Optimization modeling of vegetable pricing replenishment strategy based on machine learning

Shuchang Zhou*, Yan Dai

College of Mathematics and Statistics, Northwest Normal University, Lanzhou, 730071, China

* Corresponding Author Email: 15586482698@163.com

Abstract. Fresh supermarkets are affected by commodity attributes and transaction time, and usually need to be replenished every day. It is of great significance to study the distribution law of sales volume of each category and single product, and to give replenishment and pricing strategies for the commercial supermarket to maximize its revenue. Based on the machine learning method, this paper analyzes the sales flow records of a supermarket in the past three years, obtains the sales rules of each category, and establishes a dynamic optimization model for subsequent pricing and replenishment decisions. The steps are as follows: Firstly, data preprocessing is carried out. Secondly, the kernel density estimation method is used to analyze the distribution law of the total sales volume of each category. With the help of correlation coefficient Screening feature selection method, neural network method and regression tree method, the first 33 single vegetable variables that have an important impact on the total sales volume of vegetables and meet the given constraints are comprehensively selected. Based on these 33 variables, the revenue function of single product sales, single product pricing, single product replenishment and single product wholesale price is constructed, so that the optimization problem of the revenue function is transformed into a constrained optimization problem of a quadratic function of the cost markup rate, and the specific pricing and replenishment strategy on July 1,2023 is obtained. The model established in this study can provide some new solutions to the vegetable pricing and replenishment strategy of supermarket.

Keywords: Neural network, regression tree method, dynamic prediction, Screen algorithm.

1. Introduction

As a short life cycle product, vegetable products have the characteristics of seasonality and vulnerability. These characteristics aggravate the complexity of vegetables from transportation to sales. At the same time, in the process of circulation, their quality and quantity will decay with time, seriously damaging the profits of the purchaser. In the fresh supermarket, most of the varieties can not be sold the next day if they are not sold on the same day. Therefore, supermarkets usually make daily replenishments based on the historical sales and demand of each item. Due to the large variety of vegetables sold by the supermarket, the origin is different, and the purchase transaction time of vegetables is usually from 3:00 to 4:00 in the morning. Therefore, merchants must make replenishment decisions for each vegetable category on the same day without knowing the specific single product and purchase price.

Zhang Jinlong.et al. [1] established a joint decision model of pricing and dynamic batch replenishment for perishable new products. The joint decision of pricing and replenishment is necessary, and the optimal price has a U-shaped relationship with product diffusion rate and repeated purchase rate. Considering the time function of product decay intensity, Lu Jing [2] studied the inventory control and dynamic pricing strategy of fresh agricultural products, and constructed a time-varying decay inventory dynamic equation and system profit maximization model. Zhou Hai Jie [3] used the dynamic programming method, the optimal inventory and pricing of two cycles under different quality conditions are obtained by reverse solution. Miranda analyzed the variables affecting the determination of the sale price of vegetable which is constant over time in a supermarket qualitatively and quantitively[4].

Most of the existing literature constructs theoretical models and lacks the support of actual data. This paper obtains the detailed sales data of a supermarket in the past three years, which is used as the basis for pricing and replenishment strategies. Using the feature selection method to select a

certain number of single items that can be sold, and then constructing the optimization model, can reduce the scale of the problem, and make reasonable decisions on the relationship between the income, pricing, and sales volume of the fresh supermarket for graduate students. Reduce waste and maximize revenue under the premise of meeting customer needs.

2. Materials and methods

2.1. Data acquisition and preprocessing

The data in this paper come from actual research and open-source websites, including the commodity information of 6 vegetable categories distributed by a supermarket. The relevant data of the sales flow details and wholesale prices of each commodity from July 1,2020 to June 30,2023; recent loss rate data of each commodity.

When preprocessing the data, considering that there are few vegetable returns, it is removed as abnormal data. Considering the seasonality of vegetable items, the vacancy data of non-sales items are filled with a fixed value of zero.

