Optimal Strategy Research on Vegetable Replenishment and Pricing Based on Machine Learning Regression

Ye Liu *

School of Mathematics, Renmin University of China, Beijing, China, 100872

* Corresponding author: liuye2003@ruc.edu.cn

Abstract. The purpose of this study is to investigate the relationship between sales volume and price in the process of selling vegetables in supermarkets, and in turn, to obtain the optimal replenishment and pricing strategy for the next 7 days in supermarkets under profit maximization. By comparing six machine learning regression algorithms, namely Decision Tree, Random Forest, Plain Bayes, Support Vector Machine, XGBoost and CatBoost, the CatBoost model was found to have the best fit. After model tuning and optimization, the relationship between sales volume and price of each category of vegetables in different quarters was obtained. Finally, the objective revenue function was constructed by considering factors such as pricing, sales volume and profit margin, and the particle swarm optimization algorithm was used to determine the optimal pricing and replenishment strategies corresponding to the maximum revenue. The in-depth analysis and modeling of vegetable sales data provide important references for superstores to formulate reasonable pricing and sales plans in different time periods to enhance operational efficiency and market competitiveness.

Keywords: Machine Learning, Vegetable Pricing, Regression Analysis.

1. Introduction

In today's society, both large enterprises and small traders are faced with an important problem: how to achieve the optimal match between the quantity of goods purchased and the pricing to maximize the total profit. This core issue is of critical importance in the trading of bulk commodities in enterprises and in the daily sales of commodities in superstores. In the face of real-world situations, this core problem will be further expanded to cover a wider range of dimensions. Vegetable vendors and supermarkets, for example, need to take into account seasonal fluctuations in raw material prices to change their pricing strategies, and, more importantly, they must carefully consider how much to buy in order to cope with price declines as vegetables and foodstuffs deteriorate over time. By analyzing the sales data of the past few years, the decision making of supermarkets can be optimized. In this paper, we develop a relevant model to help supermarkets solve the problem of optimal purchasing and pricing schemes in vegetable sales.

Past studies on pricing and replenishment schemes are more abundant, Chen Jun et al. (2023) [1] and Zhao Ling et al. (2022) [2] both obtained the optimal pricing and replenishment levels of fresh agricultural and electronic products by designing heuristic algorithms; Gao Junjun et al. (2019) [3] performed real-time lifecycle learning during the sales process, and based on the results of the learning, determined the replenishment time points and quantities; Jiang Sakumei et al. (2020) [4] applied genetic algorithm to solve the joint inventory and pricing decision problem when the average profit of the cold chain is maximized; Yuyu Li et al. (2022) [5] solved the product selection and pricing strategy corresponding to maximizing the profit when the order is equipped by genetic annealing algorithm. Yuxuan Nie (2024) [6] predicted the price of fresh commodities using an ARIMA optimization model; Huan Qi (2023) [7] optimized the supply chain of two types of cold-chain food products to maximize profit by using a non-convex nonlinear constrained optimization model as well as a mixed-integer nonlinear constrained optimization model; Junxia Wang (2023) [8] provided the optimal solution for replenishment and shelf selection of low-temperature dairy products by using a particle swarm algorithm; Cui Ligang et al. (2023) [9] used an adaptive differential evolutionary algorithm to optimize the pricing of fresh products under the consideration of the effect of preservation technology; Yang Tianshan et al. (2023) [10] found the optimal green distribution technology investment, dynamic pricing and replenishment strategies for perishables based on the Pontryagin's maximum principle.

In this paper, in order to find the optimal replenishment and pricing strategy for vegetables sold in superstores, the relationship between vegetable sales and pricing will be obtained according to the unit of vegetable category. By comparing six machine learning regression algorithms, the CatBoost algorithm with the best fit is selected, and after tuning to the optimal parameters, regression is performed on each vegetable category in quarterly units to obtain the final results. In the process of revenue maximization, it is found that the revenue function includes the wholesale price of commodities, so we first use the ARIMA time series to predict the wholesale price of each category of vegetables in the next seven days, and construct the corresponding target profit function, and then use the PSO algorithm to solve for the optimal replenishment and price strategy.

