ISSN: 2664-9640

Research on Automatic Vegetable Pricing and Replenishment Decision Based on Long and Short-term Memory Network (LSTM) and Mixed Integer Linear Programming (MILP) Model

Xingxun Cao †, Yuanpeng Hu †, Xiaowei Huang †

Modern Technology College, Mianyang City College, Mianyang, China

† These authors also contributed equally to this work

Abstract

Driven by the wave of the current digital industry, the management of vegetable products in the supermarket is experiencing a revolution. In this paper, by analyzing, fitting and predicting the sales data of the supermarket, this paper provides a scientific and accurate scheme in the direction of replenishment and pricing decision. First, This paper normalized the provided data, conducted the correlation analysis for each category according to the Pearson correlation coefficient, and obtained the visual heat map based on this. This paper then used python to summarize the average sales volume of each category, with the average unit price and the average gross profit. This paper used the long-and short-term Memory Network model (LSTM) to predict daily sales and pricing across categories based on available data. Finally, This paper combine the long and short-term memory network (LSTM) and mixed integer linear programming (MILP) models to develop more accurate replenishment and pricing strategies for vegetable commodities.

Keywords

Time Series Analysis, Long and Short-term Memory Network (LSTM), Mixed Integer Linear Programming (MILP), Shopping Basket Analysis.

1. Introduction

In the modern fresh food supermarket, vegetable products pose special challenges to the replenishment and pricing strategy due to their unique limitation of preservation period and appearance changes. The preservation period of most vegetable products is short, and the appearance becomes worse with the increase of sales time. If certain items are not sold that day, they may not be sold the next day. Therefore, in order to ensure the freshness of the goods and maximize their sales revenue, the merchants need to make precise replenishment and pricing strategies for vegetable products.

In this paper, the sales patterns and interrelationships of vegetable commodities were obtained by analyzing the collected vegetable sales data. The replenishment and pricing strategy model of vegetable commodities was established by Long Short-Term Memory Network (LSTM) and Mixed Integer Linear Programming (MILP), which provides decision-making suggestions for modern fresh food supermarkets.

2. The distribution law and mutual relationship of vegetable commodities

2.1. Data preprocessing

In this paper, we collected sales data from a merchant over the last three years for individual vegetable categories and individual products, and pre-processed the data collected. We integrated the data and divided it into six independent small datasets based on category

ISSN: 2664-9640

information. Each dataset represents a specific category, namely "leafy", "cauliflower", "aquatic root", "eggplant ", "peppers", and "edible mushrooms". On the basis of ensuring data integrity, we further cleaned each dataset to ensure that each dataset was suitable for subsequent analysis and modeling.

2.2. Data analysis

2.2.1. To the category

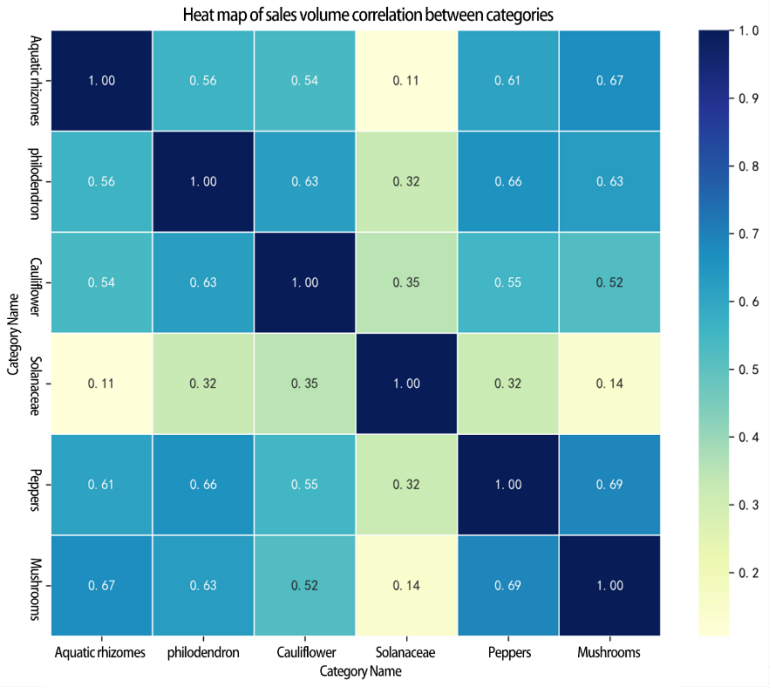
According to the Pearson correlation coefficient, the calculation formula of Pearson correlation coefficient is [1]:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (1)$$

