Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 2
CII 2023

Automated Pricing and Replenishment Decision Modeling for Supermarket Vegetables

Luoyi Zheng *, Zhiyan Lou

School of physics, Hangzhou Normal University, Hangzhou, China

* Corresponding author Email: 17815747587@163.com

Abstract. This paper analyzes the historical sales data of vegetable products in a fresh food supermarket, and explores the automatic pricing and replenishment decision-making scheme of vegetable products in this supermarket. Firstly, we integrate and clean the attached data, and then we solve the characteristics and interrelationships of individual products and categories by establishing SARIMA model and Spearman's correlation coefficient model, and then we use nonlinear regression model, nonlinear autoregressive neural network model and optimization model to realize automatic pricing and replenishment decision-making for vegetable products.

Keywords: SARIMA Model; Spearman Correlation Coefficient; NAR Neural Network; Nonlinear Programming Model.

1. Introduction

Considering the strong cyclicality, seasonality and low freshness of vegetable fresh commodities, the price fluctuation of vegetable commodities and the market demand fluctuation are very large [1], so fresh commercial supermarkets need to adjust the pricing and replenishment program of vegetable commodities frequently [2]. Usually, supermarkets will analyze and forecast the historical sales data and demand for each commodity, and adopt the "cost-plus pricing" method to formulate the dynamic pricing and replenishment plan with the highest economic efficiency [3].

2. Data Preprocessing

The data used in this paper come from Question C of the 2023 National Student Mathematical Modeling Contest. Before constructing the model, it is necessary to carry out data preprocessing on the data, including the finding and processing of missing values outliers, data standardization and other operations, to make the model construction and solving more efficient.

2.1. Data Integration and Processing

Before performing data cleansing, the data first needs to be consolidated and analyzed to get the correct classification needed. Utilize tools such as the VLOOKUP function and pivot tables in Excel to get the required data such as individual items by day and month as well as category sales volume.

2.2. Missing Value Outlier Test

There were no missing values in the current data. After further determining the outlier data using the 3σ principle of normal distribution and removing it, the mean of the normal values around the outliers was used to fill the outlier vacancies due to the existence of a relationship over time in the data set used. Figure 1 shows the normal distribution plot before and after the outliers are removed and filled for the sales volume data by day for the flower and foliage category.

It can be seen that after the data outliers are removed and populated, the smoothness of the overall data has risen considerably, laying a relatively good foundation for the next model prediction.

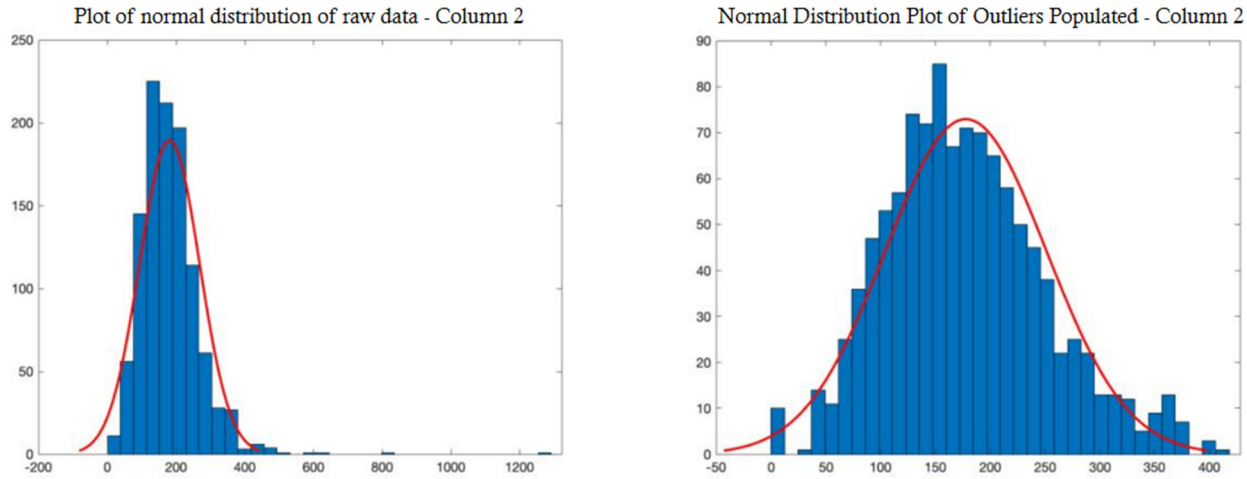

Figure 1. Normal distribution of the data before and after the treatment of outliers

