Proceedings of Business and Economic Studies, 2024, Volume 7, Issue 1
http://ojs.bbwpublisher.com/index.php/PBES
Online ISSN: 2209-265X
Print ISSN: 2209-2641

Maximizing Supermarket Profits: Data-Driven Strategies for Pricing, Sales, and Forecasting

Wenkang Li*

School of Civil Engineering and Transportation, Northeast Forestry University, Harbin 150040, China

***Corresponding author:** Wenkang Li, dljtxyyhb@163.com

Abstract: The actual circumstances of daily life are crucial for the purchasing and pricing strategies of supermarkets. Developing strategies based on these circumstances can assist businesses in ensuring profits and fostering win-win cooperation. This paper explores methods to maximize profit through purchasing and sales strategies. Initially, the relevant data for various categories of vegetables is integrated. Through histograms, their sales patterns are directly understood, highlighting the most popular vegetables. Upon analyzing each vegetable category, it becomes evident that their sales data do not conform to normal distributions. Therefore, Spearman correlation coefficients are calculated, revealing strong correlations between certain categories, such as aquatic roots and edible fungi. A line chart depicting the top ten selling vegetables indicates a noticeable periodicity. Traditional fitting methods struggle to adequately model the sales of each vegetable category and their relationship with cost-plus pricing. To address this, additional factors such as holidays, weeks, and months are incorporated using techniques like random forest regression. This approach yields cost-plus pricing dependence curves that better capture the relationship, while effectively managing noise. Regarding sales volume prediction, the original data displays significant volatility, necessitating the handling of outliers using the threshold method. For missing data, linear interpolation is employed to mitigate the impact of continuous missing values on prediction accuracy. Subsequently, Adam-optimized long short-term memory (LSTM) networks are utilized to forecast incoming quantities for the next seven days. By extrapolating from normal sales volume, market capacity is estimated, allowing for additional sales through discount strategies. This framework has the potential to increase original income by 1.1 times.

Keywords: Long short-term memory (LSTM); Pricing strategy; Decision making

Online publication: February 25, 2024

1. Background overview

Due to the short freshness period of vegetable commodities, their quality deteriorates over time, prompting supermarkets to replenish their stock daily based on demand and historical sales data. However, businesses often face the challenge of making replenishment decisions without specific information about individual products and their purchase prices. Pricing strategies for vegetables typically employ the “cost-plus pricing” method, with supermarkets often offering discounts to account for phase changes and transportation losses. Accurate market demand analysis is crucial for both pricing and replenishment decisions.

The sales volume of vegetable commodities exhibits a certain relationship with time, particularly with an abundant supply of varieties from April to October, creating a significant sales opportunity for supermarkets. Therefore, establishing a rational sales combination becomes paramount. Mathematical models are employed to address the following issues:

(1) Analyzing the correlation between different categories or individual products of vegetable commodities, and identifying the distribution patterns and interrelationships of sales volume within each category or product.

(2) Investigating the relationship between cost-plus pricing and the total sales of each vegetable category, and devising pricing strategies and daily replenishment plans for each category to maximize profits in the upcoming week.

(3) Proposing relevant data that supermarkets should collect to enhance pricing and replenishment decisions for vegetable commodities, and elucidating how these data can contribute to solving the aforementioned problems.

2. Data handling

Firstly, the table data were sorted, consolidating daily records and organizing them by vegetable type on a weekly basis. This process enables the identification of goods available for sale from June 24 to June 30 (within a week), with accompanying information on profit, pricing, and other relevant data, thereby ensuring comprehensive data processing and maintaining data integrity to the fullest extent.

2.1. Data normalization processing

It is evident that the units of measurement for each index differ, resulting in varying dimensions for different indicators. This discrepancy may skew the importance of certain indicators. To mitigate this effect and facilitate better comparison of analyzed data, enhancing model performance, data normalization is conducted. Normalizing the data confines it within the [0,1] range, with 1 representing the maximum and 0 the minimum value [1]. Here, "meaning" refers to the value corresponding to the *j*-th index, while the normalized value is derived from the maximum and minimum values of the first index.