2. Modeling and solving

2.1. Data analysis

The data used in this article was taken from the website: www.mcm.edu.cn. This website was used to obtain data related to the sales flow and wholesale prices of each vegetable item from July 1, 2020 to June 30, 2023 for a market. Before proceeding with modeling, this paper first conducts a preliminary analysis of the data through descriptive statistics of the data, and the results of the descriptive statistics of cauliflower commodities, for example, are as follows:

Table 1. Descriptive statistics of sales of cauliflower items

Category	Sample size	Maximum	Minimum	Mean	Variance	Kurtosis	Skewness	Coefficient of variation
Cauliflower	1083	186.55	0.632	36.499	522.715	4.346	1.507	0.626

From the above Table 1, it can be seen that there is a large gap between the maximum value and the minimum value of the sales of cauliflower items, which indicates that there may be outliers, and the data should be subsequently processed for outliers before proceeding with the modeling. The kurtosis and skewness coefficients of the data are both positive, indicating that the overall distribution of the data shows a right skewed and spiky pattern.

Since the data did not pass the normality test, this paper considers the IQR quartile distance method to identify outliers when choosing outliers to deal with, and considers points exceeding the upper quartile +1.5 times the IQR distance, or the lower quartile -1.5 times the IQR distance to be outliers. Due to the large sample size, data containing abnormal sales volume, abnormal sales unit price, and abnormal wholesale price are eliminated. The time series were then decomposed using the seasonal decompose function, splitting into three dimensions: trend, seasonal term and random error term. It is found that the six categories of vegetable sales have similar patterns of change with the number of days, all of which show similar trends in each of the three years, presenting the same direction of change and magnitude of fluctuation. And within the same year, the fluctuation with time is obvious, showing seasonal changes.

2.2. Catboost regression model

After pre-processing the data, in order to more accurately get the relationship between sales volume and pricing, the following two special cases need to be dealt with: 1. Discounts: some of the goods in the night due to freshness limitations on the appearance of the quality of the goods, the quality of the deterioration of the situation and have to sell at a discount, and even negative margins will occur. The objective of this question is to analyze the relationship between sales volume and cost-plus pricing, the discounted price is not equal to the original pricing, and it is not possible to know the strength of the discount, so we choose to exclude the discounted samples. 2.Returns: some of the goods returned, and at this time, the sales volume is negative, which means that not only the goods did not sell, but also because of quality problems and so on reasons to the customer to make compensation for the situation can not be avoided completely, should be retained and counted in the total sales. This situation cannot be completely avoided and should be retained and calculated into the total sales.

Further changes in the observed data show that time plays a role on both the demand and supply sides. On the demand side, different types of vegetables mature at different times, and if a certain vegetable is harvested at the right time, then customer demand for it will naturally increase. On the supply side, in the long term, the supply of vegetables is more abundant in April-October, during which time the wholesale price is lower, and the volume of supermarkets' purchases will increase. In the short term, how much supermarkets stock on the same day is affected by the sales situation of the neighboring days, if the supply of a certain commodity exceeds the demand in the previous days, it will increase the amount of the commodity stocked afterwards. In the above analysis, it is obtained that the sales of each category of vegetables show cyclical changes on an annual basis, and at the same time change on a quarterly basis within the same year. Therefore, this paper will analyze the regression analysis of each category on a quarterly basis.

Affected by a variety of factors, the sales volume and pricing of vegetables do not have a linear relationship, so this paper will respectively use decision trees, random forests, plain Bayes, support vector machines, XGBoost, CatBoost six machine learning algorithms to first make a preliminary exploration of the data laws, and then take the optimal. Set the necessary parameters, initialize the six machine learning models, take the cauliflower class data as an example for regression, the results are shown in Table 2 below.

Table 2. Regression results of different algorithms

Regression Algorithm	MSE	MAE	R^2
Decision tree	84.83	5.42	0.85
Random Forest	130.23	8.38	0.77
Plain Bayes	558.87	18.03	0.02
Support Vector Machine	510.81	16.99	0.11
XGBoost	71.92	5.03	0.87
CatBoost	69.78	4.47	0.88

Upon comparison, the CatBoost model has the smallest MSE and MAE and the largest R2. Although it has not been specifically tuned, this model has achieved the smallest error, the best fit, and the best overall result in the side-by-side comparison. Therefore, CatBoost model is selected to continue the optimization of tuning parameter.

