Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

A Decision model for merchandise replenishment based on logistic regression and ARIMA time series forecasting

Zenan Huang [1, *], Kai Wang [2]

[1] School of Mechanical and Automotive Engineering, Qingdao University of Technology, Qingdao, China, 266520

[2] School of Civil Environmental Engineering and Geography Science, Ningbo University, Ningbo, China, 315211

* Corresponding author: hznzyq1314@163.com

Abstract. In supermarkets or shopping malls that mainly deal with fresh food, vegetable commodities are generally difficult to store for a long time with a short freshness period. And the quality of vegetables and storage time present a certain relationship, the longer the storage time the worse the quality of vegetables, and even the existence of varieties if the day did not sell the next day will not be able to sell. Therefore, supermarkets need to replenish goods on the same day according to the past sales situation and the demand of residents. To address the problem, this paper fits the data to derive a linear equation for sales volume and cost-plus pricing using a logistic regression model and Adam's optimized gradient descent algorithm. Finally, a time series ARIMA forecasting model is used to give daily replenishment and specific pricing for the week ahead based on past daily sales data. In the results, we can adjust the merchandise based on predicted replenishment and pricing strategies. The model reveals the supermarket's replenishment and pricing strategy for fresh food in general, and the supermarket can adjust its replenishment strategy based on its data.

Keywords: Logistic regression; AMIRA; Decision model.

1. Introduction

Vegetables and fruits are generally highly seasonal, and during the period from one year's mass market to the next year's mass market, they generally go through a process from more to less, or even from something to nothing - to better satisfy the demand and to achieve good economic and social benefits, it is necessary to decide on the production and marketing activities according to the actual situation [1-2].

Nie Yuxuan scholars presented a study on automated pricing of fresh commodities using ARIMA predictive modeling [3]. Poongodi M and other scholars presented research on the use of ARIMA prediction model to predict the price of the virtual commodity bitcoin [4]. Literature [5] provides short-term forecasting of prices of agricultural commodities in the Indian market using ARIMA forecasting model. Literature [6] Forecasting and analyzing the prices of agricultural products such as fruits and vegetables using seasonal ARIMA forecasting models. Literature [7] utilizes an ARIMA model to analyze potato prices in one major market in each state. Literature [8] combines an ARIMA model with a gray model to predict cocoa production in six major countries. Literature [9] applies multivariate stochastic modeling, using ARIMA as a basis for evaluating the best ARDL models. Mohanty M K, Thakurta P K G and other scholars have proposed models for forecasting prices of agricultural commodities based on crop yields, residues and import values using three methods - time series methods, statistical methods and machine learning techniques [10].

In the above literature using ARIMA model for the prediction of commodity prices, the model fit is not very good, Analysis of the model-building process, found that the above literature for the relationship between the variables did not have an in-depth study, this paper through the Logistic model to clarify the relationship between the price of the commodity and the sales volume of the

relationship between the formula, and then through the prediction of the sales volume derived from the price of the combination of the two models to make the overall model of the fit is better, the model The combination of the two models makes the overall model fit better and the model accuracy higher.

2. The fundamentals of the logistic regression model and ARIMA model

2.1. The structure of logistic regression model

The structure of the logistic regression model is considered a single-layer neural network. It consists of an input layer, an output layer with only one neuron with a sigmoid activation function, and no hidden layer. The logistic regression model structure is shown in Figure 1.

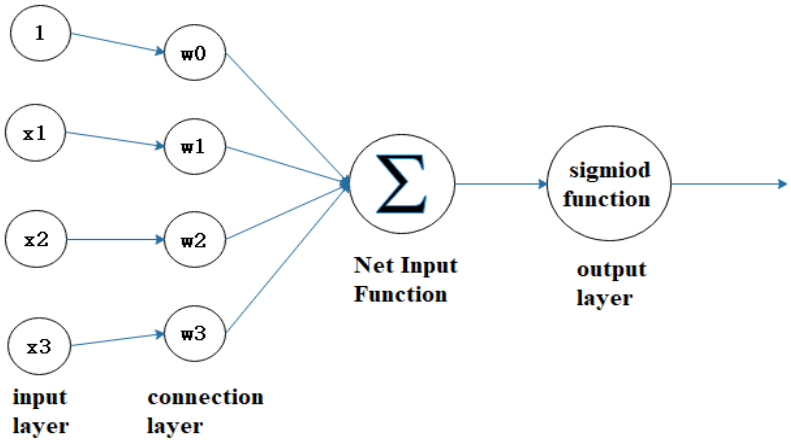

Figure 1. Logistic regression model structure

The function of the logistic regression model can be simplified into two steps, which are:

(1) The input features are linearly summed by the model weights

(2) The output is output after nonlinearly converting the summation result to a probability value between 0 and 1 by a sigmoid activation function.

