

Prediction and Replenishment Decision Making for Automatic Pricing of Vegetable Commodities Based on LSTM Models

Yijing Liang^{1,*}, Youbang Li², Xuepeng Chen³

¹Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

²Chaohu University, Anhui, 238024, China

³North China University of Water Resources and Electric Power, Henan, 450046, China

* Corresponding author

Abstract: In this paper, the problem of automatic pricing and replenishment decision-making of vegetable commodities is studied, and the LSTM model is used to predict and optimize it, and the replenishment volume and pricing strategy to maximize the revenue of supermarkets are proposed. Calculating the total annual sales volume of each category and each vegetable, the distribution law of most vegetables with strong seasonal periodicity and high sales volume from September 2022 to January 2023 can be obtained, and finally the correlation coefficient between different categories and different single products of vegetable commodities is obtained through the Pearson correlation coefficient algorithm, among which the highest correlation coefficient between different categories is mosaic and cauliflower. The coefficient was 0.75, and the highest correlation coefficient between different single items was 0.99 for green eggplant (1) and purple round eggplant. According to the cost-plus pricing algorithm and formula, the cost-plus pricing of each vegetable category is calculated, and then the relationship between the total sales of each vegetable category and the cost-plus pricing can be obtained through visual analysis. Then, the LSTM time series model was constructed to predict the total daily replenishment of each vegetable category in the coming week, and finally the linear programming model was used to maximize the revenue as the objective function to finalize the total daily replenishment and pricing strategy, and the maximum revenue was 75,624 yuan.

Keywords: LSTM model; Pearson correlation coefficient; automatic pricing; Cost analysis; Time series data.

1. Introduction

In fresh produce superstores, since most vegetables have a short shelf life and their quality deteriorates with time, if they are not sold within one day, they cannot be sold the next day. Therefore, supermarkets usually replenish their stocks on a daily basis according to historical sales data and demand [1]. However, since supermarkets sell many types of vegetables, the origins of various vegetables are different, and the vegetables are usually stocked from 3am to 4am again, merchants have to give the replenishment decision for various vegetable types for the day without specifying the specific individual items and the price of the stock [2]. Considering all aspects, it analyses and helps the superstore to make replenishment and pricing decisions for vegetable items in order to maximise its revenue.

2. Related Work

There has been some related work in the research area of forecasting and replenishment decisions for automated pricing of vegetable commodities based on LSTM models. LSTM models are effective in capturing the long term price dependencies and trends when dealing with vegetable price time series data, which helps to improve the accuracy of pricing [3]. Researchers may emphasise the superiority of LSTM over traditional time series methods and how to select appropriate model parameters for price forecasting. LSTM models can better predict future vegetable demand, thus helping to optimise inventory management and replenishment decisions. Research may include how to combine demand forecasts with actual inventory levels to avoid overstocking

or shortages and improve supply chain effectiveness [4]. The application of LSTM models in market trend analysis can provide deeper insights for developing pricing strategies. Researchers may focus on how LSTM models can be used to capture market dynamics, identify key trends and adjust accordingly in pricing [5]. The application of LSTM models in supply chain management can enable intelligent optimisation of inventory, transportation and replenishment. The study may introduce how to combine LSTM models and supply chain management principles in order to build efficient and agile supply chain systems [6]. Considering the impact of multiple factors on vegetable prices can make pricing models more comprehensive and accurate [7]. The researcher may mention multiple influencing factors including seasonality, climate, market demand, etc., and describe how to integrate these factors to build a comprehensive pricing model.

3. Model Establishment and Solution

3.1. Frequency-based analysis of the distribution pattern of sales volume of vegetables by category and individual product

3.1.1. Distribution law of sales volume of each category of vegetables

Using Python to calculate the total annual sales volume of each category of vegetables, cauliflower category in 2020 from June to December to maintain a high level, with slight fluctuations. 2021 overall declining trend. 2022 fluctuations in 2022, reaching a peak in August. 2023 is a substantial downward trend [8]. Overall cauliflower sales are subject to seasonal variations and are higher in the winter months.

Through the query of related information, it is known that cauliflower vegetables bloom in November to February each year, which is more consistent with the results of this paper.

