Automatic Pricing and Replenishment Strategy for Vegetable Products Based on XGBoost and LSTM Prediction Models

Xiaohe Guo †, Guoli Zhao †

Education College, Yunnan University of Business Management, Kunming, China

† These authors also contributed equally to this work

Abstract. Due to the short shelf life and imbalance between supply and demand of vegetable commodities, it is important to develop reasonable replenishment and pricing strategies. This paper takes the measured data of a superstore as an example, and develops a reasonable decision-making program by studying the relationship between the information of each commodity in the vegetable category, the breakdown of the flow, the wholesale price and the recent attrition rate. First, the distribution pattern and interrelationship of the sales volume of each category and single product of vegetables are studied, and the distribution, difference and correlation between them are analyzed. Then, XGBoost is used to predict the sales volume and then LSTM is used to predict the wholesale price, to get the relationship between the total sales volume of vegetable categories and the cost-plus pricing, and for the total daily replenishment and pricing strategy in the coming week, a planning model is set up to solve the problem by using genetic algorithm. Finally, in order to maximize the revenue of the superstore, a goal planning model is established to take the sales unit price and sales volume as variables, and mathematical calculation methods are used to determine the daily sales volume and sales unit price and the maximum revenue, so as to determine the replenishment volume of a single product and the optimal pricing strategy. The proposed model fits the actual needs, is practical, efficient in calculation, and can effectively solve the proposed problem.

Keywords: K-W Test, XGBoost Model, LSTM Prediction Model, Genetic Algorithm, Goal Planning.

1. Introduction

In modern fresh food superstores, vegetables have a short freshness period and their quality will gradually deteriorate with the increase in sales time. In order to ensure that customers can buy fresh vegetables, supermarkets usually replenish their stocks every day according to the historical sales and demand of each commodity. Vegetable products sold in the superstore shorter freshness period and shipping damage, poor quality of the goods to be sold at a discount, as well as vegetable pricing using the "cost-plus pricing" method, which is based on the market demand for replenishment decisions and pricing decisions [1].

In this context, the thesis analyzes the automatic pricing and replenishment strategy of vegetable commodities by taking the measured data of a superstore as an example. First of all, there may be a certain relationship between different categories or different single items of vegetable goods, based on the sales volume of each category and single item of vegetables to study the distribution pattern as well as the interrelationship. Then, based on the superstore to do replenishment plan by category, explain the relationship between the total sales volume of each vegetable category and the cost-plus pricing, and develop the total daily replenishment volume of each vegetable category in the coming week and pricing strategy, so as to make the superstore maximize the revenue. Finally, within the limited sales space of vegetable items, the superstore would like to further develop a replenishment plan for individual items, give the replenishment volume and pricing strategy for individual items, and make the superstore maximize its profitability under the premise of trying to satisfy the market's demand for vegetable items in each category.

2. Sales Volume Distribution Pattern and Correlation Analysis

The distribution of sales volume of different categories and individual items is analyzed. The sales trend of vegetable categories plotted through Excel is shown in Fig. 1. It can be seen that the sales

volume of aquatic roots and tubers, foliage and edibles are relatively high throughout the timeframe, the sales volume of cauliflower and eggplant is low for most of the timeframe but rises in some specific timeframes, and the sales volume of chili peppers is stable for most of the timeframe but there are some small rises and declines.

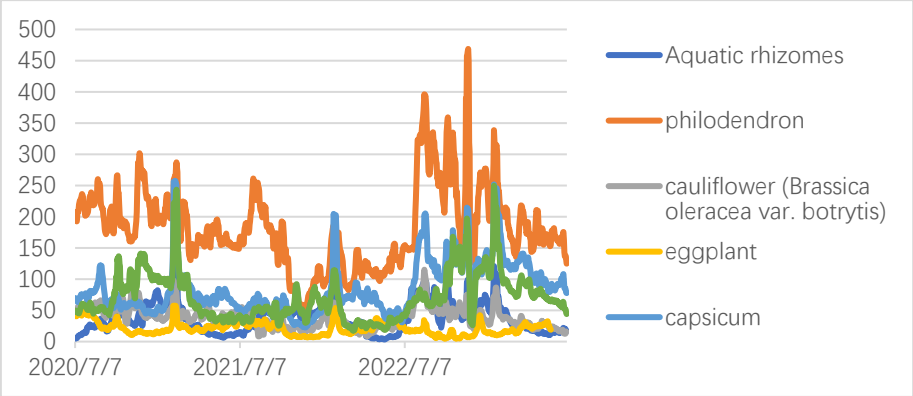

Figure 1. Distribution of sales volume by category

After the categories and the distribution of sales of each individual product into a bar chart as shown in Fig. 2, due to the excessive number of individual categories, so only the top ten sales of individual products were selected.