Here are some key points about the Pearson correlation coefficient:

The Pearson correlation coefficient is a measure of the linear relationship between two variables. Its values range between -1 and +1. +1 represents a perfect positive linear relationship, meaning that when one variable increases, the other variable tends to increase as well. 0 means there is no linear relationship. -1 represents a completely negative linear relationship, meaning that when one variable increases, the other variable tends to decrease. The closer the absolute value of the correlation coefficient is to 1, the stronger the linear relationship between the two variables. Conversely, if the correlation coefficient is close to 0, then the linear relationship between the two variables is weak.

The Pearson correlation coefficient measures only the linear relationships. If there is a nonlinear relationship between the two variables, the coefficient may not capture this relationship. Correlation does not imply causality. Even if two variables are highly correlated, it cannot be said that one variable causes a change in the other.



B

Figure 1. Heat map of the Pearson correlation coefficient matrix

Based on Figure 1, it can be concluded that there is a strong positive correlation between aquatic roots and edible fungi and pepper; cauliflower and cauliflower, pepper and edible fungi, and cauliflower and pepper, while the correlation between eggplant and other categories is low.

ISSN: 2664-9640

2.2.2. For single products

Let T be the set of time series of shopping records, where each time t_i corresponds to a shopping event.We define Δt_i as the time difference between i th and i-1 shopping events, i. e., $\Delta t_i = t_i - t_{i-1}$. To determine the boundaries of the basket, we first calculated the average time interval in the current basket Δt,

$$\bar{\Delta}t = \frac{\Delta t_{i-1} - \Delta t_{start}}{i - start} \quad (2)$$

Among them, start is the beginning index of the current shopping basket index.

Next, we define a dynamic time window W, whose size is: W= Δt + random(1,3). When the number of Δt_i>W or shopping records exceeds 5, we believe that a new shopping basket has already started.

To further determine the boundaries of the shopping basket, we used the change point detection method. Let S be a subset of time series in a shopping basket, and we use the ruptures library to look for variable points p in S to maximize the variation in S at p.

Finally, we standardized the number of items in each basket to N, where N is the largest number of basket items. For baskets with less than N, we replaced 0.

Next, we further processed the data to convert it into a binary matrix, where 1 represents the item in the shopping basket and 0 is not included. Then we used the Apriori algorithm to mine the association between items [2]. The specific association rules are follows.

After finding all the sets of frequent entries, we further generate the association rules. The association rule is a form like X⇒Y, where X and Y are sets of goods, and $X \cap Y = \emptyset$.

To evaluate the quality of the association rules, we used the lift degree (Lift) as the evaluation index.

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Confidence}(X \Rightarrow Y)}{\text{Support}(Y)} \quad (3)$$

Where Confidence is the confidence of the rule, indicating the probability that Y is also included in a shopping basket containing X. Support Is the support degree of the item set.

In this study, we set a lift threshold of 1, where only those rules with lift greater than 1 were considered. The degree of promotion is greater than that, and the correlation between goods X and Y is positively correlated, that is, customers who buy X are more likely to buy Y. Here we give the top 10 data with the highest improvement between each item. The results are shown in Table 1.

Table.1. Improvement degree between individual items (top 10)

Previous items	Later items	Elevation degree
Flammulina mushroom (part)	Purple cabbage (2)	5
Hongshan cabbage bolt lotus root assembled gift box	Living tremella	5
Hongshan cabbage bolt lotus root assembled gift box	Living tremella	5
Steak mushroom (box)	Fruit pepper (orange)	5
Steak mushroom (box)	spring cabbage	5
Pumpkin tip	Living tremella	5
Pumpkin tip	Purple cabbage (2)	5
Purple cabbage (2)	Pumpkin tip	5

ISSN: 2664-9640

Living tremella	Pumpkin tip	5
spring cabbage	Steak mushroom (box)	5

Based on this table, we can know that the distribution of the top 10 most correlated items among all the items is shown in the historical data record: a positive association between two items, that is, buying one item increases the probability of buying another item.

3. Forecast and optimize the sales situation of vegetable commodities

Analyze the relationship between total sales of each vegetable category and cost plus pricing, and predict the total daily replenishment and pricing strategy in the next week. In order to solve this problem, we adopted neural network model (LSTM) for prediction and optimization, and conducted hierarchical prediction, integration method, feature engineering and iterative optimization. Here are our detailed analysis steps [3, 4].

