Exploring the Market: Used Sailboat Price Estimates Based on Artificial Bee Colony-BP Neural Network

Mohan Wang [1,†], Bo Lu [2,†], Hao Wang [3,*,†]

[1] Paris Curie Engineer School, Beijing University of Chemical Technology, Beijing, China, 100029

[2] College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China, 100029

[3] College of Mathematics and Physics, Beijing University of Chemical Technology, Beijing, China, 100029

* Corresponding Author Email: wanghao040903@163.com

† These authors contributed equally.

Abstract. In recent years, the shipping industry's share of world trade has been increasing year by year. As an important part of the shipping market, the accurate price prediction of second-hand sailboats is of great significance to grasping the price factors and improving the social and economic benefits. To accurately predict the price of second-hand sailboats, the artificial bee colony algorithm (ABC) is used to improve the BP neural network model, to solve the problem of overfitting of BP neural network. At the same time, compared with the prediction using Hyperopt improved XGBoost algorithm, the prediction effect of ABC-BP is better, and the fitting coefficient of the prediction results can reach 0.92.

Keywords: Used Sailboats, ABC-BP Neural Network, XGBoost Algorithm.

1. Introduction

With the development of the market economy, the global shipping industry gradually recovers after about 10 years of market downturn caused by excess shipping capacity and the sharp decline in trade volume [1]. Like other ships, many sailing companies have begun to increase their profits by building new boats or buying used ones. High acquisition costs, like luxury goods, discourage many people from buying sailboats, so second-hand sailboats are a popular choice. Due to the particularity of ship assets, the factors affecting the second-hand ship price are not only related to the ship's age, hull structure, and other ship factors but also to the shock of the sailing market will have a great change. However, different from bulk carriers, river ships, and further ships, the main attribute of sailing is a sport, not to transport goods as the main way of profit. Sailboat buying and selling are usually done by sailing brokers. In addition to providing the market information of the ship at any time, the manager can also conduct market research and make market analysis reports according to the specific requirements of the principal. Therefore, it is usually necessary to find the factors that affect the price of second-hand sailboats and establish a model that can accurately predict the price of second-hand sailboats.

In the past, researches on second-hand ships mainly focus on bulk carriers, which usually use the VAR model and BP neural network model [2] to predict the price of second-hand bulk carriers. Zhou et al. divided the factors affecting the price of ships into internal factors and external factors [3]. The internal factors were factors such as the age of the ship and the equipment configuration of the ship, while the external factors were factors such as the price of new shipbuilding. Du et al. introduced the market price evaluation methods applicable to different ship types, so the market replacement method should be used to evaluate the price of second-hand sailboats [1]. Huang et al. gave a dynamic equation to express the price of used ships and further explained the factors affecting the price of dry bulk used ships in detail [4]. Xiao et al. used BP neural network to evaluate the price of second-hand bulk carriers but did not make a specific classification of ship types, and this method produced an error of nearly 20% [2].

Since BP neural network is easy to fall into the problem of overfitting, this paper mainly uses ABC-BP neural network model to predict the price of second-hand sailing boats, and the fitting coefficient can reach 0.92.

2. The fundamentals of predictive models

2.1. The structure of the BP neural network

BP neural network is an error backpropagation algorithm first proposed by Rumelhart and McClelland in 1986. Compared with early neural networks, BP neural networks can continuously adjust the connection weights and convert the input information into the expected output information [5]. Mainly propagates layer by layer through the difference between input and output, continuously reducing errors to determine the connection weights between neurons in each layer.

The BP neural network includes an input layer, a hidden layer, and an output layer. The number of neurons in the input layer is determined by the input variable, the number of neurons in the hidden layer is determined by the accuracy of the model, and the expected number of results determines the number of neurons in the output layer [6]. Neurons in the same layer have no connection, while adjacent layers of neurons are connected by weights [7]. The activation function can map the input layer to the hidden layer nonlinearly, and then map the hidden layer to the output layer nonlinearly [8].