2.3. Data Standardization (Normalization)

When carrying out the problem of testing the normal distribution of data, it is necessary to use the K-S test to determine the normal distribution pattern, so this paper uses the Z-Score standardization to unify the magnitude between different data, the following formula is used to calculate the standardized data:

$$\frac{x-\mu}{\sigma} \tag{1}$$

where x is the individual observation, μ is the mean of the overall data, and σ is the standard deviation of the overall data.

3. SARIMA Model (Seasonal ARIMA Model) Determining Cyclicality

In daily life, vegetables are seasonal, each season has different seasonal vegetables, it can be said that the data of vegetable sales over time is a seasonal change of data, using the SARIMA model to determine whether the vegetable sales are cyclical.

3.1. Cyclical Analysis of Sales Volume of Each Vegetable Item

In conducting the periodicity analysis of the sales volume of each vegetable item, the month was used as the time unit, and 12 months were used as the set seasonal cycle after excluding the data of cases with fewer than five sales per month. The p-value of the seasonal autoregressive parameter of the SARIMA model was used here to determine the significance of the seasonal parameter, and if this p-value is less than 0.05, then the data is considered to be cyclical. It can be assumed that the vegetable singles are all highly seasonal and difficult or not worthwhile to be grown in greenhouses or cultivated artificially [4-6].

3.2. Cyclical Analysis of Sales Volume by Vegetable Category

In conducting the cyclical analysis of sales volume of each vegetable category, the p-value of the seasonal autoregression parameter for the vegetable categories that conformed to the cyclicality was obtained by using month as the time unit and 12 months as the set seasonal cycle as shown in the Table 1 below.

The cyclicality of the vegetable category is not really that pronounced compared to the vegetable singles, which have a very pronounced cyclicality. However, it can also be seen that the sales volume of cauliflower and edible mushrooms are relatively stable throughout the year, without cyclical fluctuations. While flowers and leaves, eggplant, chili peppers, aquatic roots and tubers will have a certain degree of seasonality, there is a cyclical fluctuation in the sales volume.

Table 1. P-value of seasonal autoregressive parameters corresponding to cyclical vegetable categories

Vegetables	Cauliflower	Foliage	Eggplant	Peppers	Mushrooms	Aquatic Roots
P-value	0.449596	0.000299	0.004982	0.013901	0. 163944	0.03153715

4. Interrelationship Analysis

4.1. Normal Distribution Test (K-S Test)

Before the correlation coefficient calculation, it is first necessary to analyze whether the data conform to the normal distribution law. In this paper, we use the method of K-S test to carry out the analysis, and Table 2 shows the test results of the sales volume of each vegetable category after Z-Score standardization as an example.

Table 2. Z-Score Standardized Sales Volume Test by Vegetable Category

Vegetables	Cauliflower	Foliage	Eggplant	Peppers	Mushrooms	Aquatic Roots
D-value	0.0889	0.0729	0.0878	0. 1344	0. 1147	0. 1190
P-value	5.4964e-08	1.6423e-05	8.4025e-08	1.0096e- 17	5.3107e-13	5.8573e-14
Normal Distribution	no	no	no	no	no	no

Where, D value is the K-S statistic which indicates the maximum gap between the observed cumulative distribution function and the theoretical distribution. The larger the D value, the larger the gap between the data and the theoretical distribution. The P-value of the K-S test, on the other hand, is used to determine whether the data significantly deviates from the theoretical distribution. Generally, if the P-value is less than or equal to 0.05, the original hypothesis is rejected (the data conform to the law of normal distribution); if the P-value is greater than 0.05, the original hypothesis cannot be rejected [7].

According to the results of Table 2, it was found that the P-value of the K-S test for each vegetable category was less than 0.05, so the data did not conform to the law of normality. Due to the excessive amount of data for each vegetable item, it will not be repeated too much in the text, and after testing in the same way, it can be found that the data of each vegetable item also does not conform to the law of normal distribution.

