Research on Price Forecast Based on Genetic Algorithm Optimizing Neural Network

Minya Zou [1, *, #], Haixia Zheng [2, *, #]

[1]School of Business Administration, Zhongnan University of Economics and Law, Wuhan, China

[2]School of Law and Economic Sciences, Jaume I University, Castellón de la Plana, Spain

*Corresponding author: minahzou@gmail.com, al418637@uji.es

[#]These authors contributed equally.

Abstract. This paper chooses a number of indicators such as per capita GDP and economic level. First of all, the box diagram method is used to remove the outliers in the data, and then the Box-Cox transform is used to normalize the data. Then calculate the correlation coefficient between features, and finally determine 15 features for the construction of the model. On the basis of BP neural network, using GA algorithm to optimize the weight setting, a GA-BP model is established to predict the prices of different types of forecasting objects. Combined with the search ability of genetic algorithm optimization algorithm and the learning ability of BP neural network, the advantages of both are fully utilized to get better regression prediction results. The result of GA-BP is better than that of BP neural network, which shows that the prediction accuracy of the model is very high, in order to provide some reference for the prediction of other fields.

Keywords: Neural Network, Genetic Algorithm, Big Data, Price Forecast.

1. Introduction

The second-hand sailboat market is popular due to its low prices and diversity. For consumers who want to buy a second-hand sailboat, it is important to understand the price trends and influencing factors [1]. At the same time, for businesses that sell second-hand sailboats, accurately predicting the prices of sailboats can help them develop more effective sales strategies. Therefore, analysing the price changes of second-hand sailboats and their influencing factors has important practical value [2].

2. Data Pre-processing

2.1 Outliers Processing

The data in the problem, like most real-world data sets, suffers from data anomalies, so we treat the data for outliers before analysis. Commonly used outlier detection methods are the 3σ principle and the box-plot method, where the 3σ principle is used for normally distributed data. Therefore, we first counted the price distribution in the dataset and plotted its distribution as a function of probability density. To simplify the establishment of the model, we make the following assumptions:(i) The impact of speculation and monopolies on prices is not considered [3]. As the competition requires analyzing based on objective data, speculation and monopolies are unpredictable factors. Therefore, high prices will be treated as outliers in the analysis. (ii)The data selected in this article contains features that can reflect price changes. As the competition requires selecting features independently, different features may have a significant impact on the results. Therefore, selecting suitable features is crucial to achieving better results. (iii) The data used in the model is true and reliable. Only with reliable data can the results obtained from data analysis in this article be reliable.

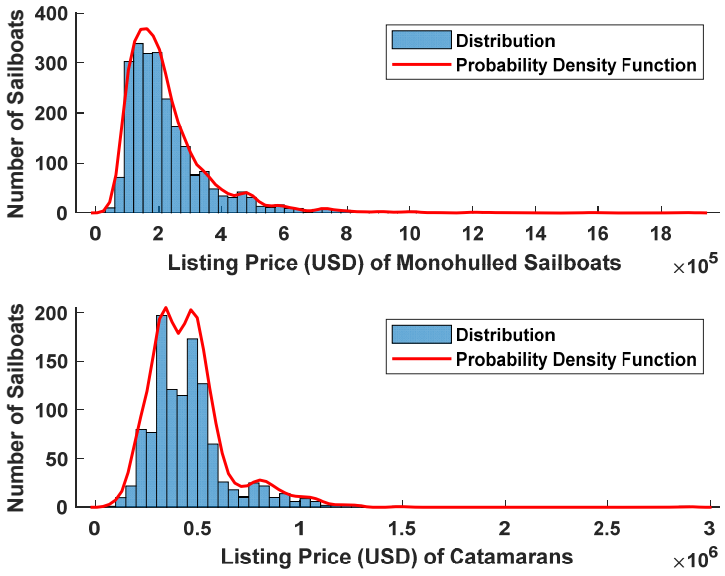

Figure 1. Price distribution of sailboats

As shown in the Figure 1, the distribution of the data is Poisson distribution, so it is not possible to remove the outliers using the 3σ principle. We use the box line plot method to filter outliers.