Specifically, the input data feature x is summed by multiplying it by the corresponding model weight w, and then the result of (ωx + b) is converted to a probability value by the activation function σ (i.e., sigmoid function) of the neurons in the output layer.

2.2. Logistic regression modeling

First, respectively, to establish a linear regression and nonlinear regression model, after the exclusion of data found that the error of the linear function obtained by the regression is less than the nonlinear function, so the establishment of linear regression equations, the specific formula is calculated as follows Equation (1)

$$y = wx + b \tag{1}$$

Where y is the sales price of goods, x is the volume of merchandise sales, ω is weights, b is misalignment.

To evaluate the predictive effectiveness of the model, it is necessary to define a loss function, which is defined as follows in equation (2)

$$L = \sqrt{(y - \hat{y})^2}, \tag{2}$$

Finally, the weight parameters of the model are continuously adjusted to minimize the value of the loss function through optimization algorithms such as gradient descent. This process is an iterative one, with each iteration updating the weight parameters based on the current gradient information.

2.3. The structure of the ARIMA model

Time series ARIMA model is a widely used statistical model for forecasting time series data. It consists of three main components: an autoregressive model (AR), a difference process (I) and a moving average model (MA). Its specific flow chart is shown in Figure 2.

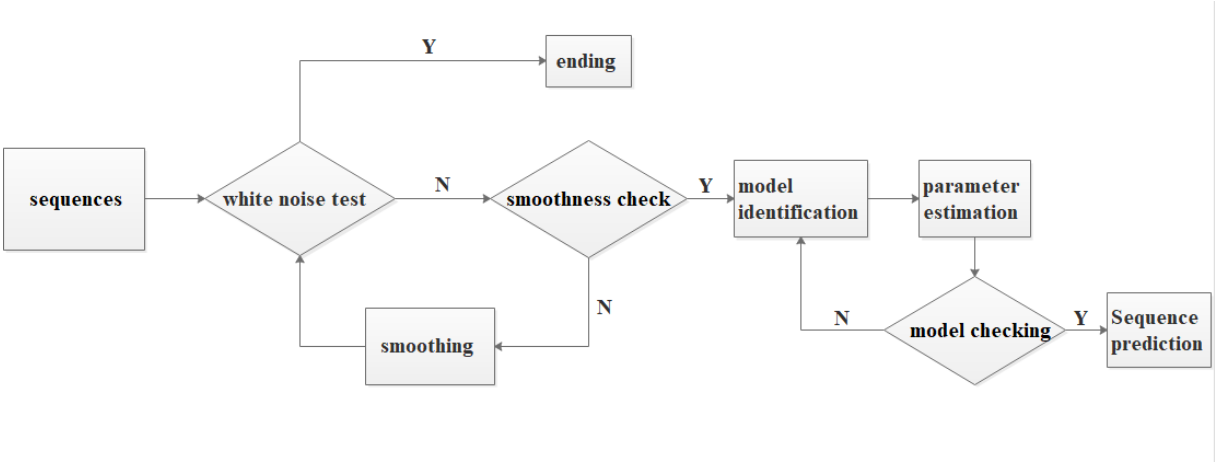

Figure 2. ARIMA model structure

2.4. Fundamentals of the ARIMA model

The ARIMA model is a widely used method for analyzing and modeling various types of time series data. The model is based on the idea that the time series to be predicted is generated by a stochastic process. If the stochastic process that generates the series does not change over time, the structure of the stochastic process can be accurately characterized and described. Using the past observations of the sequence, the future values of the sequence can be extrapolated. In the ARIMA model, the future value of the sequence is expressed as a linear function of the current and lagged periods of the lag term and the random disturbance term, i.e., the general form of the model is shown in the following equation (3)

$$Y_t = c + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + \epsilon_t + \beta_1 \epsilon_{t-1} + \cdots + \beta_q \epsilon_{t-q}, \tag{3}$$

