# Research on Vegetables Sales Profit Based on Machine Learning

Ruichen Liu [1, #], Yinxi Li [1, *, #], Mengdan Li [2, #]

[1] Jinan University - University of Birmingham Joint Institute, Jinan University, Guangzhou, China, 511436

[2] College of Cyber Security, Jinan University, Guangzhou, China, 511436

* Corresponding Author Email: Yinxi.Li0122@outlook.com

[#]These authors contributed equally

**Abstract.** The sales volume of different vegetables in supermarkets is different, and the profit is different. The research on the profit of each category of vegetables can help supermarkets predict the sales revenue more accurately to improve inventory management, purchasing plans, and sales strategy. This can help the supermarket avoid excess or insufficient inventory, improve profits, and help the supermarket identify potential sales risks. The research team obtained daily sales data for various categories of vegetables from a supermarket on the Internet for some time. First, according to the "cost plus pricing" method of vegetables, regression analysis was conducted on the total cost and total profit of various categories of vegetables, and it was found that there was a strong linear relationship, and the profit rate of various categories of vegetables was relatively fixed. They were [0.6629, 0.5915, 0.4663, 0.5121, 0.5596, 0.5381]. Since the sales volume of vegetables has seasonal characteristics, this paper established the ARIMA model to forecast the sales volume of each category of vegetables; and obtained the sales volume distribution of each category of vegetables from July 1 to 7, 2023. The BP neural network model with 22 layers of hidden layers was established, and the purchase unit price data could be obtained according to the known sales data. The R square of the model was greater than 0.7, and the goodness of fit was high. According to the known correlation, considering the impact of the loss rate on the profit and the rationality of the data, a multivariate linear programming model is established to give the replenishment volume and pricing strategy from July 1 to 7, 2023. It is concluded that the maximum profit of the supermarket in these seven days is [845.1835, 870.5028,861.6210, 858.8551, 839.4595, 864.7580, 892.9897]. Through the sensitivity analysis of the profit rate of each category of vegetables, it is concluded that the profit rate increases by 1%, the income will increase by about 1.1%, and the model is robust.

**Keywords:** BP-Neural Network, ARIMA, Regression, Multivariable linear programming.

## 1. Introduction

### 1.1. Background

The fresh food retailer industry faces fierce competition and changing consumer demands. Among them, the operating characteristics of vegetable commodities include short shelf life, variable appearance, and limited sales space. To cope with these challenges and improve competitiveness, supermarkets need to scientifically and effectively formulate replenishment plans and pricing strategies. First of all, the shelf life of vegetable products is short, and its quality will gradually deteriorate with the passage of sales time.

First of all, vegetable products have a short shelf life. As the sales time goes by, their quality will gradually deteriorate. Many varieties cannot be sold if they cannot be sold on the same day. Therefore, supermarkets need to make replenishment decisions for various vegetable categories every day based on historical sales and demand to ensure that there are sufficient products on the shelves and to meet consumer demand.

Secondly, supermarkets must also formulate appropriate pricing strategies to improve profits and market competitiveness. Usually, supermarkets will adopt the "cost-plus pricing" method and offer discounts for goods that are damaged or in poor condition. Therefore, it is very important to understand market demand, including understanding the relationship between sales volume and time of different vegetable categories, as well as supply-side factors. Supermarkets need to reasonably combine different vegetable categories within limited sales space to achieve maximum operating benefits.

The team obtained the commodity information, sales flow details, wholesale price data, and loss rate data of 6 vegetable categories distributed by a supermarket from the Internet. It is necessary to establish a mathematical model to solve the following problems and analyze the relationship between the total sales volume of each vegetable category and the cost-plus pricing. Give the daily replenishment amount and pricing strategy of the vegetable category in the next week (July 1-7, 2023) to maximize the profit of the supermarket.

