# Automatic Pricing and Replenishment Decisions of Vegetable Products Based on Machine Learning

Huifeng Zhang *, ZihengYan

School of Information and Communication Engineering, Guangzhou Maritime University, Guangzhou, China

* Corresponding Author Email: Zhang_hf_2022@163.com

**Abstract.** The objective of this paper is to address the replenishment challenge caused by the limited shelf life of vegetables and maximize the profit potential of a diverse range of vegetables, based on extensive data analysis encompassing vegetable sales rates and gross profits. In this study, machine learning techniques, including the Spielman correlation coefficient, are employed to derive the demand price regression curve and R^2 value. Subsequently, a random forest regression model is utilized for training and predicting data, while an ARIMA model based on time series data is applied for individual data processing. The expected return under current prices is calculated iteratively until convergence is achieved in terms of optimal expected returns, best prices, and ideal replenishment strategies. The stability of the data is assessed using the ADF test with satisfactory results obtained. This research yields numerous computational outcomes through machine learning methodologies. Among these findings, broccoli emerges as the most profitable vegetable with the highest purchase quantity and sales volume; meanwhile, Yunnan lettuce stands out as having the largest catch quantity. Demand exhibits a predominantly negative correlation with price; furthermore, after incorporating constraints such as "minimum packaging amount," medium-priced dishes yield maximum profitability. These findings hold significant value and practical implications for real-life problem-solving.

**Keywords:** Autoregressive Integrated Moving Average model, Random Forest Regressor, Spearman's rank correlation coefficient.

## 1. Introduction

With the continuous improvement of social productivity and the increasing market competition, the type and quality requirements of fresh vegetables, fruits, and aquatic products are getting higher and higher [1]. The freshness of products will directly affect customers' desire to buy, and the value of fresh agricultural products is reflected in the degree of freshness, the higher the degree of freshness, the higher the economic value. The higher the level of damage and corruption, the lower the value of its economy [2-4]. As a perishable product with fast deterioration, the stock loss and waste are very serious. Whether it is the inventory imprecision caused by the quantity decay or the product obsolescence caused by the quality decay, it will bring losses to retailers and cause a decline in earnings [5]. By analyzing the historical sales and demand data of each commodity, the supermarket can better understand the market demand, adjust the inventory level, formulate a reasonable price strategy, and provide high-quality vegetable commodities to consumers, to improve sales efficiency and customer satisfaction [6-7].

To solve this problem, models have been proposed in the past as tools that can manage perishable stocks in general, and food in particular, by comparing the expected maximum profit of alternative commodities, determining which commodities differ in purchase cost and shelf life, and deciding which products are more profitable [8-10]. In addition, this model is used to calculate the retailer's profit loss to solve the problem of the retailer ignoring the decline in demand rate caused by the loss of freshness, and the demand function related to price is used to assist the calculation [11]. However, this needs to be based on many experimental data and rely on many functions to assist, the engineering amount is large, the type of goods targeted is relatively narrow, and the wide variety of fresh products today, the focus should not only be on vegetables, but also consider the profits of other fresh products, and other factors affecting freshness are not thoughtful enough.

We consider the establishment of a regression simulation model to analyze the relationship between the sales volume pricing of a single product and the surplus revenue and establish a target planning model to simulate and solve the total amount of daily replenishment and pricing. Total profit and cost profit rate are calculated using the methods of (selling price - wholesale price) * total sales volume and (selling price - wholesale price) / wholesale price respectively. Therefore, wholesale prices are obtained through analysis based on available data (https://cumcm.cnki.net/), and the item on the corresponding date is selected, excluding returned goods and discounted goods. Construct its items' cost, price, volume, and cost margin. The ARIMA model was used to predict the sales volume of different products in the integrated table, and a random forest regression model was established. The features and target variables of the training set were introduced into the model for training by invoking the fit method. The goodness of fit of the model on the verification set was calculated by using the scoring method, and another random forest regression model was constructed. The entire data set is trained so that the model can finally be applied. Then, Min Max Scaler is used to normalize the trained data set and store the normalized data. Parameters (p, d, q) of the ARIMA model are determined by AIC, different parameter combinations are tried by nested loops, and AIC values corresponding to each parameter combination are stored. The parameter combination with the minimum AIC value is selected as the parameter of the model, and the minimum AIC value is printed to find the best pricing strategy and replenishment quantity under certain conditions.