2.2. Data standardization processing

Standardization, also known as score normalization or standard deviation normalization, ensures that post-standardization, the standard deviation of all features equals 1, and the mean equals 0. This transformation renders the data comparable, simplifying data analysis and interpretation.

2.3. Tabular data processing

To address the second question's requirements, daily sales volumes for each of the six vegetable categories were aggregated and processed. Ensuring cost accuracy, loss rates for each category, considering intermediary losses, were factored in [2]. While specific purchase quantities for each vegetable category couldn't be determined for the second question, rough calculations of the loss rates for each vegetable type were conducted, yielding the consumption rates presented in **Table 1**.

Considering daily discounts on vegetable sales, undiscounted data were supplemented to reflect normal market demand. For discounted items, the maximum feasible sales volume within the market's purchasing power was considered, facilitating discounted sales within this range.

The data processing techniques outlined in this paper, as elaborated in subsequent sections, will not be discussed exhaustively here.

Table 1. Consumption rates of vegetable types

Vegetable types	Attrition rate (%)
Flower vegetables	15.51
Aquatic rhizomes	13.65
Flowers and leaves	12.8
Edible fungi	9.44
Pepper class	9.24
Solanum	6.67

3. Model building

3.1. Model assumptions

(1) Only recent vegetable loss rates are considered, thus replacing loss rates at all time points in this paper.

(2) Market fluctuations are assumed to be minimal, and consumer preferences are presumed stable over short periods.

(3) Malicious competition from other supermarkets is excluded, and pricing is assumed to be relatively stable.

3.2. Problem 1: model establishment and solution

To address the first question, sales data for the six types of vegetables were initially examined to gain a preliminary understanding of their sales patterns.

Analyzing vegetable data from 2020 to 2023 revealed strong periodicity for each vegetable type, with notable correlations among their trends. Notably, foliage plants consistently emerged as the top-selling vegetable each year, as depicted in **Figure 1**, which presents line plots for the six vegetables.