2.2. Method introduction

BP neural network is back propagating, mainly composed of three parts: input layer, middle layer and output layer. The number of nodes in the input and output layers is relatively easy to determine, but the determination of the number of nodes in the hidden layer is a very important and complex problem.

2.2.1 Pearson correlation analysis[5][6]

Pearson correlation analysis is a method to measure the strength of the linear relationship between two variables. The total Pearson correlation coefficient between two variables is defined as the quotient of the product of covariance and standard deviation between two variables, which is defined as follows:

The covariance and standard deviation of the sample are estimated, and the Pearson correlation coefficient of the sample is obtained. The commonly used English lowercase letter r is represented. The expression of r is as follows:

Where $\overline{X}$ and $\overline{Y}$ denote the sample mean of the two, respectively.

2.2.2 BP neural network[7]

BP network, also known as back propagation neural network, first needs to train a part of the sample data, and constantly adjust the threshold and weight in the training process to approximate the expected output. The forward propagation of input information and the backward propagation of output error constitute the information cycle of BP network. It is divided into two processes: working signal forward transfer sub-process and error signal reverse transfer sub-process. The BP network is composed of input layer, hidden layer and output layer. The hidden layer can have one or more layers. In general, the three-layer BP network can complete any m-dimensional to n-dimensional mapping.

2.2.3 Regression trees[8][9]

The classification and regression tree (CART) model was proposed by Breiman.et al.in 1984, which is a widely used decision tree learning method. CART is also composed of feature selection, tree generation and pruning, which can be used for both classification and regression.

Suppose that X and Y are input and output variables, respectively, and Y is a continuous variable. Given a training data set $D=\left\{X_i,Y_i\right\}_{i=1}^{n}$, where $Xi=\left\{x_i^1,x_i^2,\cdots,x_i^d\right\}$ is a d-dimensional feature vector, a regression tree is generated for it.

When dividing the input space (i.e., the feature space), all values of all features in the current region are selected one by one at a time, and the optimal value (segmentation point) of one of the best

features is selected according to the square error minimization criterion. For example, by selecting the No. j feature and its value x^j for segmentation, two regions can be segmented:

$$R_1(j,s) = \{x_i | x_i^j \leqslant s\}, R_2(j,s) = \{x_i | x_i^j > s\} \tag{1}$$

According to the above proof, in each region, in order to minimize the square error, the mean value of all the data in the region is selected as the regional output value, that is, the output value of the regions $R_1(j,s)$ and $R_2(j,s)$ are

$$\hat{c}_1 = \frac{1}{|R_1(j,s)|} \sum_{x_i \in R_1(j,s)} y_i \quad , \quad \hat{c}_2 = \frac{1}{|R_2(j,s)|} \sum_{x_i \in R_2(j,s)} y_i \tag{2}$$

and the square error is :

$$\sum_{x_i \in R_1(j,s)} (y_i - \hat{c}_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - \hat{c}_2)^2 \tag{3}$$

Therefore, in order to find the optimal features and their optimal segmentation points, it is necessary to find all the features and all their segmentation points at each time of division, and select the features and their segmentation points that minimize the square error after division, that is, to solve:

$$\min_{j,s} \left[\sum_{x_i \in R_1(j,s)} (y_i - \hat{c}_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - \hat{c}_2)^2 \right] \tag{4}$$

After finding the optimal feature and the optimal segmentation point, the input space is divided into two regions in turn. Then, the above partitioning process is repeated for each region until the stopping condition is satisfied.

3. Establishment and Solution of Model

According to the actual situation, the following reasonable assumptions are made in the subsequent modeling of this paper: without considering inventory, discount sales at different times of the day, returns, transportation, rent and labor costs.

The symbols involved in this study are described in Table 1.