CatBoost is a gradient boosting algorithm that is particularly suitable for processing categorical features. It can automatically process category features, use symmetric tree structure and alignment selection techniques to reduce the risk of overfitting, and also improve the training speed through histogram optimization. Overall, CatBoost provides an efficient and adaptable gradient boosting algorithm through these features, making the model more stable and accurate.

In this paper, we use the grid search function that comes with the CatBoost regression model to find the optimal parameters. First set the value of the larger interval option, combined with the model score and the above three indicators scores continue to refine the granularity, and ultimately establish the optimal model, and each category according to the four quarterly division of the regression to get the final results. The selection process is shown in the figure 1:

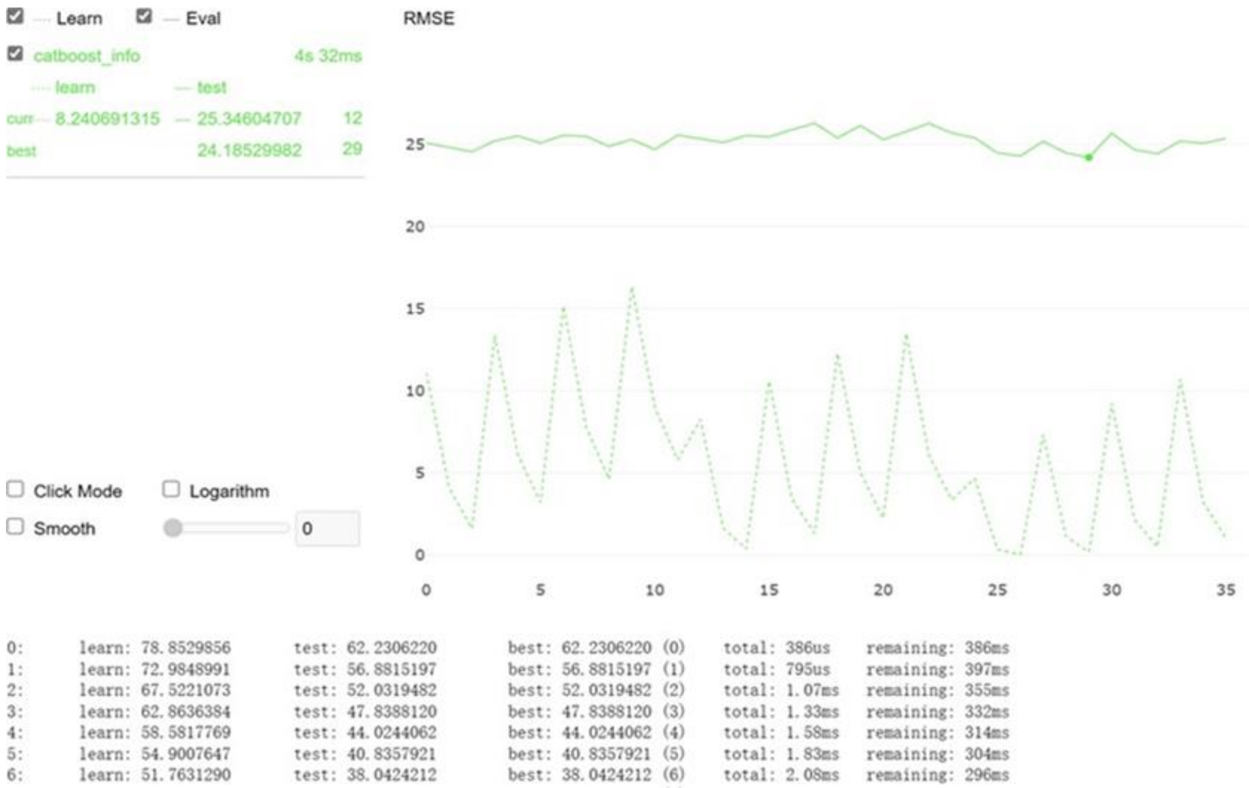

Figure 1. Diagram of CatBoost optimization selection process

2.3. ARIMA time series forecasting model

From the above, it can be seen that time also has an impact on the supply side, and therefore the feed price of each vegetable commodity also varies with time, and this variation is mainly dominated by market relations, and merchants are not free to determine the feed price. Therefore, in order to construct the objective revenue function, it is necessary to predict the future July 1-7 feed prices. In this paper, we use the ARIMA time series model to predict the wholesale prices in the coming week, while in the ARIMA time series forecasting model, we choose to use a combination of first-order differencing and ARIMA time series forecasting to forecast the total daily replenishment of each category of vegetables in the coming week.

The ARIMA model combines the three components of autoregression, differencing, and moving average to cope with the characteristics of different time series data, and to forecast and model data with some regularity but with noise and trends. The following is the expression of the ARIMA model, where y_t is the difference series, p denotes the order of the autoregressive model, d denotes the order of the difference, and q denotes the order of the moving average model.

$$y_t = \mathrm{c} + \phi_1 y_{t-1} + \cdots \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \tag{1}$$

By adjusting the parameters in the ARIMA model and combining the data from July 1, 2020 to June 30, 2023 for forecasting, the next 7 days into the price trend of each category of vegetables can be obtained, and the results are shown in Table 3 below:

Table 3. Forecast results of input prices by category

Category	Day1	Day2	Day3	Day4	Day5	Day6	Day7
Cauliflower	8.03	7.94	8.09	7.98	7.83	7.97	8.13
Foliage	2.91	3.08	3.17	3.14	2.89	3.06	3.02
Pepper	4.84	4.75	5.17	4.99	4.88	5.09	4.48
Eggplant	5.00	4.76	5.02	5.13	5.15	4.87	4.71
