Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 11, No. 3, 2023

Research on Pricing and Replenishment Strategies of Supermarket Vegetables Based on LSTM Neural Network and Goal Planning

Yishan Zhou*

School of Mathematics and Statistics, Hubei Normal University, Huang Shi, China
* **Corresponding author Email:** 2874563927@qq.com

Abstract: This paper focuses on the daily replenishment volume and pricing of vegetable commodities, especially the impact of freshness on market selling price. Through Pearson correlation analysis, this study first explored the correlation between different vegetable categories and their single products and evaluated the sales correlation between each category and the single products of the same category. Secondly, this paper considers the cost-plus pricing method to analyze the relationship between the total sales volume of vegetable products and the pricing. The LSTM neural network model is used to forecast the sales volume and wholesale price of vegetable commodities, with special attention to the forecast data from July 1 to 7. Finally, this paper establishes a goal planning model with the goal of maximizing supermarket returns. Considering the total quantity of replenishment, sales volume and cost pricing as constraint conditions, the depth-first search algorithm is adopted to solve the problem, and the daily replenishment volume and pricing strategy for vegetable products in the next week are provided.

Keywords: Vegetable Pricing; Correlation Analysis; LSTM Neural Network; Objective Planning.

1. Introduction

In the context of the current economic development, the public attaches increasing importance to nutritionally balanced dietary habits, and this trend has directly contributed to the rapid development of fresh commodities in the supermarket market. In particular, the unique perishable nature of fresh vegetables poses a challenge to the operation and management of supermarkets. The freshness of these primary agricultural products not only determines their value, but also becomes a major problem in the operation of superstores due to the short freshness period and large loss. To address this problem, this paper aims to solve the following key issues: first, analyze the correlation between different categories and different individual vegetables of the same category to reveal the interrelationship between their sales volume and the distribution pattern; second, formulate an effective replenishment plan for the category as a unit and predict the replenishment volume and pricing strategy for the coming week, with the aim of maximizing the superstore's revenue and further optimizing the superstore's operational efficiency and The aim is to maximize the revenue of the superstore and further optimize its operational efficiency and profitability.

2. Relevance analysis

2.1. Correlation analysis

This paper analyzes a sample of 251 vegetable individual items, categorizes them into six major categories, and calculates the quantity share of individual items in each category. It was found that the leafy and flowering vegetable categories accounted for more than 50% of the total, and revealed a positive relationship between the percentage of vegetable categories and the average wholesale price: the richer the number of individual items within a category, the higher the wholesale cost. In terms of selling unit price, the wholesale price and selling unit price of edible mushrooms were higher than those of other categories, showing a positive correlation between wholesale price and selling unit price. In addition, by analyzing the profit margins of different categories of vegetables, this paper finds that edible mushrooms have the largest difference between wholesale price and sales price, while aquatic roots and tubers have smaller price differences and relatively small average profit margins. Finally, through the statistics of the loss rate of different categories of vegetables, this paper points out that the loss rate of cauliflower category is the highest, and the eggplant category is the lowest, which provides an important guidance for the superstore in the development of transportation and storage strategies, and suggests that the categories with high loss rate increase the input of transportation and storage in order to reduce the loss, and the categories with low loss rate can increase the amount of purchase and the number of single products to improve the revenue.

(1) Linearity test and normality test

In Pearson correlation coefficient, whether the data is linear, and outliers have a big impact on the results, and the data needs to satisfy the normal distribution. Before calculating the coefficients, this paper carries out the linearity test and normality test by making scatter plot and Q-Q plot respectively [1].

Linearity test: the results are shown in Figure 1. below.

Where the scatterplot is presented with different categories and foliar singles, it is obtained visually that cauliflower and foliar, foliar and chili, eggplant and edible mushrooms, chili and edible mushrooms, eggplant and aquatic rhizomes and edible mushrooms and aquatic rhizomes have a linear relationship.