### 1.2. Research method

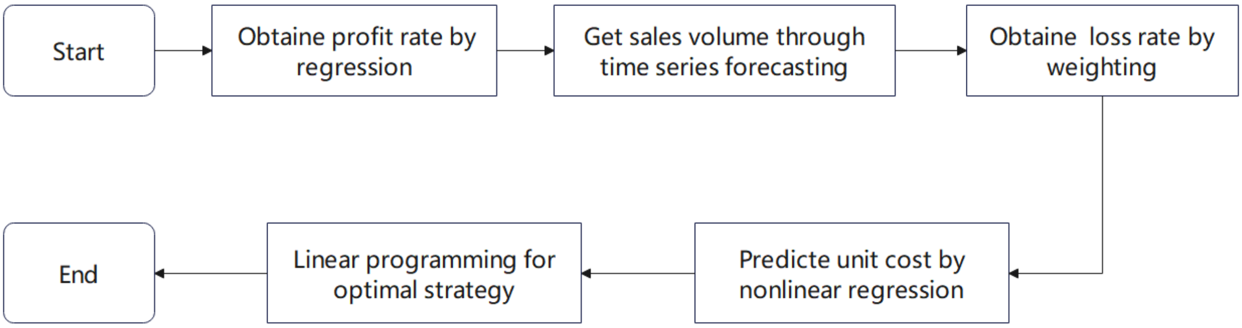


**Figure. 1** Research method

Replenishment strategy and pricing strategy are essentially the calculation of sales volume, profit rate, cost and loss, and other indicators. The research team applies regression analysis [1] to calculate the historical total profit margin for each category, uses time series [2] to predict future sales, weighted wear and tear historical data to predict attrition rates, and uses nonlinear regression to predict individual product costs. These predicted values are used as objective function parameters, constraint conditions are set, and the optimal strategy is obtained by linear programming [3]. The process was shown in Figure 1.

## 2. Data Pre-processing

### 2.1. Index selection

#### 2.1.1 Rate of profit

The price formula (1) of cost-plus pricing is.

$$\text{price} = \text{cost} + \text{cost} \times \text{rate of profit} = \text{cost}(1 + \text{rate of profit}) \tag{1}$$

The relationship between the total cost and total profit of each vegetable category was analyzed by establishing a linear regression model. The total profit is set as, and the total cost is set as (i represents the number of days, j represents the number of types) is set as the total cost. The linear regression is carried out, and the fitting shown in Figure 2 and the regression equation of Flower and leaves (2), Flower vegetables (3), Aquatic rhizomes (4), Solanaceae (5), Peppers (6), Edible mushrooms (7) are obtained:

$$P_{i1} = 0.6629\mathrm{M}_{\mathrm{i1}} + 0.0997, \mathrm{R}^2 = 0.7725 \tag{2}$$

$$P_{i2} = 0.5915\mathrm{M}_{\mathrm{i2}} - 12.8275, \mathrm{R}^2 = 0.7669 \tag{3}$$

$$P_{i3} = 0.4663\mathrm{M}_{\mathrm{i3}} - 4.6856, \mathrm{R}^2 = 0.8651 \tag{4}$$

$$P_{i4} = 0.5121\mathrm{M}_{\mathrm{i4}} + 6.4849, \mathrm{R}^2 = 0.7164 \tag{5}$$