Let the number of neurons in the input layer, hidden layer, and output layer of the neural network be M, I, and N respectively, the input of each layer is x, the actual output is y, the expected output is d, the activation function from the input layer to the hidden layer is t, and the activation function from the hidden layer to the output layer is p, and the error in the learning process can be expressed as e. The topology structure of the BP neural network can be shown in Figure 1.

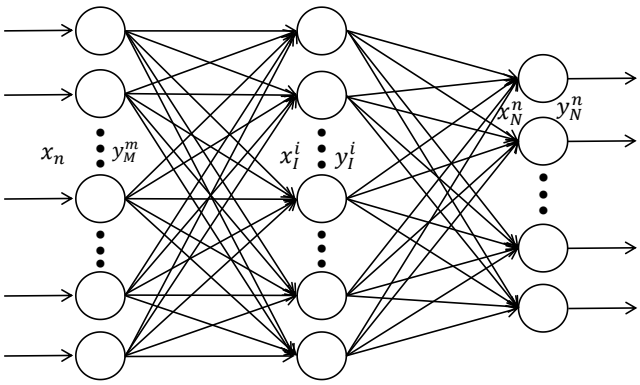

Figure 1. Neural network structure

The learning process of the BP neural network is:

Step *1: Positive Learning*

The signal propagates forward from the input layer to the output layer. If the input signal is x_n, the output of the neural network can be expressed as:

$$y_N^n(n) = p[\sum_{i=1}^{I} \omega_{in} \cdot t(\sum_{m=1}^{M} \omega_{mi} \cdot x_n)]. \quad (1)$$

The error of the nth node in the output layer can be expressed as:

$$e_n = d_n - y_n. \quad (2)$$

The total error of the forward learning process can be expressed as:

$$\mathrm{e} = \frac{1}{2}\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{e_n}^2 = \frac{1}{2}\sum_{\mathrm{n}=1}^{\mathrm{N}} (\mathrm{d_n} - \mathrm{y_n})^2, \quad (3)$$

Where $\mathrm{x_n}$ is the input signal ω_{mi} is the connection weight value from the mth neuron of the input layer to the ith neuron of the hidden layer, and t is the activation function from the input

layer to the hidden layer; ω_{in} is the weight of the ith neuron in the hidden layer to the nth neuron in the output layer, and p is the activation function from the hidden layer to the output layer.

Step 2: Back Propagation

Firstly, calculate the contribution rate of the weight value to the error, multiply it with the learning rate to obtain the correct amount of the weight value, and then further backpropagation:

$$\Delta\omega_{in}(n) = -\eta \frac{\partial e(n)}{\partial \omega_{in}(n)} = n\delta_n^i y_I^i(n), \tag{4}$$

$$\omega_{in}(n+1) = \Delta\omega_{in}(n) + \omega_{in}(n), \tag{5}$$

Where η is the learning rate, δ is the local gradient, $\Delta\omega_{in}$ is the weight correction amount.

Step 3: Adjust the Weight

$$\Delta\omega_{in}(n) = n\delta_n^i y_I^i(n), \tag{6}$$

$$\Delta\omega_i(n) = n\delta_n^i y_M^m(n). \tag{7}$$

When the error reaches the setting requirement, the training is over.

2.2. The structure of ABC-BP

Although BP neural networks have strong nonlinear mapping and fault tolerance capabilities, as neural networks, they are prone to falling into local optima and causing the model to lose its generalization ability. The artificial bee colony algorithm is a global intelligent optimization algorithm with the main purpose of solving optimization problems of multivariate functions [9]. This algorithm has few parameter settings, fast convergence speed, and can to some extent jump out of the local optimal. The combination of the BP neural network and artificial bee colony algorithm can get rid of the problems of local minimization and slow convergence of the BP neural network. It can not only use the global optimization function of the ABC algorithm [4], but also give play to the advantages of the BP neural network's strong nonlinear mapping ability for complex problems, reduce the risk of overfitting, and effectively improve the accuracy of prediction.

Artificial bee colonies mainly research hired bees and non-hired bees [10], among which non-hired bees are mainly composed of observation bees and reconnaissance bees. The three types of bees switch roles when searching for honey sources, and the process of finding honey sources by the bee colony is the process of solving the optimal result.