Cauliflower and leafy category in 2020 at a high level, 2021 was first rising and then falling, 2022 was an upward trend, 2023 was a substantial decline, but focused on, Cauliflower and leafy category is also subject to seasonal influences, the sales volume is larger in the summer, the data of August 2023 is not recorded, so it can be assumed that Cauliflower and leafy sales volume is mainly concentrated in the summer that is, July-September [9].

The sales volume of chilli category as a whole shows some fluctuations and trends. From July 2020 to February 2021, the sales volume gradually increased and reached its peak. In 2021 in July 2022 and August 2022, the sales volume again increased and peaked then continued to decline between September 2022 and June 2023 [10]. Chilli category is at its peak in August in 2020 and 2022 and in January and February in 2021 and 2023, so it can be tentatively concluded that the sales volume of chilli category is not seasonally related and the fluctuation and trend of the sales volume of chilli category may be affected by the competition in the market.

The sales volume of aquatic roots and tubers category has a significant seasonal effect between the same months of the year. It peaks in December 2020 and February 2021 and January 2023, and again in July and August 2022, so overall sales volumes are higher in winter and may be related to seasonal demand in summer. The peak sales of edible mushrooms are all in winter, which is affected by seasonality, and the rest of the season sales are relatively low, especially between April and June, which may be related to the environmental conditions for growing edible mushrooms. The eggplant category was also affected by seasonality, with a larger sales volume in May, June and July, followed by a sharp decline in August and beyond, probably related to the growing conditions and maturity period of eggplant.

Overall, most categories of vegetables are affected by seasonality, edible fungi and eggplant are affected by its larger, preliminary judgement may be related to its growth conditions and maturity period [11]. The sales volume of flowering and leafy vegetables topped the six categories of vegetables, while the sales volume of eggplant ranked at the bottom of the six categories of vegetables. It can also be seen that the sales volume of all categories of vegetables during the period from September 2022 to January 2023 was higher than the rest of the period, which may have been influenced by a combination of several factors, including time trends, market competition and other external factors.

3.1.2. Distribution pattern of sales volume of individual vegetable categories

Due to the excessive number of vegetable single-item categories, its distribution law this paper adopts another perspective, the total sales volume of each vegetable is regarded as unit one, and analysed in terms of the percentage of sales volume in each year, most of the vegetable single-item sales volume in the first two years is not as good as that in the last two years, and the total sales volume of more than 20,000 is only three kinds, respectively, Wuhu green peppers (1), net lotus root (1), and broccoli. How many sales of various vegetables each year and comparison, for example, Wuhu green pepper (1), the highest sales in 2021, 2022, a relative decrease in 2023 and again reduced, which can be derived from the distribution of sales of Wuhu green pepper (1) law: 2020 sales of 0, from 2021 onwards, year by year

decreasing [12]. The same can be derived from the distribution of the rest of the vegetable single product law. And through the calculation, it is known that it accounts for 18.38% in 2020, 29.18% in 2021, 34.25% in 2022, and 18.19% in 2023, which can be concluded that the distribution law of all vegetables in the four years of the sales volume is the first increase and then decrease.

3.2. Analysis of interrelationships based on Pearson's correlation coefficient

For the interrelationship of each category of vegetables, it is not possible to see the relationship between them directly from the data, so this paper uses the Pearson correlation coefficient to calculate the interrelationship between each category of vegetables. Pearson's correlation coefficient is an evaluation index used to measure the degree of correlation between groups of data, calculated according to the method of cumulative difference, based on the deviation of two groups of data from their respective mean values, and reflecting the degree of correlation between the two variables through the multiplication of the two deviations. The Pearson correlation coefficient is defined as:

$$K = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

In the above formula: X, Y for the two groups of data to calculate the degree of correlation, the two data have n elements; \bar{X}, \bar{Y} are the average value of the two groups of data; K is the Pearson correlation coefficient, which takes the value of -1 to 1. The degree of correlation between the two groups of data by the Pearson correlation coefficient that is, the absolute value of the K decision, K the larger the absolute value of the data that is the correlation between the two groups of data the more high. The strength of correlation is usually determined by the range of values in the following table.

