Study of Sailboat Price Forecasting Based on Random Forest Regression

Yipeng Miao[1, *], Tingting Xia[2], Yutong Yang[3], Junhe Hou[4]

[1]School of Mathematics and Computer Science, Jilin Normal University, Siping, China, 136000

[2]School of Mathematical Sciences, Harbin Normal University, Harbin, China, 150500

[3]School of Electrical Engineering, Shandong University, Jinan, China, 250100

[4]High School Attached to Northeast Normal University, Changchun, China, 130021

*Corresponding author: phillipmiao120@gmail.com

Abstract. This paper aims to solve the problem of predicting the listing price of sailboats, using random forest, decision tree and integrated learning methods for analysis. First, data cleaning is performed to remove missing data. Then, the data are analyzed to find that the make, model, size, time and regional factors of sailboats may affect their listing prices. In terms of regional characteristics, annual income per capita was extracted as one of the influencing factors. Using all available characteristics, regression trees and random forest regression models were built and R^2 was used as an evaluation criterion for model prediction accuracy. Finally, the R^2 value of the monohull sailboat listing price prediction model was calculated to be 0.84, and the R^2value of the catamaran listing price prediction model was 0.99.In addition, this paper also involves the exploration of other features. Ultimately, this paper provides an estimate of the listed price per sailboat and discusses the issue of estimation accuracy.

Keywords: Regression Decision Tree, Random forest regression, Integration Learning.

1. Introduction

With the increasing popularity of recreational sailing, the demand for sailboats has also been increasing. An important factor that affects sailboat sales is the listing price. Therefore, predicting the listing price of sailboats has become a research topic of great interest. Prediction of sailboat listing prices is a challenging task due to the large number of factors that can potentially influence the price. Machine learning techniques have been increasingly employed in the field of price prediction in recent years. In this study, we use the regression decision tree, random forest regression, and integrated learning methods to predict the listing prices of sailboats. We also explore the effects of different features on the prediction accuracy.

In recent years, there has been growing interest in using machine learning techniques to predict the prices of sailboats. A number of studies have focused on identifying the key factors that influence sailboat prices to develop more accurate prediction models. For example, in the financial direction, there are studies on predicting stock prices by combining neural networks and integration for comparison [1]; in the manufacturing industry, there are predictions by combining Bayesian optimization and neural networks [2]; in life, there are also predictions of pig prices by combining neural network models [3]. This reflects to us that machine learning has become more and more popular and widely used in all aspects of work and life. On the contrary, the research of many scholars, ranging from the prediction of soybeans and photovoltaic cells to the statistical research of the whole financial industry [4-6], most of them still favor the use of neural networks in machine learning to make predictions and produce good prediction results. However, this does not mean that the machine learning approach stops there, and one study left behind the neural networks that most people are used to using and instead used the k-nearest neighbor (k-NN) regression method to predict sailboat prices based on vessel characteristics. The authors achieved similarly good accuracy with this model, with R-squared values ranging from 0.76 to 0.79 [7].

Overall, the reviewed literature suggests that machine learning techniques, such as decision trees, random forests, neural networks, and regression models, are effective in predicting sailboat prices. The performance of these models can be improved by feature selection, hybrid methods, or other techniques that refine the predictive features [8-11].

2. Data pre-processing

2.1. Quick View of Data Set and Data Cleaning

Before performing data cleaning, it is necessary to have some understanding of the data as a whole, so the team decided to use a computer to quickly browse the original data set. The following tables are available, i.e., TABLE 1 and TABLE 2. In addition, the data used in this article are from the 2023 Spring MCM and economic data published by the United Nations.

Table 1. Monohull Sailboat Data Quick Table

Variable	Non-Null Count	Dtype
Make	2346 non-null	object
Variant	2346 non-null	object
Length (ft)	2346 non-null	int64
Geographic Region	2346 non-null	object
Country/Region/State	2343 non-null	object
Listing Price (USD)	2346 non-null	int64
Year	2346 non-null	int64

Table 2. Catamaran Data Quick Facts Table

Variable	Non-Null Count	Dtype
Make	1145 non-null	object
Variant	1145 non-null	object
Length (ft)	1145 non-null	int64
Geographic Region	1145 non-null	object
Country/Region/State	1145 non-null	object
Listing Price (USD)	1145 non-null	int64
Year	1145 non-null	object