Where Y_t is the value of the sequence we observe at time t and the target variable we want to predict, c is the constant term, which can also be thought of as the intercept in the model, it indicates that when all explanatory variables (in this case the lag and error terms) are zero, the expected value of Y_t, $\alpha_1, \alpha_2, \ldots, \alpha_p$ are the coefficients of the autoregressive component, they measure the effect of past values in the series (i.e., $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$) on current values Y_t, $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ have lagged values of Y_t(i.e., the observed values of the series at past time points), ϵ_t is the white noise error term at time t. It represents part of the variation that is not explained in the model, $\beta_1, \beta_2, \ldots, \beta_q$ are the coefficients of the moving average component. Unlike the autoregressive component, they do not measure the effect of past y-values, but rather the past error term (i.e., $\epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_{t-q}$) on the current value Y indirectly, $\epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_{t-q}$ are the error terms for past points in time. In ARMA modeling, these error terms are important factors affecting the current observations, p and q denote the lag order of the autoregressive and moving average components, respectively. Choosing the appropriate p and q is one of the key steps in building an effective ARMA model. It is usually necessary to determine these values by some statistical criteria (e.g., AIC, BIC) or graphical methods

(e.g., autocorrelation function ACF and partial autocorrelation function PACF plots), this paper uses the AIC guidelines and the ACF method for determining the p and q.

2.5. The basic steps of running an ARIMA model

Step1: Smoothness test for time series. ADF or PP test is usually used to test the unit root of the original series. If the sequence does not meet the smoothness conditions, the non-smooth time series can be transformed into a smooth time series by difference transformation or log difference transformation, and then construct ARIMA model for the smooth time series.

Step2: Determine the order of the model. With the help of some statistics that characterize the series, such as autocorrelation (AC) coefficients and partial autocorrelation (PAC) coefficients, the possible forms of the model are initially identified, and then an optimal model is selected from the available models according to the criteria for determining the order, such as the AIC.

Step3: Parameter estimation and diagnostic tests. This includes testing the significance of the model parameters, the validity of the model itself, and testing whether the residual series is a white noise series. If the model passes the test, the model setting is essentially correct; otherwise, the model form must be redefined and diagnostic tests must be performed until a model form with the correct setting is obtained.

Step4: Predictions were made using the established ARIMA model.

3. Results

3.1. Data preprocessing and descriptive statistical analysis

To better analyze the products, we have pre-processed the data by cleaning, classifying summarizing, etc., and classified the single products into six categories, and the total daily sales of the six categories in 2023 are shown in Figure 3.

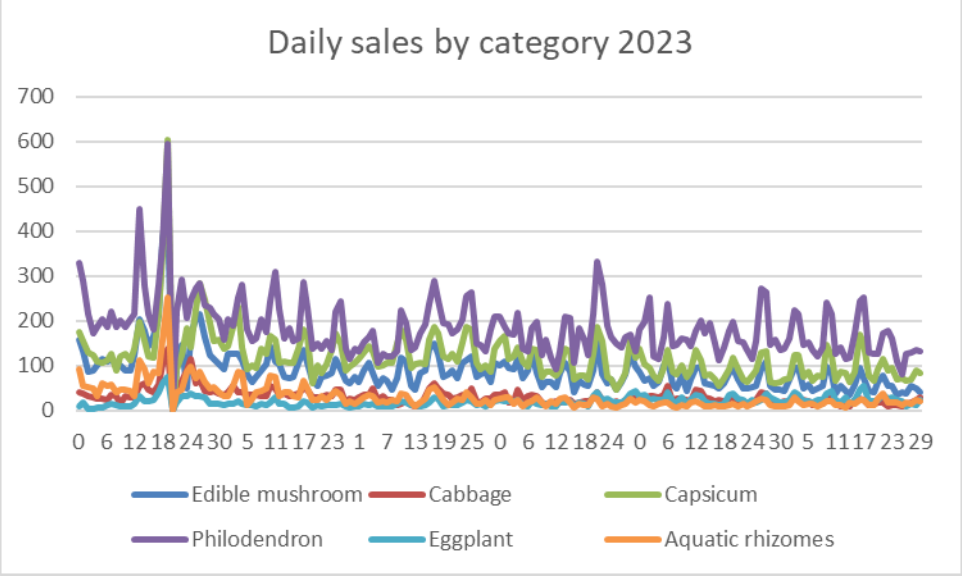

Figure 3. Six Categories 2023 Daily Sales Breakdown Chart

According to Figure 3, the daily sales of major categories have great changes over time, which also provides a data basis for our subsequent forecast analysis.

