Prediction of Second-Hand Sailboat Prices Based on GA-BP Model

Kaiyan Xiao *, Huatai Pan, Shuo Zheng

College of Information, Mechanical, and Electrical Engineering, Shanghai Normal University, Shanghai, China, 201418

* Corresponding Author Email: 1000500996@smail.shnu.edu.cn

Abstract. Second-hand sailboats now have a huge market worldwide. Accurate prediction of second-hand sailboat prices is of great significance for reducing market opacity, protecting consumer rights and improving social and economic benefits. In order to achieve accurate prediction of second-hand sailboat prices, genetic algorithm was used to optimize the parameters of the BP neural network and a GA-BP network model was constructed. The experiment shows that the prediction result of the GA-BP model for catamarans has a R^2 value of 0.709, which is significantly improved compared to the BP neural network.

Keywords: BP Neural Network, Genetic Algorithm, Prediction Model, Sensitivity Analysis.

1. Introduction

The prediction of second-hand sailboat prices can greatly assist second-hand sailboat brokers and consumers in evaluating the prices of second-hand sailboats and reduce the opacity of the market to protect consumer rights [1]. However, the price of second-hand sailboats is often influenced by many factors, including factors such as the lifespan of the sailboat itself, the size of the deck space, the length of the crossbeam, and the range of navigation, as well as factors such as the climate, economic level, and coastline length of the area where second-hand sailboats are sold [2-3]. It can be seen that predicting the price of second-hand sailboats is a difficult and complex issue.

In the prediction situation influenced by multiple factors, machine learning methods represented by BP neural network [4] often has more advantages than traditional prediction methods. The structure of BP neural network is simple and has been applied to many prediction problems [5-6]. However, the selection of hyperparameters in its network structure is relatively difficult, such as a learning rate [7]. A small learning rate can lead to slow model convergence, while a large learning rate may lead to the model not being able to converge. The traditional parameter optimization methods include grid search (GS) [8] and random search (RS) [9]. GS is a search performed in parameter space with a certain step size. A shorter step size can lead to very time-consuming training, while a longer step size can lead to poor optimization results. RS randomly samples in the feasible domain, and the effect of this method is similar to GS. It requires a larger number of sampling times to achieve good optimization results. Therefore, traditional hyperparameter optimization methods are often very time-consuming. And research [10] proposed genetic algorithms to find the optimal hyperparameters by imitating the mechanisms of selection and genetics in nature. Through distributed computing, the optimization efficiency of hyperparameters has been greatly improved. Therefore, this paper combines genetic algorithms and BP neural networks to improve the accuracy of predicting second-hand sailboat prices [11].

2. Model building

2.1. Data collection

This paper collected various data of used sailboats from the website in Table 1 and provided corresponding explanations for the features, as shown in Table 2.

Table 1. Data source

Data Set	Website Source
Sail boat data	www.sailboatdata.com
GDP per capita data	www.worldbank.org
Wind speed data	www.globalwindatlas.info
Temperature data	www.en.climate-data.org
latitude and longitude data	www.latlong.net

Table 2. Explanation of features

Feature	Explanation	Unit
Beam	the width of the boat	ft
S.A	the total area of all sails on the boat	ft^2
Draft	the maximum depth of water that the boat's hull occupies when it is afloat	ft
Displacement	the weight of water displaced by the boat when it is afloat	lb
Fuel	the capacity of the fuel storage on the boat	$gals$
Water	the capacity of the fresh water storage on the boat	$gals$

2.2. Feature selection

This paper selected ‘Beam’, ‘S.A’, ‘Draft’, ‘Displacement’, ‘Fuel’, ‘Length’ and ‘Water’ as the features of the sailboat variant. At the same time, ‘GDP per capita’ was included as a representative of regional prosperity level into consideration. In addition, two additional features were constructed: ‘Service life’ and ‘Make value’. Among them, ‘Service life’ indicates the number of years the ship has been used, and ‘Make value’ indicates the brand value of the manufacturer.