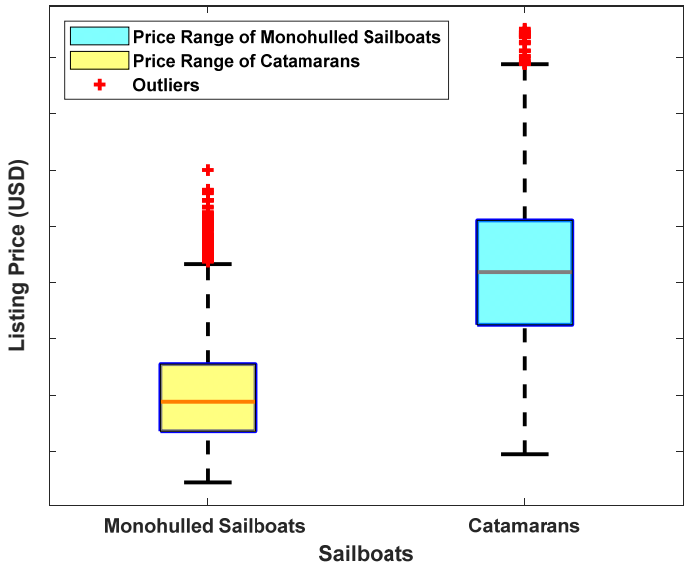

Figure 2. Price outliers for Monohulled Sailboats and Catamarans

The outliers we filtered using the box line plot method are shown in Figure 2, after which we marked them out in the original data. We arrange the prices of each sailboat in order and after that we modify the outliers using linear interpolation.

$$y = y_0 + \frac{(x - x_0)\, y_1 - (x - x_0)\, y_0}{x_1 - x_0} \tag{1}$$

Where x, y denote the outlier data after sorting in order. Then, the original data are replaced with the calculated values of the interpolated values. The data before and after the outliers processing are shown in Figure 3.