3.1. Data exploration

We first explored the data of each vegetable category, and calculated the total sales volume, the average selling price, the average cost, and the average gross profit. To better understand the relationship between total sales and cost plus pricing, we used visual methods such as scatter plots. We define the cost-plus pricing as the difference between the selling price of a commodity and its cost, specifically as the selling unit price (yuan / kg) minus the cost. Scatter plot of the relationship between total sales volume and cost-plus pricing is shown in Figure 2.

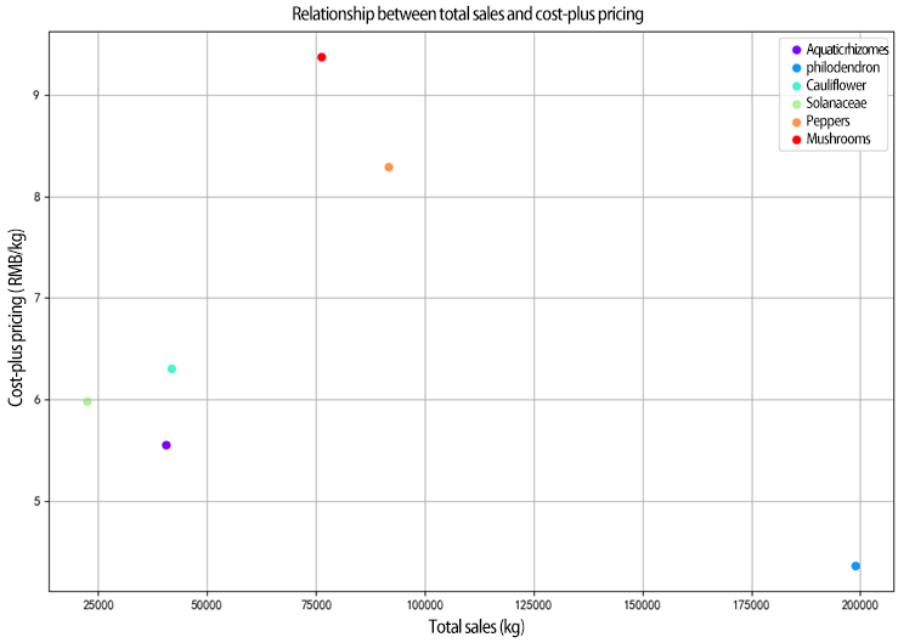


Figure 2. Scatter plot of the relationship between total sales volume and cost-plus pricing

Table 2 calculates the total sales volume, average sales price, average cost and average gross profit of each category. The specific results are shown below.

Table.2. Total sales, average selling price, average cost and average gross profit

Classification name	Sales volume (kg)	Sales unit price(Yuan / kg)	Average cost (Yuan / kg)	Average gross profit(Yuan / kg)
Aquatic rhizomes	40633.751	9.689851143	4.135451593	1.841362592
Flowers and leaves	198798.128	6.317257386	1.957201138	1.297931454

ISSN: 2664-9640

Flower vegetables	41813.103	9.1385376	2.831928313	1.513803266
Solanum	22452.456	8.695458595	2.714480074	1.546950462
Pepper class	91701.595	10.57821641	2.284255886	1.345601559
Edible fungi	76176.727	12.03694146	2.657087726	1.524329092

3.2. Forecast total daily sales

In order to prevent the results of the subsequent correlations, we normalized the assignment of different categories of information above, and because there were no outliers in the assigned data, the maximum and minimum normalization method was chosen. Let the assignment result of different information in each type of index be y_{ij}, then the normalization calculation formula is

$$y'_{ij} = \frac{y_{ij} - \min y_j}{\max y_j - \min y_j} \tag{4}$$

The data of a specific category is selected, and the sales volume is summarized according to "full time" to facilitate the training of LSTM model.

Create a time window

Let D be the time-series dataset, among $D = \{d_1, d_2, \cdots, d_n\}$. To create a time window for the LSTM model, we define a time window $w = \{d_i, d_{i+1}, \cdots, d_{i+6}\}$, Where i is the starting index of the window. Each window w contains 7 days of data, i. e., $|w|$ =7, and was used to predict sales of d_{i+7} on day 8.