3.2. Analysis of model results

In the logistic regression model, the gradient descent algorithm with Adma optimization is used for 10,000 iterations to obtain the image of the loss function L with the gradient as shown in Figure 4

below, and then the final output of the linear relationship between the sales volume of each category and the cost-plus pricing is given in the following equation (4).

$$Y = \begin{bmatrix} -0.010890868 \\ 0.003126739 \\ -0.013843117 \\ -0.00403201 \\ 0.010987624 \\ -0.005413778 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}^T + \begin{bmatrix} 12.00107765 \\ 9.046160698 \\ 14.77716255 \\ 11.8351202 \\ 13.68172932 \\ 10.4261198 \end{bmatrix}^T , \tag{4}$$

WHere $x_p = 1, 2, 3, 4, 5, 6$ represent edible mushrooms, cabbage, capsicum, philodendron, eggplant, and aquatic rhizomes.

Take edible mushroom as an example to get its linear equation as following equation (5), and make a linear diagram as following Figure 4, other categories also have similar relationship diagrams.

$$y_1 = -0.00955474 \times x_1 + 12.501405 , \tag{5}$$

Through the linear equation after the logistic regression and the image we can get the relationship between the product markup pricing and the product sales, to facilitate the subsequent replenishment strategy and pricing strategy development.

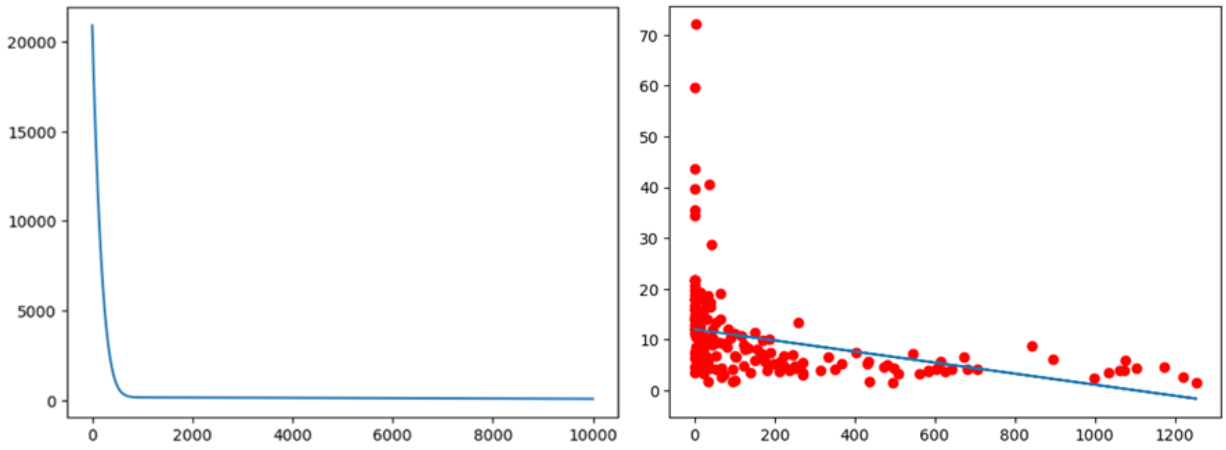

Figure 4. Plot of the number of edible mushroom iterations versus the loss function and a linear plot of the logistic regression of sales volume versus price

From the analysis of Figure 3, the trend of daily sales in the six categories is the same, and this data is substituted into the ARIMA time series prediction model to predict the daily sales from July 1, 2023 to July 7, 2023.

Firstly, based on the sales data, the daily sales for the coming week are predicted and each category is replenished according to the daily sales according to the assumed production and sales balance principle. The selling price is then estimated as pricing based on a regression model. The following cauliflower category is an example to show the solution, the rest of the categories of daily replenishment and pricing strategy are like the solution, the process is no longer described in the battle as the result of the solution.