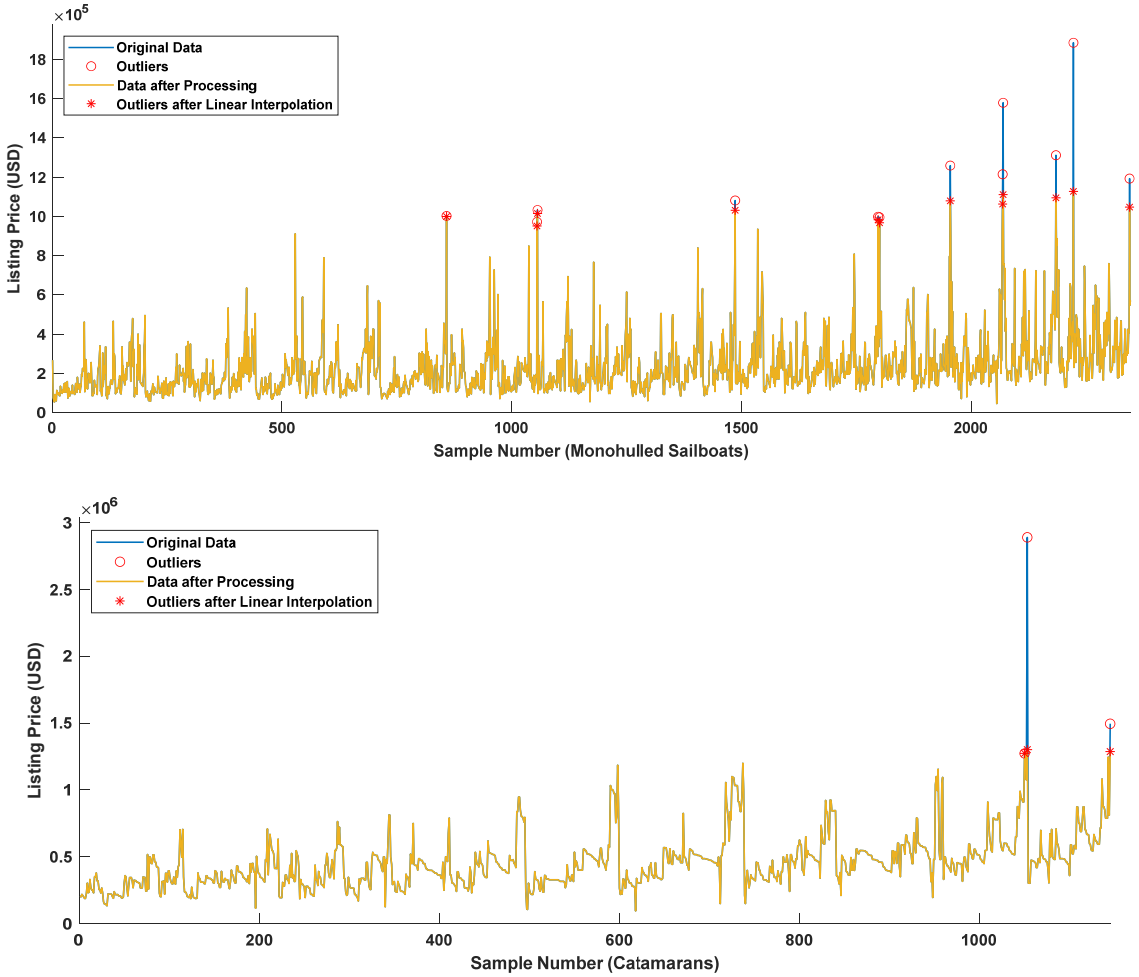

Figure 3. Sailboat price outlier removal

2.2 Data Visualization & Data Pre-processing

The data we used to estimate the indicators came from multiple databases, including the World Bank, International Association of Goods Transport and Trade, World Economic Forum, and the official statistical agency, as well as data published by sailboat manufacturers. For tolerable missing data, we used multiple linear interpolation to reasonably fill in the missing values caused by data omissions, ensuring the smooth progress of data processing and analysis, and obtaining relatively accurate results while maintaining sufficient information. For categorical data such as MAKE, we quantified it using the average selling price of sailboats from that brand. The processed results of the selected data are shown in Figure 4.

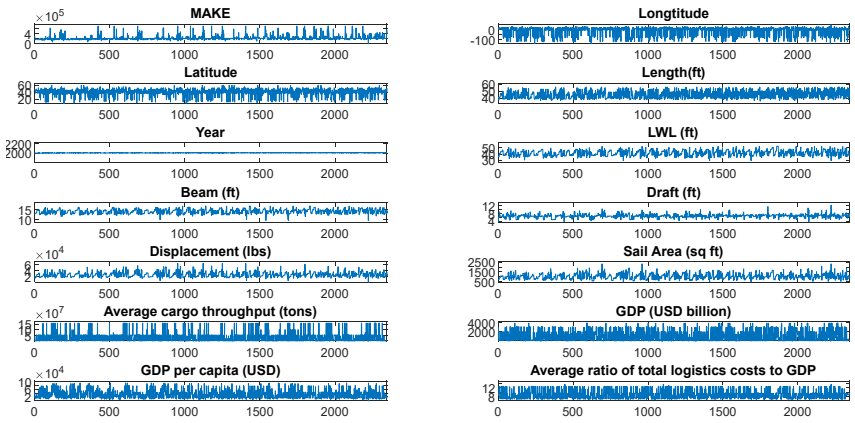

Figure 4. Indicators data chart

To prevent overfitting of the established regression prediction model and improve the generalization ability of the model, it is necessary to explore and analyze the distribution of data in the training set and the test set, striving to ensure the consistency of the data set distribution and present a roughly normal distribution. Taking three features as an example, their data distribution is shown in Figure 5.