Construction method of ‘Service life’:

$$Server\ life_i = 2023 - year_i \tag{1}$$

Construction method of ‘Make value’:

$$Make\ value_i = \frac{\sum_i^{n_i} v_{ij}}{n_j} \tag{2}$$

Where, $year_i$ is the year of construction of the ship i, n_i is the number of sailboats belonging to make i, V_{ij} is the listing price of the sailboat j in make i.

2.3. Data Pre-processing

The dataset contains missing values, outliers or other anomalies due to the long span of sailboat manufactured time, multiple manufacturers involved and various complex basic configuration information. These factors could potentially affect subsequent data mining and model building processes. Therefore, preliminary data cleaning on the dataset was necessary.

This paper used box plots to observe the price distribution of different variants and remove those outliers that are overly extreme as shown in Figure 1 and Figure 2.

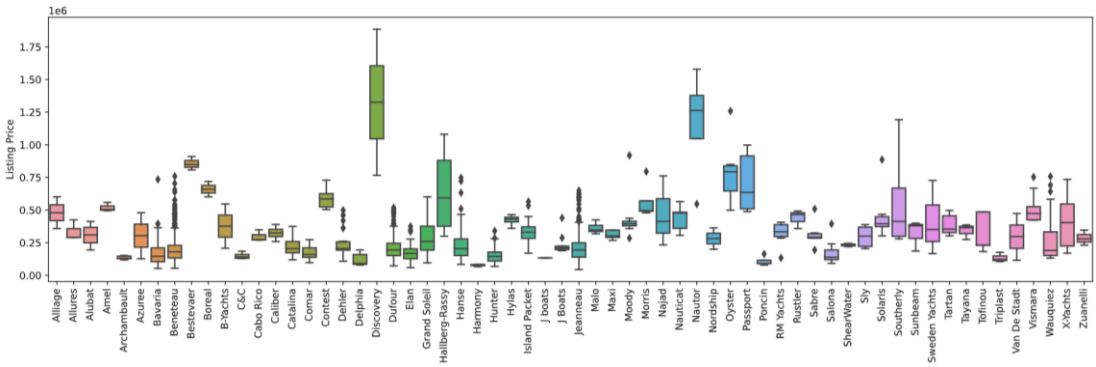

Figure 1. Box plot of monohulled sailboat listing price and make

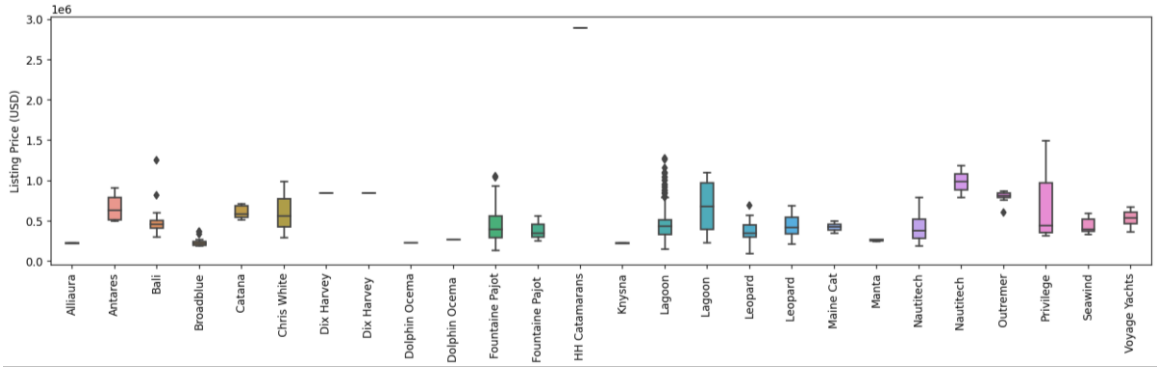

Figure 2. Box plot of catamaran listing price and make

After removing abnormal data, correlation matrices based on covariance were calculated to avoid selecting features that contain highly similar information as shown in Figure 3 and Figure 4.