$$P_{i5} = 0.5596\mathrm{M}_{i5} + 12.8424, \mathrm{R}^2 = 0.8249 \tag{6}$$

$$P_{i6} = 0.5381\mathrm{M}_{i6} + 2.4705, \mathrm{R}^2 = 0.8682 \tag{7}$$

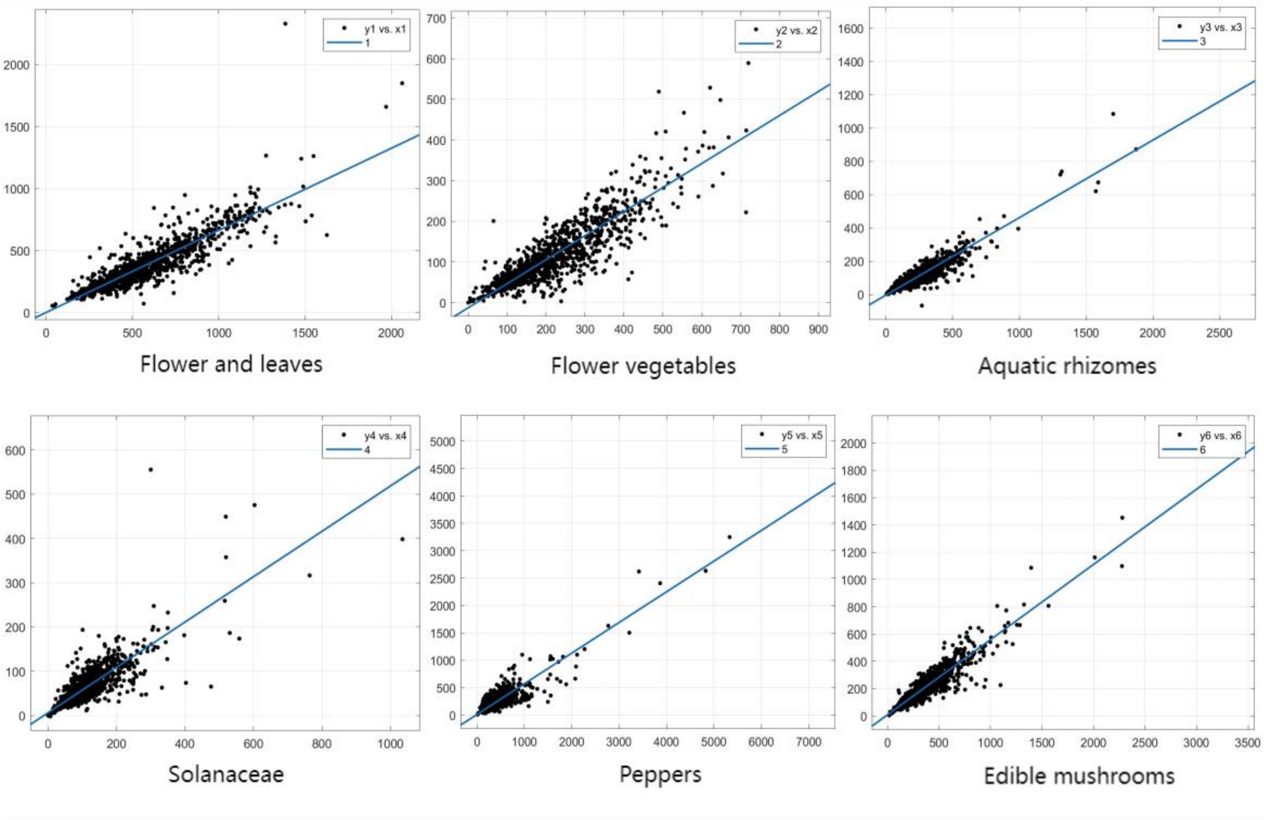


**Figure. 2** Total cost and total profit linear fitting

### 2.1.2 Quantity of sale

Since the shelf life of general vegetable commodities is relatively short, to maximize profits, in an ideal situation, the daily purchase quantity should be exactly equal to the sales volume, so that production and sales can reach a balance. It can be concluded that the replenishment strategy is that the replenishment quantity equals the predicted sales volume. Take the ARIMA model [4] and take advantage of its underlying time patterns by fitting historical data. The ARIMA model can provide a prediction of future sales values.

### 2.1.3 Loss ratio

Due to the certain loss of vegetables in the process of transportation, storage, and sales, to predict the loss of vegetables, the formula (8) is used according to the sales volume and loss rate of each product in each category in the past week.

$$SR_j = \frac{\sum_k S_{jk} \times n_{jk}}{w_j} \quad j = 1,2,3,4,5,6 \tag{8}$$

Where is the loss rate of category j, is the loss rate of item k in category j, is the sales volume of item k in category j, and is the sales volume of category j, so the weighted loss rate of each category can be calculated as Table 2.

**Table. 1** Weighted attrition rate of 6 categories.

| Categories | Flower and leaves | Flower vegetables | Aquatic rhizomes | Solanaceae | Peppers | Edible mushrooms |
|---|---|---|---|---|---|---|
| Weighted attrition rate | 0.085 | 0.076 | 0.075 | 0.074 | 0.13 | 0.107 |

# 3. Model Building

## 3.1. ARIMA models predict sales in the coming week.

ARIMA (Autoregressive Integrated Moving Average Model) is mainly composed of three parts: autoregressive model (AR), differential process (I) and moving average model (MA). It attempts to extract the time series patterns hidden behind the data through the autocorrelation and difference of the data, and then use these patterns to predict future data.

The AR part is used to deal with the autoregressive part of the time series, which considers the influence of observations in the past several periods on the current value, the formula shown in (9).