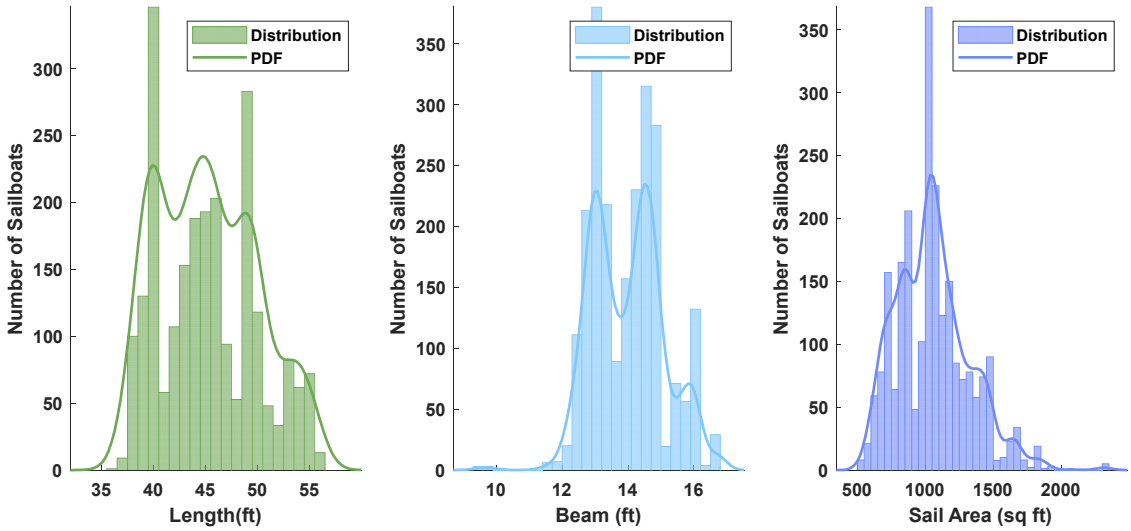

Figure 5. Data distribution of the three features

The distribution of the three features shows a trend towards a normal distribution. However, to ensure the reliability and effectiveness of the study, we need to process the features to meet the requirements of a normal distribution. We use box-cox transformation to achieve this goal. Its formula is as follows.

$$y(\lambda)=\begin{cases}\dfrac{y^{\lambda}-1}{\lambda}, \lambda \neq 0 \\ \ln y, \lambda = 0\end{cases} \tag{2}$$

We plotted the data distribution histograms of the three features after processing, and the results are shown in Figure 6. As can be seen from the figure, the distribution of the feature data after the box-cox transformation is close to normal and can be used to build the model.

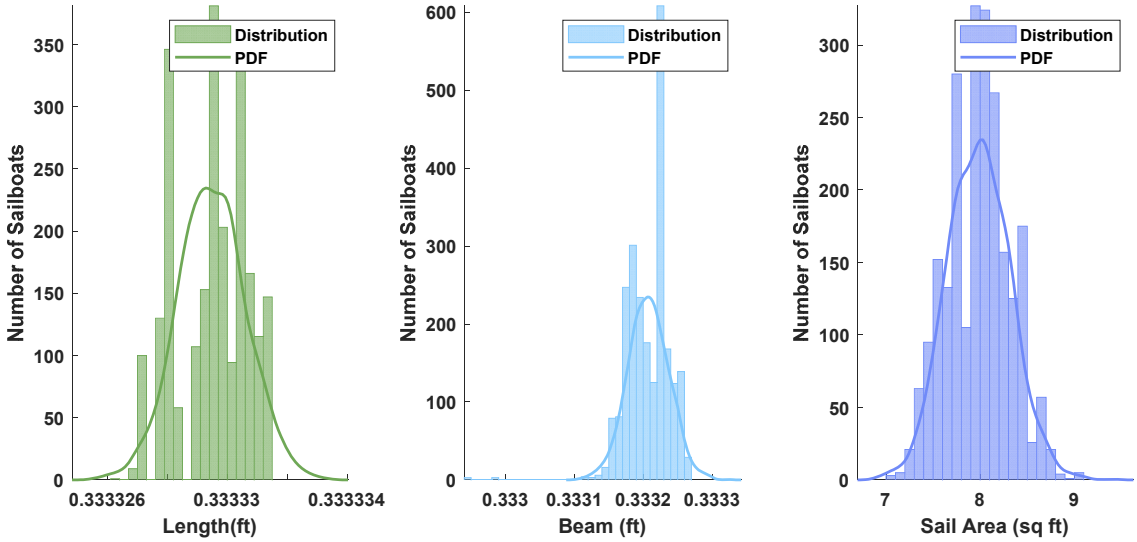

Figure 6. The distribution of the three features after Box-Cox transformation

3. Establishment and solution of the model

3.1 Feature Selection

We first preliminarily select sample features, remove features that are not very relevant to the label or are overly correlated with each other. First, we calculate the correlation coefficient r between the values of the features, and the calculation formula is as follows:

