Prediction Model of Sailing Ship Price Based on Random Forest Regression

Zhaobo Lin [*,†], Xinyue Huang [†], Jiajun Chen [†]

Maynooth International Engineering College, Fuzhou University, Fujian, China, 350108

* Corresponding Author Email: 534903759@qq.com

† These authors contributed equally.

Abstract. The sailboat market is one of the emerging sunrise industries. Many people choose to buy a used sailboat rather than a brand new one. In this paper, various physical variables of sailboats are collected and Pearson correlation coefficient matrix and random forest model are used to filter the important variables and build a model to predict the price of sailboats based on random forest regression. A random forest regression model is used to build a model about the relationship between the independent and dependent variables to predict the prices of sailboats in the market. After training the model, it is used to make predictions on new data. The test set was tested and the results showed that the model was 80% accurate with absolute squared error and mean squared error values of only 0.05 and 0.01. It is concluded that the main factors affecting the price of used sailboats were country/region/state, draft (maximum) (meters), breed, width (meters), length (feet), year, and brand (increasing importance). Our model was assessed as a good predictor of sailboat prices with a small margin of error.

Keywords: Random forest, sailboat, markup, prediction.

1. Introduction

In modern society, luxury goods are everywhere: clothing, accessories, jewelry, watches, and even cars and sailboats. These items cover the characteristics of high-end and designer products, mainly in terms of quality, rarity, high price and non-essential. The sailing market is one of the emerging sunrise industries [1-2], it can promote the development of the marine economy, tourism and leisure industry and foster new industries promoting the development of the sailing market [3]. Since sailboats are an expensive luxury, many people consider buying used sailboats before buying a new one. A wide variety of sailboats are available in the second-hand sailboat market, ranging from small sailboats to large luxury sailboats, with prices ranging from thousands to millions of dollars.

More reasonable prices for used sailboats are a benefit to sailboat consumers. Since new sailboats are expensive, many owners sell their sailboats after only a short time, leaving the market with a large selection of used sailboats to choose from. Buyers can find the best sailboat for them by carefully researching the stock prices of the sailboats they want to buy on the market and asking experienced sailboat brokers.However, selling a used sailboat is not so easy for the seller. Like many luxury items, the value of a sailboat can change with age and market conditions. Sellers also need to have their sailboat accurately valued and find the right buyer in the marketplace.

2. Random Forest Model

The random forest (RF) [4] proposed by Breiman is an integrated learning algorithm combining decision tree and feature selection, which solves the problem of traditional decision trees with complex classification rules that easily lead to local optimal solutions, and is commonly used for feature variable selection, classification and outlier detection [5-6]. It has the advantages of high accuracy, good interpretability, low risk of overfitting and good noise tolerance [7-8], and has obtained many good results in the field of computer vision [9-10] and data mining [11-13]. The classification and regression trees (CART) algorithm is used to train the binomial decision trees to construct weak classifiers, and the results of each classifier are output by the majority voting method.

The number of features is generally the square root of the total number of variables [14], and the upper limit of the number of decision trees is generally set to 1000 [15].

3. Experiments

3.1. Data Description

Physical characteristics of sailboats are crawled from the website, including information such as beam, draft, displacement, rigging, hull type and ballast, whether the sailboat was classified as a monohull or catamaran, the boat's manufacturer, length, geographic location, year, and price; and combined these data into a table. The text type data is quantified by means of label encoding, which facilitates the processing of the data by machine learning algorithms. Then, box plot checks are used to check for outliers and z-cores are used to detect outliers, which are finally removed from the table. Two highly correlated variables (displacement and ballast) were found and removed by calculating the Pearson correlation coefficient matrix [16], as shown in Figure 1.