4.2. Spearman's Correlation Coefficient

Because of the large data set, the correlation coefficient was chosen in this paper to analyze the interrelationship between vegetable sales. Currently, both Spearman's correlation coefficient and Pearson's correlation coefficient can be used to measure the relationship between two variables, but because the data are not normally distributed data, and according to the previous distribution law to get the data there is a nonlinear relationship between the components, so the more appropriate Spearman's correlation coefficient to measure the interrelationships between the sales volume of vegetables was chosen [8-9].

(1) Interrelationship of sales by vegetable category

Figure 2 shows a heat map of the Spearman correlation coefficient matrix visualized using Python. The color of each cell indicates the magnitude of the Spearman correlation coefficient between the two variables, with darker colors indicating stronger correlations.

Generally, the absolute value of the correlation coefficient is located at 0.6 to 0.8 indicating that the two variables are strongly correlated variables. Observing the heat map, it can be found that the correlation coefficients of cauliflower and foliage, foliage and chili pepper, foliage and edible mushrooms, and edible mushrooms and aquatic roots and tubers are all greater than 0.6, which has a strong positive correlation. And among them, cauliflower and many kinds of vegetables have positive correlation, which can be seen that the residents of this region will have a large part of the probability

to choose cauliflower and leafy vegetables when buying vegetables in the supermarket, and cauliflower and leafy vegetables are extremely popular.

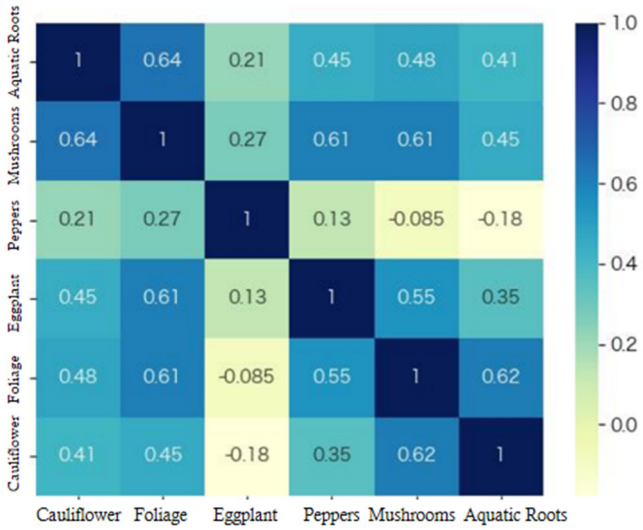

Figure 2. Heat Map of Spearman Correlation Matrix for Sales by Category

(2) Interrelationship of sales of individual vegetable items

Observing the data, it can be found that the number of days of sales of certain vegetables is too small, which will greatly affect the test of the correlation coefficient to infinitely converge to 1 or equal to 1. Therefore, the Pandas library in Python is utilized to eliminate the sales of vegetable single items with less than 10 days of sales before calculating the correlation coefficients, deriving the correlation coefficient matrix, and after selecting the five pairs of vegetable single items that have the highest positive and negative correlation respectively, the two correlation coefficient matrix heat maps of Figure 3 and Figure 4 are plotted.

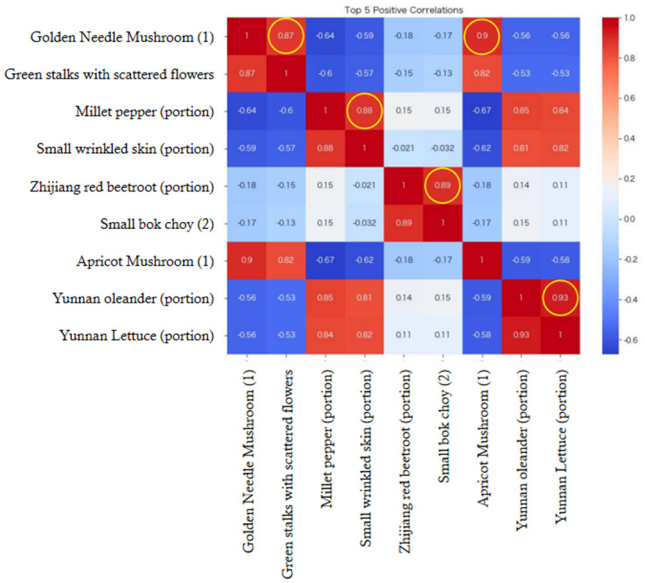

Figure 3. Heat map (1) of Spearman's correlation coefficient matrix for sales of each individual item