The normality of the vegetable sales data was tested. The Shapiro-Wilk test was chosen as the test method because the sample size for each individual item was less than 5000. The original hypothesis of this test is that the data meets normal

distribution; if the test result shows significance (P<0.05), the original hypothesis is rejected and the data is considered not to meet normal distribution; otherwise, the original hypothesis is accepted, and the data is considered to meet normal distribution. The results of Shapiro-Wilk test for sales volume of six vegetable categories showed that the significance level of all categories was higher than 0.05 at P. This indicates that the test results are not significant and therefore the original hypothesis cannot be rejected, i.e., the data all satisfy normal distribution. Combined with the Q-Q plot, the actual sales data as the X-axis, the quartile of the data when assuming normality as the Y-axis, make a scatter plot, the higher the overlap between the scatter and the straight line the more it obeys the normal distribution, and the larger the difference between the scatter the less it obeys the normal distribution, as shown in Figure 2.

Figure 1. Scatterplot of categories and individual products

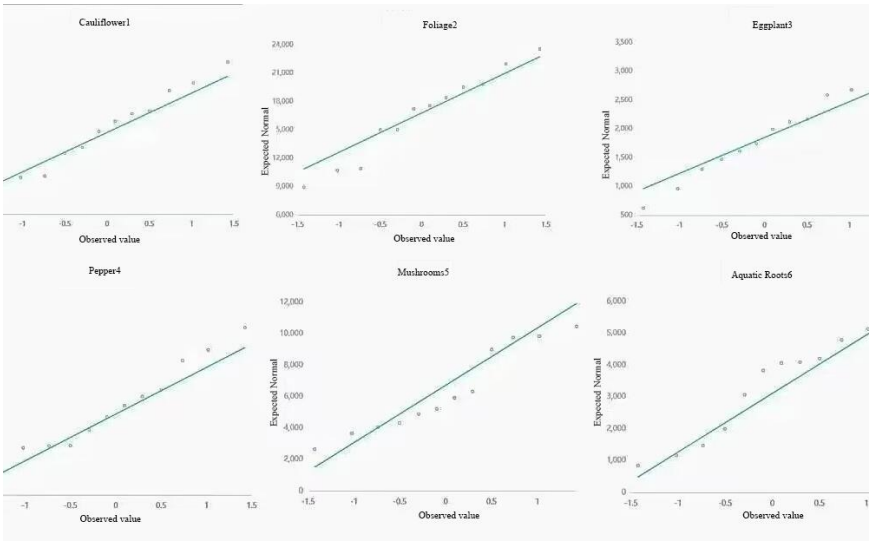

Figure 2. Q-Q diagram

Since there is a certain correlation between the sales volume of vegetable goods and time, the time variables in the tabular data are quarterly discretized as January to March for the first quarter; April to June for the second quarter; July to September for the third quarter; and October to December for the fourth quarter, for a total of 12 quarters.

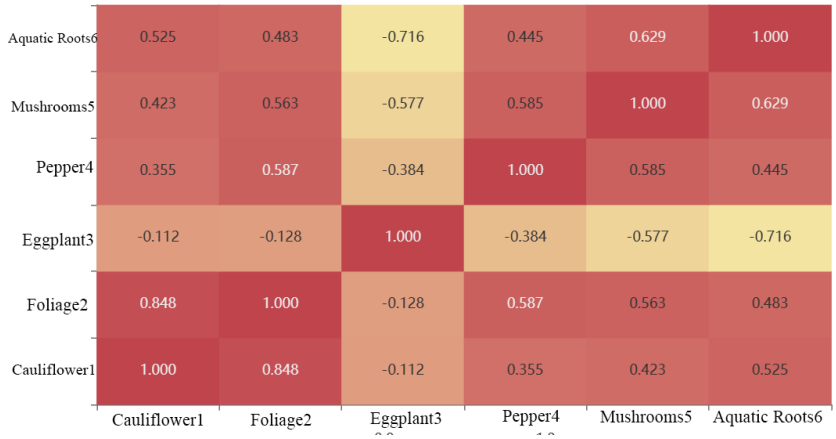

Figure 3. Heat map of correlation coefficients

As shown in Figure 3, the results indicated that among the six major categories, cauliflower and cauliflower, cauliflower and pepper, eggplant and edible mushrooms, pepper and edible mushrooms, eggplant and aquatic rootstocks and

edible mushrooms and aquatic rootstocks were significantly correlated.