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + \alpha_t \tag{9}$$

The I part is used to make the non-stationary time series stable. Through the first-order or second-order differential processing, the trend and seasonal factors in the time series are eliminated.

The MA part is used to process the moving average part of the time series, which takes into account the influence of past prediction errors on the current value. The formula is shown in (10).

$$X_t = \alpha_t - \theta_1 \alpha_{t-1} - \theta_2 \alpha_{t-2} - \cdots - \theta_q \alpha_{t-q} \tag{10}$$

After the time series is stabilized by the d-order difference, the ARMA model is established for analysis. After the parameter estimation of the model, the reversibility of the data transformation makes the model parameter estimation result adapt to the data before the stabilization. The model established through this process is called the integrated ARMA model, that is, the ARIMA(p,d,q) model.[5]

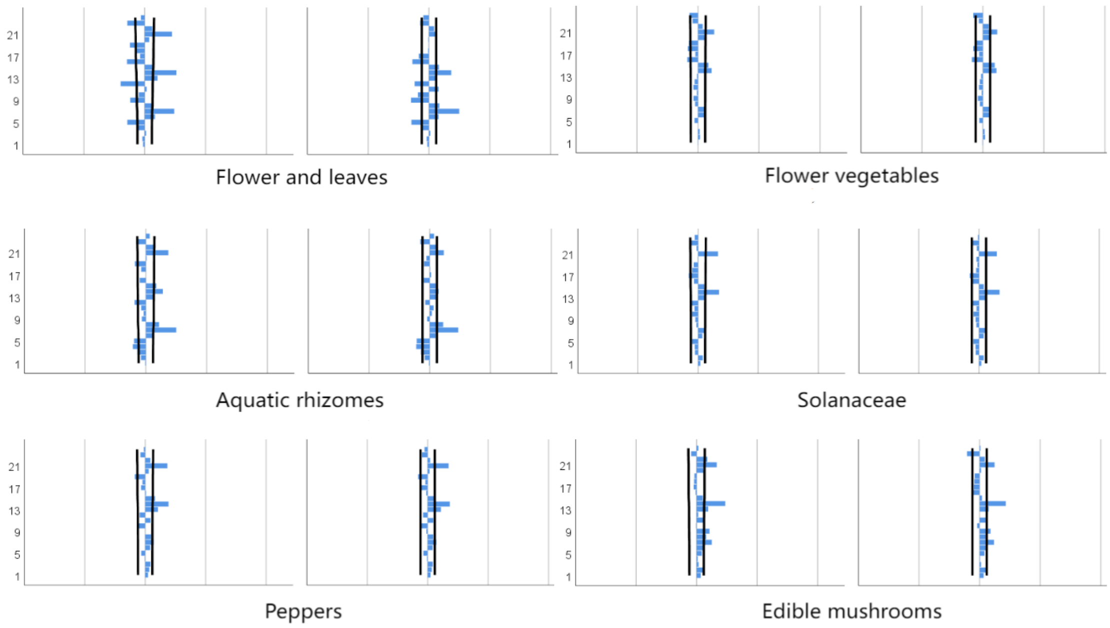


**Figure 3** ARIMA models' ACF (left) and PACF (right) of 6 categories.

Observing the ACF Figure 3, the p value is selected if it is significantly truncated to zero after the lag p; observe the PACF plot If the PACF plot [6] is significantly truncated to zero after the lag q, the q value is selected [7]. The model parameters and goodness of fit of each category are obtained in Table 3, The fitting effect diagram is shown in Figure 4.

**Table. 2** ARIMA models' parameters and goodness of fit

| Categories | Model | $R^2$ | Categories | Model | $R^2$ |
|---|---|---|---|---|---|
| Flower and leaves | ARIMA(4,1,5) | 0.733 | Solanaceae | ARIMA(0,1,9) | 0.605 |
| Flower vegetables | ARIMA(1,0,9) | 0.616 | Peppers | ARIMA(0,0,8) | 0.841 |
| Aquatic rhizomes | ARIMA(1,1,3) | 0.700 | Edible mushrooms | ARIMA(0,0,16) | 0.836 |