Numerical range	Relevance
0.8 - 1.0	Very strongly correlated
0.6 - 0.8	Strongly correlated
0.4 - 0.6	Moderately correlated
0.2 - 0.4	Weakly correlated
0.0 - 0.2	Very weak or no correlation

Using this method, we can calculate the correlation coefficient between different categories of vegetables, the correlation coefficient between different categories of vegetables, that is, the correlation coefficient between different categories of vegetables, the highest correlation coefficient of leafy and cauliflower, correlation coefficient of 0.75. Correlation coefficient of the sales of a single product we use the same kind of category for the calculation of the correlation coefficient, the correlation between the sales of a single product: cauliflower vegetables, correlation coefficient of the largest is: green peduncle scattered flowers - branch river green peduncle scattered flowers, maximum of -0.48. For eggplant vegetables, the largest correlation coefficient is: green aubergine (1)-purple round eggplant, with a maximum of 0.99. For leafy vegetables, the largest correlation coefficient is: red oak leaf-red coral (coarse leaf), with a

maximum of 0.97. For aquatic root vegetables, the largest correlation coefficient is: lotus root in Honghu Lake-water chestnuts, with a maximum of 0.63. For chilli peppers, the largest correlation coefficient is: green pepper-red thread pepper, with a maximum of 0.63. For edible mushroom vegetables, the largest correlation coefficient was 0.83. For edible mushroom vegetables, the largest correlation coefficient was 0.85 for: enoki mushroom (1) - flat mushroom.

To sum up, there are correlations between different categories or individual items of vegetable commodities, and the distribution pattern of sales volume is affected by factors such as seasons, holidays and abnormal weather, while the interrelationships between individual items reflect people's preferences and mixing habits when purchasing vegetables.

3.3. Analysis of the relationship between total sales of vegetable categories and cost-plus pricing

The so-called cost-plus pricing method refers to the government in the natural monopoly industry products pricing, based on the actual costs reported by the enterprise and audited by the government, plus the government to determine the industry's profit margin, as the price of its products or services. This pricing method is simple, intuitive, the biggest advantage is conducive to protect investors' investment recovery and return on capital [13]. As a pricing policy, it has a strong guiding effect on attracting investors to enter capital-intensive industries, especially when the industry needs a lot of capital to start up quickly. There are two common pricing models for this method, one of which is used in this paper:

$$P = C(1 + w) \quad (2)$$

Where P denotes cost-plus pricing, C denotes average cost, and w denotes cost-plus rate.

The total sales volume of each vegetable category is calculated using the results of the first question, and the cost-plus pricing of each category is calculated using the method described above to show that the total sales volume is not the only determinant of cost-plus pricing, and that the relationship between the total sales volume and cost-plus pricing for different categories is not entirely consistent. For example, the fact that the foliage category has the highest

sales volume but relatively low cost-plus pricing could mean that the foliage category has a high market demand and is highly competitive, so cost-plus pricing is relatively low. Aquatic roots and tubers have relatively low sales volume but high cost-plus pricing, which may be due to the fact that it is a specialised category with relatively low supply and some market demand, hence high cost-plus pricing. The chilli category and edible mushrooms have relatively high total sales and medium cost-plus pricing. This may indicate that these categories have high demand in the market and are in relatively sufficient supply, hence cost-plus pricing is more balanced.

3.4. LSTM model building and solving

To predict the total daily replenishment of each vegetable category in the coming week (1-7 July 2023), the volatility of the historical time series of total sales is smoother, so the time series model ARIMA is used, but after the test, the ARIMA model predicts the results are not very perfect, and even the results of the predicted value does not occur, after re-analysis and review of the information, it was found that the ARIMA model requires a data time sequence is shorter so that the prediction results will be better. So LSTM model is used because LSTM is suitable for data with complex time dependence, the time series of data is long, and we can provide enough historical information to train LSTM model.