2.2. Distribution patterns.

This paper presents a time-series analysis of sales data for individual vegetable products, summarizing data from the third quarter of 2020 to the second quarter of 2023, covering a total of 12 quarters within the last three years, categorized by vegetable category. This analysis aims to explore trends in vegetable sales data across different quarters. In order to visualize these changes, a line graph was produced (Figure 4):

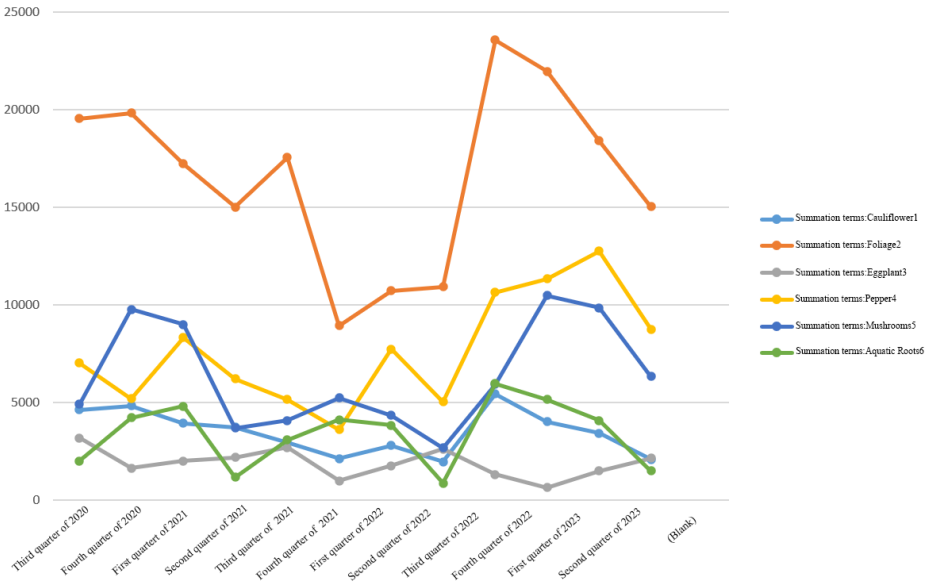

Figure 4. Quarterly Sales by Category

According to the information in the image: the total sales volume of each category of vegetables mostly started to rise from the second quarter and was in a state of growth in the second and third quarters, and then declined in the fourth quarter, and the decline continued in the first quarter of the coming year.

Six categories of vegetables in the second and third quarters of the sales volume relative to other time periods fast growth, the highest total sales volume; in the first and fourth quarters of the sales volume fell back, a gradual downward trend. From the supply side, vegetable supply varieties are more abundant from April to October, indicating that their sales volume fulfills the law of positive correlation with supply.

3. Category replenishment and pricing strategy optimization

3.1. Modeling and solving

The LSTM model, also known as long and short-term memory network model, is essentially a specific RNN (recurrent neural network) model, which solves the problem of RNN short-term memory by adding gates based on RNN model, so that the recurrent neural network can truly and effectively utilize the long-range temporal information. LSTM adds input gates on the basis of the RNN structure (Input Gate), Output Gate (Output Gate), Forget Gate (Forget Gate) three logic control units, can better solve the long-range dependence, gradient disappearance and other problems. The specific principle is shown in Figure 5 below [2].