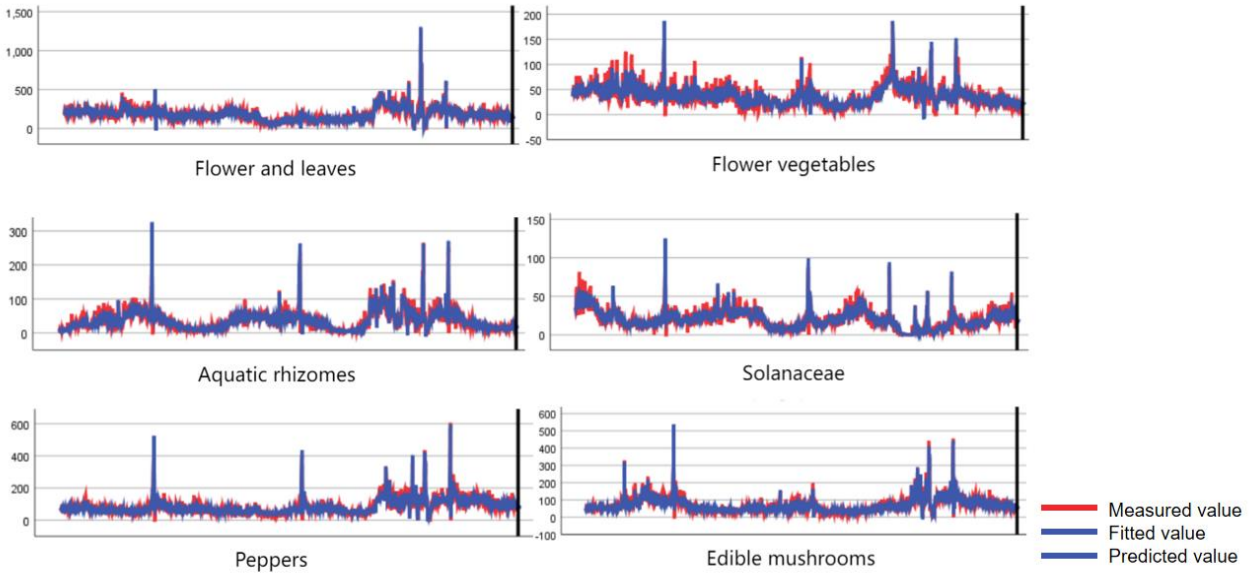


**Figure. 4** ARIMA models' fitting diagrams of 6 categories.

### 3.2. BP-Neural network-the total sales forecast of each vegetable category

BP-neural network [8]错误!未找到引用源。 is an artificial neural network with a back propagation function, which is composed of the input layer, output layer, and hidden layer. The hidden layer often has several layers.

The dimension of input data of the BP neural network model [9]is 7, which is the sales volume and unit cost of 6 kinds of vegetables in the past week. They are used as input data set (dimension 6) and target data set (dimension 1) respectively to form a multi-input and single-output model. After parameter adjustment and repeated training, the number of hidden layers is set to 22 layers to obtain the model, the structure shown in Figure 4.

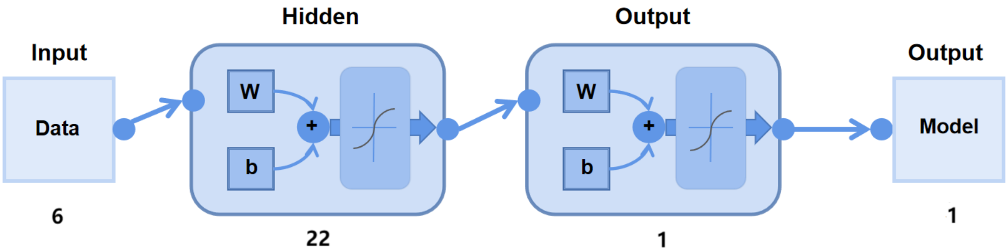


**Figure. 5** BP-Neural network's structure

At the same time, to prevent the model from overfitting, the team will randomly divide the sample data. The ratio of training, testing, and verification data is 70: 15 : 15, and the team uses mean square error to measure network performance. When the number of training times reaches 1000, the iteration stops.