LSTM model is a special variant of Recurrent neural network (Recurrent neural network (RNN)), with a "gate" structure, through the logical control of the gate unit to decide whether to update the data or choose to discard, to overcome the RNN weights have too much influence, easy to produce gradient disappearance and the shortcomings of the explosion, so that the network can be better and more efficient. It overcomes the shortcomings of RNN, which has too much influence on weights and is prone to gradient disappearance and explosion, so that the network can converge better and faster, and can effectively improve the prediction accuracy.

The model is built by first normalising the original data, scaling the sales volume data to between 0 and 1, and then dividing the data into training data and target data. For each time step, we use the data from the past 7 time steps as input features and the data from the next time step as target output. The shape of the training data is then adjusted to fit the input requirements of the LSTM model. Then a model containing two LSTM layers and one densely connected layer is constructed, and finally the model is trained. The specific results are shown in the following table:

Table 2. Total daily replenishment of each vegetable category in the coming week

Date	Aquatic Roots and Tubers	Foliage	Cauliflower	Eggplant	Chilli	Mushrooms
2023/7/1	19.624	132.887	21.838	19.470	76.818	49.072
2023/7/2	15.461	113.108	19.198	16.818	67.600	41.700
2023/7/3	12.622	106.280	17.504	15.204	63.380	39.319
2023/7/4	12.884	109.215	17.499	14.449	62.232	40.216
2023/7/5	15.031	117.314	19.431	16.077	65.021	43.122
2023/7/6	17.542	124.392	21.558	17.131	68.764	45.751
2023/7/7	18.660	127.291	22.616	18.294	67.674	46.646

3.5. Optimisation model establishment and solution

Optimisation model, based on the requirements of the topic, this paper uses a linear programming model to establish and

solve:

Decision variables:

- 1) Replenishment quantity of each category: x_i , indicating the replenishment quantity of the i th vegetable category
- 2) Pricing: p_i , represents the pricing of the i th vegetable

category.

Objective function:

Maximise the revenue of the superstore:

$$M = \max\left(\sum_{i=1}^6 (p_i * x_i) - \sum_{i=1}^6 (c_i * x_i)\right) \quad (3)$$

Where M denotes the superstore revenue and c_i denotes the cost of the i th vegetable category.

Constraints:

1) Market demand constraint

2)

$$x_i \leq d_i \quad (i = 1,2,..6) \quad (4)$$

where d_i denotes the market demand for each sellable

individual item.

2) Strategic Constraints

$$p_{imin} \leq p_i \leq p_{imax} \quad (i = 1,2,..6) \quad (5)$$

In summary the model is built as follows:

$$s.t \begin{cases} x_i \quad (i = 1,2 \dots 6) \\ p_i \quad (i = 1,2 \dots 6) \\ M = \max\left(\sum_{i=1}^6 (p_i * x_i) - \sum_{i=1}^6 (c_i * x_i)\right) \\ x_i \leq d_i \quad (i = 1,2 \dots 6) \\ p_{imin} \leq p_i \leq p_{imax} \quad (i = 1,2 \dots 6) \end{cases} \quad (6)$$

After solving the results can be obtained as follows:

Table 3. Pricing strategies for each vegetable category for the coming week

Date	Aquatic Roots and Tubers	Foliage	Cauliflower	Eggplant	Chilli	Mushrooms
2023/7/1	9.054	13.911	20.959	15.601	15.485	98.144
2023/7/2	7.508	12.755	17.361	12.502	12.336	78.601
2023/7/3	7.509	11.752	15.522	10.754	10.175	71.175
2023/7/4	7.413	12.043	15.521	10.655	10.583	73.401
2023/7/5	8.278	12.942	17.286	12.241	11.398	80.864
2023/7/6	9.632	13.736	19.161	13.101	12.592	89.453
2023/7/7	9.809	13.938	19.604	14.148	12.428	92.949

4. Conclusion

Different vegetable categories show different price trends in the coming week, with some showing less volatility and others showing a more pronounced upward trend. Less volatile vegetable categories may be more stable, while more volatile categories may require more flexible inventory management and pricing strategies. There may be certain vegetable categories that are affected by seasonal and market factors, which may account for the price fluctuations. Further analysis and understanding of the market situation is necessary for more accurate conclusions. Based on the price trend, one may consider increasing inventory moderately when price increases are more pronounced and decreasing inventory moderately when prices are relatively stable or declining in order to optimise inventory management. More in-depth analysis of market trends is essential to understanding the causes of price fluctuations, optimising pricing strategies and making more targeted replenishment decisions.