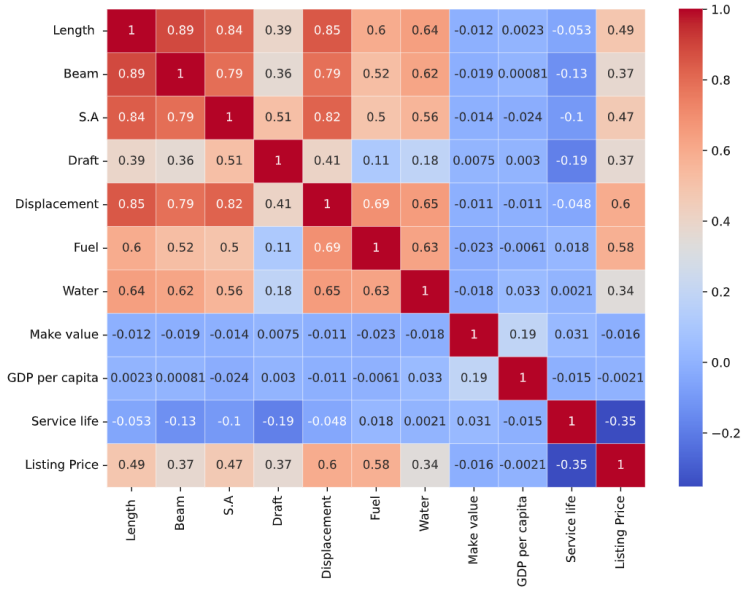

Figure 3. Correlation coefficient matrix for monohulled sailboat

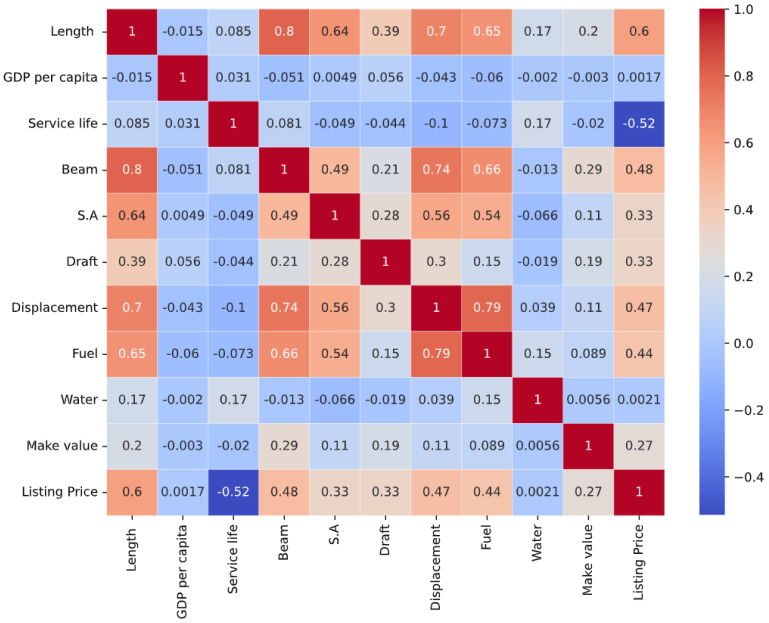

Figure 4. Correlation coefficient matrix for catamaran

It can be seen in Figure 3 that the correlation coefficient of 'Length' and 'Beam' is greater than 0.85, which are highly correlated, so some features of monohulled sailboat need to be removed.

2.4. Feature removal

This paper used XGBoost to rank the feature importance of a single ship. The core of the tree model's ability to obtain the importance of each feature lies in the calculation of information gain. Information gain is the extent to which uncertainty is reduced under a certain conditional split, and its formula is:

$$IG(X,Y) = H(Y) - H(Y|X) = \sum_{x,y} p(x) \cdot p(y|x) \cdot logp(y|x) - \sum_{y} p(y) \cdot logp(y) \quad (3)$$

The empirical entropy $H(Y)$ represents the uncertainty of the dataset Y, while the empirical conditional entropy $H(Y|X)$ represents the uncertainty of the dataset Y under the given condition of feature X. The difference between the two is the information gain, which represents the degree to which the uncertainty of the dataset Y decreases under the influence of the feature X. From above, the feature ranking of monohulled sailboat can be obtained as shown in Figure 5.

