Research on Computer Forecast Model Using BP Neural Network and Pearson Correlation Coefficient

Ji Li¹, *, Baihan Wang² and Haorui Li³

¹Institute of statistics and applied mathematics, Anhui University of Finance & Economics, Bengbu, Anhui 233000
²College of Information and Electrical Engineering, China Agricultural University, Beijing 100000
³School of Software, Taiyuan University of Technology, Jinzhong, Shanxi, 030600

*Corresponding author: jili2018@aufe.edu.cn

Abstract. Driven by such special needs as "individuation, fashion and beauty", the production model of the new retail enterprise is gradually moving towards multiple varieties and small batches, which makes the variety of ornaments and toys in the retail stores in the shopping malls more dazzling, and at the same time, makes inventory management in the retail industry more difficult. How to give an accurate demand forecast based on the historical sales data with complex levels and various categories at the regional level, small category level and even the store skc (single monochrome) level is a problem that most new retail enterprises need to focus on and think about. First of all, we study the impact of various related factors on the sales volume of the target skc during the four holidays of National Day, Double Eleven, Double Twelve and New Year's Day in 2018. We first reviewed the literature and data to identify the influencing factors. We believe that factors such as product characteristics, inventory information and holidays will affect the sales volume. Secondly, we test the correlation coefficient of Pearson and Spearman on the sales volume and each factor, and obtain the correlation strength between each factor and the price, we think that those with strong correlation have a greater impact on the sales volume. Finally, we get the specific impact of each factor on the sales volume through regression fitting between each relevant factor and the sales volume. Secondly, we need to forecast the sales volume of a given region's target subcategory every month for three months, give the average absolute percentage error between each and the forecast value, and forecast the top ten subcategories of sales volume at the same time, first, we analyze the data and adopt the neural network model to forecast, give the monthly forecast data, and consider the impact of relevant factors, so that the results are more accurate. Finally, the monthly forecast value of MAPE is given.

Keywords: Pearson correlation coefficient, Spearman rank correlation coefficient, BP neural network model.
1. Introduction
In recent years, China has been developing in various aspects, among which the consumption pattern in the consumption market has changed from "content-oriented" to "customer-oriented". [1] In the new retail industry, customer demand has also changed from "practicality" to a combination of "practicality" and "aesthetics". Demand drives production, and the production model of the new retail industry has gradually changed from single-variety [2], large-scale production to multi-variety, small-scale production to meet the needs of customers to a greater extent. However, at the same time, it causes difficulties in inventory management, additional expenses and complicated sales data. In order to better improve the management of enterprises and reduce the cost of sales, the new retail enterprises are facing the problem of how to solve the "accurate demand forecast". [3]

2. A Model of the Impact of Holidays on Sales Volume
This question requires to find out the impact of various related factors on the sales volume of the target skc during the four holidays of National Day, Double Eleven, Double Twelve and New Year's Day in 2018. [4] We first reviewed the literature and data to identify the influencing factors. We believe that factors such as product characteristics, inventory information and holidays will affect the sales volume. Secondly, we test the correlation coefficient of Pearson and Spearman on the sales volume and each factor, and obtain the correlation strength between each factor and the sales volume, we think that the strong correlation has a greater impact on the sales volume. Finally, we get the specific impact of each factor on the sales volume through regression fitting between each relevant factor and the sales volume.

After consulting the literature, [5] we have concluded three first-level factors that affect the sales volume, namely, sales product characteristics, inventory information and holidays. We further divided them into secondary factors and the results are shown in Table 1.

| category                  | serial number | factor           |
|---------------------------|---------------|------------------|
| Sales product characteristics | 01            | Year_id          |
|                           | 02            | Tiny_class_code  |
|                           | 03            | Tag_price        |
| Inventory information     | 04            | ie               |
| festival and holiday      | 05            | National Day     |
|                           | 06            | Double Eleven    |
|                           | 07            | Double Twelve    |
|                           | 08            | New Year's Day   |

Taking 2018 National Day, Double Eleven, Double Twelve and New Year's Day as examples, we use Pearson correlation coefficient and Spearman rank correlation coefficient method to test the relationship between different factors of sales product characteristics and inventory information and sales volume.