The modeling process of the ABC algorithm mainly consists of four stages: the initialization stage, the hiring bee stage, the observing bee stage, and the reconnaissance bee stage [11]. Define the size of the bee population and set control parameters during the initialization phase. For the search process, firstly, the honeybee is employed to find and evaluate the honey source and find a new honey source, and then observe the honeybee to further select the honey source according to the probability calculated by the fitness value provided by the honeybee, and optimize the honey source, that is, the objective function. At the same time, to avoid falling into local optima, when the iteration limit has been reached and the ideal optimization has not been reached, all bees are transformed into reconnaissance bees, abandoning the original solution, and searching for and generating a new optimal solution.

The steps to improve the BP neural network using the ABC optimization algorithm are as follows:

Step 1: Establish a BP neural network

Step 2: Initialize the artificial bee colony algorithm

If the number of hired bees is N_e, the number of observed bees is N_o, and the number of honey sources is N_s, then the three satisfy the following relationship:

$$N_e = N_o = N_s. \tag{8}$$

Assuming the maximum number of iterations is MCN, the optimization limit is lim, and the D-dimensional vector $Xi_{(i=1,2,\cdots,Ns)}$ represents the weight and threshold of the BP neural network. As

the BP neural network, if m, i, and n are the number of neurons in the input layer, hidden layer, and output layer, then dimension D satisfies:

$$D = (m + 1) \cdot i + (i + 1) \cdot n. \tag{9}$$

Step *3: Calculate the fitness values of all solutions*

The fitness function can be expressed as:

$$f(X_i) = \begin{cases} 1, MSE_i = 0 \\ \frac{1}{MSE_i + 1}, MSE_i > 0 \end{cases}, \tag{10}$$

$$MSE_i = \frac{1}{N_s} \sum_{i=1}^{N_s} (e_i)^2, \tag{11}$$

Where MSE_i is the mean square error of the BP neural network.

Step *4: Observing bees in search of new honey sources*

Observe the bees to find new honey sources, and use a greedy algorithm to reserve the new and old honey sources with the largest fitness. Set α_p as the honey source, and l_i, u_i represents the upper and lower limits of the honey source. The new honey source can be expressed as:

$$\alpha_{pi} = l_i + rand(0,1) \cdot (u_i - l_i). \tag{12}$$

Step 5: Calculate the probability of each possible solution appearing

Observing bees to further search for existing neighborhoods based on possible solutions. The probability of being selected as a honey source by observed bees can be expressed as:

$$P_i = \frac{f(X_i)}{\sum_{i=1}^{N_s} f(X_n)}. \tag{13}$$

Step *6: Determine the optimal solution*

When the threshold of optimization times is reached, the reconnaissance bee instructs to "abandon the honey source", and then generates a new solution again and replaces the old solution.

Step *7: Determine the number of cycles*

If the number of iterations is greater than MCN, the training ends; If the number of iterations is less than MCN, return to step 4 to continue training.

2.3. The structure of XGBoost

XGBoost is an optimized distributed Gradient enhancement library, which achieves a machine learning algorithm under the Gradient Boosting framework [12]. XGBoost avoids overfitting by regularization and early stopping while maintaining efficient processing tasks and automatically handles anomalies such as missing values [13]. Unlike algorithms such as VAR, which are not supported by theory, XGBoost's high interpretation helps users gain insight into how the model makes its reasoning and predictions.

Each iteration of the XGBoost system builds a new decision tree, learns new functions, and fits the last predicted residuals by adding trees and growing a tree through feature splitting [14]. After the training of k trees, the prediction of a sample score is completed. In other words, each tree is placed on a leaf node according to the characteristics of the sample, and each leaf node corresponds to a score, and the combined scores of all leaf nodes are the predicted value of the sample [15].