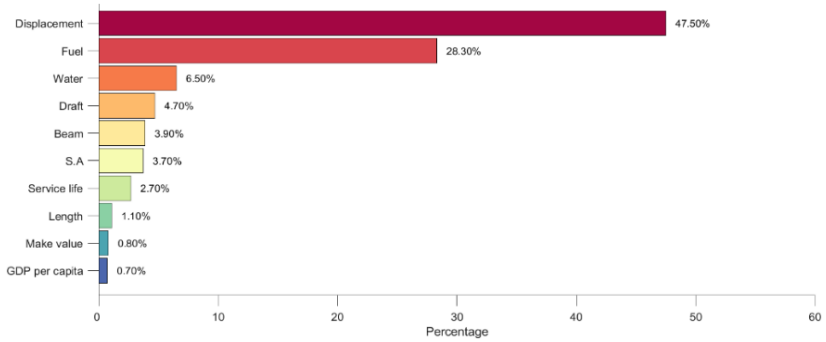

Figure 5. Monohulled sailboat features importance bar chart

According to the correlation matrix and Figure 5, it can be inferred that the correlation coefficients between the feature 'Length' and the features 'Displacement' and 'Beam' are 0.85 and 0.89, respectively. However, the importance of the 'Length' is lower in the selected feature set. Therefore, 'Length' was removed from the dataset of monohulled sailboat.

2.5. BP neural network

BP neural network can learn the mapping relationship between a large number of features and dependent variables. At the same time, it adopts the steepest descent method, that is, continuously adjusts the weight and threshold of the network through backpropagation, so as to minimize the sum of squared errors of the network.

Therefore, a simple network structure based on the number of features of the data was built, as shown in the Figure 6, which can predict the price of monohulled sailboat and catamaran.

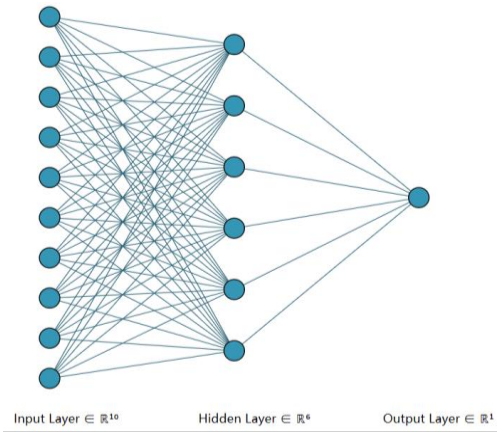

Figure 6. BP neural network structure

Where the transfer function of the hidden layer is:

$$tansig(x) = \frac{2}{1+e^{-2x}} - 1 \tag{4}$$

And the transfer function of the output layer is:

$$logsig(x) = \frac{1}{1+e^{-x}} \tag{5}$$

2.6. GA-BP neural network

The selection of network structure, initial connection weights, and thresholds of BP neural networks has a significant impact on network training, but it is difficult to accurately obtain them. Genetic algorithm simulates the process of natural selection and increasing fitness in biological evolution, continuously optimizing the fitness of the population through operations such as selection, crossover, and mutation, and ultimately obtaining the optimal solution. In this regard, this paper used genetic algorithm to optimize the initial connection weights and thresholds of BP neural networks, as shown in Figure 7.