## 2. Model

### 2.1. Random Forest Regression Model

Random forest regression is an ensemble learning-based algorithm that performs regression tasks by building multiple decision trees and integrating their predictions. In a random forest, each decision tree is independent and random, such as the randomness of sample sampling and the randomness of feature sampling, and is trained on randomly selected subsamples, which can effectively reduce the risk of overfitting. The random forest averages or weights the predictions of multiple decision trees to get the final regression result. The principle of the random forest regression model is shown in Figure 1.

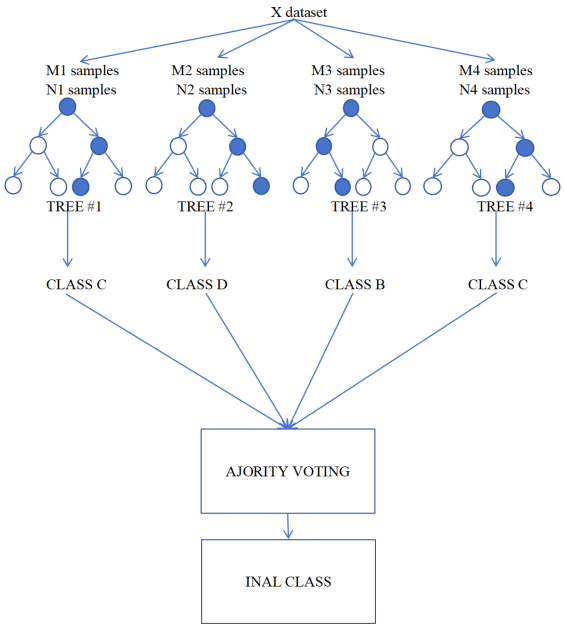


**Figure 1.** The principle of the random forest regression model

The basic principle of random forest regression generally consists of four points, which are:

(1) Random selection of samples: A subset of samples is randomly selected from the original training set to form a subsample set. This allows each decision tree to be trained on a different set of samples, increasing the diversity of the model.

(2) Randomly selected features: For each node of each decision tree, only a subset of randomly selected features is considered when selecting the best partition features. This can prevent certain features from having too much influence on the whole model, thus improving the robustness of the model.

(3) Build a decision tree: Build a decision tree on each subsample set using some kind of decision tree algorithm, such as the CART algorithm. In the process of decision tree growth, the best partition features are usually selected recursively to divide the data set into the least impure subset.

(4) Integrated prediction: For a new input sample, the final regression result is obtained by averaging or weighted averaging the predictions of multiple decision trees.

This implementation adopts the exhaustive method, that is, traversing each feature and all the values of each feature, and finally finding the best segmentation variable and segmentation point from it, the calculation formula is as follows:

$$G(x_i, v_{ij}) = \frac{n_{left}}{N_s} H(X_{left}) + \frac{n_{right}}{N_s} H(X_{right}), \tag{1}$$

where $x_i$ is the Shred variable, $v_{ij}$ denotes a segmentation value of the segmentation variable, $X_{left}$ and $X_{right}$ represents a training sample set divided into left and right child nodes, H(X) is a function used to measure node impurity, such as impurity, function, and criterion, and different impurity functions are generally adopted for classification and regression tasks. The training process function formula of a node is as follows:

$$(x^*, v^*) = argmin_{x,v} G(x_i, v_{ij}), \tag{2}$$

The formula for a function of some points is as follows:

$$G(x, v) = \frac{1}{N_s}\left(\sum_{y_i \in X_{left}} (y_i - \bar{y}_{right})^2\right). \tag{3}$$

where $n_{left}$ is the number of training samples of the left sub-node after segmentation, $n_{right}$ denotes the number of training samples of the right sub-node, and Ns represents the number of all training samples of the current node.