And LSTM model construction and training

Let X be the input data, its expression is (m,7,1), where m is the number of samples. We constructed an LSTM model M, which consists of lower layers.

-LSTM layer,

$$f_{LSTM}(X) = LSTM(X) \tag{5}$$

Where the LSTM function represents the operation of the LSTM layer, and its activation function is the ReLU.

-Fully connected layer,

$$f_{Dense}(f_{LSTM}) = Dense(f_{LSTM}) \tag{6}$$

Thus, the output of the model is given for the

$$y_{pred} = f_{Dense}(f_{LSTM}(X)) \tag{7}$$

The loss function L is defined as the mean square error (MSE)

$$L(y_{true}, y_{pred}) = \frac{1}{m}\sum_{i=1}^{m}(y_{true,i} - y_{pred,i})^2 \tag{8}$$

Where, y_{true} is the true sales value and y_{pred} is the predicted value of the model.

Using the Adam optimizer, we iteratively update the model weights to minimize the loss function L.

Based on this model, we predict the total daily sales of each vegetable category in the next week, and give the replenishment strategies. Considering the possible seasonality and trends, we ensure that the model can capture these patterns. The projected sales volume of each type of vegetable is shown in Table 3.

Table.3. Forecast sales volume of each category

	Flowers and leaves	Flower vegetables	Aquatic rhizomes	Solanum	Pepper class	Edible fungi
Monday	152.8874	19.79373	20.42285	23.27763	95.03952	54.13958
Tuesday	110.9889	14.36459	14.82115	16.89291	68.97154	39.29186

ISSN: 2664-9640

Wednesday	96.13208	12.45177	12.83722	14.63165	59.73913	34.03059
Thursday	113.6106	14.70391	15.17126	17.28895	70.60079	40.21797
Friday	124.9617	16.17432	16.68839	19.02115	77.66087	44.23977
Saturday	138.9546	17.98401	18.55562	21.14839	86.35021	49.18967
Sunday	135.4558	17.53159	18.08881	20.61733	84.16786	47.95622

3.3. Pricing strategy

The average selling price per day is first calculated for each category, the data is normalized, and the same training method of LSTM model is used to predict the selling unit price of each category to give the suggested pricing for the coming week. We also consider gross margins, market competition and consumer acceptance to adjust the pricing strategy. Considering that the coming week is a complete week with no holidays, the impact of weekdays and rest days on sales volume cannot be ignored. Ensuring the profitability of a superstore is important, but satisfying the purchasing needs of consumers is always a priority. Over-ambitious returns at the expense of actual sales needs can lead to a number of problems, such as merchandise exceeding its shelf life, insufficient warehouse space, or an inability to meet the purchasing needs of customers. This may not only lead to immediate loss of sales, but may also cause consumers to lose trust in the superstore, thus damaging the superstore's customer base in the long run. Finally, the pricing strategy for each category is obtained as shown in Table 4.

Table.4. Pricing strategy for each category

	Flowers and leaves	Flower vegetables	Aquatic rhizomes	Solanum	Pepper class	Edible fungi
Monday	0.631624	0.487051	0.491309	0.573781	0.64409	0.584361
Tuesday	0.647752	0.522402	0.524329	0.573377	0.60185	0.597598
Wednesday	0.657828	0.523525	0.537349	0.576475	0.589621	0.598996
Thursday	0.645114	0.510706	0.522049	0.572993	0.602605	0.598604
Friday	0.640525	0.504297	0.513282	0.571804	0.610352	0.596339
Saturday	0.638909	0.494764	0.501243	0.571689	0.625462	0.581398
Sunday	0.638388	0.496583	0.503928	0.57154	0.616203	0.592846

4. Make the replenishment plan for the single product

First, the LSTM model is used to budget the sales volume of each item on July 1st. Next, a mixed linear programming (MILP) model is used to develop the best capture and pricing strategy[6, 7].

Decision variable:

- x_i, order volume of goods i, where i belongs to the collection of all goods.
- p_i, price of commodity i.

Parameter:

- s_i, forecast sales volume of merchandise i.
- c_i, The cost of the commodity, i.
- min_display_amount, Minimum display amount of each item.

Objective function:

Maximize total profits, $\max \sum_i (p_i - c_i) \times s_i$

Restrictions:

ISSN: 2664-9640

The total order volume for the goods shall be greater than or equal to 27, $\sum_i \geq x_i \geq 27$.

The total order volume of the goods shall be less than or equal to 33, $\sum_i \leq x_i \leq 33$.