The construction process of the XGBoost model can be written as:

$$A_f = 2020 - A_i, \tag{14}$$

$$S_s = S_t - S_l. \tag{15}$$

The loss function is used to reveal training errors:

$$Obj(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) + \sum_{i=1}^{t} \Omega(f_i). \tag{16}$$

Regularization defines complexity:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2. \quad (17)$$

Therefore, the objective function of XGBoost can be expressed as:

$$L(\phi) = \sum_i l(y_i') - y^i + \sum_k \Omega(f_t). \quad (18)$$

XGBoost utilizes a greedy algorithm to traverse the feature partition points of all features and uses the objective function value after splitting to gain more than the objective function of the single leaf node. Each time, take the best split/build further based on the last split.

3. Results

3.1. Data preprocessing

By reading literature and combining with the special nature of sailing ships, we mainly explore make, length, service age, LWL, beam, draft, displacement, sail area, longitude, latitude, global GDP, GDP, GDP per capita, area, population Possible impact on second-hand sailboat prices [4]. And our data preprocessing mainly includes search, conversion, transformation, protocol, standardization, cleaning, integration, data set division, and other steps. We have searched the data describing the nature of sailing on the official website of Sailing, the World Bank, and the National Bureau of Statistics of the United States, including GDP and all sea areas in the region.

For the data conversion, because the model in this paper is a data-driven prediction model, which requires converting original non-digitized data (such as text, images, etc.) into digital or numerical data for subsequent analysis and modeling. For the list price, we use MySQL to convert a string to a number; The variant was quantified with factors such as longitude, LWL, draft, and displacement. Similarly, Geographic Region and Country/Region/State are quantified by their latitude, longitude, GDP, area...etc. We quantify consumer recognition of the make by the number of sailboats produced, and the more the population of used sailboats, the higher the consumer recognition, provided quality is assured. Therefore, we use the number of sailboats included in make instead of make for the study.

For the data protocol and data standardization, previously, the effect of make on the price of used sailboats was measured by the effect of the total number of made in the table on the price of used sailboats. However, because the number of sailboats made varies greatly from one sailboat to another, some important factors are lost in direct data processing. Therefore, we first carried out k-means clustering analysis, and then divided the quantity data of all sailboats into four categories and reassigned them with “1,2,3,4”, to reduce all quantity data to a manageable scale.

For the data cleaning, since the data of second-hand sailboats are not time series, the validity of the outlier test cannot be guaranteed. This article focuses on the handling of missing values. Similarly, non-time series data make it impossible to use mean value substitution, interpolation, and other methods to deal with missing values. Therefore, all missing values are eliminated in this paper: the initial data amount is 3491, and the effective data amount is 3207.

For data integration, this article uses SQL language to link different table data. Because there are too many variables involved in this paper, to facilitate searching and identification, all factors affecting price changes are arranged and numbered according to the priority of “ship age” and the second priority of “make” We used 2200 data as the training set and 1007 data as the test set, with a training set to test set ratio of nearly 2:1.

The preprocessing of all data can be shown in Figure 2.

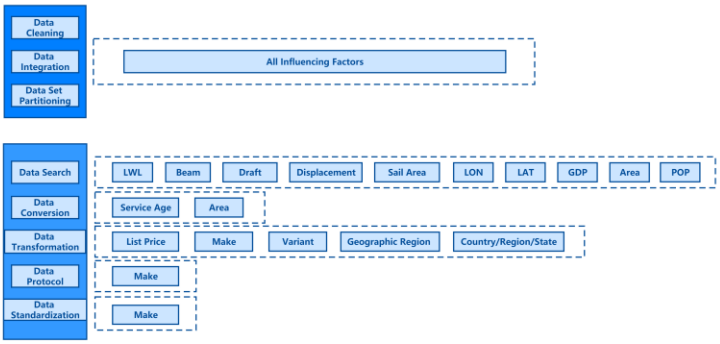

Figure 2. Data preprocessing process

3.2. Analysis of experimental results of the BP neural network

The results of the prediction with the BP neural network are shown in Figure 3. When using BP neural network to predict the price of second-hand sailboats, viewing the regression results in the neural network toolbox can obtain in Figure 3. The regression coefficients between the predicted price and the marked price on the training set, testing set, and validation set are 0.9079, 0.8466, and 0.8876, respectively. After multiple training sessions, the model's performance is excellent, with regression coefficients ranging from 0.85 to 0.9. The model's performance meets the basic requirements, but the regression coefficient does not exceed 0.9.