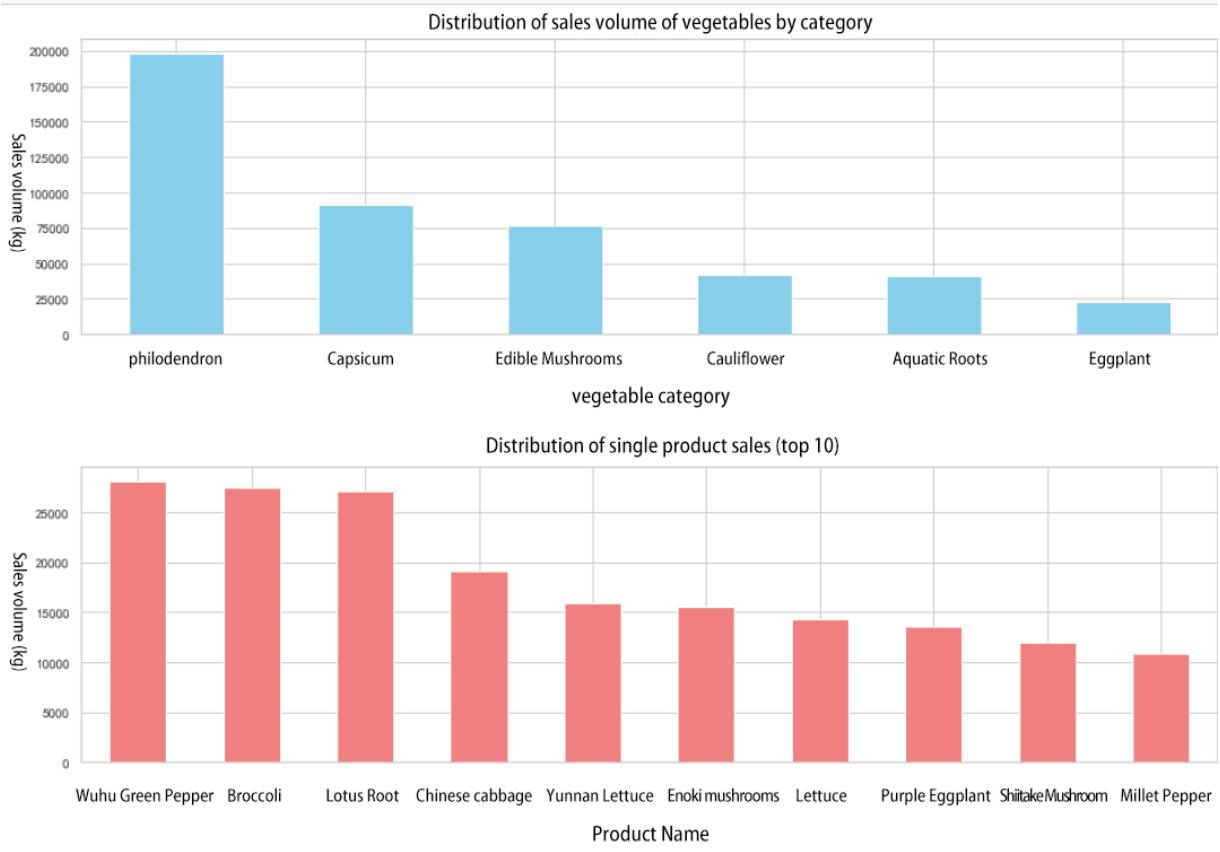

Figure 2. Histogram of sales by category and by individual product (top 10)

For the distribution of sales of vegetables by category, you can see that the highest sales of flowers and leaves, far more than other categories, eggplant sales are the lowest sales of all categories.

As for the distribution of sales volume of individual products, it can be seen that "Wuhu Green Pepper (1)" has the highest sales volume, followed closely by "Broccoli" and "Net Lotus Root (1)", and among the top 10 products, the sales volume varies, but the difference is not particularly large.

Among the top 10 items, there are some differences in sales volume, but the difference is not particularly large.

Differences between categories

The variance between the categories was analyzed. The variance chi-square test [2] was performed using SPSS and it was found that the p-value was less than 0.05, hence the variance was not chi-square and did not show normal distribution. The results are shown in Table 1.