### 2.2. Basic principles of ARIMA time series model

The known data in this paper contains a large amount of sales time, so the ARIMA model is used to predict wholesale prices based on time series data. First, relevant data are screened out in the data set through classification names and the feature column to be predicted, namely weighted wholesale price, is obtained. ADF test is conducted to test the stability of the data. The combination of parameters with the minimum AIC value is selected as the parameters of the final model. The autocorrelation coefficient diagram and partial autocorrelation coefficient diagram of the original data and the differential data were drawn.

The ARIMA model is mainly composed of three parts, namely, the autoregressive model (AR), differential process (I), and moving average model (MA), whose meanings are as follows:

(1) AR represents Auto Regression, an autoregressive model that builds a mapping relationship between historical data and future data to obtain a stationary sequence. P-order autoregressive process formula definition:

$$y_t = \mu + \sum_{i=1}^{p} Y_i y_t - 1 + \epsilon_t \tag{4}$$

where γ is an autocorrelation coefficient, indicating the correlation between time i and time t. The greater the correlation, the greater the coefficient, and the greater the influence of the y value at time i on the y value at time t.

The p-order indicates that time t is related to the y value of the preceding p moments, and the size of p can be determined by calculating the autocorrelation coefficient. If the autocorrelation coefficient of time i is less than 0.5, then p ends at time i+1. This section deals with the autoregressive part of the time series, which considers the effect of observations from several past periods on the current value.

(2) I stand for Integration, and the time series must be stationary to build an econometric model. The unit root test is carried out on the time series. If the time series is non-stationary, it needs to be transformed into a stationary series by difference. After several differences, it is called several-order single integers. This part is used to make the non-stationary time series reach the stationary state, and the trend and seasonal factors in the time series are eliminated by the first or second-order difference processing.

(3) MA stands for Moving Average (Moving Average) model. The error term in the autoregressive model will add up, so the MA term is added to the prediction model formula to eliminate the random fluctuations in the prediction, and the q order is defined by the MA model formula:

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_t - 1 \tag{5}$$

This section focuses on the moving average component of time series analysis, which considers the impact of past prediction errors on current values.

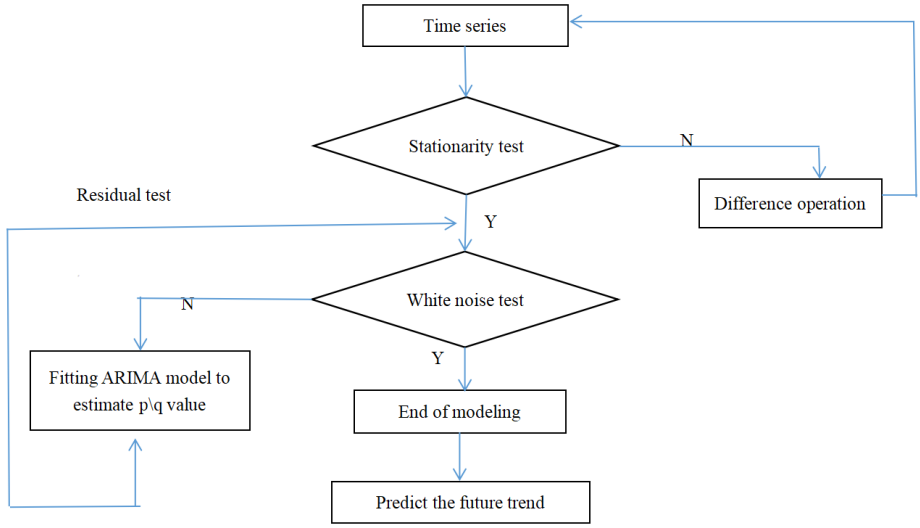


**Figure 2.** ARIMA time series regression algorithm flow chart

The ARIMA model is represented as ARIMA (p, d, q), where p denotes the number of autoregressive terms, q represents the number of moving average terms, and d signifies the number of differencing operations performed to achieve stationarity. The ARIMA model assumes that label values exhibit fluctuations around long-term trends influenced by historical labels and short-term variations influenced by random events over time; these trends themselves may not be constant. In essence, the ARIMA model aims to uncover hidden patterns within data through autocorrelation and differencing techniques to forecast future data. The Figure 2 above shows the operation flow of ARIMA time series regression algorithm.