It is easy to see that for the single sailboat dataset, the variable Country/Region/State has missing values compared to the other variables, which means that special attention needs to be paid to the handling of this variable during the data cleaning phase. Since this variable is object type data this makes it difficult to achieve the filling of missing values without affecting the data analysis, and since the dataset is large enough and the percentage of missing values is Because the data set is large enough and the missing values are relatively small, the team decided to simply delete the rows where the missing values are located. For the double-hulled sailboat data, it is easy to see that there are no missing values, but the variable Year is an object type data, and since the time effect may be of concern in the subsequent analysis, the team processed it as an int64 type data in order to facilitate the subsequent analysis.

2.2. Descriptive Analysis and Correlation Analysis

In order to fully understand the data, the team decided to perform descriptive and correlation analyses on the numerical variables in the two cleaned datasets.

The results of the descriptive analysis for both datasets are given below, i.e. ,TABLE 3 and TABLE 4. Where 25%, 50% and 75% denote the lower quartile, median and upper quartile, respectively.

Table 3. Descriptive analysis of a single sailboat dataset

Variable	mean	std	min	25%	50%	75%	max
Length (ft)	4.761024	4.761024	36.0	40.5	45.0	49.0	56.0
Listing Price (USD)	229295.922322	147786.989337	45000.0	139671.0	193144.0	267244.0	1885229.0

Table 4. Descriptive analysis of the catamaran sailboat dataset

Variable	mean	std	min	25%	50%	75%	max
Length (ft)	43.749520	4.020278	37.5	39.5	43.6	45.8	56.0
Listing Price (USD)	455138.437555	202132.435722	95000.0	327712.0	431654.0	521397.0	2890000.0

The heat map of correlation coefficients for the two datasets is given below i.e. Figure 1.

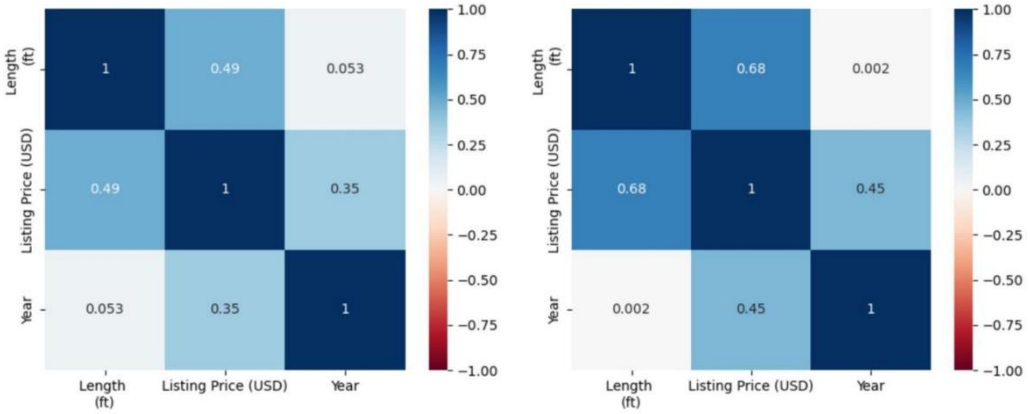

Figure 1. Heat map of correlation coefficients for each of the two data sets

From the above, it is easy to see that there is a strong correlation between the listing price of either sailboat and its size, which may play a good role in the subsequent analysis.

2.3. Decision Factor R2

R^2 is used to measure the degree of fit of the model to the data, with the values ranging from 0 to 1. The calculation formula of R^2 is:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{1}$$

where SS_{res} represents the residual sum of squares and SS_{tot} represents the total sum of squares. When R^2 is close to 1, it means that the model fits the data well, and when R^2 is close to 0, it means that the model fits the data poorly.

Considering the R^2 can effectively evaluate the fit of the regression model to achieve a better fit, and in this paper, it plays a great role as an important evaluation index.

2.4. Variance Inflation Factor

Variance Inflation Factor (VIF) is used to detect the collinearity between independent variables. For the ith independent variable, its VIF value is:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2}$$

where R_i^2 is the R^2 of the regression model built by taking the ith independent variable as the dependent variable, and the remaining independent variables as the independent variables. The larger the VIF value, the stronger the collinearity relationship between the ith independent variable and

the other independent variables. Generally, when the VIF value is greater than 5 or 10, the corresponding independent variable should be removed to improve the reliability of the model.