Refer to the concepts of "complementary goods" and "substitute goods" in economics: Complementary goods refer to two goods with complementary functions, which are mutually reinforcing; substitute goods refer to two goods with similar functions, which can be substituted for each other. Observing the two heat maps above, we can find that the positive correlation of complementary goods is strong, and the negative correlation of substitute goods is strong. In addition,

people with different tastes will have a relative concentration in purchasing vegetables, which will lead to a strong positive correlation for goods with similar tastes and a strong negative correlation for goods with opposite tastes [10].

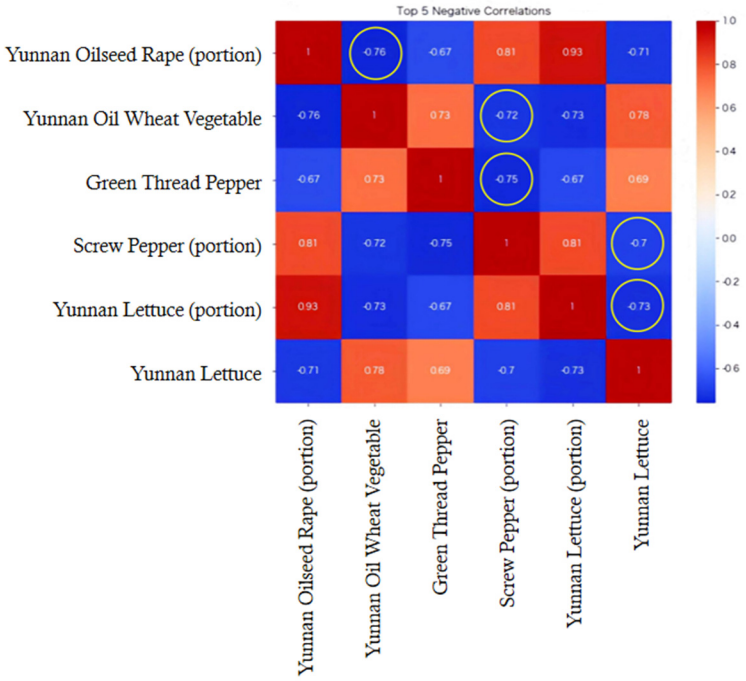

Figure 4. Heat map (1) of Spearman's correlation coefficient matrix for sales of each individual item

5. Analysis of the Relationship between Total Sales and Cost-Plus Pricing

5.1. Analysis of Nonlinear Regression Model Construction

Cost-plus pricing is a method whereby a merchant, to ensure a profit, multiplies the unit cost (i.e., the cost margin) by the profit margin that the merchant wishes to obtain in order to obtain the selling price of the unit, so the most important aspect of this pricing method is to determine the cost margin. This question is about the relationship between the total sales volume of each category and its relationship, so we can utilize the data in Appendix 1, 2, and 3 to consolidate and process the data, and ultimately get the total sales volume by category each day, as well as use the following formula to get the cost profit margin α:

$$\alpha = \frac{p-q}{q} \tag{2}$$

Where p is the selling price per unit of goods, q is the cost price per unit of goods.

However, the cost margin obtained by this formula is divided by individual products, so it is also necessary to calculate the proportion of total sales of each single product to the corresponding category on the same day according to different categories, and then obtain the cost margin of each category on a daily basis according to the weighted sum. The cost margin for each category for each day is obtained by weighting and summing the total sales of each item by category.

$$\alpha_i = \sum_{k=1}^{max} \frac{S_{ik}}{S_i} \cdot \alpha_{ik} \tag{3}$$

From this, we can get the total sales and cost margin data of each category every day, and based on this, we can develop a regression model to explore the relationship between the two.