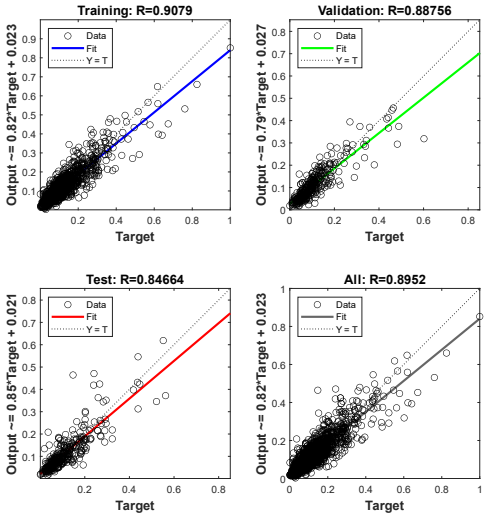

Figure 3. Predicted results of BP neural network

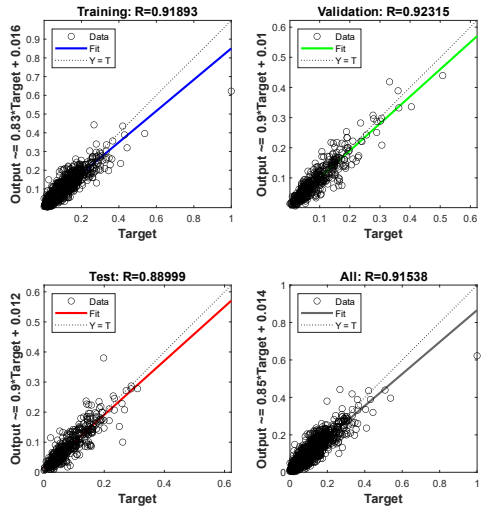

Figure 4. Predicted results of ABC-BP

3.3. Analysis of experimental results of the ABC-BP model

Using ABC algorithm to improve BP neural network. Looking at the regression results in the neural network toolbox, Figure 4 shows that the regression coefficients between the predicted price and the labeled price on the training set, test set, and validation set are 0.9190, 0.8900, and 0.9232, respectively. In the ABC-BP training results, the regression is basically above 0.9, with a small deviation.

3.4. Analysis of experimental results of the XGBoost

Using PyTorch to call the XGBoost model, the training parameter values are n_ estimators=1000, learning_rate=0.05, gamma=0, subsample=0.7, colsample_bytree=1, max_ Depth=7, 20% of the dataset was extracted as the test set. After multiple iterations, the XGBoost model met the requirements, and the predicted values of all data were close to the true values (Figure 5). The regression coefficient of the test set was calculated to be 0.87.