Considering VIF can effectively improve the fit and reliability of the regression model. In practical applications, it is necessary to conduct data analysis and processing according to specific situations to obtain more accurate and reliable results. In this article, this indicator plays a crucial role.

3. Feature Extraction

In order to quantify nominal variables so that they can be used in regression analysis, it is necessary to extract features from nominal variables.

For the variable Mark, a computer traversal of the data set showed that each element is basically a "word" or "word + separator + word" where the separator can be a space or a short horizontal line (-). The team disassembled each element, then counted the frequency of each letter and sorted them in alphabetical order, thus constructing a mapping relationship where each element has a unique ordered array to which it corresponds, an example of which is given below for the manufacturer Maine Cat.

$$\{\textbf{Maine Cat}\} \xrightarrow{Frequency\ Count} (2,0,1,0,1,0,0,0,1,0,0,0,1,1,0,0,0,0,0,1,0 \cdots,0,1)_{1\times 27} \quad (3)$$

Since the upper and lower case letters are somewhat reflective of the position of the letters and can better reflect the characteristics of the variables, the team took into account the difference between upper and lower case letters when extracting the features. Finally for Mark in general it was possible to find a unique sparse matrix to correspond to it. This completes the feature extraction of the variable Mark. The figure shows the sparse matrix solved for the Mark of a single sailboat as an example, i.e., Figure 2.

	A	B	C	D	E	F	G	H	I	J	...	r	s	t	u	v	w	x	y	z	Separate characters
0	1	0	0	0	0	0	0	0	0	0	...	0	0	1	1	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	...	0	0	1	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	...	0	0	1	1	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	...	0	0	1	1	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	...	1	0	0	2	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1771	0	0	0	0	0	0	0	0	0	1	...	0	0	0	1	0	0	0	0	0	0
1772	0	0	0	0	0	0	0	0	0	1	...	0	0	0	1	0	0	0	0	0	0
1773	0	0	0	0	0	0	0	0	0	1	...	0	0	0	1	0	0	0	0	0	0
1774	0	0	0	0	0	0	0	0	0	1	...	0	0	0	1	0	0	0	0	0	0
1775	0	0	0	0	0	0	0	0	0	0	...	0	1	1	0	0	0	0	0	0	1

Figure 2. Sparse matrices

For the variable Variant, the approach is more or less the same as above. The only difference is that instead of counting the numbers in the model number, it is considered to be a purely numerical value, which is not described here due to space constraints. This is still an ordered array for each model number, which completes the extraction of the variable Variant.

For the extraction of regional features, as the features of each region are too complicated, only the features relevant to the question should be of interest. This was done by extracting the regional characteristics from the region's annual income per capita (AIC).

The subsequent model fitting was based on the data sets of the two sailing boats for which the nominal extraction was done. In addition, the team has provided links to the files of the processed data in the Appendix.

Again, for the convenience of subsequent analysis, heat maps of the correlation coefficients for some of the processed data are presented below, i.e. Figure 3.

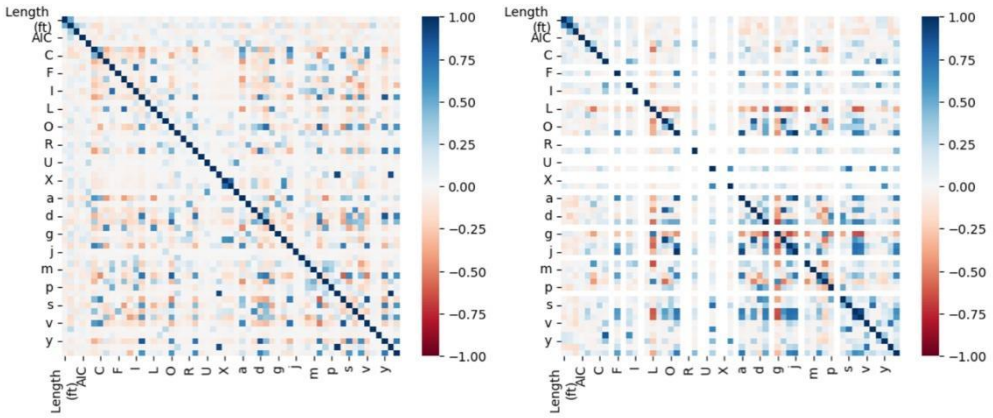

Figure 3. Heat map of correlation coefficients after feature extraction for monohull (left) and catamaran (right) sailboats