Pearson correlation coefficient is used to measure the strength of the linear correlation between two variables X and Y, and the formula is:

\[
\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} = \frac{E(X - \mu_X)(Y - \mu_Y)}{\sigma_x \sigma_y} \\
\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}
\]

Spearman correlation coefficient is a measure of the strength of two correlations, which can be obtained but has a wider application range than Pearson.
The calculation results of correlation coefficient between each factor and sales volume are shown in Table 2:

| factor          | Pearson correlation coefficient | Spearman correlation coefficient |
|-----------------|---------------------------------|----------------------------------|
| Year_id         | -0.075                          | -0.015                           |
| Tag_price       | 0.734                           | 0.599                            |
| Tiny_class_code | -0.653                          | -0.242                           |
| ie              | 0.835                           | 0.811                            |

According to the correlation coefficient $0.8 < |r| < 1$, we believe that the two variables have a strong correlation, when $0.6 < |r| < 0.8$, we believe that the two variables are strongly correlated, when $0.4 < |r| < 0.6$, it is said that the correlation between the two variables is moderate, $0.2 < |r| < 0.4$, the two variables are weakly correlated, when $0 < |r| < 0.2$, we generally think that the two variables do not want to be closed. Therefore, we eliminate factors with correlation coefficient $|r| < 0.5$.

We regressed different relevant factors with sales volume to obtain the relationship between sales volume and price and inventory as follows:

$$\ln Q = \ln P + a \times Q_i + C$$

3. Sales volume forecast model

In the neural network model, the neural network can be divided into feed forward neural network, feedback neural network and self-organizing magic neural network according to the distinction of the interconnection modes of the neural networks.

This topic adopts the BP neural network, because in the neural network model, the neural network model with feedback has stronger revisability, in order to improve the accuracy of this model, this topic selects the feed forward neural network as the model construction foundation, and its structure diagram is shown in Figure 1:

![BP Neural Network](image)

**Figure 1.** BP Neural Network.

Next, the model is solved based on genetic algorithm:

Step 1: Calculation of Neuron Input and Output.

The data of each input layer and each output layer are defined separately, after calculating the inputs and outputs of neurons of each layer, it can be concluded that:

(1) The implicit layer input vector is:
$$h_i(k) = \sum_{i=0}^{7} w_{hi} x_i(k) \quad h = 1,2,\ldots,13$$

(2) The implicit layer output vector is:
$$h_o(k) = f(h_i(k)) \quad h = 1,2,\ldots,13$$

(3) The input vector of the output layer is:
$$y_i_o(k) = \sum_{h=0}^{13} w_{oh} h_o(k) \quad o = 1$$

(4) The output vector of the output layer is:
$$y_o_o(k) = f(y_i_o(k)) \quad o = 1$$

Step 2: Calculation of Partial Derivative of Error Function to Neurons in Output Layer.
After obtaining the input and output function, the partial derivative of the error function to each neuron in the output layer needs to be obtained through mathematical operation, the specific solution is:

$$\frac{\partial e}{\partial w_{oh}} = \frac{\partial e}{\partial y_i_o} \frac{\partial y_i_o}{\partial w_{oh}}$$

Among them, the partial derivative can be known through calculation:

$$\frac{\partial e}{\partial y_i_o} = \frac{1}{2} \sum_{o=1}^{1}(d_o(k) - y_o_o(k))^2$$

$$\frac{\partial y_i_o}{\partial w_{oh}} = - (d_o(k) - y_o_o(k)) y_o_o(k)$$

$$\frac{\partial y_i_o}{\partial h_i} = \frac{\partial \sum_{h=0}^{7} w_{oh} h_o(k)}{\partial h_i} = h_o(k)$$

Step 3: Calculation of Partial Derivative of Error Function to Neurons in Hidden Layer.
When calculating the partial derivative of the error function to each neuron in the hidden layer, the connection weights from the hidden layer to the output layer, the function of the output layer and the function of the hidden layer need to be used for calculation and solution, the specific solution calculation is as follows:

$$\frac{\partial e}{\partial w_{oh}} = \frac{\partial e}{\partial y_i_o} \frac{\partial y_i_o}{\partial w_{oh}} = - \delta_o(k) h_o(k)$$

$$\frac{\partial e}{\partial h_i} = \frac{\partial \sum_{h=0}^{7} w_{hi} x_i(k)}{\partial h_i} = x_i(k)$$

$$\frac{\partial e}{\partial h_i} = \frac{\partial \sum_{o=1}^{1}(d_o(k) - f(\sum_{h=0}^{7} w_{ho} h_o(k))^2)}{\partial h_o(k)} \frac{\partial h_o(k)}{\partial h_i}$$