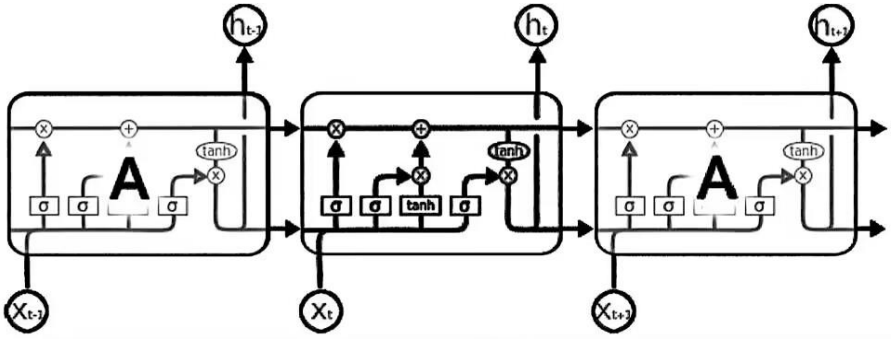

Figure 5. Flowchart of LSTM processing

The LSTM model improves the simple nodes of a traditional neural network into the form of memory units, from which correlations between data over long-time spans can be learned. Its memory unit exists in each neural network unit, the state at moment t is c_t (cell state), which is used to save the state information of the current LSTM and pass it to the next moment LSTM, while the current LSTM receives the cell state c{t-1} from the previous moment and generates c_t together with the signal input xt received by the current LSTM, which is activated by activating the unit function sigma or tanh on the input gate it, the forgotten gate ft, and the input gate o_t, which enables the identification and alteration of the cell blocks inside the model. Each of the three types of gates within each cell block is activated or not using

an activation unit function to decide whether to activate it, which is used to control the information retention and transmission of the LSTM, i.e., the transmission state is controlled by the gating state. The inputs are transformed by a nonlinear function and superimposed with the memory cell states processed by the forgetting gate to form a new memory cell ct. Eventually, the memory cell states are nonlinearly operated and controlled by the output gates to form the output h1 of the LSTM cell, which is computed by the following formula [3].

$$i_t = \sigma\left(w_{x_i} + W_{hi}h_{t-1} + W_{Ci}c_{t-1} + b_i\right) \quad (1)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \quad (2)$$

$$c_t = f_t c_{t-1} + i_t\, tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (4)$$

LSTM network cell output:

$$h_t = o_t\, tanh(c_t) \quad (5)$$

Where W and b are the weight coefficient matrix and the bias vector of the input gate, respectively, and the sigma function is the activation function. This function transforms the output function value to a value between (0, 1) with the following formula:

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (6)$$

The tanh is the activation function of c_t, which can map the output function value to [-1,1], and has a better inclusiveness for neural networks, and its formula is as follows:

$$tanh = \frac{(e^{x}-e^{-x})}{(e^{x}+e^{-x})} \quad (7)$$

After forecasting, the future sales volume and wholesale price forecasts for the six vegetable categories can be obtained.

3.2. Optimization of models

The requirement is to plan replenishment on a category-by-category basis so that the superstore maximizes its revenue in the coming week. This question uses a goal planning approach to solve the optimal replenishment total and pricing strategy and models a deep search algorithm to do so [4].

Variable Definition: There are a total of 6 vegetable categories, where the total amount of replenishment on day t for category j is, and the pricing strategy is (real).

Objective Function: According to the question of maximizing the revenue of the superstore, the objective function can be defined as maximizing the total revenue. The total revenue of the supermarket is the sum of the revenue of the six categories of individual products, and the revenue of individual products can be calculated by the relationship between the total sales volume and the cost-plus pricing. denotes the revenue of the j_{th} category of vegetables on day t, which is calculated as the product of sales volume and revenue, and the formula is:

$$P_{tj} = S_{tj}(v_{tj} - c_{tj}) \quad (8)$$

Constraints:

(1) Commodity cost pricing is to determine the price of goods by the variable cost per unit of product, plus a certain percentage of fixed costs and profit per unit of product. The unit price of goods is expressed as follows.

$$v_t = c_t(1+r) \quad (9)$$

Here r is the merchandise markup rate, and c denotes the total cost per unit of merchandise. Supermarkets need to control their costs by needing to limit the ratio of cost to total sales for the total amount of replenishment in each category.