It is easy to see that none of the independent variables are overly strongly correlated with each other, so they can be used boldly in subsequent regression analyses. In addition, the variance inflation factor (VIF) between the variables was calculated by the team and is not repeated here due to space constraints, although it is worth noting that its value is sufficient to demonstrate that there is no serious multicollinearity between them. The subsequent discussion can now be pursued with confidence and boldness.

4. Model Overview

4.1. CART Regression tree

CART regression tree is a nonparametric regression method based on decision tree, which is mainly used to analyze the relationship between continuous dependent variables and multiple independent variables. The model of CART regression tree can be expressed as:

$$f(x) = \sum_{m=1}^{M} c_m \cdot I(x \in R_m) \tag{4}$$

where $f(x)$ represents the predicted value of the dependent variable given input x, M is the number of leaf nodes, R_m is the region of the mth leaf node, $I(x \in R_m)$ equals 1 when x falls within the region of the mth leaf node, and 0 otherwise. c_m represents the predicted value of the mth leaf node, which can be calculated by least squares or other optimization methods.

CART regression tree recursively divides the samples into subsets and fits a constant value in each subset. In order to improve the generalization ability of the fitted model, it is necessary to prune the tree by adjusting the number of leaf nodes or reducing the number of splits, in order to avoid overfitting and improve the model's generalization ability.

In practical applications, CART regression tree can be used to model the relationship between independent variables by splitting the dataset, and can handle issues such as missing and outlier values. Moreover, CART regression tree can provide useful insights into the relationship between independent and dependent variables, making it widely employed in data exploration and mining.

4.2. Bagging Regression

Bagging regression is a machine learning method based on ensemble learning, which can improve overall prediction performance by combining multiple basic regression models. The core idea is to build multiple training sets through bootstrap sampling (with replacement) and train a basic regression model on each of them.

In Bagging regression, the predicted result of the ensemble model can be obtained by averaging the predicted results of all basic models:

$$y_{ensemble} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{5}$$

where $y_{ensemble}$ represents the predicted result of the ensemble model, N is the number of basic models, and y_i represents the predicted result of the ith basic model.

Researchers can use different algorithms for each basic regression model, such as linear regression, decision tree regression, neural network regression, and so on. It is worth noting that when selecting basic regression models, they should meet two conditions. First, they should capture the main features of the data, in other words, they should have good prediction performance. Second, the prediction errors of the basic regression models should be random, because if all the basic regression models have the same or similar prediction errors, the prediction result of the ensemble model will be invalid.

In Bagging regression, each basic regression model has equal weight, so they all contribute equally to the predicted result of the ensemble model. The final predicted result is the average of the predicted results of multiple basic regression models. Therefore, Bagging regression usually has good prediction performance and low variance.

4.3. Random Forest

The random forest regression model adopts the idea of Bagging ensemble learning, which combines multiple CART regression tree models for prediction, in order to improve prediction performance. Specifically, the random forest regression model can be represented as the average of all basic regression model predictions:

$$f(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \tag{6}$$

where $f_t(x)$ is the prediction value of the tth tree, and T is the number of trees.

The construction process of the random forest regression model mainly includes the following steps. Firstly, multiple different bootstrap samples are constructed using the bootstrapping method. Then, a random forest is constructed for each bootstrap sample, with the feature selection randomized. Next, predictions are made for each random forest, resulting in multiple sets of basic regression model predictions:

$$f_t(x) = h_t(x, \theta_t) \tag{7}$$

where h_t is the basic regression model of the tth tree, and θ_t is the set of hyperparameters. Finally, the predicted results of all basic regression models are averaged to obtain the final predicted result.

The advantages of the random forest regression model are that it can effectively avoid overfitting problems, and has good generalization and prediction performance. The model introduces randomness, and by adjusting the hyperparameters of the regression trees, the prediction performance and stability of the model can be further improved. In summary, random forest regression is an efficient, robust, and easy-to-understand nonparametric regression method. This research has applied it to sailboat price prediction, achieving good results and obtaining credible effects.