Table 1. ANOVA Test Results

		Levin statistics	Degree of freedom 1	Degrees of freedom 2	significance
sales volume	Based on average values	269.976	5	6288	.000
	Based on median	240.476	5	6288	.000
	Based on median with adjusted degrees of freedom	240.476	5	3322.324	.000
	Based on clipped mean	254.755	5	6288	.000

It was therefore decided to group the different categories. Based on the sales volume to analyze whether there is any difference within, if the overall show a degree of difference, can be analyzed for the category between two.

The overall degree of difference between the different categories was analyzed using the Kruskal-Walli's test in SPSS and the results are shown in Table 2.

Table 2. Summary of hypothesis testing

	original hypothesis	inspect	significance	strategic decision
1	The distribution of sales is the same in the categories of the category.	Independent Sample Crucial-Wallis Test	.000	Reject the original hypothesis.

The result of Kruskal-Walli's test shows that the p-value of the statistic is 0.000, so we can reject the original hypothesis that there is a statistically significant difference between the sales volume of different categories. The significance between two of the specific categories is shown in Table 3.

This illustrates the fact that the sales trend of each individual item can be affected by a number of factors that lead to a high degree of uncertainty in daily sales, and in order to test the significance of the Kruskal-Wallis, it was chosen to test it using a chi-square distribution.

$$\chi^2 = k \times (n-1) \times W \tag{1}$$

Here k refers to the number of different sales dates and n refers to the number of individual items.

For the consistency test for each category, we can get the following χ^2 values and p values.

The results of the consistency test are shown in Table 3. From the results, so the p-value of all categories is equal to 1, which means that we do not have enough evidence to reject the original hypothesis that the ordering of the sales volume of each single product on different dates is random and there is no significant consistency. Therefore, we cannot prove that there is significant consistency in the ordering of sales volume of individual items under different categories on different dates. That is, the sales pattern of the vegetable market may be influenced by a number of unpredictable factors that make the sales pattern lack consistency from day to day.