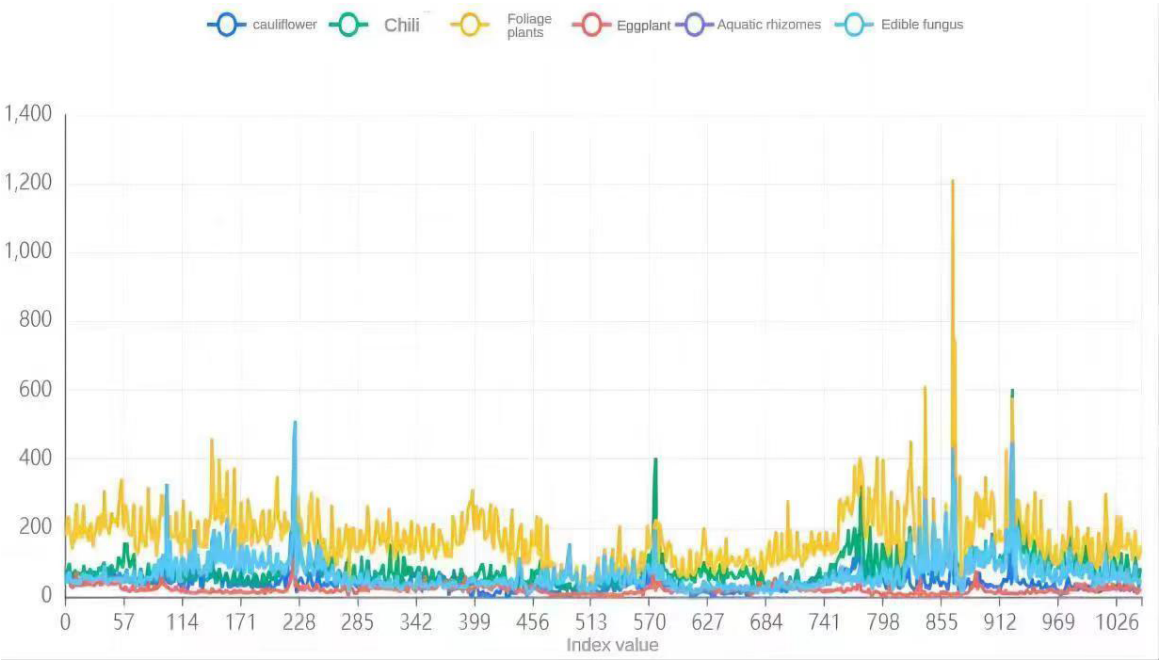

Figure 1. Line plots for the six vegetable types

The data underwent Shapiro-Wilk testing (for small sample sizes, typically below 5,000) or Kolmogorov-Smirnov testing (for larger sample sizes, typically above 5,000). A non-significant result ($P > 0.05$) indicated conformity to normal distribution; otherwise, it is suggested nonconformity, as detailed in **Table 2**.

Table 2. Normality test

Vegetable types	Upper quartile	Average value	Standard deviation	Kurtosis	Shapiro-Wilk test	Kolmogorov-Smirnov test
Flower vegetables	32.37	36.5	1.505	4.33 5	0.908 (0.000***)	0.089 (6e-8)
Flowers and leaves	160.8	170.8	2.991	28.1 47	0.849 (0.000***)	0.067 (0.000107)
Edible fungi	50.38	64.37	3.22	18.6 94	0.753 (0.000***)	0.133 (3e-17)
Pepper class	67.96	79.36	3.562	22.7 55	0.737 (0.000***)	0.147 (5.7e-21)
Solanum	18.643	20.79	1.8	6.64 5	0.885 (0.000***)	0.103 (1.87e-10)
Aquatic rhizomes	50.38	64.37	3.22	18.6 94	0.753 (0.000***)	0.133 (3.6e-17)

Given the non-normal distribution, Spearman correlation coefficient tests were conducted between each vegetable category. Additionally, individual product analysis was performed, yielding sales rankings and top-ten sales line charts. The correlation coefficient heatmap in **Figure 2** illustrates these findings.

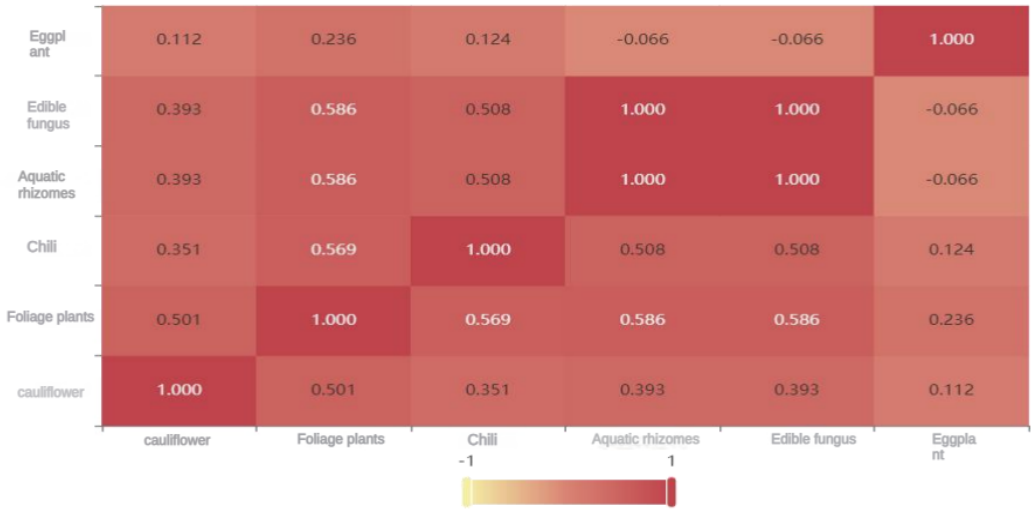

Figure 2. Heat map of correlation coefficient

Furthermore, individual product analysis led to the compilation of sales rankings and the creation of top-ten sales line graphs. These graphs depicted the sales trends of the top ten vegetable commodities, revealing similarities in their periodicity and indicating relatively consistent periodic fluctuations.