Table 1. Symbol description

symbol	description	unit
t	Day t	day
k	The k th vegetable category	nil
i	The i th vegetable product	nil
$L_t^{(k)}$	The income value of the kth category on day t	yuan
$P_{i,t}^{(k)}$	The fixed value of the i th item of the k th category on day t	yuan
$Q_{i,t}^{(k)}$	Sales volume of category k vegetables on day t of item i	kg
$C_{i,t}^{(k)}$	Wholesale cost price on day t of class k, item i	yuan
$S_{i,t}^{(k)}$	Replenishment quantity of category k vegetables on day t of item i	kg
L_t	The total income of all sales items on day t	yuan
w_i	The loss rate of item i	nil

3.1. Problem analysis

Based on the historical sales flow data of a supermarket from July 1,2020 to June 30,2023, this paper first analyzes the distribution law of sales volume of different categories of vegetable commodities in supermarkets, and mines the data to obtain the distribution density map of category sales volume. Considering the seasonality of vegetables, when making the replenishment decision on July 1, only items with sales data in the previous two weeks are considered as marketable items. Taking into account the effectiveness of the shelf space, select 33 items from the available items that contribute more to the sales of the respective categories. In order to make the supermarket vegetable categories complete, it is also necessary to ensure that the daily sales of the items in the six categories are roughly evenly distributed. With the help of correlation coefficient Screening feature selection method, neural network method and regression tree method, the single products with large contribution rate to the sales volume of each category are obtained. By debugging the threshold of correlation coefficient of feature selection, the importance of neural network normalization and the importance of regression tree, 33 single products can be selected based on these single products, and the revenue function of sales volume, pricing, replenishment volume and wholesale price of 33 single products can be constructed. According to the actual situation, the constraints are set reasonably, and the optimization problem of the revenue function is transformed into the constraint optimization problem of the quadratic function of the cost markup rate. According to the obtained optimal cost markup rate, the pricing and replenishment strategy of the replenishment goods on July 1 is given.

3.2. The distribution law and relationship of sales volume of each category

Using R software, the kernel density estimation diagram of the sales data of the six vegetable categories of flowers and leaves, cauliflower, aquatic rhizomes, eggplants, peppers and edible fungi is given, as shown in Figure 1. Figure 1 shows the distribution of sales of 6 kinds of vegetables.

Figure 1 shows that during the period from July 1, 2020 to June 30, 2023, the sales distribution of the six categories of cauliflower, eggplant, flower and leaf, edible fungi, peppers, and aquatic rhizomes generally meets the normal distribution law.

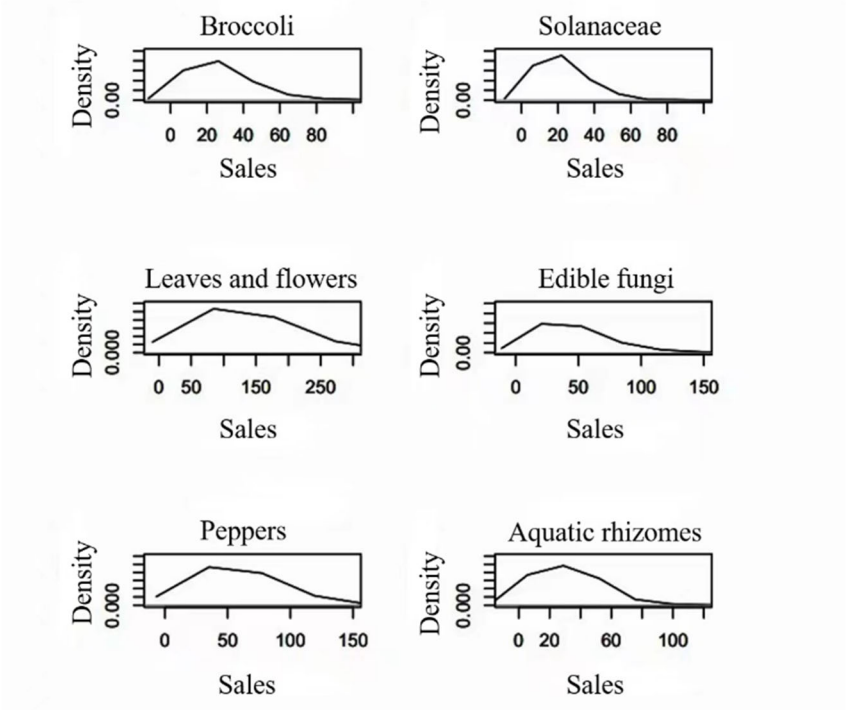

Figure 1. Kernel density estimation chart of total sales volume of each category

Next, the relationship between the various categories is studied, and the Pearson correlation analysis is given.