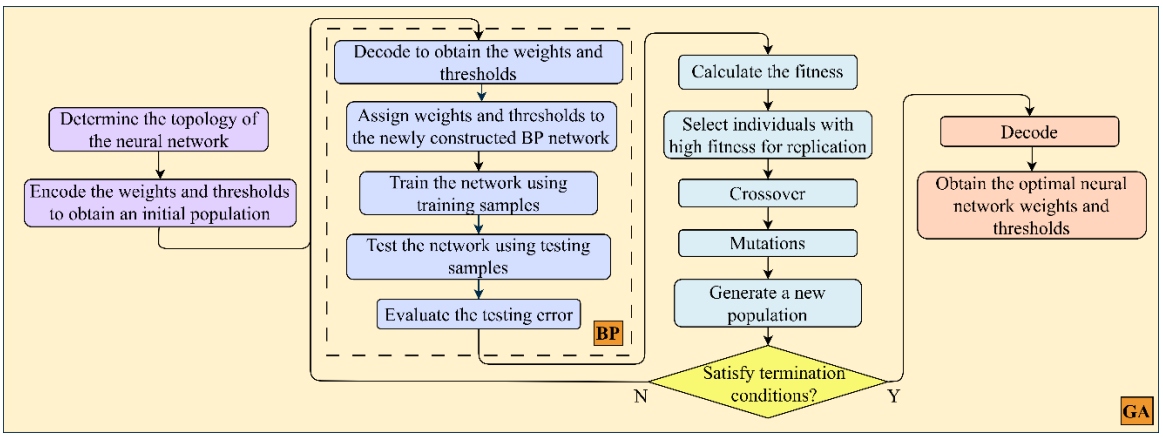

Figure 7. GA-BP neural network flow chart

When optimizing BP neural networks using genetic algorithm, the parameters of the BP neural network are encoded into a chromosome, where each gene represents the value of a parameter. By continuously updating the chromosomes of the population, i.e. continuously updating the parameters of the BP neural network, in order to obtain better regression prediction results. The specific steps are as follows:

(1) Population Initialization:

Determine the population size and encode each individual in the population, each population individual is composed of the weights and thresholds of each node in the neural network. The weights and thresholds of population individuals are randomly selected within the value range of $[0, 1]$.

(2) Determine the fitness function:

The reciprocal of the absolute error sum between the predicted value y_k and the actual value o_k of the BP neural network as the fitness function F.

$$F = \frac{1}{\sum_{i=1}^{N}\sqrt{\sum_{k=1}^{m}(y_k^i - o_k^i)^2}} \tag{6}$$

Where, N is the number of samples, m is the dimension of the output variable.

(3) Network training and testing:

Decode the individual population, and use the decoded weights and thresholds as the initial weights and thresholds of the neural network. Using the batch gradient descent training method to adjust the weights and thresholds of the network according to the negative gradient direction of the network parameters.

(4) Selection:

Using the roulette wheel method for selection operation, individuals with high fitness have a higher probability of inheriting to the next generation, and individuals with low fitness have a lower probability. The probability p_i of each individual i being selected is:

$$p_i = \frac{F_i}{\sum_{i=1}^{c} F_i} \tag{7}$$

Where, c is the number of individuals in the population, and F_i is the fitness value of individual i.

(5) Crossover:

Two paired individuals cross over in a certain way with probability p_c. Exchange some of the genes to form two new individuals. Using the real number crossover method, the k_1th and k_2th individuals at the jth gene crossover operation method is as follows:

$$\begin{cases} g_{k_i j} = g_{k_1 j} \cdot \alpha + g_{k_2 j} \cdot (1 - \alpha) \\ g_{k_2 j} = g_{k_1 j} \cdot \alpha + g_{k_1 j} \cdot (1 - \alpha) \end{cases} \tag{8}$$

Where, the jth genes of individuals k_1 and k_2 are $g_{k_i j}$ and $g_{k_2 j}$ respectively, and α is a random number in the range of $[0, 1]$.

(6) Mutation:

Select the jth gene g_{ij} of the ith individual to mutate to increase the diversity of the population. Choose a relatively small mutation probability p_γ. The specific steps are as follows:

$$g_{ij} = \begin{cases} g_{ij} \cdot \alpha + (g_{ij} - g_{max}) \cdot \theta \cdot \left(1 - \frac{s}{s_{max}}\right), \beta \geq 0.5 \\ g_{ij} \cdot \alpha + (g_{min} - g_{ij}) \cdot \theta \cdot \left(1 - \frac{s}{s_{max}}\right), \beta < 0.5 \end{cases} \tag{9}$$

Where, g_{max}, g_{min} are the upper and lower bounds of the gene g_{ij}, respectively, s is the current iteration number, s_{max} is the maximum iterations number, θ and β are random numbers in the range of $[0, 1]$.