# 3. Results

## 3.1. Preprocessing and exploration of existing data

Sales flow data analyzed the distribution law of categories and items, and drew corresponding bar charts, density charts, and box charts for the sales frequency visualization of six categories, as shown in Figure 3 below.

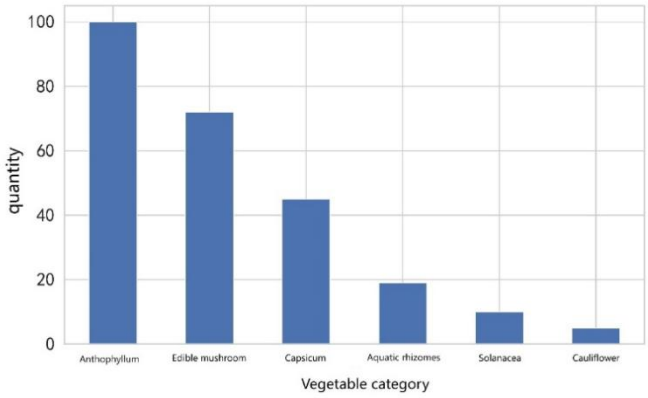


**Figure 3.** Frequency of sales in six categories

Since only the date of sale is provided in the original data it is impossible to judge whether the goods scanned are from the same order. To reduce the impact of such interference on the association rules of this study, the APRIORI algorithm is implemented by dividing the sales interval in this paper, and the frequent item set is generated iteratively by using the support meter as a metric. The support degree of sales frequency of single product, multi-product, and combined single product was obtained, as shown in Table 1 below.

**Table 1.** Sales frequency support

| Number | Frequent item set | Support count |
|---|---|---|
| 35 | Yellow heart | 7410 |
| 39 | Oyster mushroom | 6754 |
| 60 | Agaricus bosporus | 6392 |
| 24 | Green eggplant | 6335 |
| 77 | Solanum japonicum | 5850 |
| 5 | Sweet cabbage | 5778 |

As for the changes in the sales volume of each category over time, this paper divided the time into time periods and used the Seaborn library for data visualization. It was observed that the sales volume increased to different degrees at about 10 PM, which was related to the short shelf life of vegetables and the need for discounts to sell them, as shown in Figure 4 below.

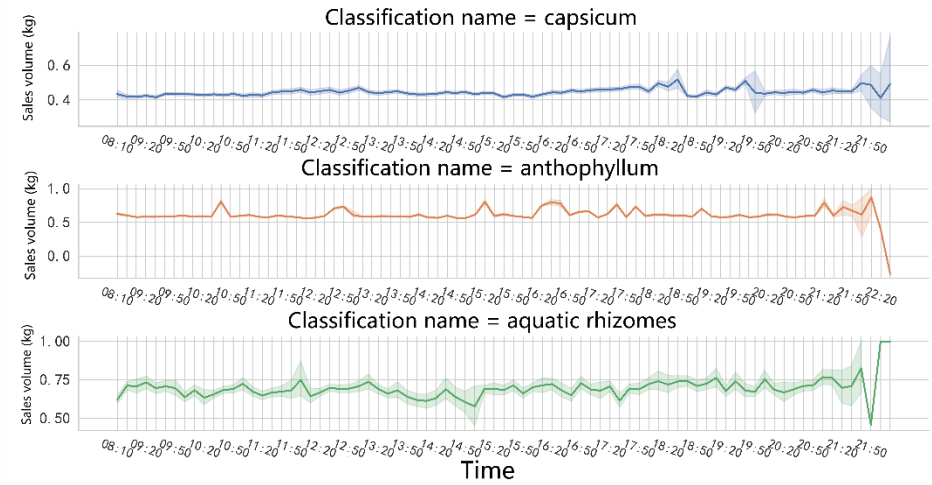


**Figure 4.** The relationship between vegetable price and time

For the sales of a single product and combination, a relatively real sales situation can be obtained by extracting frequent item sets and strong association rules from shopping basket data through the APRIORI algorithm.