The Pearson correlation coefficient heat maps of 6 vegetable categories are shown in Figure 2.

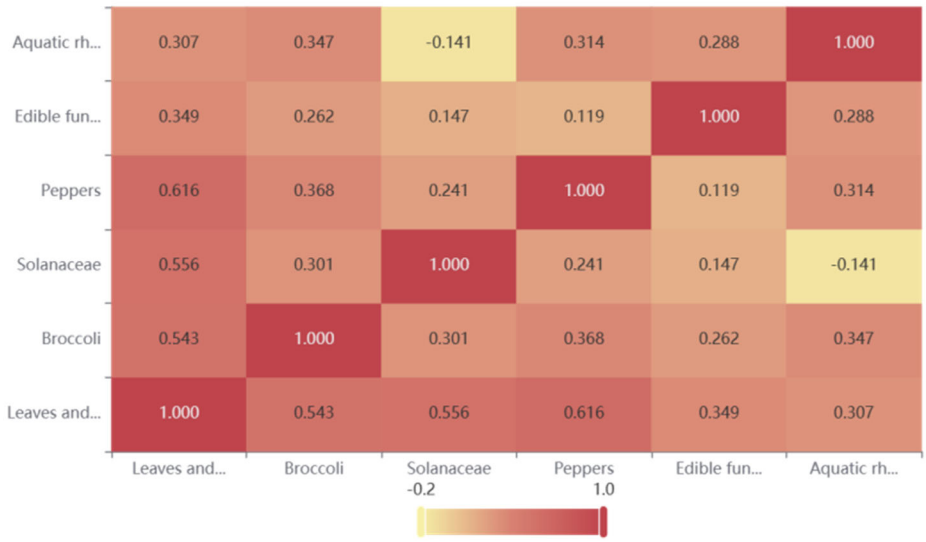

Figure 2. Pearson Correlation Coefficient between Sales Volumes of Various Categories

It can be seen from Figure 2 that all Pearson correlation coefficients are positive, indicating that there is a certain positive correlation between the sales volume of the six categories. The sales of edible fungi had a high positive correlation with peppers, mosaics, aquatic rhizomes and cauliflowers. In addition to the low correlation with eggplants, chili peppers have a high correlation with other categories. The correlation between eggplant and other categories was weak. In addition to the low correlation with eggplants, aquatic rhizomes have a high correlation with other categories. The correlation between flowers and leaves and cauliflower and eggplant was weak, and the correlation with other categories was high.

3.3. Filter out the available items for replenishment.

Firstly, 251 items should be screened for features. The data of the total daily sales of each category and the total sales of each item in each category from July 1,2020 to June 30,2023 were sorted out. Pearson correlation analysis, BP neural network model and regression tree model were carried out to obtain the contribution rate of each item to its sales in each category. By setting a threshold, the number of items obtained by synthesizing the three models is controlled between 27 and 33.

3.3.1 Feature scanning screening based on Pearson correlation coefficient.

Through calculation, and set the threshold value of 0.4, the selected various types of items are shown in Table 2 below, a total of 44.