Utilizing random forest regression, dependence maps for weeks and months were generated. **Figure 3** illustrates the weekly buying trends, showing similar consumption patterns from Monday to Thursday, increasing on Fridays, and peaking on weekends. Regarding months, July and August emerged as optimal months for vegetable purchases, with heightened consumer activity during these periods.

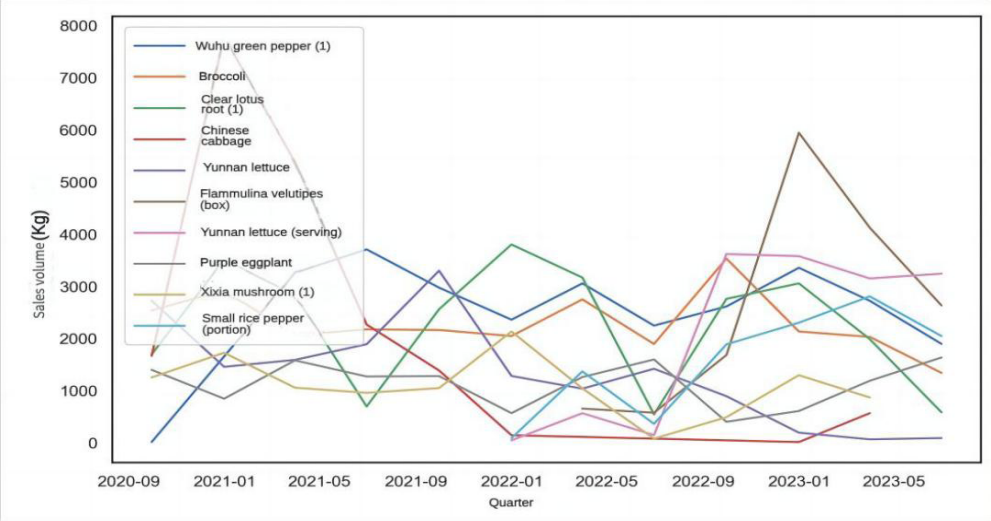

Figure 3. Trends of the top ten items in vegetable sales

3.3. Problem 2: model establishment and solution

3.3.1. Analysis of the relationship between total sales volume and cost-plus pricing

To analyze the relationship between pricing and total sales volume, preliminary steps involve data preprocessing and noise analysis. Particularly during holidays or epidemic periods, pricing and profit may exhibit anomalies, necessitating data screening and processing. As the data often deviate from normal distribution, the boxplot method is employed for outlier detection [3]. Continuous missing values are directly removed to maintain data integrity.

Economic issues, including profit and sales parameters, inherently involve real-life considerations. Following outlier removal, regression analysis is conducted, with logistic regression chosen for its efficacy in economic scenarios, enhancing result accuracy. To further reduce noise interference and tailor analysis to specific circumstances, proportion coefficients are introduced. Additionally, volatility is factored into the regression to enhance accuracy. Due to the extensive data, detailed analysis is omitted [4].

Despite attempts to fit multiple functions, conventional regression analysis yields unsatisfactory results, prompting the utilization of random forest analysis. In tree regression, maximizing information entropy at each node is crucial for variable distinction. Additional variables, such as week, month, and holiday effects, are incorporated to minimize error. The resultant algorithm demonstrates a favorable fit, as evidenced by residual mapping.

A factor dependence diagram is then constructed to illustrate the impact of pricing on sales volume. Notably, the diagram exhibits a “B” type function, indicative of a realistic scenario where high prices deter purchases. The F-test ranks the impact of commodity pricing on sales volume, with epidemic conditions showing minimal influence. The factor-dependence diagram is shown in **Figure 4**.