### 3.2. Analysis of experimental results of the Random Forest Regression Model

The purpose of using the random forest regression model in this study is to train and predict the data. First, each date is converted into a string format, and the feature column of working day, holiday, season, and classification name coding is added to the data box for subsequent iterative calculation. The feature column x and target column y are defined, in which the feature column includes weighted sales price, classification name, weighted wholesale price, whether working day, holiday or not, and spring, summer, autumn, and winter. The weighted data set was divided into a training set and a validation set according to the specified ratio (test size=0.2), and then the random forest regression model was used for model construction. The model is trained using the training set by invoking the fit method. Calculate the predicted score on the validation set using the score method. Finally, a random forest regression model is reconstructed and trained using the entire weighted data set, and the model is trained on the entire data set for subsequent overall predictions.

The random forest regression model was used to train the set data to fit the ARIMA model. The fitted model predicted the wholesale price for the next 7-time steps, and the predicted value was reverse-normalized to obtain the prediction result of the original data. Each class name corresponds to a training and prediction process for the ARIMA model. An example of category and item correlation matrix heat is shown in Figure 5.

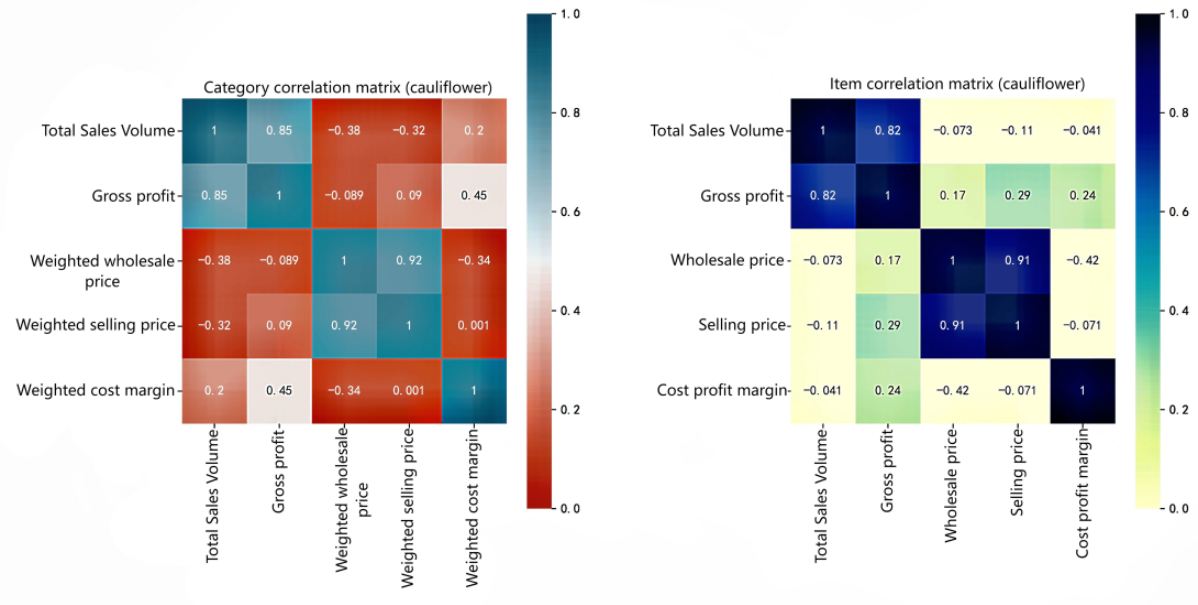


**Figure 5.** Category and item correlation matrix

By looping through each date, calculating working day, holiday, and season codes, and adding the results to the corresponding list, you get a predicted wholesale price that includes date characteristics. By traversing the predicted wholesale price of each sample, its eigenvalue is obtained, and the search is conducted to determine the best replenishment volume, pricing strategy, and projected revenue. The pricing range is generated, and each price is traversed, the current price and original characteristic value are obtained, the replenishment quantity is predicted, and the expected income under the current price is calculated, that is, the replenishment quantity is multiplied by the price difference. If the current estimated income is greater than the best-estimated income, the best-estimated income, the best price, and the best replenishment quantity are updated. The best-expected income, best price, and best replenishment volume obtained by traversing the last update are the daily replenishment total amount and pricing strategy of the supermarket with the largest return in the coming week.