The order volume of each item shall not be lower than its minimum display amount, $x_i \geq$ min display_amount $\forall i$.

The commodity price should be greater than or equal to 0, $p_i \geq 0 \forall i$.

5. Model evaluation and generalization

5.1. Model evaluation

(1) Data integrity, our model is based on data from 2020 to 2023, which gives us a comprehensive perspective on the trends and patterns of vegetable sales. This long-term data set provides us with sufficient information to capture seasonality, trends, and other potential patterns.

(2) Model accuracy, by using time series analysis, Apriori algorithm, long-and short-term memory network (LSTM), mixed integer linear programming (MILP) and other prediction models, we can compare their prediction accuracy and choose the best model. Cross-validation and other evaluation techniques were used to determine model accuracy and robustness.

(3) Multiple retail formats, the model is not only applicable to large supermarkets, but can also be extended to other retail formats, such as convenience stores, specialty stores and online retail platforms.

(4) Cross-category applications, although the focus of this study is on vegetable products, the core logic of the model can be applied to other commodity categories, such as meat, commodities, or electronics.

5.2. Model advantages

(1) Time series prediction, using LSTM, the model can capture long-term dependencies in time series data to more accurately predict future sales trends.

(2) Processing large-scale data, both LSTM and MILP are able to handle large-scale datasets to ensure the stability and accuracy of the model in practical application.

(3) Constraint optimization, through mixed integer linear programming (MILP), the model can find the optimal replenishment and pricing strategy under the premise of satisfying multiple constraints.

(4) Adaptation, the model can be automatically adjusted to new data to maintain the best predictions.

5.3. Model shortcomings

Since the data is from 2020-2023, the outbreak of COVID-19 in 2020 may have a certain impact on the data, the model of word data may consider too many other factors in the model, and there may be some disadvantages if more accurate predictions are needed.

According to deviation, due to the impact of the epidemic, the data from 2020-2023 may be greatly different from the previous data, which may cause the prediction of the model to deviate from the actual situation.

Short-term fluctuations affect long-term forecasts, the outbreak is a short-term emergency, but it may have an impact on long-term data. If simply based on data from this period, predictions may lead to inaccuracy in long-term predictions.

5.4. Model promotion

Real-time data integration, considering the impact of other emergencies, merchants should consider integrating real-time data into the model to respond quickly to market changes quickly.

ISSN: 2664-9640

Multiple factors, in addition to sales data, other factors should also be considered, such as weather, holidays, promotional activities, etc., which may affect consumers' purchase decisions.

User-friendly interface, give merchants a user-friendly interface that allows them to easily enter data, run predictions and view results. This will encourage more merchants to use this model.

Continuous training and support, provide continuous training and support to merchants to ensure that they can take full advantage of the functions of the model and adjust them as needed.

Integration with other systems, consider integrating the model with other merchant systems (e. g., inventory management, supply chain management, etc.) to automate decision making.

References

[1] PAN Pengcheng,LIU Hui,WANG Renming. Adaptive density clustering combined data cleaning for LSTM wind power prediction[J/OL]. Journal of Power Systems and Automation:1-8[2023-09-16].DOI:10.19635/j.cnki.csu-epsa.001341.

[2] Yanping Guo,Yun Gao,Wen Jing. Research on correlation analysis of air pollutants based on Apriori algorithm[J]. Software Engineering,2023,26(09):8-11+32.DOI:10.19644/j.cnki.issn2096-1472.2023.009.002.

[3] Cheng Juanjuan. An empirical study on the relationship between research and teaching in colleges and universities--an analysis based on Pearson's correlation coefficient[J]. Science and Technology in Chinese Colleges and Universities,2022(10):46-52.DOI:10.16209/j.cnki.cust.2022.10.016.

[4] Shi-Ei Wang,Jian Zhou. Application of time series mathematical model in tax analysis[J]. Science and Technology Square,2011(07):150-154.

[5] Jia Runda, Li Zhiqi, Zhang Shulei and so on. Coordinated optimization of the dense dehydration process based on mixed integer linear programming [J/OL]. Control and decision-making: 1-7 [2023-09-10].

[6] Zhang Yi, Zhang Chaoyang, Hu Yunqing, etc. Study on the optimization of integrated energy-saving operation curve of urban rail transit train based on mixed integer linear planning [J]. Control and Information Technology, 2021 (06): 43-50.