(7) Best result selection:

After each population iteration, the individual with the maximum fitness is selected as the optimal individual in the population. Then repeat step **4)**, **5)**, **6)** until the maximum number of iterations is met. After the iteration, the optimal individual in the population is decoded and used as the initialization weights and thresholds to obtain the optimized BP neural network.

3. Results

3.1. The establishment of simulation model (model solving)

In the process of the model solving, the structure of the BP neural network can refer to Figure 6, and the various parameters of the genetic algorithm are set as Table 3.

Table 3. GA-BP neural network hyperparameter

Hyperparameters	Value
Optimization times	5
Population size	10
Maximum number of iterations	50
Crossover probability	0.3
Selection probability	0.05

The prediction model is implemented in MATLAB software. This paper compares traditional BP neural network with GA optimized BP neural network, using R^2 as the prediction performance evaluation indicator. The data collected in the experiment will be divided into training and testing sets according to 0.7: 0.3. The results are shown in Table .4, and the fitting effect is shown in Figure

8 and Figure 9. Considering the randomness of GA, this paper selects the best result from the five experiments of GA-BP.

Table 4. Comparison of Prediction Results between BP and GA-BP

Model	Sailboat	R^2	
		Train	Test
BP	Monohulled sailboat	0.595	0.405
	Catamaran	0.697	0.567
GA-BP	Monohulled sailboat	**0.766**	**0.648**
	Catamaran	**0.883**	**0.709**

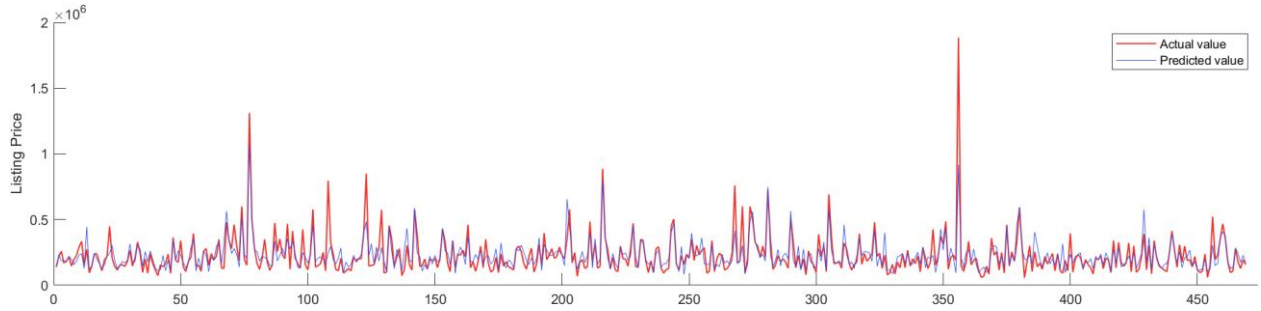

Figure 8. Monohulled sailboat prediction result of GA-BP

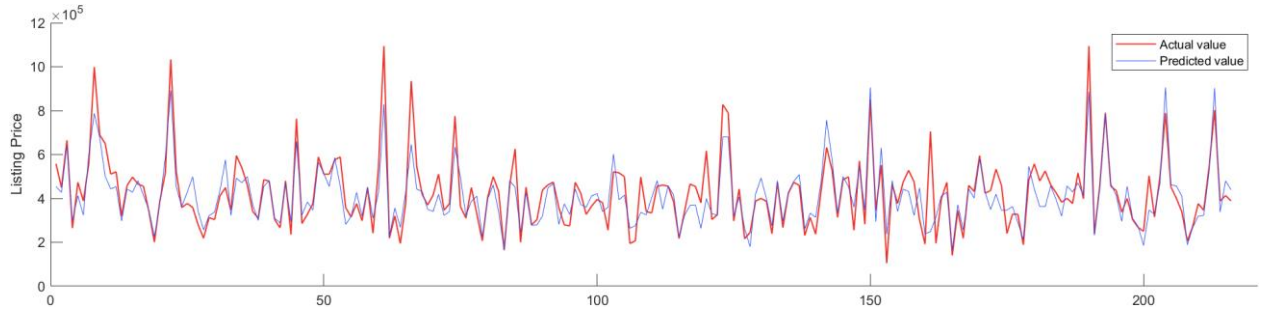

Figure 9. Catamaran prediction result of GA-BP

From the R^2 values in Table.4, it can be seen that the GA-BP model has improved prediction accuracy compared to the BP model in both monohulled sailboat and catamaran.