### 3.3. Multivariable linear programming model-profit maximization

The programming goal is to maximize revenue, the objective function is shown in (11).

$$\max \sum_{j=1}^{6} y_{ij} \tag{11}$$

for any $i \in \{1096，1097，1098，1099，1100，1101，1102\}$
for any
Constraints and correlations are (12)

$$
\text{s.t.}\begin{cases}
y_{ij} = P_{ij} - \left(1 - \frac{1}{1-SR_j}\right) \times M_{ij} \\
P_{i1} = 0.6629 + 0.0997 \\
P_{i2} = 0.5615M_{i2} - 12.8275 \\
P_{i3} = 0.4663M_{i3} - 4.6856 \\
P_{i4} = 0.5121M_{i4} + 6.4849 \\
P_{i5} = 0.5596M_{i5} + 12.8424 \\
P_{i6} = 0.5381M_{i6} + 2.4705 \\
M_{ij} = w_{ij} \times m_{ij} \\
0 \le w_{i1} \le 1265.473 \\
0 \le w_{i2} \le 186.155 \\
0 \le w_{i3} \le 296.792 \\
0 \le w_{i4} \le 118.931 \\
0 \le w_{i5} \le 604.231 \\
0 \le w_{i6} \le 511.136
\end{cases} \tag{12}
$$

Among them, it is the sales volume of the j category on the ith day, and the cost unit price solution model of the j category on the ith day. Considering the loss rate, the purchase strategy and pricing strategy can be obtained to solve the problem.[10]

## 4. Results

### 4.1. Future sales forecast of various vegetables

Through the observation of the fitting diagram and the model, the ARIMA model can better predict the sales of various types of vegetables. The results are as Table 4.

**Table. 3** Future sales forecasts of various vegetables

| | Flower and leaves | Flower vegetables | Aquatic rhizomes | Solanaceae | Peppers | Edible mushrooms |
|---|---|---|---|---|---|---|
| July 1,2023 | 134.309 | 23.494 | 19.377 | 23.19 | 76.158 | 43.945 |
| July 2,2023 | 148.82 | 22.163 | 17.235 | 20.418 | 78.488 | 48.094 |
| July 3,2023 | 140.792 | 20.864 | 18.102 | 19.41 | 83.172 | 53.255 |
| July 4,2023 | 147.737 | 21.14 | 17.695 | 18.563 | 81.602 | 56.567 |
| July 5,2023 | 140.856 | 21.714 | 17.886 | 17.528 | 80.022 | 54.734 |
| July 6,2023 | 143.159 | 23.076 | 17.796 | 18.682 | 86.518 | 54.782 |
| July 7,2023 | 144.817 | 24.816 | 17.838 | 20.237 | 89.126 | 54.192 |

### 4.2. Future unit cost forecast of various vegetables

The fitting accuracy of the model is shown in the Figure 6. The parameters of the training set, the test set and the verification set are all greater than 0.73. The predicted curve is similar to the actual curve, indicating that the model has a good fitting effect.

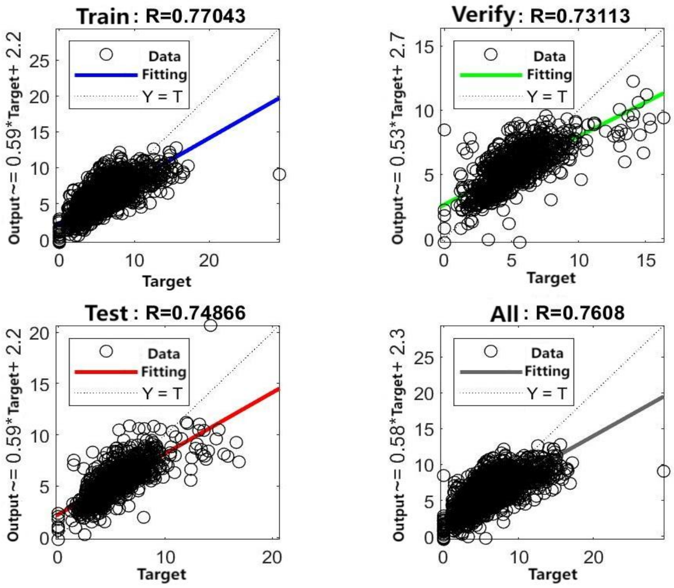


**Figure. 6** BP-Neural network's fitting outcome

In the case of known sales of each vegetable category, the team obtained the unit cost of each vegetable category in Table 5.