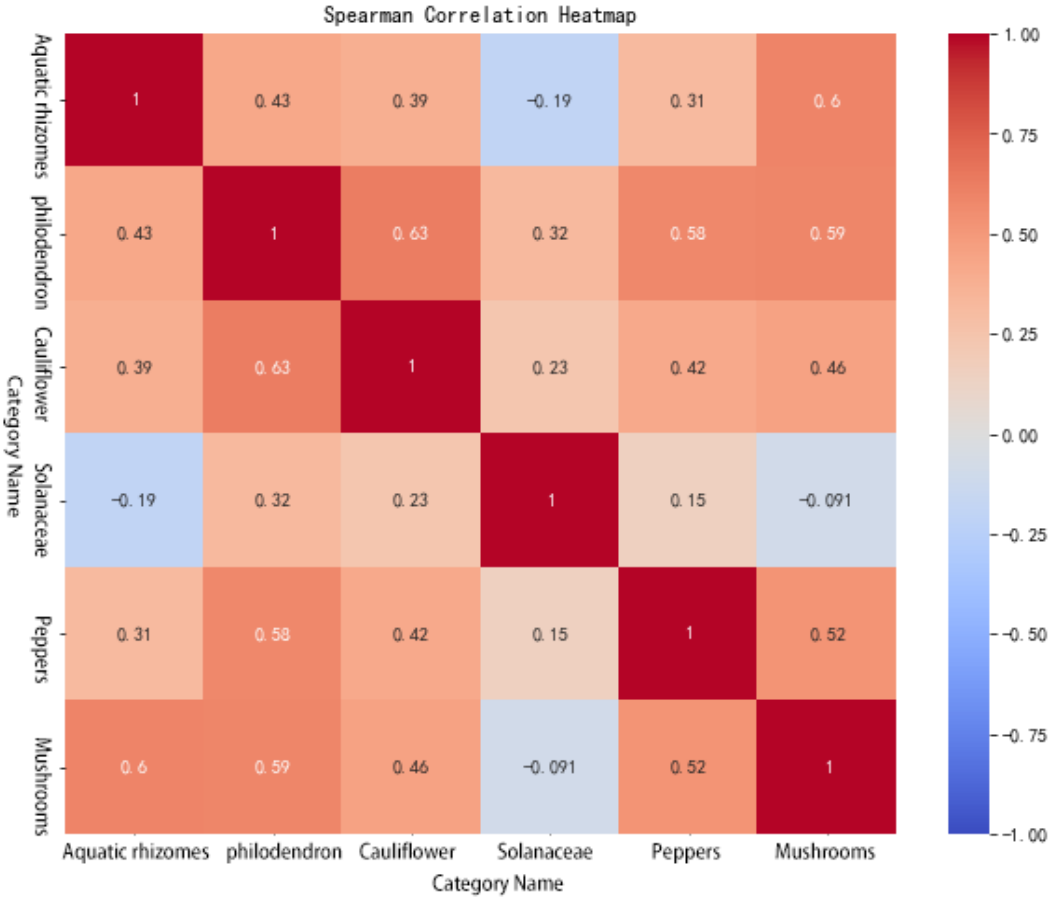

Figure 3. Heat map of Spearman's correlation coefficient

The correlation between the categories was analyzed. The data [3] was imported into SPSSPRO and Spearman's correlation test was carried out on the basis of categories and the results are shown in Fig. 3.

From the heat map of Spearman's correlation coefficient, we can conclude that.

Moderate to strong positive correlations existed between aquatic rhizomes and phloem, aubergine, and chili; phloem showed high positive correlations with phloem and aubergine; phloem showed significant positive correlations with chili; and positive correlations between aubergine and chili were strong.

Overall, the correlation between most of the variables is positive, which means that in most cases the increase or decrease in sales volume for these vegetable categories is synchronized.

For each individual product, calculating the Spearman correlation between the individual products can result in a strong negative correlation for black porcini mushrooms (box)-seven-colored peppers (servings), a strong positive correlation for dragon's tooth vegetables-yellow heart vegetables (2), a strong negative correlation for dragon's tooth vegetables-black porcini mushrooms, and a strong positive correlation for black porcini mushrooms-black porcini mushrooms (box), which are shown in the supporting materials due to the large amount of data.

The variability between the individual items was analyzed. The sales volume was corresponded and later imported into SPSSPRO to calculate the Kendall's Tau coefficient and p-value between the single items overall, which resulted in a Kendall's Tau coefficient of -0.06234936890311886 and a p-value of 0.0, so that it can be assumed that there is a statistically significant difference in sales volume between the different single items.

3. Vegetable Replenishment and Pricing Model for Each Category

The data were grouped by date of sale, calculated to obtain the mean values of unit sales price, wholesale price, and wastage rate and the summation value of sales volume, combined and calculated the average price difference per kilogram as a percentage of the wholesale price.

Sales forecasting based on XGBoost. XGBoost updates the model by minimizing the loss function, which is solved using gradient boosting. In each iteration, XGBoost computes the first and second order derivatives of the loss function equal to the predicted value of the previous model, and then constructs a new decision tree based on these derivatives.

$$obj(D) = \sum_{i=1}^{n} l(y_i . \hat{y}_i^t) + \sum_{i=1}^{t} \Omega(f_i) \tag{2}$$

Among them:

y_i is the true sales volume of the ith sample.

$\hat{y}_i^t$ is the pre-sale sales volume of the ith sample after the tth iteration.

l is the loss function a common choice is the squared loss function

$$l(y, \hat{y}) = (y - \hat{y})^2 \tag{3}$$

Ω is a regularization term to avoid overfitting.

The predictive model is represented after the seventh iteration:

$$\hat{y}_i^t = \sum_{j=1}^{t} f_j(x_i) \tag{4}$$

Among them:

$xi = (x_{1i}, x_{2i})$ is the feature vector of the ith sample, where x_{1i} is the selling unit price, x_{2i} is the wholesale price, and f_j is the jth decision tree.

Since we want to analyze each category, we choose edible mushrooms as an example first, in order to explore the relationship between sales unit price, wholesale price, and sales volume, we take sales unit price and wholesale price as independent variables, and sales volume as dependent variable, and choose XGBoost model for fitting, and the model parameters use the default constants, so that we can predict the sales volume of the same day according to the sales unit price and the wholesale price, and since XGBoost can't be used like the traditional model to get a definite equation, so the model is evaluated by testing the data prediction accuracy, and its model evaluation is shown in Table 3.

Table 3. Evaluation of XGBoost model for edible mushrooms

Classification name	edible mushroom
Mean absolute percentage error	16.972
r2_score	0.939006393

Similarly, the model evaluation for all categories is obtained as shown in Table 4.