### 3.3. Analysis of results of ARIMA model experiment

Given that the data in this article contains many sales times, the ARIMA model is used to predict wholesale prices based on time series data. The expected revenue can be calculated based on the current pricing strategy and demand, which is obtained by multiplying the demand by the difference between the pricing strategy and purchase cost. If the current expected revenue exceeds the best expected revenue, update the best expected revenue, best pricing strategy, and optimal replenishment quantity while maintaining a total inventory-33 items to meet requirements. Selecting data for sales date "2023-07-01" and sorting it in descending order based on the "expected revenue" column will allow us to identify the top 33 records with highest expected revenues. Through iterative search, we can determine the optimal pricing strategy and replenishment volume that maximize supermarket profits. The autocorrelation diagram of cauliflower autocorrelation is shown in Figure 6. At the same time, the optimal replenishment and pricing strategy table based partly on ARIMA time series model is shown in Table 2. The Table 2 shows the statistics of holidays that are not working days and consider seasonal changes.

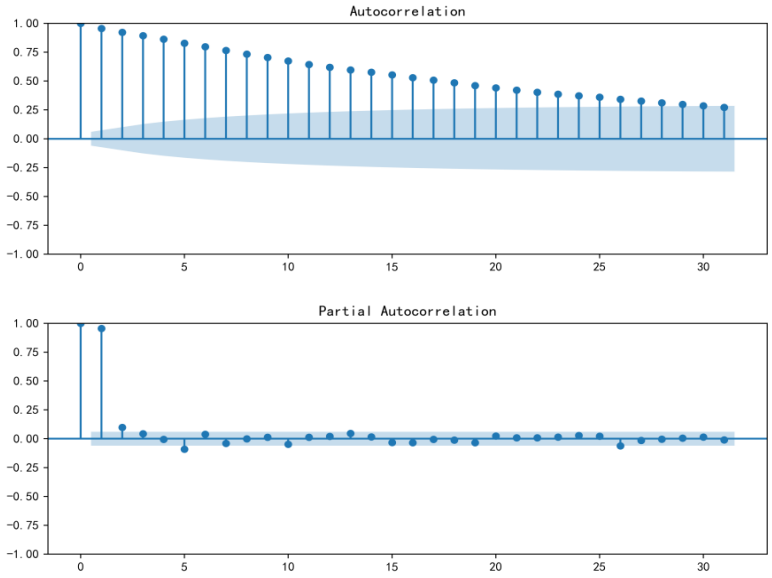


**Figure 6.** Cauliflower autocorrelation partial autocorrelation graph

**Table 2.** Partial optimal replenishment and pricing strategy

| Item coding | Date | Weighted wholesale price | Item coding | Replenishment volume | Pricing strategy | Expected income |
|---|---|---|---|---|---|---|
| 1.029E+14 | 2023-07-01 | 7.800 | 10 | 33.746 | 14.500 | 226.097 |
| 1.029E+14 | 2023-07-01 | 3.504 | 26 | 54.560 | 6.955 | 188.331 |
| 1.029E+14 | 2023-07-01 | 6.461 | 2 | 39.117 | 10.491 | 157.656 |
| 1.029E+14 | 2023-07-01 | 4.064 | 28 | 32.420 | 8.096 | 130.712 |
| 1.029E+14 | 2023-07-01 | 3.344 | 22 | 40.362 | 6.216 | 115.913 |
| 1.029E+14 | 2023-07-01 | 15.278 | 0 | 7.337 | 30.516 | 111.801 |