Mushrooms	7.31	7.26	7.29	7.52	6.69	7.67	8.17
Aquatic Roots	13.36	13.01	13.04	13.02	12.66	12.34	11.42

2.4. Particle swarm algorithm single-objective optimization model

Since most vegetable commodities cannot be sold every other day, when there is a replenishment quantity that is larger than the actual sales quantity on the same day, the superstore will use the strategy of discounting to sell the excess commodities. In order to maximize the revenue, this paper assumes that the superstore can sell all the commodities by discounting, in which case the relationship between the replenishment volume and the sales volume exists as follows:

$$x = \frac{SaleQ}{1-\eta} \tag{2}$$

Where *SaleQ* denotes the sales volume of the day, x denotes the replenishment volume of the day, and η denotes the average wastage rate. Vegetable commodities in the sale of the attrition rate is affected by a variety of factors, in a short period of time without sudden changes in the situation, this paper adopts the method of taking the average of the attrition rate in the past three years of data to estimate the attrition rate of commodities in the next 7 days. Then the revenue function is constructed as follows:

$$P = x \cdot (1-\eta) \cdot y - x \cdot cost \tag{3}$$

Where P represents the total revenue, y represents the pricing, and *cost* is the inlet price (the result of ARIMA prediction in the previous section).

In this paper, we will take the cauliflower goods as an example to solve the optimal strategy of replenishment and pricing for the next 7 days, and the results are obtained by finding the average attrition rate and predicting the incoming price using the ARIMA model as:

Average attrition rate: η = 0.1048272694≈0.105

Wholesale Prices: x = [8.03, 7.94, 8.09, 7.98, 7.83, 7.97, 8.13]

Since there is only one objective function, the particle swarm algorithm for single-objective optimization is used in this paper. The particle swarm optimization algorithm is a simplified algorithmic model inspired by the regularity of birds' foraging behavior, which has a fast convergence speed, a small number of applied parameters, and its algorithm itself is simple and easy to implement. The main idea is that each particle in the particle swarm searches for the maximum value of the objective function along many different routes in the space, each particle will search along the direction of its own search and record the data, at the same time, each route will share its own data, so that the whole swarm of particles can know where the maximum value is in the cycle, at the same time, the particles will adjust the next search route according to the existing data, and then after traversing the data, the final maximum value is obtained. the final maximum value. Before solving the optimization problem, this paper will add the following constraints to the objective function in combination with the actual sales situation:

1. The overall goal of economic activity is to maximize profits, so to ensure positive returns need to meet the pricing higher than the wholesale price. That is: commodity pricing $y \geq$ purchase price *cost*

2. When pricing, you need to take into account the competition in the whole market, and you can't set the profit margin too high, i.e., you can't price it too high, and here we use the maximum profit margin of the past three years after removing outliers to limit the upper limit of pricing. That is: $y \leq cost(1+\beta)$, where β is the maximum profit margin in the past three years, after substituting the value of cauliflower goods the expression is: $y \leq 1.58cost$.

3. In the sales process, due to the space constraints of the superstore, each replenishment cannot be too much. However, it is too complicated to calculate the space occupied by the goods, this paper takes the maximum sales volume of the superstore over the past three years as a constraint, and requires that the replenishment volume of each category cannot exceed the historical maximum sales volume of each category when discussing the subcategories as a limitation of the maximum capacity of the superstore. That is: the replenishment quantity $x \leq$ previous years maximum sales volume *SaleQMax*, after substituting the value of the cauliflower category goods the expression is: $x \leq 186.55$.

Considering the above three constraints together, the optimization problem of the objective return function in this paper is represented as shown below:

$$\max\{p = x\cdot(1-\eta)\cdot y - x\cdot cost\} \tag{4}$$

$$s.t.\begin{cases} \mathrm{y} \ge cost \\ y \le 1.58\cdot cost \\ x \le 186.55 \end{cases} \tag{5}$$

In addition, we note that in reality, the pricing of vegetable commodities is either integer or one decimal, so we need to re-optimize the expression of the objective function, the pricing will be retained one decimal and then brought to the objective function, so that the maximum gain obtained corresponds to the pricing retained one decimal can simultaneously satisfy the restrictions on the maximum gain and the number of decimal digits.

The PSO particle swarm algorithm is used to solve this single-objective optimization problem for each category separately, take cauliflower category as an example, set the initial number of particles to 150, the acceleration coefficient is set to 2, the inertia weight is set to 0.9, and the replenishment and pricing corresponding to the maximum revenue in the next seven days are obtained after 1000 iterations as shown in Table 4 below:

Table 4. Sales strategies for cauliflower items for the next 7 days

Date of Sale	Wholesale Price	Optimal Pricing	Optimal Volume	Maximum Revenue
2023-07-01	8.03	12.7	55.568	163.92
2023-07-02	7.94	12.5	55.639	159.76
2023-07-03	8.09	12.7	55.568	160.18
2023-07-04	7.98	12.6	55.639	162.08
2023-07-05	7.84	12.3	54.74	153.10
2023-07-06	7.97	12.6	55.639	162.98
2023-07-07	8.13	12.7	55.568	158.20

Considering the sales strategy of the six categories of goods, we can briefly summarize the replenishment strategy and pricing strategy of the superstore in the next 7 days as follows:

1. Replenishment strategy: taking into account the vegetables are mass consumers just need commodities, should be fully integrated with the market demand, more into the leafy, chili vegetables, the rest of the proportion to maintain a relatively balanced. Percentage of replenishment of goods by category is shown in figure 2.

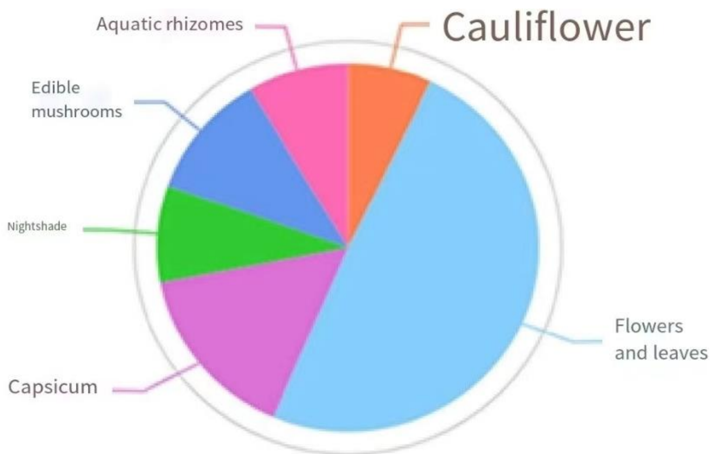

Figure 2. Percentage of replenishment of goods by category

2. Pricing strategy: Prices should be maintained as stable as possible to control price fluctuations. According to the cost of goods and market supply and demand to develop a reasonable, friendly prices, the main thin margins and more sales.