5. Analysis of Results

The study analyzed two different kinds of sailboat datasets separately based on the algorithm described above.

First, before formally solving, the study agreed that 75% of the original data would be used as the training set and the other 25% as the test set in the subsequent discussion, i.e., before building the model, the team relied on the Sklearn module in Python to randomly split the original data into a training set and a test set in the proportions described above.

Secondly, the study fitted two weaker regressors based on decision tree regression to the training set of both sailboat data. After training the models, the R2 value for the monohull sailboat listing price prediction model was calculated to be 0.40243 and the R2 value for the catamaran sailboat listing price prediction model was 0.77755 using the data from the test set. it is easy to see that this is a very poor result. Therefore, in order to get better results, our team introduced the random forest algorithm to fit the training set again.

Finally, two models were trained separately based on the random forest algorithm for both training sets, and then the models were scored using the test set. Finally, the R2 value of the monohull sailboat listing price prediction model was calculated to be 0.83621, and the R2 value of the catamaran listing price prediction model was 0.99072. Obviously, fitting the model based on random forest regression is a good choice.

In this paper, the researcher firstly performs feature extraction for nominal type variables, in which the method borrows the idea of bag-of-words model but differs from the bag-of-words model to complete, or extracts the representative features of variables directly for the problem demand. The completed data are then fitted with a random forest model, and finally the model is scored with appropriate metrics to assess its merit. In addition, for researcher's idea, the two listed price prediction models will be of greater use in the subsequent analysis.

6. Conclusion

In summary, this study successfully developed an effective sailboat listing price prediction model using the random forest algorithm and demonstrated the feasibility of using a nominal variable feature extraction approach. This study provides a meaningful reference for the field of sailboat sales and marketing, which can be used to help decision makers make more accurate price predictions and develop more effective sales strategies.

However, there are some shortcomings in this study. First, the relatively small sample size of the data set may affect the accuracy of the model. Second, this study only considered a small number of characteristics and failed to consider more factors that may affect sailboat prices, which may also have some impact on the predictive power of the model.

Future studies can further explore more factors that may affect sailboat prices and incorporate these factors into the model to improve the accuracy of the model. In addition, a larger data set could be used to validate the reliability of the model and apply the model to other types of boat listing price predictions.

References

[1] Huang, Yutang. (2022). A comparative study on stock forecasting based on neural networks and their integrated machine learning [D]. Central University for Nationalities, 2022.000272.

[2] Xu Wangli, Shi Tingchun, Chen Hongyu, Yue Xiuyan. (2023) A quality prediction method for additive manufacturing molded parts based on the combination of Bayesian superchannel optimization algorithm and BP neural network [J/OL]. Computer Integrated Manufacturing Systems:1-16 [2023-05-15].

[3] Yipeng Lu. (2020). Hog price prediction based on neural network combinatorial model [D]. South China Agricultural University,2020.000334.

[4] Wang Chengyu. (2021). Soybean futures price forecasting based on LSTM neural network [D]. Chongqing University, 2021.003613.

[5] Xingcai. (2023). Research on PV module price prediction based on BP neural network[J]. Shanghai Energy Conservation,2023(03):355-361.

[6] Wang, Bin. (2020). Stochastic neural network financial price forecasting model and statistical analysis [D]. Beijing Jiaotong University,2020.003181.

[7] Park, H., Jung, N., Kim, S., & Han, S. (2020). Hybrid model for prediction of used sailboat prices [J]. Journal of Coastal Research, 99(sp1), 189-193.

[8] Zhang, J., Qi, C., Wu, T., & Chen, J. (2020). Prediction of sailboat prices via feature engineering and machine learning algorithms [J]. Sustainability, 12(17), 7037.

[9] Zhao, H., Wang, W., & Huang, H. (2019). A deep learning method for predicting sailboat prices [J]. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 165-174).

[10] Zhou, C., & Liu, C. (2020). A hybrid model for the prediction of sailboat prices based on a modified particle swarm optimization algorithm [J]. Applied Sciences, 10(15), 5053.

[11] Zhu, Y., Luo, Y., & Li, Y. (2021). Prediction of sailboat prices based on Bayesian linear regression [J]. Ocean Engineering, 222, 108499.