To delve into the correlation between commodity pricing and sales volume, an F-test was conducted, wherein x_1 represents commodity pricing, and x__1 denotes the impact of the pandemic. The results indicate that commodity pricing exerts a significant influence on sales volume, whereas the pandemic situation has a negligible impact on commodities.

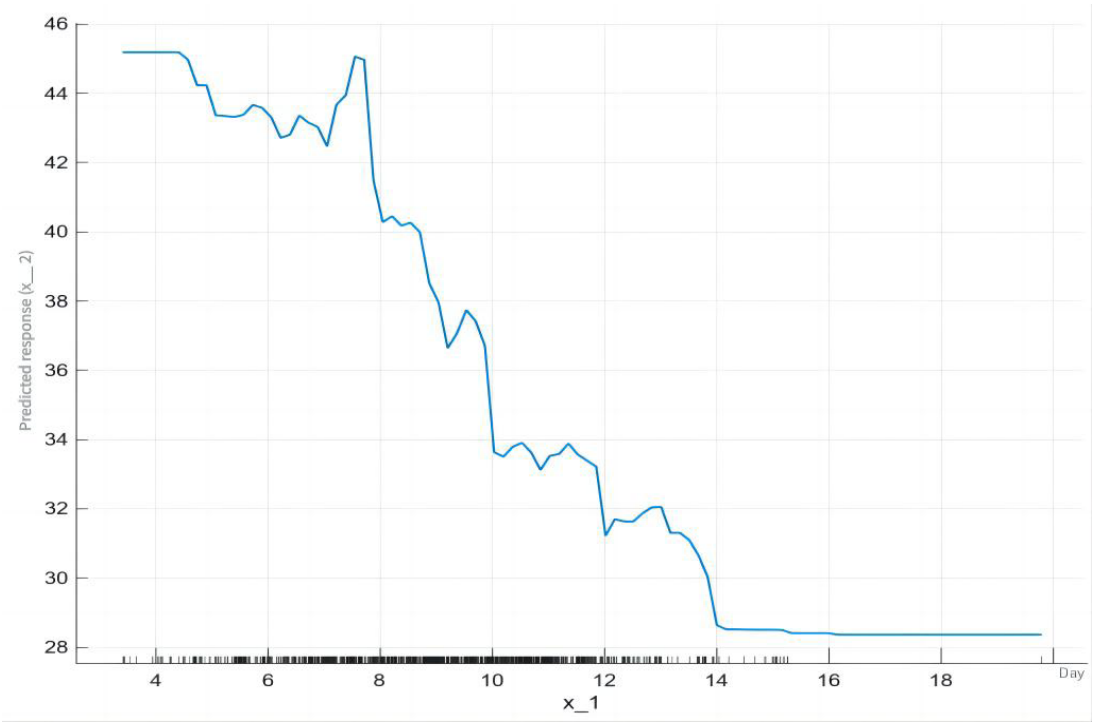

Figure 4. Factor dependence diagram for cauliflower

ANOVA is employed to verify the significance of the overall regression model, assessing whether independent variables significantly affect the dependent variable [6]. The test is based on the following principles:

(1) Hypothesis test: null hypothesis (H_0): all independent variables are equal to zero, indicating no effect of independent variables on the dependent variable.

(2) Alternative hypothesis (H_1): Not all coefficients are zero, implying that at least one independent variable has a significant effect on the dependent variable.

Where SSR represents the sum of regression squares, SSE is the sum of residual squares, *P* is the number of independent variables, and *n* is the sample capacity. The p-value of the resulting F-statistic is then calculated. If the *P*-value is less than the significance level, the null hypothesis can be rejected, and the regression model can be considered significant, indicating that at least one independent variable affects the dependent variable significantly.