3. Conclusion

With the rapid development of social economy and the improvement of people's living standards, China's consumer market has shown continuous active and steady growth. And fresh products as the immediate consumer goods to protect people's livelihood and maintain social stability, the market size is huge. Vegetables, as one of the indispensable foods in people's daily life, have high-frequency and refined consumption characteristics, and the market prospect is broad. However, due to the natural attributes of vegetables such as cyclicality, seasonality, perishability and vulnerability, coupled with the many varieties of vegetables and the different origins of the purchase cost of different, so a reasonable pricing and replenishment decisions for businesses to achieve the business objectives of lower costs and maximize profits is critical. This paper is based on a scientific and systematic mathematical method to study and analyze the sales process of vegetables, and optimize the pricing and replenishment strategy of vegetables in the coming week through the sales data of the supermarkets over the past week, so as to help the supermarkets to obtain the maximum profit. In this paper, we discuss vegetables by category to get the distribution law that vegetables have a yearly cycle and change quarterly in a year, and use CatBoost regression to get the relationship between sales volume and selling price of vegetables in each category in the four quarters, and construct the target revenue function under the assumption that merchants are able to sell all the commodities through discounts, and then finally use particle swarm algorithm combined with the regression results to get the maximum revenue under the corresponding replenishment strategy. Finally, the particle swarm algorithm is used to combine the regression results to obtain the replenishment quantity and pricing corresponding to the maximum return, thus providing a strong reference for the merchants' replenishment choice and pricing strategy in the coming week.

References

[1] Jun Chen, Sha Kang. Joint decision making for pricing and inventory replenishment of agricultural products for dual-channel sales [J]. Industrial Engineering, 2023, 26 (03): 39 - 46.

[2] Ling Zhao, Zhixue Liu. Joint replenishment and pricing strategy considering customer returns and fixed costs [J]. Operations Research and Management, 2022, 31 (06): 105 - 110+124.

[3] Junjun Gao, Yu Chen. Dynamic inventory and pricing integrated decision making based on life cycle learning[J]. Journal of Shanghai University (Natural Science Edition), 2019, 25 (05): 807 - 816.

[4] Yingmei Jiang, Jinjin Mou. Joint decision making of inventory and pricing of fresh processed products based on perceived freshness [J]. Highway Transportation Science and Technology, 2020, 37 (03): 151 - 158.

[5] Yuyu Li, Bo Huang, Hui Huang. Product selection and pricing strategies for build-to-order assembly manufacturers based on heterogeneous demand [J]. Computer Integrated Manufacturing Systems, 2022, 28 (07): 2263 - 2272.

[6] Yuxuan Nie. Research on automatic pricing and replenishment strategy of fresh commodities based on ARIMA prediction optimization model--taking vegetable commodities as an example [J]. Commercial Exhibition Economy, 2024 (05): 19 - 22.

[7] Huan Qi. Research on optimization model and algorithm of cold chain food supply chain management strategy empowered by blockchain [D]. Central South University, 2023.

[8] Junxia Wang. Joint decision making for dairy shelf replenishment and zonal display considering different replenishment modes[D]. Chongqing Jiaotong University, 2023.

[9] LiGang Cui, Yali Li, Jinxing Liu, et al. Joint decision making of multi-product replenishment and pricing considering investment in preservation technology [J]. Industrial Engineering and Management, 2023, 28 (03): 17 - 26.

[10] Tianshan Yang, Gonglin Yuan. Research on dynamic pricing and replenishment strategy of perishable goods considering green distribution technology investment [J]. China Management Science, 2023, 31 (11): 279 - 287.