(2) The calculation of the maximum demand E_i of the superstore, you need to consider the total daily sales volume of each category and its corresponding average depletion rate, through the total sales volume to predict the demand. In order to maximize profits, then you need to consider the minimum loss rate, so you need to ask for the average loss rate of each of the six categories of vegetables, respectively, the six categories of goods for the inverse weighting that loss rate is high single product, then its weight is small, in order to achieve the goal of controlling the rate of loss, to obtain greater benefits [5].

The formula for calculating the maximum demand E_i is as follows:

$$E_i = \frac{S_{t_i}}{1-e_i} \quad (10)$$

Where E_i is the average discount rate of vegetables in category i, which is calculated as follows?

$$e_i = \frac{1}{n}\sum_{i=1}^{n} e_{ij} \quad (11)$$

(3) Restocking the total amount of constraints: the total amount of replenishment of each vegetable category should meet the supply constraints, i.e., the sum of the sales volume and the replenishment volume should be the maximum demand, to achieve the relationship between supply and demand equilibrium, i.e., there is:

$$E_i = S_{tj} + G_{tj} \quad (12)$$

At the same time, the range of the replenishment quantity should be greater than or equal to 0, with a certain restricted range, and set to an integer.

In summary, the final optimized integer model is established as follows:

$$\max P_{sum} = \sum_{t=1}^{7}\sum_{j=1}^{6} P_{tj}$$

$$s.t.\begin{cases} v_t = c_t(1+r) \\ r \in [0.05, 0.20] \\ E_i = \dfrac{S_{t_i}}{1-e_i} \\ e_i = \dfrac{1}{n}\sum_{i=1}^{n} e_{ij} \\ P_{tj} = S_{tj}(v_{tj} - c_{tj}) \\ E_i = S_{tj} + G_{tj} \\ 0 \le G_{tj} \le S_{tj} \end{cases} \quad (13)$$

(4) Depth-first algorithm (DFS) is a method of searching or traversing a tree along the vertical direction of the tree. Starting from the root node, it follows a path and keeps going until it encounters an obstacle or fails to reach the goal point, whereupon it returns to the previous level for a new path and retraces its steps until it finds the end point, as shown in Figure 6.

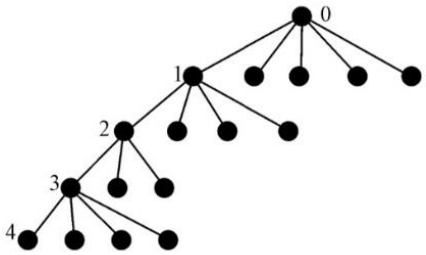

Figure 6. Depth search tree

The operation steps are shown in Figure 7:

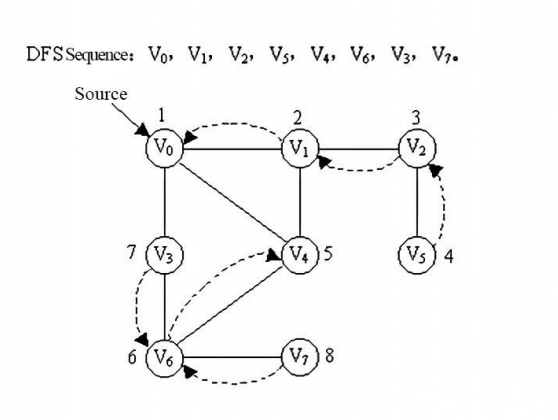

Figure 7. Algorithmic flow

(5) Deep Search Algorithm Implementation

The search algorithm uses two stack tables, one OPEN and one CLOSE, to manage the nodes, which is more suitable for steady state node management. According to the total sales of different vegetable categories and the wholesale price of the profit changes, the dynamic generation of node data structure, the nodes are associated with pointers, the establishment of the search tree. As shown in Figure 8.