When predicting the supply of each vegetable, considering the data's periodic nature and its correlation with past data, this paper employs LSTM time series prediction. LSTM is adept at capturing periodic data and effectively handles long-term dependence issues using a gating mechanism. Utilizing data from 2020 to the present, the model requires corresponding data within the same year to enhance prediction accuracy. For instance, analyzing the sales volume of edible fungi over a period of time reveals a strong periodic pattern.

3.3.2. Forecast of the total replenishment amount

The LSTM network's forgetting gate and input gate play crucial roles in selectively remembering or discarding information. The prediction principle entails forecasting the next unit based on each step's value, iteratively extracting information. Initially, values are inputted into the network for prediction, compared with specific values, and subjected to loss function calculation. Subsequently, feedback output adjusts weights iteratively to refine the network:

$$i = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (1)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2)$$
$$c_t = f_1 c_{t-1} + i_t \, tanh\,(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (3)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \quad (4)$$
$$h_t = o_t tanh(c_t) \quad (5)$$

where i, f, c, and o denote the input gate, forgetting gate, cell state, and output gate, respectively. However, LSTM also presents drawbacks, including high training complexity, parameter adjustment difficulty, strong data dependence, and challenges in result interpretability.

This paper addresses these shortcomings through various strategies: identifying outliers, noise reduction processing, and standardizing data to diminish data dependence and enhance computational efficiency. Parameter adjustment complexity is mitigated through Cartetes product group optimization to determine optimal parameters. Regarding result interpretability, the focus lies on prediction outcomes rather than parameter criteria typical of general time-series predictions. To alleviate training complexity, Adam optimization LSTM is employed. Adam is a stochastic optimization algorithm akin to an optimized version of the gradient descent method.

In summary, LSTM time series prediction optimized with the Adam algorithm proves effective for solving this problem. Due to space constraints, only the prediction effect for a specific vegetable class is exemplified, with the remaining results detailed in the catalog. Notably, the data exhibits significant volatility, particularly during significant holidays, necessitating outlier processing to reduce noise. This paper adopts the three times standard deviation method for noise removal, with an additional refinement step excluding one-eighth of the maximum and minimum values to stabilize the data. Linear interpolation is then employed for data supplementation. Post-noise reduction, prediction accuracy significantly improves. **Table 3** presents the prediction and evaluation indicators before and after noise reduction.

Table 3. Prediction and evaluation indicators before and after noise reduction

Comparison	R^2	MAE	MAPE	MSE
Before	0.55	12.67	0.69	431.40
After	0.69	6.30	0.52	73.44

Abbreviations: MAE, mean absolute error; MAPE, mean absolute percentage error; MSE, mean squared error.

It is evident at a glance that the extreme variance values have significantly decreased after noise reduction. The processed data still retains its periodic characteristics and demonstrates robust temporal attributes.

Mean Absolute Percentage Error (MAPE), an indicator used to assess prediction model accuracy, measures the average relative error between predicted and actual values. A smaller MAPE indicates higher prediction accuracy. A perfect model has a MAPE of 0%, while values exceeding 100% denote inferior models.

A smaller Mean Squared Error (MSE) value signifies more accurate predictions and better alignment with observed values. It reflects the difference between actual observations, predicted values, and the number of observations. **Table 4** presents the predictions based on data from the following days.

Table 4. Predictions for six vegetable categories based on the data of the following days

Flower vegetables	Flowers and leaves	Pepper class	Solanum	Edible fungi	Aquatic rhizomes
25.24909	121.6161	75.48999	23.32983	51.76167	13.8084335
25.63047	120.4669	74.35873	24.19651	52.73499	14.7551184
25.59111	118.5735	73.63853	24.96006	53.94331	15.4822264
25.45245	117.1962	73.15794	25.51992	55.30052	16.1141644
25.3984	116.5814	72.82406	25.92678	56.71239	16.6774807
25.4958	116.5492	72.57187	26.22749	58.12201	17.1898079
25.7385	116.8345	72.35823	26.45825	59.49626	17.6661777

3.3.3. Pricing Strategy

From the table, the purchasing quality can be assessed and instances where certain vegetables are bought in excess can be identified. In such cases, lowering prices becomes necessary to facilitate sales during these periods. Assuming that without price reductions, sales would stagnate, particularly during normal market demand, we can increase goods availability and offer discounts during nighttime hours.