5.2. Nonlinear Regression Modeling

First of all, according to the real law, the relationship between total sales and cost margin is unlikely to be a simple linear relationship, so the first consideration is non-linear regression analysis. In the process of using polynomial regression, it is found that the regression effect is not good in the case of utilizing all the days of data, and the R^2 is very small, which is far from meeting the regression conditions. And after careful observation of the data, it was found that the concentration of the data in the graph with sales volume as the independent variable and cost margin as the dependent variable is very high, which seriously affects the regression results in Figure.5.

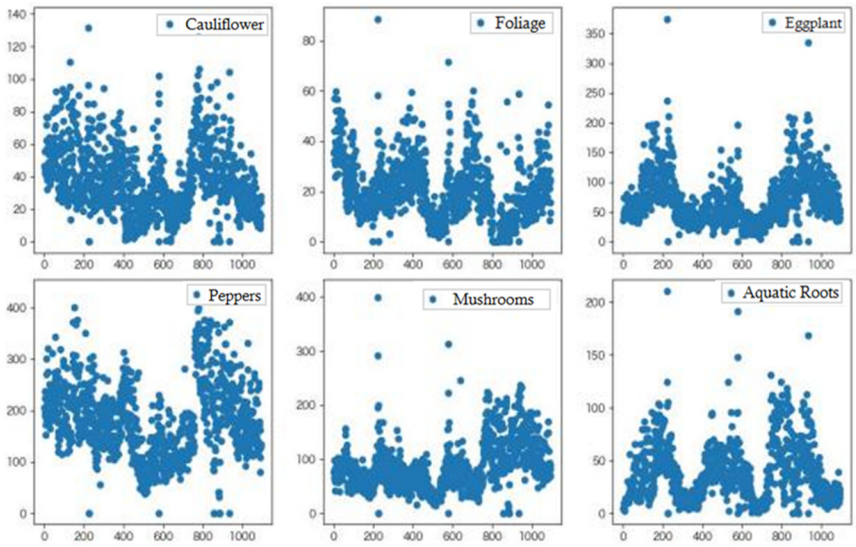

Figure 5. Scatterplot of Profitability-Sales Volume by Category

Therefore, it was decided to use the dimensionality reduction method of concentrating the adjacent data (taking the mean) to form the main influence nodes and eliminating the less influential nodes that are not concentrated, to reduce the amount of data and avoid the excessive concentration of data, and to obtain a more optimal regression model. Table 3 shows the regression model obtained after taking the main influence node as the independent variable at intervals of 0.1 for the cost margin, and averaging the concentrated sales as the new dependent variable (the numbers after the categories represent the range of values taken for the cost margin).

Table 3. Parameters associated with the nonlinear regression model for each category

/	Cauliflower (0.2~0.9)	Foliage (0.4~0.9)	Eggplant (0.3~1.2)
R^2	0.804588265	0.804986021	0.967318539
Formulas	$y=97.60-379.94x+689.77x^2-356.36x^3$	$y=1381.63-5558.16x+8344.40x^2-4078.55x^3$	$y = 61.32 -200.32x+307.91x^2-145.56x^3$
/	Peppers (0.3~1. 1)	Mushrooms (0.3~0.8)	Aquatic Roots (0. 1~1)
R^2	0.962618521	0.946556283	0.990852189
Formulas	$y = 16.38+ 176.52x-159.64x^2+73.45x^3$	$y = -544.27 + 3396.49x-6007.38x^2+ 3412.52x^3$	$y = -35.95+ 95.27x - 254.67x^2+133.36x^3$

6. Total Daily Replenishment and Pricing Strategy for Each Vegetable Category for the Coming Week

6.1. Nonlinear Autoregressive Neural Network (NARX) Predictive Modeling

The total daily replenishment for each category is mainly related to the customer demand, while the customer demand for the next 7 days can be predicted by the customer's past purchases. Considering that only one data set of past purchases exists and the set is not significantly linear, the prediction is

done using a nonlinear autoregressive neural network model (NARX). The results obtained with a set of 70% training set, 15% validation set, 15% test set and with R-value greater than 0.9 for all three sets are shown in Figure 6 and Table 4.