**Table. 4** Future unit cost forecast of various vegetables

| | Flower and leaves | Flower vegetables | Aquatic rhizomes | Solanaceae | Peppers | Edible mushrooms |
|---|---|---|---|---|---|---|
| July 1,2023 | 3.7294 | 6.8158 | 7.7106 | 5.6897 | 5.2042 | 5.2254 |
| July 2,2023 | 3.5372 | 6.8488 | 7.6102 | 5.6358 | 4.8396 | 5.0584 |
| July 3,2023 | 3.4926 | 7.0250 | 7.4497 | 5.9269 | 4.8338 | 4.6899 |
| July 4,2023 | 3.4450 | 7.0236 | 7.4113 | 5.7456 | 4.6016 | 4.6216 |
| July 5,2023 | 3.4422 | 6.9444 | 7.2400 | 5.8995 | 4.7281 | 4.6488 |
| July 6,2023 | 3.4341 | 6.8264 | 7.2640 | 5.9853 | 4.6528 | 4.6439 |
| July 7,2023 | 3.4714 | 6.7115 | 7.4220 | 5.9256 | 4.6599 | 4.7190 |

### 4.3. Determination of replenishment and pricing strategy

The linear programming model is solved by the known data, and the future replenishment and pricing strategies for vegetables are obtained in Table 6 and Table 7.

**Table. 5** Future replenishment strategy (kg)

| | Flower and leaves | Flower vegetables | Aquatic rhizomes | Solanaceae | Peppers | Edible mushrooms |
|---|---|---|---|---|---|---|
| July 1,2023 | 146.7874 | 25.4375 | 20.9422 | 25.0515 | 87.5481 | 49.2090 |
| July 2,2023 | 162.6466 | 23.9964 | 18.6271 | 22.0569 | 90.2266 | 53.8550 |
| July 3,2023 | 153.8727 | 22.5899 | 19.5642 | 20.9680 | 95.6111 | 59.6342 |
| July 4,2023 | 161.463 | 22.8888 | 19.1243 | 20.0530 | 93.8063 | 63.3429 |
| July 5,2023 | 153.9427 | 23.5103 | 19.3307 | 18.9350 | 91.9900 | 61.2904 |
| July 6,2023 | 156.4596 | 24.9849 | 19.2335 | 20.1816 | 99.4575 | 61.3441 |
| July 7,2023 | 158.2717 | 26.8689 | 19.2788 | 21.8614 | 102.455 | 60.6835 |

**Table. 6** Future prising ptrategy (YUAN/kg)

| | Flower and leaves | Flower vegetables | Aquatic rhizomes | Solanaceae | Peppers | Edible mushrooms |
|---|---|---|---|---|---|---|
| July 1,2023 | 6.2024 | 10.301 | 11.064 | 8.8830 | 8.2852 | 8.0934 |
| July 2,2023 | 5.8827 | 10.321 | 10.887 | 8.8395 | 7.7114 | 7.8317 |
| July 3,2023 | 5.8086 | 10.565 | 10.664 | 9.2962 | 7.6932 | 7.2600 |
| July 4,2023 | 5.7294 | 10.571 | 10.602 | 9.0373 | 7.3340 | 7.1522 |
| July 5,2023 | 5.7248 | 10.461 | 10.354 | 9.2907 | 7.5345 | 7.1954 |
| July 6,2023 | 5.7113 | 10.308 | 10.387 | 9.3975 | 7.4050 | 7.1878 |
| July 7,2023 | 5.7733 | 10.164 | 10.620 | 9.2806 | 7.4118 | 7.3039 |

### 4.4. Analysis of the model

The model contains important parameter profit rate. If the profit rate fluctuates, it will affect the whole result. Taking the data on July 1, 2023, as an example, if the profit rate fluctuates, the impact on the overall return is shown in Table 8.

**Table. 7** The impact of profit rate fluctuations on earnings

| 98% | 99% | 100% | 101% | 102% |
|---|---|---|---|---|
| 826.996356 | 836.0899658 | 845.1835756 | 854.2771853 | 863.3707951 |

It shows that if the profit margin is increased by 1 %, the income of this category will increase by about 1.1 %. The model has little influence on the model, and the model is robust.

## 5. Conclusions

There is no one-size-fits-all answer to how to determine the best replenishment and pricing strategy in a given situation, as different factors such as demand patterns, product characteristics, market conditions and environmental regulations may influence the optimal decision. However, some general principles and methods can be applied to help evaluate and compare different strategies.