Addressing the second question, the discount relationship can inform the development of a new pricing plan. This plan assumes correct pricing during the day but necessitates adjustments to nighttime sales strategies to maximize profits.

The quantity of excess discounted goods is unpredictable and linked to normal market demand. If daily imports exceed market demand, they must be sold at a discount. Otherwise, they risk being unsold or selling at inappropriate prices. Thus, a new discount strategy is devised in the second question to ensure mall profitability. This involves determining the wholesale quantity of a specific vegetable.

Regarding outlier treatment, significant increases in supply during special events are observed. Excluding outliers is determined by normal range division based on the purchase curve obtained via LSTM, as depicted in the figure below.

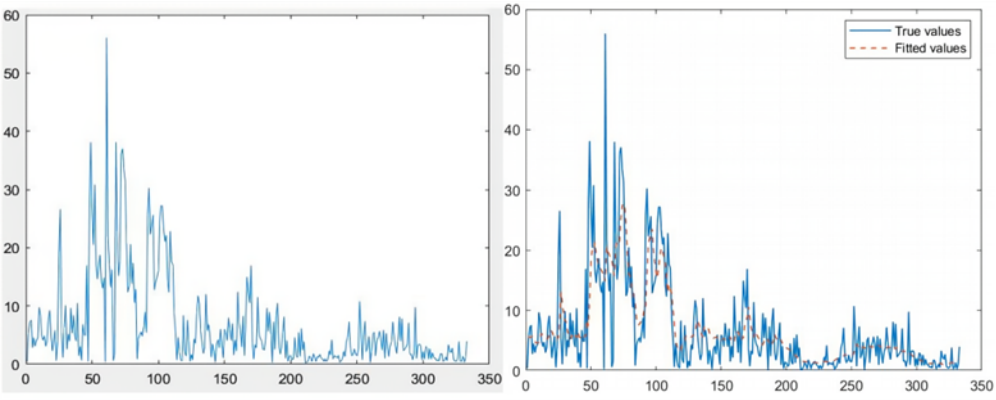

Figure 5. Actual purchase curve (left) and market demand curve (right)

The fitted curve represents the saleable amount within the normal range. Merchants accelerate goods sales through discount processing, organizing daily discounts excluding holidays to ascertain the relationship between discount rates and sales volume. When goods are overstocked, and maximum sales are reached, losses may be incurred. Therefore, studying market demand maximums is crucial. Based on fitting curve analysis

and actual scenarios, it's estimated that market demand will not exceed 20% of the fitting curve. Consequently, daily purchases are calculated as follows: *total vegetable supply = market demand* × (*1+ relative tolerance* (*i.e,* 0.2) ÷ (*1*–lossrate).

The purchase quantities for each vegetable type, accounting for their respective loss rates, are detailed in **Table 5**. When examining the relationship between discount and sales, it's noted that the cost-profit margin typically falls within the 0.2-0.4 range. To expedite sales and prevent loss, this paper sets the cost utilization rate at 0.3. Consequently, the cost-profit margin during sales is approximately 0.5, a notable reduction.

Table 5. Purchase quantity of each vegetable type

Flower vegetables	Flowers and leaves	Pepper class	Solanum	Edible fungi	Aquatic rhizomes
35.86094	167.3616	99.81048	29.99657	68.58879	19.18948489
36.40261	165.7801	98.31476	31.11091	69.87852	20.50508637
36.34671	163.1745	97.36253	32.09265	71.47965	21.51554335
36.14977	161.2792	96.72711	32.8125	73.27807	22.39374323
36.073	160.4331	96.28567	33.33562	75.14893	23.17658001
36.21134	160.3888	95.95223	33.72226	77.0168	23.88855759
36.55604	160.7814	95.66976	34.01896	78.8378	24.55056542

In summary, maximizing daily import supply and offering reduced-price nighttime vegetables can roughly increase profits by about 10%.