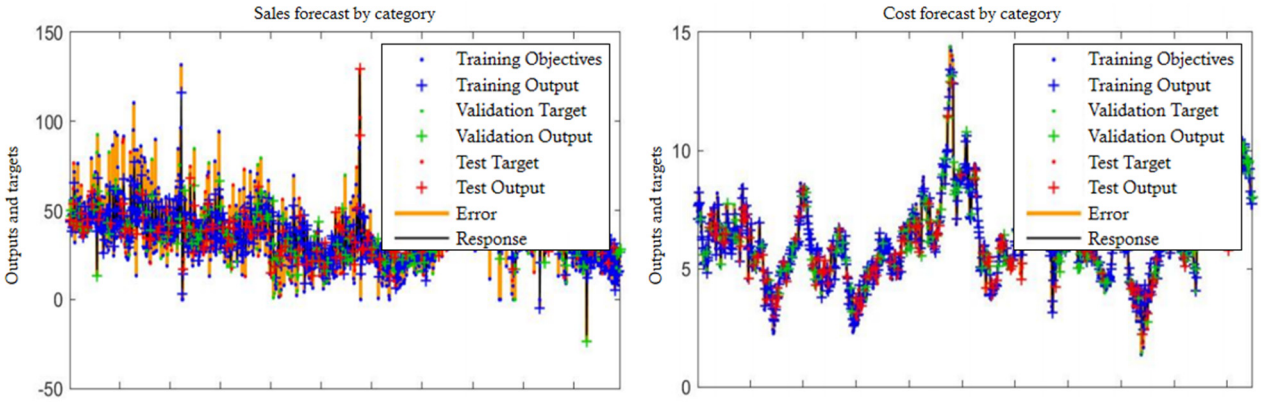

Figure 6. Chart of NARX forecast results by category

Table 4. NARX Sales Forecast by Category

Vegetables	Cauliflower	Foliage	Eggplant	Peppers	Mushrooms	Aquatic Roots
2023/7/ 1	37.92951	201.8056	31.49418	71. 16871	45.5425	14.28951
2023/7/2	34.24756	199.9477	30.40914	70.27701	47.41594	18.36825
2023/7/3	33.43057	192.9928	31.58712	65.26718	45.07234	18.07238
2023/7/4	31. 18621	190.0913	30.49302	63.56835	44.28584	17.4452
2023/7/5	29.21665	185.2505	28.95531	61.09094	44.24111	17.6655
2023/7/6	29. 12605	182.9522	29.32847	57.73269	43.41552	17.66378
2023/7/7	28.38372	181.5584	29. 18392	55.98395	43.46592	16.86995

Similarly a nonlinear autoregressive neural network model (NARX) is used to predict the cost of each category. With 70% training set, 15% validation set and 15% test set and R-value greater than 0.9 for all three sets, the results are shown in Table 5 below.

Table 5. NARX Cost Projections by Category

Vegetables	Cauliflower	Foliage	Eggplant	Peppers	Mushrooms	Aquatic Roots
2023/7/ 1	7.79301	4.430693	3.030926	5.718836	6.746899	7.952206
2023/7/2	7.769308	4.403739	2.999467	5.551791	6.553653	8.364544
2023/7/3	7.686843	4.479196	3.075727	5.544193	6.327671	8.408461
2023/7/4	7.625519	4.520373	3. 105108	5.556888	6.22972	8.505475
2023/7/5	7.555432	4.56031	3. 13243	5.592152	6.216834	8.349919
2023/7/6	7.486876	4.591549	3.153459	5.646372	6.245172	8.166899
2023/7/7	7.423755	4.61953	3.177012	5.693309	6.295435	7.981799

6.2. Analysis of Maximum Revenue Pricing Strategies

(1) Establishment of the objective function

In real life, there is a loss rate of vegetables during transportation and storage, and the weighted sum is used to obtain the loss rate of each category in Table.6.