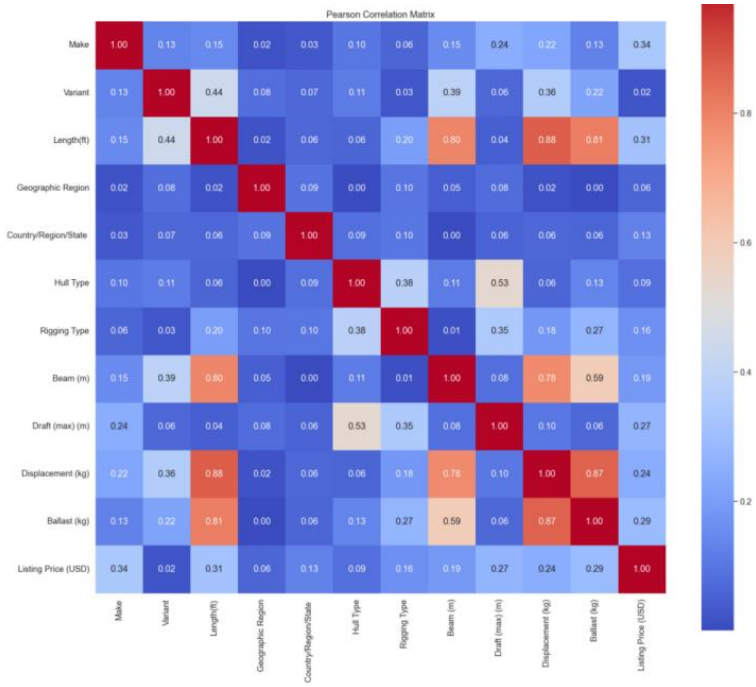

Figure 1. Results of importing the data into the Pearson correlation coefficient matrix

After that, a random forest model was used to assess the importance of the variables and rank the importance of each variable (as shown in Figure 2). The three variables with the lowest importance (rigging type, hull type, and geographic area) were manually removed.

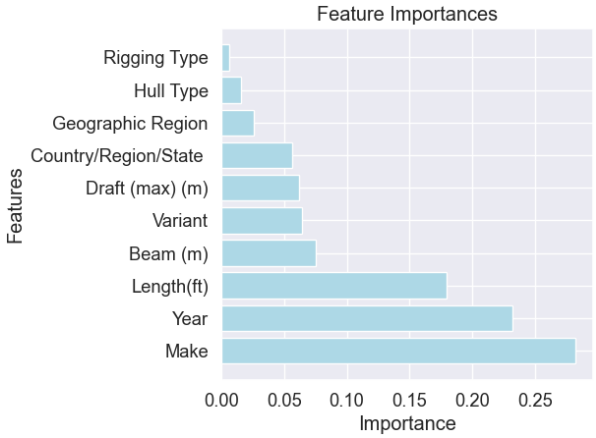

Figure 2. Variable Importance Ranking

The data required for the first modeling problem (including Country/Region/State, Draft(max)(m), Variant, Beam(m), Length(ft), Year, Make) is got by the above operation .

3.2. Prediction of sailing ship price using random forest regression models

The data is divided into dependent variable y (price) and independent variables x (including country/region/state, Draft(max)(m), Variant, Beam(m), Length(ft), Year, Make) and normalize the data. Normalization formula: normalizes the original data to between 0 and 1. After normalization, the data set is divided into a training set and a test set, where the training set accounts for 80% and the test set accounts for 20%.

After getting the training and test sets, a random forest regression model is built since the data type makes continuous data, and a random forest regression model containing 100 trees is built. Random forest is an integrated learning algorithm, the basic principle of which is to construct multiple decision trees by randomly selecting data and features, and then make predictions by integration.

The model is used to make predictions on the test set, after training on the training set. The mean_absolute_error, mean_squared_error, and r2_score functions are used to calculate the evaluation metrics of the model, namely mean absolute error (MAE), mean squared error (MSE), and R-squared value (R^2), respectively.

Finally, scatter plots are plotted with the true and predicted values of the test set as the x-axis and y-axis (as shown in Figure 3) and plotted diagonal lines where the true and predicted values are equal.

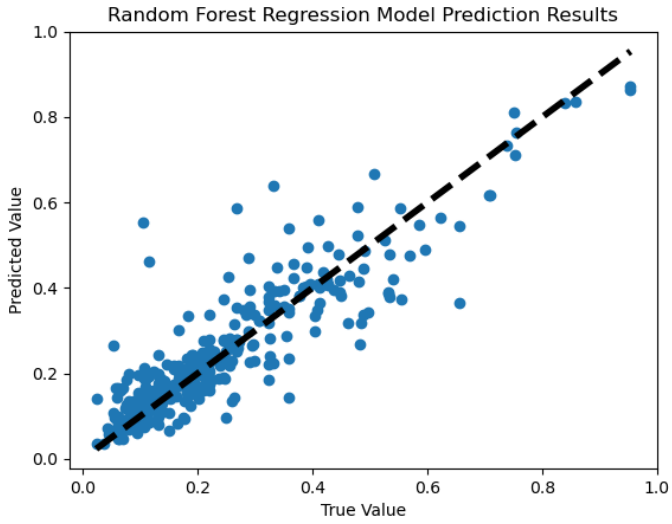

Figure 3. Scatterplot of prediction results

As shown in Figure 4, the accuracy of the model reached 80%, and the values of absolute squared error and mean squared error were only 0.05 and 0.01, which indicated that the model performed well.