5. Discuss

The LSTM model excels in handling time-series data, and is suitable for data with complex temporal dependencies and data with long time series, as well as for sales over time, which is more in line with the situation of this question. LSTM can selectively remember or forget previous information through a gating mechanism, which makes it able to effectively deal with the problem of long-term memory. In sales forecasting, sales over the past few days or weeks may have an important impact on future replenishment decisions, and LSTM can capture such long-term dependencies. LSTM has been widely used in the field of NLP, including tasks such as language modelling, text categorization, named entity recognition, machine translation, and sentiment analysis. It

can capture long-term dependencies of text sequences and is very effective for processing time-series data in natural language.

References

- [1] Priyadarshi, R., Panigrahi, A., Routroy, S., & Garg, G. K. (2019). Demand forecasting at retail stage for selected vegetables: a performance analysis. *Journal of Modelling in Management*, 14(4), 1042-1063.
- [2] Rahman, M. M., Nguyen, R., & Lu, L. (2022). Multi-level impacts of climate change and supply disruption events on a potato supply chain: An agent-based modeling approach. *Agricultural Systems*, 201, 103469.
- [3] Chen, Q., Qian, J., Yang, H., & Wu, W. (2022). Sustainable food cold chain logistics: From microenvironmental monitoring to global impact. *Comprehensive reviews in food science and food safety*, 21(5), 4189-4209.
- [4] Xiao, M., Zhou, Q., Lu, L., Tao, X., He, W., & Zhou, Y. (2022). Applying Deep Learning-Based Personalized Item Recommendation for Mobile Service in Retailor Industry. *Mobile Information Systems*, 2022.
- [5] Hafez, M. M., Fernández Vilas, A., Redondo, R. P. D., & Pazó, H. O. (2021). Classification of retail products: from probabilistic ranking to neural networks. *Applied Sciences*, 11(9), 4117.
- [6] Raj, M., Gupta, S., Chamola, V., Elhence, A., Garg, T., Atiqzaman, M., & Niyato, D. (2021). A survey on the role of Internet of Things for adopting and promoting Agriculture 4.0. *Journal of Network and Computer Applications*, 187, 103107.
- [7] Ramírez-Zerpa, N., Spelz, R. M., Yarbuhi, I., Negrete-Aranda, R., Contreras, J., Clague, D. A., ... & González-Fernández, A. (2022). Architecture and tectonostratigraphic evolution of the Pescadero Basin Complex, southern Gulf of California: Analysis of high-re

- [8] Yang, M., Lim, M. K., Qu, Y., Ni, D., & Xiao, Z. (2022). Supply chain risk management with machine learning technology: A literature review and future research directions. *Computers & Industrial Engineering*, 108859.
- [9] Singh, S., Mishra, S., Singh, T., & Thakur, S. (2023). A Study on Machine Learning-Based Water Quality Assessment and Wastewater Treatment. In *Artificial Intelligence Applications in Water Treatment and Water Resource Management* (pp. 146-162). IGI Global.
- [10] Meng, D., Iqbal, N., & Zhao, S. (2022). Natural resources environmental quality and economic development: Fresh analysis. *Resources Policy*, 79, 102948.
- [11] Akın Ateş, M. (2022). Sustainability in the textile industry: a quantitative analysis of the role of supply chain disclosure and supply chain complexity.
- [12] Jiang, Y., & Li, X. (2023). Automatic Pricing and Replenishment Decision-Making for Vegetable Commodities Based on Bi-directional Long Short-Term Memory Recurrent Neural Networks and Markov Prediction Models. *Academic Journal of Science and Technology*, 7(3), 69-73.
- [13] Guo, X., & Zhao, G. (2023). Automatic Pricing and Replenishment Strategy for Vegetable Products Based on XGBoost and LSTM Prediction Models. *Highlights in Science, Engineering and Technology*, 70, 615-623.