Table 6. Weighted attrition rate by category

Vegetables	Weighted attrition rate (%)
Cauliflower	4. 163716371
Foliage	9.63717448
Eggplant	8.031499169
Peppers	8.602611
Mushrooms	10.3226375
Aquatic Roots	16.57654766

Then the objective function is as follows,

$$Z = \sum_{i=1}^{6} \left[C_i \cdot (1 + x_i) \cdot S_i - C_i \cdot \frac{S_i}{1 - w_i} \right] \tag{4}$$

(2) Establishment of constraints

The first is the requirement to satisfy customer demand, i.e., the requirement that actual daily sales must be at least greater than the forecasted daily sales

$$S_i \geq \text{predict}_i \tag{5}$$

It is actually the actual sales that have to match the relationship between the sales and the cost margin obtained.

$$S_i \geq f(x_i) \tag{6}$$

Where the above equations satisfy i = 1,2, ... ,6;

Thus, the final optimization model is obtained as follows:

$$\begin{aligned} Z &= \sum_{i=1}^{6} \left[C_i \cdot (1 + x_i) \cdot S_i - C_i \cdot \frac{S_i}{1 - w_i} \right] \\ \text{s.t} &\begin{cases} S_i \geq \text{predict}_i \\ S_i \geq f(x_i) \end{cases} \end{aligned} \tag{7}$$

(3) Total daily replenishment and maximum revenue pricing strategy results

The above optimization model will be solved by nonlinear programming using the fmincon function in MATLAB to get the total daily replenishment and maximum pricing under the optimal strategy, and in calculating the mean value to get the average profit of the supermarket is about 1136 yuan, and the profit is larger roughly in the range of 1500 yuan to 2000 yuan, which shows that the model predicts the result is good, and basically conforms to the maximum revenue pricing strategy.

7. Maximum Yield Pricing Strategy

7.1. Simple Exponential Smoothing (SES) Model to Forecast Sales Volume

Since this time only need to predict the vegetable sales of each single product on July 1, and the data set used is the sales of June 24-30, so the sample data volume is small and the prediction period is short. Observation of the data can be found, the data as a whole when smooth, there is no obvious trend and seasonality, so this paper uses a simple exponential smoothing model to predict the sales on July 1st.

It was necessary to obtain data on the wholesale price of each individual item and the relationship of each item to cost-plus pricing. Based on the screening of saleable varieties from 6/24/23 to 6/30/23, it was assumed that as long as there was sales volume data during the week, the individual item was considered saleable, and the final screening yielded 49 items.

7.2. Modeling of Nonlinear Programming

(1) Decision variables

The first decision variable is the markup rate x_j for each individual item, the second decision variable is the sales volume S_j for each individual item, and a third 0-1 variable has been added to reflect whether or not the category was selected.

(2) Establishment of the objective function

In addition to changing the category to a single item, the objective function that maximizes net profit is obtained by subtracting costs from revenues by adding a 0-1 variable y_j that represents whether or not the single item is selected (where j = 1,2, ... ,49).

$$Z = (1 + x_j) \cdot S_j \cdot y_j - C_j \cdot \frac{S_j}{1-w_j} \cdot y_j \tag{8}$$

(3) Establishment of constraints

First of all, the same is required to meet customer demand, the minimum display quantity is greater than 2.5kg, so the need S_i is greater than 2.5kg, the number of items purchased should be between 27 and 33, the sales of individual items to meet and the markup rate of the function, where the above formula meets j = 1,2, ... ,49, and ultimately get the nonlinear programming model is as follows.

$$Z = (1 + x_j) \cdot S_j \cdot y_j - C_j \cdot \frac{S_j}{1-w_j} \cdot y_j$$
$$\text{s.t.} \begin{cases} S_j \geq 2.5 \\ 27 \leq \sum_{j=1}^{49} y_j \leq 33 \\ S_j \geq \text{predict} \\ S_j \geq g(x_j) \end{cases} . \tag{9}$$