Table 2. Vegetable single products selected based on Pearson correlation coefficient

Pearson correlation coefficient extraction, the threshold is 0.4	
Leaves and flowers	Sweet cabbage, yellow cabbage (2), flowering Chinese cabbage, Shanghai green, small cabbage, Yunnan lettuce, yellow cabbage (1), Yunnan lettuce, fast vegetable
Broccoli	broccoli、green stem scattered flowers, Zhi jiang green stem scattered flowers
Solanaceae	purple eggplant (2), green eggplant (1), long-line eggplant
Peppers	red pepper (1), millet pepper, red pepper, green pepper (1), green pepper, green pepper, Wuhu green pepper (1), colorful pepper (1), screw pepper, red lantern pepper (1), combined pepper series, lantern pepper (1)
Edible fungi	White Hypsizygus marmoreus(bag), Cordyceps flower (bag), Flammulina velutipes (1), Pleurotus eryngii (bag), Pleurotus ostreatus, Xi Xia mushroom (1), Fresh Auricularia (1), Pleurotus cornucopiae (bag), Flammulina velutipes (bag) (3), Pleurotus geesteranus, Pleurotus eryngii (bag), Cordyceps flower (box) (2), Tremella fuciformis (flower)
Aquatic rhizomes	lotus root (1), Hong hu lotus root (powder lotus root), water chestnut, Hong hu lotus root (crisp lotus root)

3.3.2 bp neural network model

Through the training of neural network, the normalized importance of all single product sales in each category to the total sales of this category can be obtained. The threshold is set to 33.1 %. The selected single products are shown in Table 3, a total of 51.

Table 3. Vegetable single products selected based on BP neural network

BP neural network normalized importance extraction, the threshold is 33.1 %.	
Leaves and flowers	Malan head, Shanghai green, spinach, Hong Shan brassica campestris lotus root assembled gift box, yellow cabbage (1), day lily, Artemisia argyi, Zhi jiang red cabbage moss (parts), Chinese cabbage, double ditch Chinese cabbage, noodles, rape moss, sugar cabbage, Chinese cabbage, milk cabbage (parts), mustard, spring vegetables, perilla (parts)
Broccoli	broccoli, Zhi jiang green stem scattered flowers, purple cabbage (2), green stem scattered flowers
Solanaceae	purple eggplant (2), purple eggplant (1), long-line eggplant, green eggplant (1)
Peppers	green pepper (1), red pepper, red lantern pepper (1), green pepper, screw pepper, green and red pepper combination (parts), green pepper (parts), colorful pepper (1), red pepper (parts)
Edible fungi	Xi Xia mushroom (1), crab mushroom (bag), fresh agaric (1), Pleurotus Eryngii (bag), mushroom hot pot package (portion), black Termitomyces, Pleurotus Eryngii (portion), Flammulina velutipes (1), Tricholoma matsutake (box), white Hypsizygus marmoreus (bag)
Aquatic rhizomes	Hong hu lotus root (crisp lotus root), water chestnut, net lotus root (1), Hong hu lotus root (powder lotus root), fresh lotus root, red lotus root

3.3.3 Regression tree

Through analysis, the importance of all single product sales in each category to the total sales of this category can be obtained. The threshold is set to 0.141, and the selected single products are shown in Table 4, a total of 38.

Table 4. Vegetable single products selected based on regression tree

Regression tree model importance extraction, the threshold is 0.141	
Leaves and flowers	Chinese cabbage, flowering Chinese cabbage, Yunnan lettuce, Yunnan lettuce, yellow cabbage (1), yellow cabbage (2), red coral (coarse leaves), Chinese cabbage (parts), field chrysanthemum (parts), sugar cabbage, small cabbage (parts), Shanghai green
Broccoli	broccoli, green stem scattered flowers, Zhi jiang green stem scattered flowers
Solanaceae	purple eggplant (2), long-line eggplant, green eggplant (1)
Peppers	millet pepper (portion), Wuhu green pepper (1), red lantern pepper (1), red lantern pepper (portion), colorful pepper (portion), green pepper, pickled pepper (boutique)
Edible fungi	Flammulina velutipes (1), White Hypsizygus marmoreus (bag), Flammulina velutipes (bag) (1), Pleurotus ostreatus, Crab mushroom (bag), Steak mushroom (box), Pleurotus cornucopiae (bag), Xi Xia mushroom (portion), Sparassis crispus (bag), Sea mushroom (bag) (1)
Aquatic rhizomes	lotus root (1), Hong hu lotus root (powder lotus root), water chestnut

The common single products selected by the above three methods are : sweet cabbage, Shanghai green, Chinese cabbage, yellow cabbage (1), broccoli, green stem scattered flowers, Zhi jiang green stem scattered flowers, purple eggplant (2), green eggplant (1), long-line eggplant, green pepper, colorful pepper (1), red lantern pepper (1), white jade mushroom (bag), golden mushroom (1), Xi Xia

mushroom (1), crab mushroom (bag), net lotus root (1), Hong hu lotus root (powder lotus root), water chestnut, Hong hu lotus root (crisp lotus root).