Table 4. XGBoost model evaluation

Classification name	capsicum	philodendron	Aquatic rhizomes	edible mushroom	cauliflower (Brassica oleracea var. botrytis)	eggplant
Mean absolute percentage error	14.546	12.018	29.903	16.972	20.867	19.543
r2_score	0.944069259	0.930800258	0.945941588	0.939006393	0.935911997	0.948313095

The model fit can be seen to be more excellent through r^2.

Wholesale price prediction based on LSTM. Wholesale price prediction using a Long Short-Term Memory (LSTM) network [4] is a typical time series problem.The core structure of the LSTM consists of three main gate structures: forgetting gates, input gates, and output gates, as well as a cellular state, which allow the LSTM to store, modify, and output information.

According to the structure of LSTM network, let W_f, W_i, W_c, W_o, b_f, b_i, b_c, b_o be the model with corner labels. The formula for each LSTM cell is:

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{5}$$

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{6}$$

$$C_t = \tanh(W_c[h_{t-1}, x_t] + b_o) \tag{7}$$

$$C_t = f_t C_{t-1} + i_t C_t \tag{8}$$

$$O_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{9}$$

$$h_t = O_t \tanh(C_t) \tag{10}$$

Style:

f_t --Sigmoid activation function

x_t --Current inputs to the Memorizer Section module

i_t --Input Threshold

C_t -- t Candidate values for the cell state at the moment

C_t -- *cell*

For the prediction task, lagged features (lag features) of historical wholesale prices need to be constructed as input. For example, lagged data from the past day, week or month can be created as new features. Simply put, this is to turn a single series of data into a regression data of $X -> Y$. As shown in the figure below, a step of 2 means 2 X (there are as many X 's as there are steps) and one Y (this will not change), which simply means that the data from days 1 and 2 will be used to predict day 3, and the data from days 2 and 3 will be used to predict day 4, and so on.

The formula for the lag characteristic is expressed as follows:

$$X_{t-n} = f(X_t, X_{t-1}, \cdots, X_{t-n+1}) \tag{11}$$

Where X_t denotes the eigenvalue at time t , $X_\{t-n\}$ denotes the eigenvalue at time $t-n$, and f denotes the generating function of lagged features. Through the construction of lagged features, the dynamic change information of the time series can be incorporated into the features to improve the prediction accuracy of the model.

The model uses a three-layer LSTM network structure for the initial training, but also in the introduction of the dropout strategy, and set it to 0.5, that is, each training randomly discard the general parameters, to further prevent the emergence of the phenomenon of overfitting, the prediction results are shown in Table 5.

Table 5. LSTM model evaluation

Classification name	capsicum	philodendron	Aquatic rhizomes	edible mushroom	cauliflower (Brassica oleracea var. botrytis)	eggplant
MSE	1.5826	0.1066	0.6251	0.4443	1.0687	0.4007
RMSE	1.258	0.3266	0.7907	0.6666	1.0338	0.633
MAE	0.9416	0.2474	0.5166	0.5746	0.5055	0.5053
R2	0.6438	0.7887	0.9124	0.8519	0.8431	0.9054

As can be seen from the above figure, the aquatic rhizomes category has the highest r^2 value: 0.9124, which means that the model is able to explain 91.24% of the variance in this category, and the r^2 values for all categories exceed 0.64, which indicates that the model has a relatively good predictive ability for all categories.

In summary, the model performs particularly well in predicting floral and foliage categories and aquatic rhizomes, but for chili peppers, the model effect performs more generally, and in general, the accuracy of wholesale price prediction based on LSTM is more excellent.

When constructing the superstore revenue solution in the objective planning model, the constant term attrition rate is needed as a parameter to construct the planning model and solve it with genetic algorithm. Objective function: (sales volume × ((100 - attrition rate)/100)) × (sales unit price - wholesale price), that is:

$$\max = \sum_{i=1}^{7} y_i \times \left(\frac{100-l}{100} \right) \times \left(x_i - h_i \right) \tag{12}$$

Where y_i is the daily sales volume, which can be predicted by selling unit price and wholesale price using XGBoost, h_i represents the wholesale price, which can be predicted by LSTM model, and l represents the attrition rate, which uses the historical average attrition rate.

Decision variable: x_i indicates unit sales price

Constraint: the selling unit price is greater than the wholesale price, i.e.

$$x_i > h_i, i = 1, 2, \cdots, 7 \tag{13}$$

Thereafter the total daily replenishment and pricing strategies for all vegetable categories for the coming week are obtained by solving the above planning problem using iterative implementation of dynamic programming using genetic algorithm [5].

