Plot).

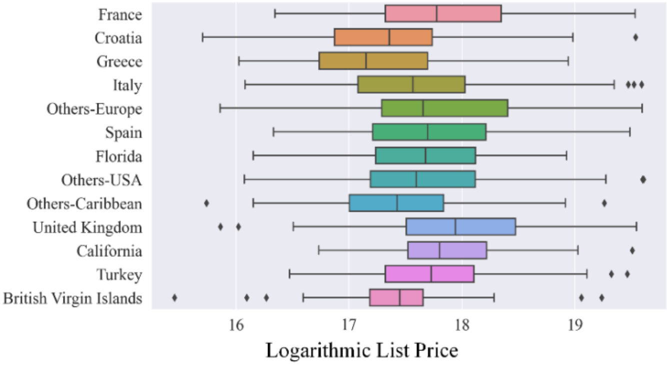

Figure 6: Price According to Specific Geographical Areas (Box Plot).

3.3 Hong Kong (SAR) Regional Effect Model

The modeling process of random forest is as follows:

Figure 7 shows the results of predicting the price of Hong Kong sailboats using the listed prices of sailboats in the dataset (True Price).

The paired t-test statistic can be given by the following formula: [10]

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{3}$$

Table 3: Prediction results for each model of Monohulled Sailboats.

Item	F	P	R^2	Adjustment of R^2
Intercept	10085.807	0.000***	0.823	0.772
Variant	17.289	0.000***		
GeographicRegion	43.439	0.000***		
Variant GeographicRegion	14.149	0.000***		

[a] ***, **, * represent 1%, 5%, 10% level of significance respectively.

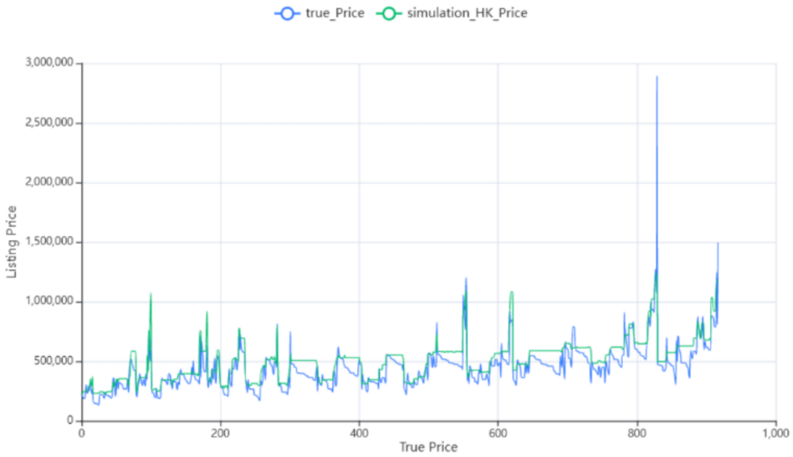

Figure 7: Simulation Hong Kong Price for Catamarans.

```
Algorithm 1 Random Forest Algorithm
Input: Training set D = {($x_1$, $y_1$), ($x_2$, $y_2$), ($x_3$, $y_3$), . . . , ($x_n$, $y_n$)};
Process: Regression Tree $f(x)$.
In the data space where the training set is located, a binomial decision tree is constructed by recursively dividing each region into two sub-regions and determining the output values on each sub-region.
(1) Choose the optimal tangent variable $j$ with tangent point $s$ and solve for
$\min_{j,s}[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2]$
Iterate over the variable $j$, scan the intersection point $\boldsymbol{s}$ for a fixed intersection variable, and select the pair ($j$,$s$) that minimizes the value of the equation above.
(2) Divide the region with the selected pair ($j$,$s$) and decide on the corresponding output value:
    $R_1(j,s) = \{x|x^{(j)} \leq s\}, \quad R_2(j,s) = \{x|x^{(j)} \rangle s\}$
    $\widehat{c}_m = \frac{\sum_{x_i \in R_m(j,s)} y_i}{N_m}, \quad x \in R_m, \quad m = 1, 2$
(3) Repeat steps (1) and (2) for both sub-areas until the stop condition is met.
(4) Divide the input space into $M$ regions $R_1$, $R_2$, ... . $R_n$, generating a decision tree:
    $f(x) = \sum_M \widehat{c}_m \mathbf{I}(x \in R_m)$
(5) For each new data point, let each decision tree predict the outcome.
(6) Aggregate the predictions from all decision trees to make the final prediction.
(7) Evaluate the model's performance using appropriate metrics and adjust hyperparameters if necessary.
```

where S_1^2, S_2^2 are sample variances, n_1, n_2 are sample size, $\overline{X_1}$, $\overline{X_2}$ are sample mean.