Through the expert method, the above three tables are used as expert scoring and screening, and the single items voted as 2 in the three experts are continued to be selected : yellow cabbage (2), cabbage, Yunnan lettuce, Yunnan lettuce, Hong hang pepper, Qing hang pepper (1), Wuhu green pepper (1), screw pepper, Pleurotus eryngii (bag), fresh fungus (1), Hypsizygus marmoreus (package), mushroom hot pot package (parts).

Through the above steps, a total of 33 items were selected, and the subsequent replenishment and pricing strategies were analyzed based on the 33 items.

3.3.4 Determine the pricing and replenishment quantity for the selected single product

According to the actual situation, the planning equation is established :

$$\eta_t = \sum_{i=1}^{n^*} (P_{i,t} Q_{i,k} - C_{i,t} S_{i,t}) = A_t \alpha^2 + B_t \alpha + D_t \tag{5}$$

There is a relationship :

$$\begin{cases} P_{i,t} = C_{i,t}(1+\alpha) \\ Q_{i,t} = \overline{\beta} P_{i,t} + \overline{r}_i \\ S_{i,t} = Q_{i,t}(1+\omega_i) \end{cases} \tag{6}$$

The constraints are :

$$\begin{cases} Q_{i,t}(1+\omega) \geqslant 2.5 \\ S_{i,t} \geqslant Z_{i0} \\ Z_{i0} = 2.5 \end{cases} \tag{7}$$

The final planning function is derived as follows :

$$S_{i,t} = (1+\omega_i)\overline{\beta}_i C_{i,t}\alpha + (1+\omega_i)\left(\overline{\beta}_i C_{i,t} + \overline{r}_i\right) \tag{8}$$

According to the selected 33 items and the final planning function, Table 5 gives the replenishment quantity and pricing strategy on July 1.

Table 5. Replenishment quantity and price on july1st

Category	Fixed value	Replenishment volume
Green stem scattered flowers	10.065	21.076
Broccoli	15.126	18.442
Zhi jiang green stem scattered flowers	11.302	17.14
Chinese cabbage	10.501	5.702
Yellow Cabbage (2)	8.925	8.083
Yellow cabbage (1)	6.815	2.5
Shanghai Qing	9.482	5.515
Sweet cabbage	10.628	8.342
Pakchoi	21.651	2.5
Yunnan lettuce	11.498	19.196
Yunnan lettuce	11.249	9.312
Red lantern pepper (1)	21.373	2.5
Red Hang pepper	29.362	2.5
Screw pepper	16.759	7.852
Colorful peppers (1)	27.167	2.5
Green pepper (1)	16.42	2.5
Green pepper	13.779	2.5
Wuhu green pepper (1)	12.401	29.149
Green eggplant (1)	11.528	2.5
Long line eggplant	16.308	3.768
Purple eggplant (2)	12.176	12.685
White jade mushroom (bag)	9.176	8.466
Pleurotus cornucopiae (package)	9.364	6.84
Flammulina velutipes (1)	11.01	7.417
Mushroom hot pot package (portion)	14.477	9.609
Xi Xia mushroom (1)	26.339	8.918
Fresh Auricularia (1)	15.886	3.611
Crab mushroom (bag)	5.418	14.088
Pleurotus eryngii (bag)	7.323	9.655
Water chestnut	8.453	7.413
Hong hu lotus root (crisp lotus root)	17.196	2.5
Hong hu lotus root (powder lotus root)	15.754	2.5
Net lotus root (1)	12.036	21.457

4. Discussion

If the inventory cost and transportation cost are taken into account[10], other factors such as supermarket management cost and rent are ignored, the above model can be further optimized to be more in line with the actual situation.