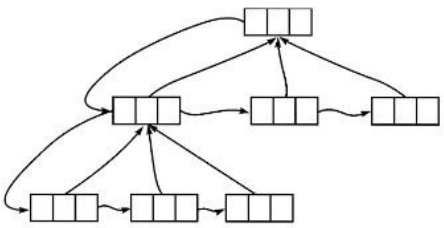

Figure 8. Data structure chained lists to build a depth search tree.

The optimized prediction results can be obtained by the above model.

In addition to algorithm optimization, to make more effective replenishment and pricing decisions for vegetable goods, this paper suggests further superstores collecting relevant data in the following seven areas: sales data, supply chain data, seasonal and weather data, competitor data, consumer behavior data, market trend data, and promotion effect data. The analysis of these data can help superstores deeply understand market demand, optimize supply chain management, cope with the impact of seasonal and weather factors, and formulate effective competitive countermeasures.

By analyzing sales data and consumer behavior data, hypermarkets can more accurately determine market demand and make replenishment decisions accordingly. Understanding supply chain data helps supermarkets build better relationships with suppliers, ensuring a stable supply of vegetable commodities and more competitive prices. Collecting seasonal and weather data can help supermarkets predict the supply and demand of vegetable commodities and adjust replenishment plans and pricing strategies in time to adapt to changes in market demand. Understanding competitors' data, such as product assortment, pricing strategies and promotional activities, can help supermarkets assess the competitive situation in the market and formulate targeted promotional and pricing strategies to improve their competitiveness.

4. Conclusions

The LSTM and depth-first algorithms used in this paper have their own strengths and weaknesses. The LSTM effectively mitigates the gradient vanishing or explosion that may occur in long sequence problems through various gate functions, and outperforms the traditional RNN in dealing with long term dependency problems. As a nonlinear model, it is suitable for building complex deep neural networks. However, LSTM still faces challenges in dealing with very long sequences and is relatively inefficient to train due to the complexity of its internal structure, which is computationally intensive and time-consuming, especially when the time span is large and the network is deep. On the other hand, depth-first algorithms have advantages over generalized-first algorithms in terms of time-consumption and speed, but its search incompleteness may lead to a dead loop or failure to find an optimal solution. Therefore, the limitations of these models need to be carefully considered in practical applications and be adjusted and optimized according to the specific situation to better adapt to the needs of practical problems [6].

References

[1] Gao Shuping; Li Xiaofang; Song Guobing; Zheng Han; Guo Fangbin. Fault routing method for low-voltage DC microgrids based on Pearson correlation coefficient and generalized S-transform[J]. Power System Protection and Control, 2023, 51(15):120-129.

DOI:10.19783/j.cnki.pspc.221965.

[2] Xue, Yang; Li, Jinxing; Yang, Jiangtian; Li, Qing; Ding, Kai. Short-term prediction of photovoltaic power based on similar day analysis and improved whale algorithm optimized LSTM network model[J/OL]. Southern Grid Technology,1-9 [2023-11-22]

http://kns.cnki.net/kcms/detail/44.1643.TK.20231120.1141.007.html.

[3] WANG Ting; XUAN Shibin; FU Mengdan; ZHOU Jianting. Anomaly detection based on similarity between memory cells and multi-scale structures [J/OL]. Microelectronics and Computers, 2023, (08):28-36[2023-11-22]

https://doi.org/10.19304/J.ISSN1000-7180.2022.0539.

[4] Chao Liu. Depth-first search algorithms in teaching artificial intelligence - the classical maze problem as an example[J]. Experimental Teaching and Instrumentation, 2021, 38(09):66-68.

DOI:10.19935/j.cnki.1004-2326.2021.09.027.

[5] Zou, Lilin; Wang, Zhanqi. Characterization of faceted geographic landscapes based on inverse weighted Voronoi diagram[J]. Geography and Geographic Information Science, 2012, 28(02):24-26.

[6] SONG Wei; HAN Jiahu; LI Feng; CAO Qingfeng. Hammerstein nonlinear model identification under colored noise interference[J]. Journal of Shaanxi University of Science and Technology, 2023, 41(05):189-194+202. DOI:10.19481/j.cnki.issn2096-398x.2023.05.002.