$$r_{ij} = \frac{\sum_{k=1}^{n}\left(f_{ik}-\overline{f_i}\right)\left(f_{jk}-\overline{f_j}\right)}{\sqrt{\sum_{k=1}^{n}\left(f_{ik}-\overline{f_i}\right)^2}\sqrt{\sum_{k=1}^{n}\left(f_{jk}-\overline{f_j}\right)^2}} \tag{3}$$

After obtaining the correlation coefficient matrix, the corresponding heat map is drawn as shown below:

Figure 7. Correlation coefficient matrix heatmap

From Figure 7, it can be seen that the correlations between various features are relatively small. Therefore, we can use them directly for mathematical modeling.

3.2 BP Neural Network

The BP (backpropagation) neural network algorithm is a typical multi-layer feedforward neural network. It uses the error backpropagation algorithm for learning and training. It is currently the most widely used and the most intuitively structured neural network with the easiest-to-understand working principle. It is highly flexible and has strong data fitting ability, and can learn and store a large number of input-output pattern mappings [4]. During the training process, the neural network continuously adjusts the weights and thresholds between the input layer and the hidden layer, and between the hidden layer and the output layer. Training stops when the neural network output value is consistent

with the target value or when the iteration limit is reached [5]. In addition, this algorithm also uses the steepest descent method to continuously adjust the weights and thresholds of the network through backpropagation, so as to minimize the sum of squared errors of the network.

3.3 Genetic Algorithm

Genetic algorithm, as a random search optimization method, is an algorithm to obtain the optimal solution according to the evolutionary survival law of the survival of the fittest in nature. In the system, it mainly uses coding technology to simulate the data set as a natural population based on the obtained initial solution, and then evolves the population through selection, mutation, crossover and other methods. Due to the screening of the natural law of ' natural selection and survival of the fittest, the obtained individuals are usually the best quality, and the result reflected is also the optimal solution of the data set [6].

There are three main replacement methods of genetic algorithm, namely selection, crossover and mutation. In the iterative process, a group of candidate individuals are retained. Through multiple iterations, the fitness of the population can be improved, and the optimal state can be basically achieved, and the optimization and calculation of the problem solution can be completed. With its good performance of optimization data, we can often use it to solve complex local optimization problems. The contents of the three replacement methods are as follows:

(1) Selection

The fitness ratio selection method, also known as the roulette method, refers to the probability that a unit individual enters the next generation as the ratio of its fitness value to the individual fitness value in the entire population.

$$f_i = \frac{k}{F_i} \tag{4}$$

$$P_i = \frac{f_i}{\sum_{i=1}^{n} f_i} \tag{5}$$

Among them, F_i represents the individual fitness value, k is the fitness coefficient, f_i represents the selection probability of the selection method, P_i is the selection probability of the individual, and n is the total number of individuals in the population. If the value P_i is larger, it indicates that the individual is more likely to be retained.

(2) Crossover

Crossover operation, similar to the genetic mode in nature, refers to the selection of a pair of parents from the population for reproduction. The main operation is to carry out a certain proportion of gene exchange between a pair of parents, so that the offspring can effectively leave some characteristics of the parent on the original basis. Through this repeated exchange, until the population has a suitable size and a higher fitness population.

The arithmetic crossover method, as a method of generating new individuals from random individuals, uses a linear combination to form new individuals. The specific methods are as follows:

$$\begin{cases} a_1 = \gamma a + (1-\gamma) b \\ b_1 = \gamma b + (1-\gamma) a \end{cases} \tag{6}$$

Here, the random constant $\gamma \in [0,1]$.