3.2. Analysis of experimental results (Sensitivity analysis)

This paper conducted a sensitivity analysis on the GA-BP model. In the process of predicting listing price, there are two hyperparameters, the maximum number of iterations and the crossover probability, in the genetic algorithm directly affect its optimization ability. So, adjusting the values of this two hyperparameters in the ranges of $[26, 28, 30, 32, 34]$ and $[0.2, 0.3, 0.4, 0.5, 0.6]$ respectively, and using the correlation coefficient R to judge the prediction effect of the model, as shown in Figure 10.

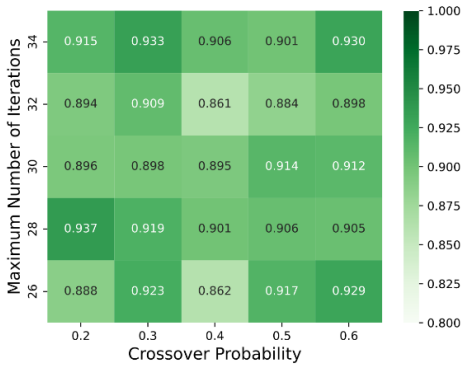

Figure 10. Sensitivity analysis heat map

It can be seen from Figure 10 that as the parameters change, the prediction effect of the model fluctuates between 0.861 to 0.937. Considering the randomness of the genetic algorithm, it can be seen that the model is stable.

4. Conclusion

The massive data in the global second-hand ship market provides a basis for the establishment of second-hand sailboat price prediction models, but traditional BP neural network has the problem of difficult to determine the initial connection weights and thresholds, leading to poor prediction performance of the models. This paper uses genetic algorithm to optimize the initial connection weights and thresholds of the BP neural network, which to some extent improves the accuracy of the model. Meanwhile, the experimental results indicate that the GA-BP neural network model has good robustness and has certain practical application value.

References

[1] Chantelat P, Vignal B. 'Intermediation'in used goods markets: Transactions, confidence, and social interaction [J]. Sociologie du travail, 2005, 47: e71-e88.

[2] Nam H S, De Alwis N, D'agostini E. Determining factors affecting second-hand ship value: linkages and implications for the shipbuilding industry [J]. WMU Journal of Maritime Affairs, 2022, 21 (4): 493-517.

[3] Thalassinos E I, Politis E. Valuation model for a second-hand vessel: Econometric analysis of the dry bulk sector [J]. Journal of Global Business and Technology, 2014, 10 (1).

[4] Werbos P. Beyond regression: New tools for prediction and analysis in the behavioral sciences [J]. PhD thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA, 1974.

[5] Yu S, Zhu K, Diao F. A dynamic all parameters adaptive BP neural networks model and its application on oil reservoir prediction [J]. Applied mathematics and computation, 2008, 195 (1): 66-75.

[6] Cui K, Jing X. Research on prediction model of geotechnical parameters based on BP neural network [J]. Neural Computing and Applications, 2019, 31: 8205-8215.

[7] Yu S, Zhu K, Diao F. A dynamic all parameters adaptive BP neural networks model and its application on oil reservoir prediction [J]. Applied mathematics and computation, 2008, 195 (1): 66-75.

[8] Tang F, Wu Y, Zhou Y. Hybridizing grid search and support vector regression to predict the compressive strength of fly ash concrete [J]. Advances in Civil Engineering, 2022, 2022: 1-12.

[9] Bischl B, Binder M, Lang M, et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2023, 13 (2): e1484.

[10] Holland J H. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence [M]. MIT press, 1992.

[11] Ding S, Su C, Yu J. An optimizing BP neural network algorithm based on genetic algorithm [J]. Artificial intelligence review, 2011, 36: 153-162.