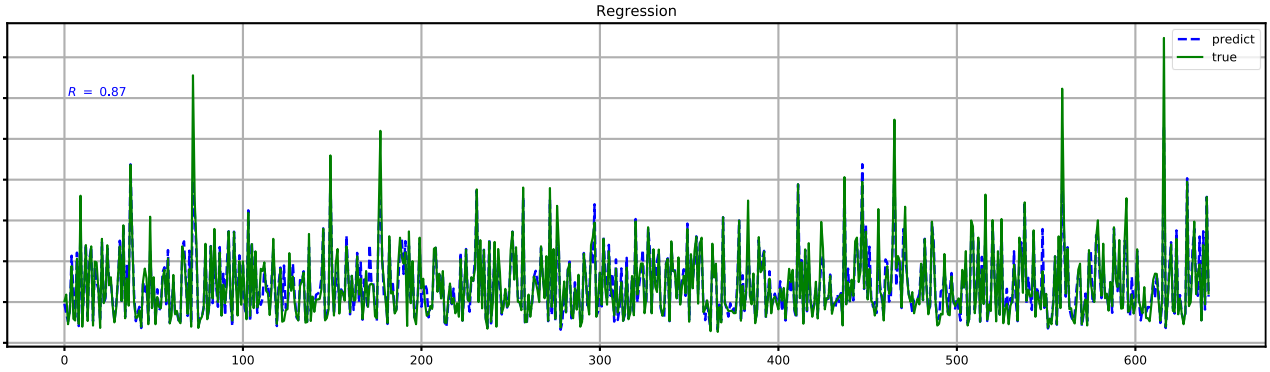

Figure 5. XGBoost predictions

By comparing the results, the prediction result of the ABC-BP neural network is the best, and the fitting coefficient of the verification set can reach 0.92. The ABC-BP model has better prediction performance and relatively small errors, which can completely meet the demand, the prediction speed is fast, and the operation is convenient.

4. Conclusion

The massive data on the characteristics of sailing boats and the fluctuations of the sailing market provide a basis for the analysis of factors affecting the price of sailing boats and the price prediction model of second-hand sailboats. Models used in the past often fail to make good predictions of such multi-input, nonlinear problems, and there are serious losses in time and computing power. This paper makes full use of the artificial bee colony algorithm to improve the problem that BP neural networks are prone to overfitting and improves the speed and efficiency of BP neural network operations. The experimental results show that compared with the BP neural network and XGBoost model, the ABC-BP neural network model has higher accuracy and robustness, and has high application value for the prediction of second-hand ship prices.

However, at the same time, due to the inability of the dataset to be considered massive, the accuracy of the model needs further improvement. In the future, models can be improved by improving data quality.

References

[1] Du K, Jiao F. Influencing factors in the evaluation of second-hand ship price market method [J]. Water Transport Management, 2020, 42(11): 1-4+7.

[2] Xiao Q, Zhang Y. Research on second-hand Ship Price Evaluation based on BP neural network [J]. Ship engineering, 2013 (2): 100-103.

[3] Zhou C. The influence of the US dollar index on asset price changes of second-hand bulk carriers [D]. Shanghai University of Finance and Economics, 2021.

[4] Huang F, Liu D, An T, et al. Port container throughput forecast based on ABC optimized BP neural network [C] // IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2020, 571(1): 012068.

[5] McClelland J L, Rumelhart D E, PDP Research Group. Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models [M]. MIT press, 1987.

[6] Cao L, Wang M. Forecast and Analysis of Domestic Iron Ore Shipping Pricing Based on BP Neural Network [C] // Advances in Artificial Systems for Logistics Engineering. Springer International Publishing, 2021: 76-84.

[7] Liu E, Li J, Zheng A, et al. Research on the Prediction Model of the Used Car Price in View of the PSO-GRA-BP Neural Network [J]. Sustainability, 2022, 14(15): 8993.

[8] Yu J, Tan Y. Analysis of the Effect of BP Neural Network on Function Approximation [J]. Advances in Applied Mathematics, 2019, 8(8): 1453-1456.

[9] Karaboga D, Basturk B. On the performance of artificial bee colony (ABC) algorithm [J]. Applied soft computing, 2008, 8(1): 687-697.

[10] Badia-Miró M, Carreras-MarÍn A, Huberman M. Smooth sailing: market integration, agglomeration, and productivity growth in interwar Brazil [J]. European Review of Economic History, 2023, 27(1): 45-69.

[11] Choong S S, Wong L P, Lim C P. An artificial bee colony algorithm with a modified choice function for the traveling salesman problem [J]. Swarm and evolutionary computation, 2019, 44: 622-635.

[12] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms [J]. Artificial Intelligence Review, 2021, 54: 1937-1967.

[13] Abdikan S, Sekertekin A, Narin O G, et al. A comparative analysis of SLR, MLR, ANN, XGBoost and CNN for crop height estimation of sunflower using Sentinel-1 and Sentinel-2 [J]. Advances in Space Research, 2023, 71(7): 3045-3059.

[14] Zhang P, Jia Y, Shang Y. Research and application of XGBoost in imbalanced data [J]. International Journal of Distributed Sensor Networks, 2022, 18(6): 1-10.

[15] Sagi O, Rokach L. Approximating XGBoost with an interpretable decision tree [J]. Information sciences, 2021, 572: 522-542.