Let $X_{i,t}^{(k)}$ denote the remainder of the day t of the item i in class k ; let $Z_{i,t}^{(k)}$ denote the transportation cost on day t of the item i of the category k ; still using the 33 single products screened by the three models of feature scanning, neural network and decision tree, the predicted revenue function of each single product on day t can be redefined as :

$$\overline{\eta}_t = \sum_{i=1}^{n^*} [P_{i,t} Q_{i,t} - (C_{i,t} + Z_{i,t}) S_{i,t}] \tag{9}$$

While:

$$\begin{cases} P_{i,t} = C_{i,t}(1+\alpha) \\ Q_{i,t} = \hat{\beta}_i P_{i,t} + \hat{r}_i \\ X_{i,t-1} + S_{i,t} = Q_{i,t}(1+\omega_i) \end{cases} \tag{10}$$

And satisfy the constraint conditions:

$$S_{i,t} = Q_{i,t}(1+\omega_i) - X_{i,t} \geqslant Z_{i0}, Z_{i0} = 2.5 \tag{11}$$

The quadratic equation with $\overline{\eta}_t$ as α can be obtained, so as to optimize the solution.

It can be seen from this that if the specific values of transportation cost and inventory cost can be given, the model established will be more in line with the actual situation.

5. Conclusions

In this paper, machine learning method, kernel density method and chart method are used to analyze and model. The kernel density estimation method is used to analyze the distribution law of the total sales volume of each category. With the help of correlation coefficient Screening feature selection method, neural network method and regression tree method, 33 single-product vegetable variables that have the most important influence on the total sales volume of vegetables are comprehensively selected. Based on these 33 variables, the income function of single-product sales volume, single-product pricing, single-product replenishment volume and single-product wholesale price is constructed, and the optimization problem of the income function is transformed into a constrained optimization problem of a quadratic function of cost markup rate. Based on the obtained optimal cost markup rate, the pricing and replenishment strategy of the replenishment single product on July 1 is given. In addition, we further consider the inventory factors and transportation cost factors that affect the pricing and replenishment strategies of various vegetable categories and individual products into the income function, establish the corresponding optimization model, and give the information data collected by the proposed supermarket. It is of great significance to the formulation of the replenishment pricing strategy.

References

[1] Zhang Jin long, Wu Xiang, Xu Haoxuan. Joint Decision Model for Pricing and Replenishment of Deteriorating New Products [J]. Journal of Systems Engineering, 2018,33 (01): 79-89.

[2] Lu Jing. Research on inventory control and dynamic pricing of fresh agricultural products [D]. Tian Jin University, 2019.

[3] ZHOU Hai Jie. Ordering and pricing decisions for perishable products under partial loss sales [D]. Nanjing University of Aeronautics and Astronautics, 2021.

[4] Miranda S. A qualitative and quantitative analysis of vegetable pricing in supermarket[J]. IOP Conference Series: Materials Science and Engineering,2017,215(1).

[5] XU Wei chao. Review of correlation coefficient [J]. Journal of Guangdong University of Technology, 2012,29 (3): 12-17.

[6] Helmy R ,Steven B ,Augie W , et al. Image encoding selection based on Pearson correlation coefficient for time series anomaly detection[J]. Alexandria Engineering Journal,2023,82.

[7] Yang Liu. Solar power prediction algorithm based on improved BP neural network[J]. Computer Informatization and Mechanical System,2023,6(6).

[8] Scientific Platform Serving for Statistics Professional 2021.SPSSPRO. (Version 1.0.11) [Online Application Software]. Retrieved from https://www.spsspro.com.

[9] Zhou Zhihua. Machine learning [M]. Tsinghua University Press, 2016.

[10] Yu Dong Ju. Research on quality and safety management system of vegetable supply chain [D].Shandong University, 2020.