By solving the above five steps, we can get the sales unit price and sales quantity corresponding to the maximum superstore revenue of the next 7 days for each category, because the data of the topic and for giving its warehouse storage, here the sales quantity as the daily replenishment, sales unit price as the pricing strategy of the category, the six categories, and finally we can get the total revenue of 41,398.93 yuan for 7 days.

4. Vegetable Single Product Replenishment Strategy Development

Sales volume prediction based on XGBoost. For the prediction of sales volume, the sales unit price and wholesale price are used as independent variables, and the sales volume is used as the dependent variable to construct the XGBoost model, and the model adopts the default constants, and the results of evaluating the model to find the mean value are shown in Table 6. It can be seen that the fitting effect of the model is more excellent.

Table 6. Model evaluation mean results

Mean absolute percentage error	33.589434
r2_score	0.835926

Wholesale price prediction based on LSTM. The LSTM prediction method was extended to the data and the average model evaluation is shown in Table 7.

Table 7. LSTM average model evaluation

MSE	2.812097
RMSE	0.873253
MAE	0.552019
R2	0.642807

It can be seen that the average single item R2 are over 0.6428, which means that the model is able to explain 64.28% of the variance in this type of data, indicating that the model has good predictive power for all data.

Constructing a goal plan for maximizing superstore revenues. Objective function:

$$\max w = y \times \left(\frac{100 - l}{100} \right) \times \left(x - h \right) \tag{14}$$

Where y is the sales volume of the goods, which is predicted by the XGBoost model based on the sales unit price x and the wholesale price h , and l is parameterized by using the historical average attrition rate, which can be found as l is 9.807879.

Constraint: the selling unit price is higher than the wholesale price, i.e. $x > h$.

The number of sales must be greater than 2.5, i.e. $\sum_{i=1}^{33} y_i > 2.5$.

Decision Variables: the main decision variable in the model is the unit sales price of each individual item, denoted as x_i . This is the output of the model and represents the recommended sales price on day i.

The input parameters of the model include the wholesale price (h) and the attrition rate (l) on July 1st. Separately, there is a key input data, the daily sales quantity predicted based on the sales unit price and wholesale price, which is predicted by the advanced XGBoost machine learning model [6].

By solving the above steps, we can get the corresponding sales unit price and sales quantity for each single product when the superstore revenue is maximum on July 1. Since the data of the topic is not to give its storage capacity, the sales quantity here is used as the daily replenishment quantity, and the sales unit price is used as the pricing strategy of the category, which is solved for the 33 categories, and we can finally get the total revenue of 2949.485669 yuan on July 1st.

5. Conclusion

In this paper, we study the relationship between the information of each item in the vegetable category, the flow details, the wholesale price, and the recent wastage rate to develop a rational decision-making program. Firstly, the distribution pattern and interrelationship of the sales volume of each category and single product of vegetables are studied, and the distribution, difference and correlation between them are analyzed. Then, XGBoost is used to predict the sales volume and then LSTM is used to predict the wholesale price, to get the relationship between the total sales volume of vegetable categories and the cost-plus pricing, and for the total daily replenishment and pricing strategy in the coming week, a planning model is set up to solve the problem by using genetic algorithm. Finally, in order to maximize the revenue of the superstore, a goal planning model is established to take the sales unit price and sales volume as variables, and mathematical calculations are used to determine the daily sales volume and sales unit price and the maximum revenue, so as to determine the replenishment volume of a single product and the optimal pricing strategy. The proposed model is highly practical and computationally efficient.

References

[1] XU Yiting, MU Yueying, ZHAO Yousen. Analysis of vegetable price changes and influencing factors in Beijing. China Vegetable, 2017 (11): 62 - 69.

[2] Li Jiansheng, Li Jianghong, Liu Xiaoning et al. Application of Kendall's W analysis method in medical data processing and its implementation in SPSS. Modern Preventive Medicine, 2008 (01): 33+42.

[3] Wang Xiaoyan, Li America. Introduction to rank correlation coefficient and Spearman's rank correlation coefficient. Journal of Guangdong Light Industry Vocational and Technical College, 2006 (04): 26 - 27.

[4] Wang X, Wu J, Liu C et al. Fault time series prediction based on LSTM recurrent neural network. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44 (04): 772 - 784.

[5] Jigen-Lin. A review of genetic algorithm research. Computer Applications and Software, 2004 (02): 69-73.

[6] Lu Weifeng. A review of research on convergence analysis of PSO algorithm [J]. Computer Knowledge and Technology, 2016, 12 (29): 181 - 182+190.