The principle is to use quantitative models and data analysis to support decision-making. To get the most bang for your buck, you first need to find the best forecast for your pricing strategy and replenishment. The profit margin of each category of vegetables is relatively fixed and has a strong linear relationship with the total cost. The profit margin is [0.6629, 0.5915, 0.4663, 0.5121, 0.5596, 0.5381]. Vegetable sales have seasonal characteristics and can be predicted through the ARIMA model. Obtain the sales distribution of various categories of vegetables from July 1 to July 7, 2023. The purchase unit price data can be obtained through the BP neural network model with high fitting goodness. The model has 22 hidden layers and R-squared is greater than 0.7. The replenishment quantity and pricing strategy can be given by a multivariate linear programming model. The maximum profit of the supermarket in these seven days is [845.1835, 870.5028, 861.6210, 858.8551, 839.4595, 864.7580, 892.9897]. The profit margin of each category of vegetables has a positive impact on revenue. For every 1% increase in profit margin, revenue will increase by about 1.1%. The model is robust.

From the entire forecasting process, we can see how sales profit margin is affected. The use and forecasting of ARIMA model, BP neural network model, and multivariable linear programming model can be used as one of the basis for decision-making. In the subsequent use of this type of model, we have been inspired as follows: For special factors such as seasons and holidays that do not entirely come from time, we can try to construct a quadratic temporal treatment of the residuals and add corresponding to the formula. variable representation; for the establishment of models to increase revenue, while reducing costs, we should also understand the economic environment and make corresponding assessments. The more factors considered, the more complex the evaluation model used will be. How to abstract the most simplified The most adaptable machine learning methods will contribute to the treatment of such problems.

## References

[1] Soto, T. (2021). Regression Analysis. In: Volkmar, F.R. (eds) Encyclopedia of Autism Spectrum Disorders. Springer, Cham.

[2] Yperman, J., Becker, T., Valkenborg, D. *et al* (2020). Machine learning analysis of motor evoked potential time series to predict disability progression in multiple sclerosis. *BMC Neurol* **20**, 105.

[3] Garajová, E., Rada, M., Hladík, M. (2021). Integer Programming Reformulations in Interval Linear Programming. In: Cerulli, R., Dell'Amico, M., Guerriero, F., Pacciarelli, D., Sforza, A. (eds) Optimization and Decision Science. AIRO Springer Series, vol 7. Springer, Cham.

[4] Sun, Y., Zhao, Z., Ma, X., Du, Z. (2019). Short-Timescale Gravitational Microlensing Events Prediction with ARIMA-LSTM and ARIMA-GRU Hybrid Model. In: Li, J., Meng, X., Zhang, Y., Cui, W., Du, Z. (eds) Big Scientific Data Management. BigSDM 2018. Lecture Notes in Computer Science (), vol 11473. Springer, Cham.

[5] Wang, M., Pan, J., Li, X (2022). ARIMA and ARIMA-ERNN models for prediction of pertussis incidence in mainland China from 2004 to 2021. In: *BMC Public Health* **22**, 1447.

[6] Li, PhD, H. (2023). Time-Series Analysis. In: Numerical Methods Using Kotlin. Apress, Berkeley, CA.

[7] Xu, G., Xu, Y. (2021). EPFSTO-ARIMA: Electric Power Forced Stochastic Optimization Predicting Based on ARIMA. In: Deng, R., *et al.* Information Security Practice and Experience. ISPEC 2021. Lecture Notes in Computer Science (), vol 13107. Springer, Cham.

[8] Han, JX., Ma, MY. & Wang, K (2021). Product modeling design based on genetic algorithm and BP neural network. In: Neural Comput & Applic 33, 4111–4117.

[9] Huang, Y., Gao, J. (2022). Prediction of Online Learning Resource Demand Based on BP Neural Network. In: Fu, W., Sun, G. (eds) e-Learning, e-Education, and Online Training. In: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 454. Springer, Cham.

[10] Parody, A., Charris, D., Viloria, A., Cervera, J., Hernandez-P, H. (2020). A Method for the Prediction of the Shrinkage in Roasted and Ground Coffee Using Multivariable Statistics. Sengodan, T., Murugappan, M., Misra, S. (eds) Advances in Electrical and Computer Technologies. Lecture Notes in Electrical Engineering, vol 672. Springer, Singapore.