(4) Solving nonlinear planning models

Table 7. Optimal Pricing Strategies for July 1

Assortment	Profitability	Set a price	Assortment	Profitability	Set a price
Yunnan lettuce (portion)	0.8794	5.04	Yunnan oilseed rape (Brassica campestris L.)	0.8917	7.89
Yunnan oil wheat (portion)	0.8866	4.37	Broccoli	0.8995	12.52
baby Chinese cabbage (mini-sized variety)	0.8895	6.15	Zijiang Qingdian Scattered Flowers	0.8976	9.19
Spinach (portions)	0.8875	4.53	Lotus root(1)	0.8947	10.60
Spinach	0.8943	11.92	Wild lotus root	0.8971	18.67
Shanghai Youth	0.8923	8.42	water caltrop or water chestnut (genus Trapa)	0.8969	17.73
cabbage (round cabbage most commonly found in Chinese medicine)	0.8881	5.44	Kogua (1)	0.8964	15.40
Yunnan lettuce	0.8929	9.57	long term eggplant	0.898	12.52
Small bok choy (1)	0.8898	6.43	Green eggplant(1)	0.8969	8.28
sweet potato tip	0.8905	6.83	Peppers (portions)	0.897	4.85
snow fungus (Tremella fuciformis)	0.8913	7.56	king cobra or chili (Naga jolokia)	0.8998	13.31
Screw peppers (portions)	0.5072	4.22	Green pepper (portion)	0.899	3.81
Ginger, garlic & millet pepper combo (Small portion)	0.8987	3.20	Xixia mushroom(1)	0.6912	23.61
Red Pepper (2)	0.8999	20.02	Agaricus bisporus(box)	0.4153	4.93
Small wrinkled skin (portions)	0.8985	3.30	Seafood mushroom(Pack)	0.4	3.16
Green & red pepper combo (Servings)	0.899	4.45	Fresh fungus (portions)	0.5382	2.33
Colorful peppers (2)	0.8999	24.76			

Solving the above model yields the results of the optimal pricing strategy on July 1 as shown in the Table 7, which yields a maximum return of $1,213.4 for the superstore.

8. Conclusion

This paper discusses the automatic pricing and replenishment decision-making for vegetable commodities under different conditions, and establishes an optimization model to maximize the revenue of selling vegetables in supermarkets for historical sales data, selling price, and wastage rate,

which is a reference value for supermarkets in pricing and replenishment of other commodities in different categories. From the perspective of promotion, a demand prediction model based on machine learning or deep learning, a pricing model based on optimization theory, and a replenishment decision model based on inventory theory can be further constructed, and then a data-driven decision-making system can be built on the basis of the above models, which is capable of automatically collecting data, predicting the demand, calculating the optimal price and replenishment, and reporting the decision-making results to the decision makers, thus reducing the complexity of decision making. The decision-making system can be extended not only in the commercial field, but even to other fields such as logistics and medical care, which is highly applicable and conducive to improving the efficiency and accuracy of decision-making.

References

[1] ZHANG Yan, MOU Jinjin, WANG Shuyun. Optimization decision of fresh produce supply chain with price control in superstores[J]. China Management Science,2023,31(10):266-275.

[2] PAN Xiaofei, ZHANG Tao. Optimization of preservation efforts and pricing in fresh produce supply chain considering loss aversion[J]. Highway Transportation Science and Technology, 2023,40 (05):228-236.

[3] LI Panxia, LI Yaxuan. Transformation and Development of Entity Supermarkets under the Background of Digital Economy[J]. National Circulation Economy, 2023, (03):16-19.

[4] PAN Xiaofei, XIE Zhiheng, WANG Shuyun. Optimized decision-making of fresh food superstore preservation efforts and pricing considering loss aversion[J]. Highway Transportation Science and Technology,2022,39(06):177-185+190.

[5] ZHANG Yuzheng. Research on information sharing incentives in low temperature milk supply chain of Y dairy industry based on smart contract[D]. Jimei University,2023.

[6] LI Mingyu. Research on factors influencing the sales performance of liquor in large-scale supermarkets[D]. Liaoning University of Engineering and Technology, 2023.

[7] ZHANG Chunyan. Research on vegetable community distribution considering demand grading under sudden epidemic[D]. Beijing Jiaotong University,2022.

[8] WANG Caifeng. Research on fresh produce supply chain coordination considering demand information asymmetry under the perspective of fair concern[D]. Chongqing University of Commerce and Industry, 2021.

[9] WANG Hejiao. Research on the problems and countermeasures of operation management of A super group [D]. Shenyang University of Science and Technology, 2022.

[10] ZHANG Mingyue. Research on the evaluation of the effect of "agricultural super docking" based on supply chain partners [D]. Shandong Agricultural University, 2018.