Step 4: using the output layer to correct the connection weights.
By using the output of each neuron to modify the connection weight, the specific calculation is as follows:

$$\Delta w_{oh}(k) = - \mu \frac{\partial e}{\partial w_{oh}} = \mu \delta_o(k) h_o(k)$$

$$w_{oh}^{N+1} = w_{oh}^{N} + \mu \delta_o(k) h_o(k)$$

Step 5: Use the input layer to correct the connection weights.
By using each neuron input to modify the connection weights, the specific calculation is as follows:

$$\Delta w_{hi}(k) = - \mu \frac{\partial e}{\partial w_{hi}} = \delta_h(k) x_i(k)$$

$$w_{hi}^{N+1} = w_{hi}^{N} + \mu \delta_h(k) x_i(k)$$
Step 6: calculation of global error.

\[ E = \frac{1}{2m} \sum_{k=1}^{m} \sum_{o=1}^{1} (d_o(k) - y_o(k))^2 \]

Step 7: Judgement.

In the judgment, it is necessary to compare the precision with the maximum number of times budgeted, when the precision and the maximum number of times budgeted do not meet the requirements, the cycle of the above steps is required until the requirements are met.

The neural network regression graph obtained from this topic is shown in Figure 2 below:

![Figure 2. Expression Image.](image)

The weights and thresholds obtained at this time are the relationship values between the levels under the current accuracy of this neural network model, through the weights and thresholds between the levels, the remaining 20% of the data can be predicted, and the accuracy of this model can be tested.

In order to calculate the MAPE value for each field in a given area, we need to use the following formula provided in the question:

\[
MAPE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n * y_i} = \sum_{i=1}^{n} \frac{1}{n} \frac{|y_i - \hat{y}_i|}{y_i}
\]

\[
APE = \frac{|y_i - \hat{y}_i|}{y_i}
\]

Where \( y_i \) represents the true value, \( \hat{y}_i \) represents the predicted value, APE represents the percentage error, and \( n \) represents the number of indicator sets. Finally, the three-month MAPE forecast values are 5.4%, 6.1%, and 4.6% respectively.
4. Conclusion
In this paper, the simulation and actual data are combined, the established model is close to the reality and has strong universality, the specific impact of various factors on sales volume can be obtained.

The new retail industry is different from the previous retail industry. People tend to pursue fashion in their demand for goods, focusing on "personality and beauty", generally, market participants in the new retail industry rely on their experience in predicting sales volume of certain small categories, or on their knowledge of production and transportation capacity in a small range, their tracking of futures, and their judgment of fashion. However, due to the one-sidedness of the information obtained, market participants cannot accurately judge the popular trend of a certain category, so accurately predicting the sales trend of a small category can enable enterprises to make preparations in advance, reduce production costs of enterprises and reduce operating risks.

The new retail industry has a high standard in the demand for diversified products, which is characterized by a large variety but small batch, which is different from the previous inventory management in the retail industry. The new retail industry requires more refinement and more specific classification of products, speeding up the reform and operation of inventory management. In this way, the inventory can be adjusted to make the products more targeted for purchase and shipment, actively adapt to the changes of the times and consumer demand, and improve the work efficiency.

Accurately predicting the demand for new retail products will also help the relevant government agencies to formulate policies. In the past, some of the legal terms or trading volume of the retail industry needed to be adjusted to the market conditions. This not only promotes the formulation of policies and protects the interests of manufacturers, but also brings more rights and interests to consumers and helps the favorable development of the new retail industry.

Accurate demand forecast for new retail products can help consumers and other groups to make trend judgment, especially for some groups in the design industry, which can find business opportunities or obtain inspiration for creation, it is a kind of survey and forecast for the current fashion trend. Open, transparent and effective market transactions are conducive to building trust between consumers and businesses and promoting the development and reform of the new retail industry.

References
[1] phase peak. The Theoretical Framework and Research Paradigm of "New Retail" [J], China's Circulation Economy, 2018 (1): 3-11.
[2] Zhao Shumei, Xu Xiaohong. "New Retail" Meaning, Model and Development Path [J], China's Circulation Economy, 2017 (05): 14-22.
[3] Du Ruiyun, Jiang Kan, New Retail: Connotation, Development Motivation and Key Issues [J], Price Theory and Practice, 2017, 000 (002): 139-141.
[4] Zhao Binyu, Research on the development path of logistics industry in the new retail era [J], Business Story, 2018.
[5] Yang Hui, New Retail Circle Theory: A New Exploration of Retail Format Development Theory [J], China's Circulation Economy, 2002 (06): 57-59.