The results of the paired samples t-test showed that there is a significant difference between the word Price paired with Simulation HK Price. The magnitude of its difference Cohen's d value is: 0.675 and the magnitude of the difference is moderate. For monohulled sailboats, the results are somewhat different. There is a significant difference between the Price and Simulation HK Price pairs. The Cohen's d value is 0.061.

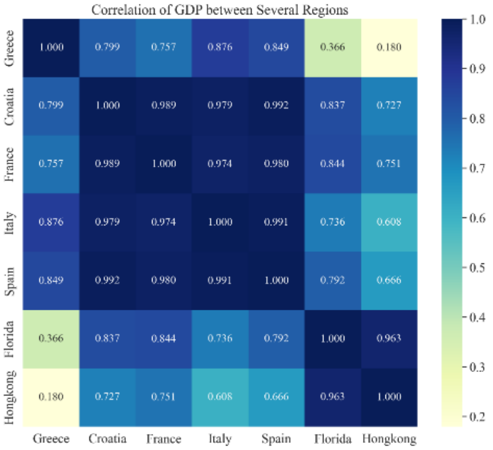

Figure 8: Correlation of GDP between Several Regions

3.4 Other features of the dataset

Figure 8 displays some additional features of this dataset. There is a strong correlation between GDP in some regions is found.

4 CONCLUSIONS

According to the model in this paper, the price of a sailboat is influenced by many factors such as brand, model, length, width, draught, displacement, sail area, LWL, GDP, year, and geographic region, among others. After trying different models, this paper has obtained a relatively accurate one. Based on this high-precision model, the paper predicts the listing price of sailboats in Hong Kong. By exploring the impact of geographical regions and sailing types on the listing price, the paper finds that both factors have a significant influence on the price. Furthermore, based on these findings, the paper introduces a correlation between the prices of sailboats in the Hong Kong market and those in other regions, and it is found that the extent of regional influence differs between monohulls and catamarans. The paper also reveals some interesting results, such as the fact that the price of sailboats tends to increase over time, while the economic situation in certain regions is also relevant in terms of indicators.

REFERENCES

[1] Hoerl, R.W. (2020) Ridge regression: a historical context. Technometrics, 62(4): 420-425.
[2] Pasha, G.R., Shah, M.A. (2004) Application of ridge regression to multicollinear data. Journal of research (Science), 15(1): 97-106.
[3] Arashi, M., Roozbeh, M., Hamzah, N.A., Gasparini, M. (2021) Ridge regression and its applications in genetic studies. Plos one, 16(4): e0245376.
[4] Chen, Z., Hu, J., Qiu, X., Jiang, W. (2022) Kernel ridge regression-based TV regularization for motion correction of dynamic MRI. Signal Processing, 197: 108559.
[5] Liaw, A., Wiener, M. (2015) randomForest: Breiman and Cutler's random forests for classification and regression. R package version, 4: 14.
[6] Guyon, I., Elisseeff, A. (2006) An introduction to feature extraction. In Feature extraction: foundations and applications. Springer Berlin Heidelberg, Berlin, Heidelberg.
[7] Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T. (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 8(1): 1-21.
[8] Winer, B.J., Brown, D.R., Michels, K.M. (1971) Statistical principles in experimental design. McGraw-Hill, New York.
[9] Henson, R.N. (2015) Analysis of variance (ANOVA). Brain Mapping: an encyclopedic reference. Elsevier: 477-481.
[10] Donna, L.M., William, J.W., Rudolf, J.F. (2022) Chapter 2 - Probability and Sampling Distributions. In: Donna, L.M., William, J.W., Rudolf, J.F. (Eds.), Statistical Methods (Fourth Edition). Academic Press, New York. pp. 65-122.