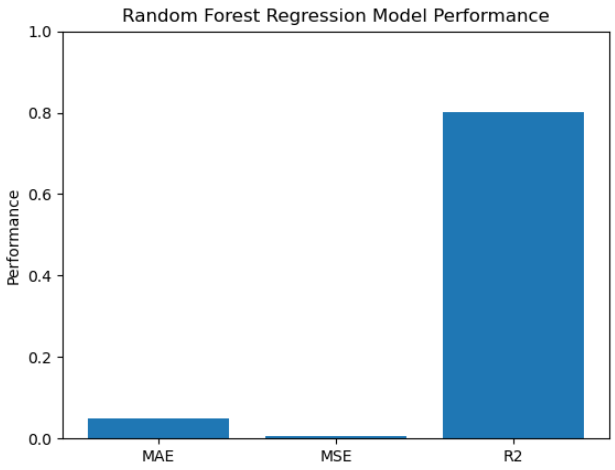

Figure 4. Model Performance

The mentioned data show that the main factors influencing the price of used sailboats are country/region/state, draft (max) (m), species, width (m), length (ft), year, brand (increasing importance) and our model is assessed to be a good predictor of sailboat prices with a small margin of error.

4. Conclusion

In this work, a dataset of sailboats sold in Europe, the Caribbean, and the United States was analyzed by feature selection and Machine Learning prediction methods, and then a sailboat price prediction model was developed and evaluated for its effectiveness. Pearson correlation coefficient analysis and the Random Forest algorithm were used to filter the factors related to sailboat prices, retaining the seven most influential factors out of the original 10. Random Forest regression model gave prediction results for the corresponding dataset after factor selection, and the mean absolute error (MAE), mean squared error (MSE), and R-squared value (R2) showed better prediction results for the Random Forest model, with smaller values for MAE and MSE and a value of 0.8 for R2. In the future, a plan is to obtain a larger dataset, possibly including different regions of GDP and climate conditions, and try more unused machine learning models to optimize the reliability of the predictions.

References

[1] Xu Weiyu. Research on government regulation of Wenzhou private yacht market [D]. College of Public Administration, East China University of Political Science and Law, 2016.

[2] Su Fang. Eternal Luxury, Immortal Fashion: Historical and Cultural Analysis of Luxury Consumption [D]. Wuhan University, 2013.

[3] Zhao Chang. Research on yacht marketing model [D]. Dalian Maritime University, 2013.

[4] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.

[5] Breiman L. Mach. Learn., 2001, 45(1): 5.

[6] Jiang Ping, LU Haoxiang, LIU Zhenbing. Identification of Random Forest combined with CatBoost by near infrared spectroscopy [J]. Spectroscopy and Spectral Analysis, 2022, Vol. 42 (7): 2148-2155.

[7] BIAU G, SCORNET E. A random forest guided tour [J]. Test, 2016, 25(2): 197-227.

[8] BOULESTEIX A L, JANITZA S, KRUPPA J, et al. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012, 2(6): 493-507.

[9] KONTSCHIEDER P, FITERAU M, CRIMINISI A, et al. Deep neural decision forests [C] Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Dec 13-16, 2015. Washington: IEEE Computer Society, 2015: 1467-1475.

[10] COOTES T F, IONITA M C, LINDNER C, et al. Robust and accurate shape model fitting using random forest regression voting [C] // LNCS 7578: Proceedings of the 12th European Conference on Computer Vision, Firenze, Oct 7-13, 2012. Berlin, Heidelberg: Springer, 2012: 278-291.

[11] BIFET A, HOLMES G, PFAHRINGER B, et al. New ensemble methods for evolving data streams [C] // Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, Jun 28-Jul 1, 2009. New York:ACM, 2009: 139-148.

[12] XIONG C M, JOHNSON D M, XU R, et al. Random forests for metric learning with implicit pair-wise position dependence [C] // Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, Aug 12-16, 2012. New York: ACM, 2012: 958-966.

[13] Xia Xiaoqiu, Chen Songcan. Improved Two-view Random Forest [J]. Computer Science and Exploration, 2022, Vol. 16 (1): 144-152.

[14] Zhang L, Liu Z, Ren T W, et al. Identification of seed maize fields with high spatial resolution and multiple spectral remote sensing using random forest classifier. Remote Sens, 2020, 12(3): 362.

[15] Tang Shulan, Meng Yong Wang Guoqiang, et al. Extraction of metamorphic minerals by multi-scale segmentation and random forest [J]. Chinese Journal of Engineering Science, 2022, Vol. 44 (2): 170-179.

[16] DeGhett VJ (2014) Effective use of Pearson's product-moment correlation coefficient: an additional point. Anim Behav 98: e1–e2. https://doi.org/10.1016/j.anbehav.2014.10.006.