(3) Variation

Variation refers to the replacement of the original gene at a specific position by the opposite gene, which can also form a new chromosome. This method can greatly promote the increase of diversity among populations, make the genetic algorithm more capable of local random search, and make the convergence to the most solution faster in the field of optimal solution [7]. The specific operation is as follows:

$$h(g)=\beta\left(1-\frac{g}{G_{\max}}\right) \tag{7}$$

$$a_{ij}=\begin{cases}a_{ij}+\left(a_{ij}-a_{\max}\right)\times h(g) & 1\geq r>0.5\\ a_{ij}+\left(a_{\min}-a_{ij}\right)\times h(g) & 0\leq r\leq 0.5\end{cases} \tag{8}$$

Among them, a_{ij} is the j th mutated gene of the selected i th individual, a_{min} is the lower bound of gene a_{ij}, a_{max} is the upper bound of gene a_{ij}, β and r are random numbers, g is the number of iterations, and G_{max} is the maximum number of iterations.

At the same time, there are many different parameters in the genetic algorithm, such as the size of the population, the fitness function, the crossover rate, the mutation rate, the maximum number of iterations, the probability of operator implementation, etc. The effective setting of their values also plays a particularly important role in optimizing the population.

3.4 Optimize BP Neural Network Based on Genetic Algorithm (GA-BP)

Even though BP neural network can effectively connect the interaction between input and output information by virtue of its good adaptability and self-learning ability, so as to facilitate data classification and fitting tasks, it is also very easy to fall into the situation of over-fitting, which makes the prediction value error too large. Because the genetic algorithm is very dependent on the random population and the parameters of each operator are too tentative, this paper intends to introduce the genetic algorithm when considering the optimization problem of BP neural network. It is expected to solve the problem of different optimal solutions of BP neural network through the expansibility and global search ability of genetic algorithm to further improve the prediction accuracy.

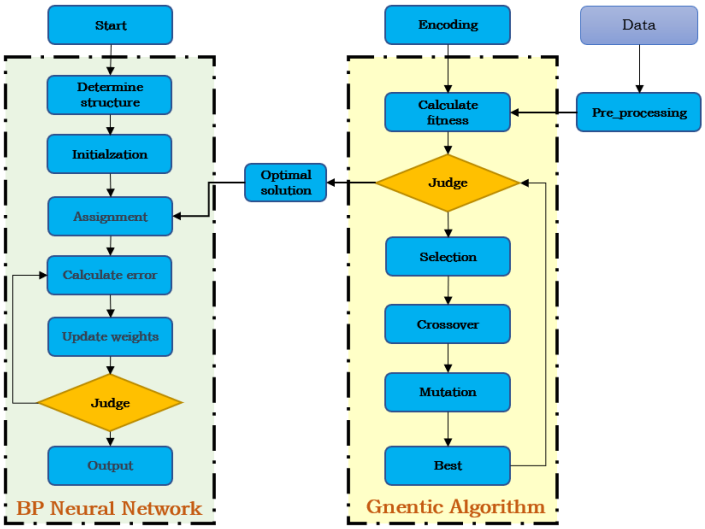

Figure 8. Genetic algorithm to optimize BP neural network main algorithm flow chart

Based on GA genetic algorithm, the prediction model of BP neural network is optimized. In essence, the solution set is expanded by the universality of the search target of genetic algorithm, and then the best scheme of BP neural network prediction is found by the positioning search method, so as to optimize the final prediction result.

The power load forecasting model based on BP network optimized by genetic algorithm is as follows:

Step 1. Construction of initial BP neural network structure: The BPNN topology is determined by the number of input and output parameters, and the weights and thresholds of the neural network are initialized to determine the length of the unit individual coding in the genetic algorithm.

Step 2. Genetic algorithm to optimize the population: according to the training error of BP neural network, set the population fitness function; for the data set initialization coding, the weights and thresholds of the initial population and the unit individual are obtained; the individual weights and thresholds are assigned to the fitness function to obtain the fitness of all individuals. Perform three replacement methods of genetic algorithm: selection, crossover and mutation to find the optimal fitness individual.

Step 3. Acquire the optimal initial value of BP neural network: Analyze whether the optimal fitness individual satisfies the iterative condition, if not, repeat step 2 until the condition is satisfied, and the output structure is used as the optimal initial value of the network.