4. Conclusion

4.1. Seasonal data

(1) Pricing decision: Seasonal data aids supermarkets in predicting vegetable price fluctuations across different seasons. By utilizing holiday data to adjust prices, businesses can optimize profits during seasonal peaks and troughs. Moreover, seasonal data guides supermarkets in organizing promotional activities tailored to each season, attracting more customers.

(2) Replenishment decision: Seasonal data assists supermarkets in anticipating vegetable demand variations throughout the year. Supermarkets can adjust their replenishment plans according to seasonal demand patterns, ensuring adequate inventory during peak seasons while reducing stock during off-peak periods. This approach minimizes vegetable waste and surplus, thereby reducing losses from expired produce.

4.2. Market competition data

(1) Pricing decision: Monitoring competitors' price data enables supermarkets to gauge the market price level for similar vegetables. This insight aids supermarkets in determining whether to price their products lower or higher than competitors, thereby developing more suitable pricing strategies. Supermarkets can also draw inspiration from competitors' promotional activities to devise their marketing strategies.

(2) Replenishment decision: Market competition data reveals the assortment of vegetables offered by competitors, informing decisions regarding the addition or removal of products. Additionally, competitive data provides insights into competitors' inventory levels, empowering supermarkets to adjust their inventory to meet market demand effectively.

4.3. Purchase and sales time data

(1) Pricing decision: Purchase and sales time data help supermarkets identify optimal supply modes and understand the freshness period of various goods categories. By tailoring pricing strategies to different sales periods, supermarkets can increase profits by adjusting prices during high and low sales periods.

(2) Replenishment decision: Purchase and sales time data enable supermarkets to prevent the wastage of vegetable goods due to expiration. By planning purchase times based on sales data, supermarkets can better manage inventory and minimize losses. Analyzing sales cycles and peak sales periods allows stores to replenish goods appropriately, avoiding inventory shortages or surpluses. Moreover, analyzing sales models across different periods aids in developing more accurate replenishment plans.

4.4. Customer purchase behavior data

(1) Pricing decision: Customer purchase behavior data provides insights into customer acceptance of price levels, enabling supermarkets to optimize pricing strategies to maximize demand and increase sales. Additionally, understanding customer preferences and purchasing habits facilitates personalized pricing and loyalty discounts, enhancing customer loyalty and satisfaction.

(2) Replenishment decision: By analyzing customer purchase behavior, supermarkets can accurately predict demand. For regularly purchased products, supermarkets can align inventory with purchase history to prevent stockouts.

Disclosure statement

The author declares no conflict of interest.

References

[1] Chen R, 2018, Research on the Category and Pricing Optimization Problems Based on the Discrete Selection Model, thesis, Tsinghua University. https://doi.org/10.27266/d.cnki.gqhau.2018.000767

[2] Shao Y, 2018, The Scientificity of Clothing Retail Management Under the Independent Management Mode. New West, 2018(23): 78 + 85.

[3] Xiao C, 2024, Evaluation Method of Hydropower Project Resettlement Effect Based on Random Forest Regression Algorithm. Water Conservancy Technical Supervision, 2024(1): 163–166.

[4] Yang S, Huang J, Yuan J, 2024, Spatiotemporal Prediction Algorithm for Mushroom Growth State Based on Improved LSTM. Journal of Agricultural Machinery, 2024: 1–14. http://kns.cnki.net/kcms/detail/11.1964.S.20240130.1204.004.html

[5] Chen H, Zhang J, Xue S, 2022, Modeling of the Backpack Problem and Experimental Case Design Based on 1st Opt. China Education Technology and Equipment, 2022(12): 133–140.

[6] Qian J, Wang B, Zheng J, et al. 2015, Quantum Evolution Solving Algorithm for the Multiple Quadratic Backpacking Problem. Journal of Computer Science, 38(8): 1516–1529.