## 4. Conclusions

The objective of this paper is to address the significant issue of inventory loss and waste in fresh vegetable products, which are highly perishable and prone to rapid deterioration. By mitigating supply shortages or excessive waste, this study aims to maximize supermarket interest. Specifically, sales flow records and market demand data are normalized to derive weighted cost profit margins, total sales volumes, total profits, weighted wholesale prices, and weighted sales prices. To predict optimal daily replenishment quantities and pricing strategies for maximizing supermarket returns in the upcoming week, this paper employs Spearman correlation coefficient analysis to establish a demand-price regression curve with an associated R^2 value. Subsequently, a random forest regression model is utilized for training and forecasting purposes while an ARIMA model processes time series data for different product classifications individually. This approach yields a table containing each category's name alongside its corresponding forecasted wholesale price. Furthermore, by conducting

a search algorithm within a determined pricing range based on estimated revenue calculations at various prices; the best replenishment quantity, pricing strategy, and projected revenue can be determined. Given limited shelf space availability during the June 24-30th 2023 period along with a controlled total product count ranging from 27-33 items; individual product replenishment quantities and pricing strategies for July 1st are provided accordingly ensuring that each ordered quantity meets the minimum display requirement of 2.5 kg per item. To handle large amounts of data effectively in this study context; an Augmented Dickey-Fuller (ADF) test is employed initially to assess data stability whereby first-order differencing is applied if instability is detected followed by reverse-normalization of predicted results. According to the data, broccoli is the most profitable dish, and it is also the largest purchase quantity and sales volume. Yunnan lettuce is the largest dish needed in a single product. This shows such sales and categories. The total number is very large, indicating that the people are commonly used at home and the demand is huge. Therefore, we can reflect the market demand for different kinds of dishes through these data, to carry out more scientific pricing and replenishment strategies, to maximize the benefits of supermarkets under different constraints.

## References

[1] Kaya O, Bayer H. Pricing and lot-sizing decisions for perishable products when demand changes by freshness [J]. Journal of Industrial and Management Optimization, 2020, 17(6): 3113-3129.

[2] Kumar S, Mahapatra R P. Design of multi-warehouse inventory model for an optimal replenishment policy using a rain optimization algorithm[J]. Knowledge-Based Systems, 2021, 231: 107406.

[3] Kaya O, Bayer H. Pricing and lot-sizing decisions for perishable products when demand changes by freshness [J]. Journal of Industrial and Management Optimization, 2020, 17(6): 3113-3129.

[4] Liu L, Zhao Q, Gonzalez E D R S, et al. Sourcing and production decisions for perishable items under quantity discounts and its impacts on environment [J]. Journal of Cleaner Production, 2021, 317: 128455.

[5] Mashud A H M, Roy D, Daryanto Y, et al. A sustainable inventory model with imperfect products, deterioration, and controllable emissions [J]. Mathematics, 2020, 8(11): 2049.

[6] MaihamiR, Ghalehkhondabi I, Ahmadi E. Pricing and inventory planning for non-instantaneous deteriorating products with greening investment: A case study in beef industry [J]. Journal of Cleaner Production, 2021, 295: 126368.

[7] Kayikci Y, Demir S, Mangla S K, et al. Data-driven optimal dynamic pricing strategy for reducing perishable food waste at retailers[J]. Journal of Cleaner Production, 2022, 344: 131068.

[8] Kirci M, Isaksson O, Seifert R. Managing perishability in the fruit and vegetable supply chains [J]. Sustainability, 2022, 14(9): 5378.

[9] Xu C, Liu X, Wu C, et al. Optimal inventory control strategies for deteriorating items with a general time-varying demand under carbon emission regulations [J]. Energies, 2020, 13(4): 999.

[10] Zhang Y, Wang Z. Joint ordering, pricing, and freshness-keeping policy for perishable products: single-period deterministic case [J]. IEEE Transactions on Automation Science and Engineering, 2020, 17(4): 1868-1882.

[11] Avinadav T, Herbon A, Spiegel U. Optimal inventory policy for a perishable item with demand function sensitive to price and time [J]. International Journal of Production Economics, 2013, 144(2): 497-506.website