Step 4. According to BP neural network model prediction training: re-assign network weights and thresholds, calculate iterative errors, judge error accuracy, and perform model prediction analysis.

3.5 Results & Analysis

3.5.1 Determine Evaluation Metrics

To compare the performance of the two models, we used two metrics commonly used in time series forecasting, SMAPE and R-squared.

SMAPE (Symmetric Mean Absolute Percentage Error) is a measure of the magnitude of the error d of the observed and predicted values. It is a second-order moment of the error that contains the variance of the estimate and its deviation, a measure of the quality of the estimate, and its formula is defined as follows.

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{\left|\hat{y}_i - y_i\right|}{\frac{\left(\left|\hat{y}_i\right| + \left|y_i\right|\right)}{2}} \tag{9}$$

R-squared is generally used in regression models to evaluate the degree of agreement between predicted values and actual values. R-squared is defined as the ratio of the regression sum of squares caused by the variable x to the total sum of squares of y. It is also known as the coefficient of determination. The formula is expressed as follows:

$$R^2 = \frac{\Sigma\left(\hat{y}_i - \bar{y}\right)^2}{\Sigma\left(y_i - \bar{y}\right)^2} \tag{10}$$

3.5.2 The Results of Monohulled Sailboats and Catamarans

We randomly selected 80% of the samples for training GA-BP, and used the remaining 20% of the samples to test the effectiveness of GA-BP. The results of the training and testing sets for the Monohulled Sailboats are shown in Figure 9, respectively. It can be seen from the figures that the GA-BP model has a good fitting effect on the data.

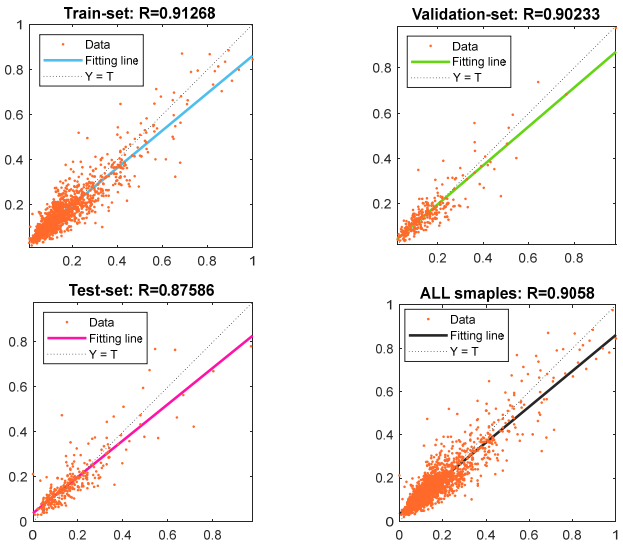

Figure 9. Regression trend lines for each dataset

From the figure, it can be observed that the regression trend lines are generally positively correlated with the data.

To describe the fitting effect of GA-BP on the data more objectively, we calculated the SMAPE and R-squared values for both the training and testing sets. The SMAPE and R-squared values for the training set were 4.7225% and 0.81986, respectively, while those for the testing set were 4.7989% and 0.84015, respectively.

Similarly, we used GA-BP to train and test the Catamarans. The results are shown in Figure 13, which shows that GA-BP has better predictive performance for Catamarans than for Monohulled Sailboats.

To objectively demonstrate this, we calculated the SMAPE and R-squared values for the training and testing sets separately. The SMAPE and R-squared values for the training set were 3.111% and 0.85509, respectively, while those for the testing set were 3.2632% and 0.84378, respectively. Therefore, in the dataset of this competition, GA-BP is more suitable for predicting Catamarans.

3.5.3 Comparison of GA-BP and BP

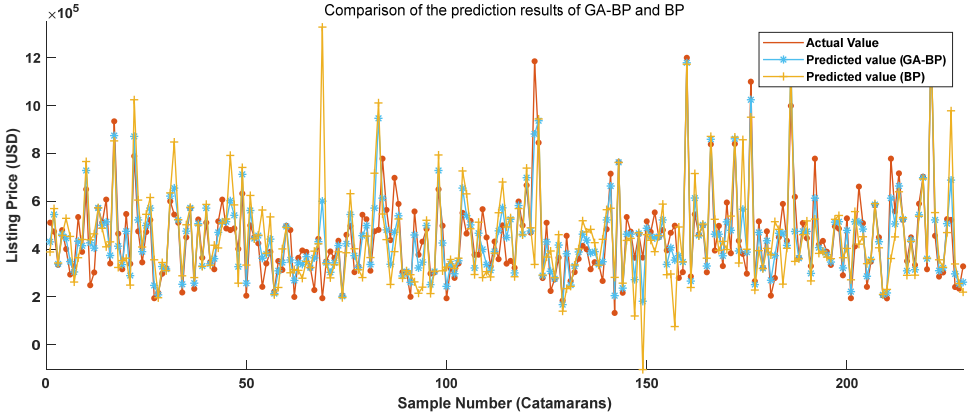

Figure 10. Comparison of GA-BP and BP

Figure 10 shows the results of training the catamaran test set with GA-BP and BP, respectively. From the figure, we can see that BP has a large prediction error for individual samples. We calculated the SMAPE and MAPE of training with GA-BP and BP for Monohulled Sailboats and Catamarans respectively, and the results are shown in the table. It is easy to see that optimizing BP with GA algorithm is a neural network, which has a good improvement on the regression results.

Table 1. Comparison of GA-BP and BP

	Train-Set		**Test-Set**	
Metrics	SMAPE	R-squared	SMAPE	R-squared
GA-BP	4.799%	0.8402	3.111%	0.8551
BP	3.305%	0.8317	3.279%	0.8149

4. Conclusions

On the basis of BP neural network, this paper uses GA algorithm to optimize the weight setting, establishes a GA-BP model to predict the prices of different types of forecasting objects, combines the search ability of genetic algorithm optimization algorithm and the learning ability of BP neural network, and makes full use of the advantages of both, so as to get better regression prediction results. The results show that the R2 and SMAPE of the training set are 0.8402 and 4.799% respectively, while the R2 and SMAPE of the test set are 0.8551 and 3.111%, respectively. The R2 of BP neural network in training set and test set is 0.8317 and 0.8149 respectively. The result of GA-BP is better

than that of BP neural network, which shows that the prediction accuracy of the model is very high, in order to provide some reference for the prediction of other fields.

References

[1] Matthew Hemley. Royal Opera House cuts ties with BP after 33 years[J]. The Stage,2023(5).

[2] Tang Xinxin, Yue Yuanhe, Shen Yansong. Prediction of separation efficiency in gas cyclones based on RSM and GA-BP: Effect of geometry designs[J]. Powder Technology,2023,416.

[3] Wang Zirui, Wu Jing, Wang Haitao, Wang Huiyuan, Hao Yukun. Optimal Underwater Acoustic Warfare Strategy Based on a Three-Layer GA-BP Neural Network[J]. Sensors,2022,22(24).

[4] Zheng Yaze, Tang Lin, Liu Shiying, Zhou Jiakai. Optimization of Aluminum Alloy Rifled Barrel ECM Process Parameters Based on GA-BP Algorithm[J]. Journal of Physics: Conference Series,2022,2383(1).

[5] Yi Hongjie, Zhang Ke, Ma Kun, Zhou Lijian, Tang Futong. Prediction of Natural Rubber Customs Declaration Price Based on Wavelet Decomposition and GA-BP Neural Network Group[J]. Mathematics,2022,10(22).

[6] S Nithya V, Rai Reena, Boppe Appalaraju, Chaithra V. Evaluation of the role of BIOCHIP mosaic based indirect immunofluorescence and ELISA BP 180 and BP 230 autoantibodies in the diagnosis of bullous pemphigoid patients[J]. Indian Dermatology Online Journal,2022,13(6).

[7] Wang Dao Wen,Ni Li,Jiang Hualiang. Answer for questions of repeated measurements of variance analysis and distribution test of data - Authors' reply.[J]. Frontiers of medicine,2022,16(4).