Imported daily sales data of the cauliflower category for 6 months in 2023 into SPSSPRO for time projection analysis

1) ADF test analysis

Table 1. ADF Inspection Form

ADF Inspection Form							
variant	Order of Difference	t	P	AIC	threshold value		
					1%	5%	10%
cabbage	0	-2.067	0.258	1349.721	-3.469	-2.878	-2.576
	1	-9.096	0.000***	1344.998	-3.469	-2.878	-2.576
	2	-9.19	0.000***	1362.271	-3.47	-2.879	-2.576

Note: ***, **, * represent 1%, 5%, and 10% significance levels, respectively.

Based on the variable cauliflower class, the results of the test of this series are shown in Table 1:

The significance p-value of 0.258 at the difference of order 0 does not present significance at the level to reject the original hypothesis and the series is an unsteady time series.

The significance p-value is 0.000*** at the level of difference of order 1, which presents significance and rejects the original hypothesis that the series is a smooth time series.

The significance p-value is 0.000*** at the difference of order 2, which presents significance at the level and rejects the original hypothesis that the series is a smooth time series.

2) Model Testing

Secondly, due to the human subjectivity in determining the parameters of ARIMA through autocorrelation analysis and partial autocorrelation analysis, it was chosen to find the optimal parameters automatically based on the AIC information criterion, and the table of model test parameters is shown in Table 2:

Table 2. ARIMA model (2, 1, 2) test table

item	notation	worth
	Df Residuals	175
sample size	N	181
Q-statistic	Q6 (P-value)	0.006 (0.940)
	Q12 (P-value)	11.089 (0.086*)
	Q18 (P-value)	13.862 (0.310)
	Q24 (P-value)	18.998 (0.392)
	Q30 (P-value)	19.461 (0.727)
Information guidelines	AIC	1461.905
	BIC	1481.063
goodness of fit	R^2	0.405

Note: ***, **, * represent 1%, 5%, and 10% significance levels, respectively.

The system is based on the AIC information criterion to automatically find the optimal parameters, the model results for the ARIMA model (2, 1, 2) test table, based on the variable: cauliflower class, from the analysis of the results of the Q statistic can be obtained: the Q6 does not present significance at the level, cannot be rejected the model's residuals for the hypothesis of the white noise sequence, at the same time the model's goodness of fit R^2 is 0.405, the model performance is average, the model is Satisfy the requirements

3) Predicted results

Finally, the daily sales of cauliflower vegetables in the coming week were obtained based on the time prediction model, and the time series prediction fitting of the cauliflower category is shown in Figure

5 (Blue curve: true value, green curve: fitted value, yellow curve: predicted value), and the time series prediction table is shown in Table 3:

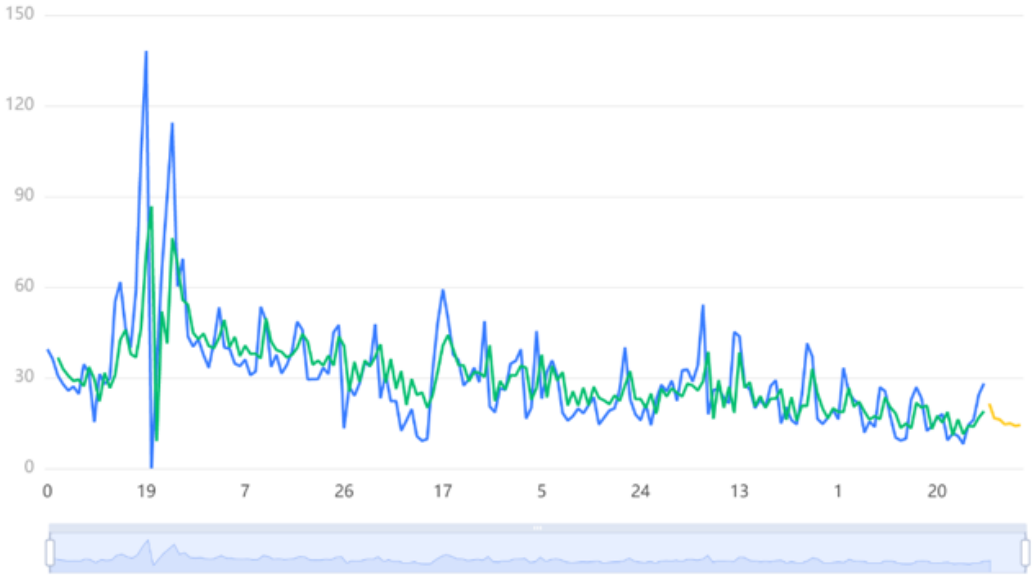

Figure 5. Fitted plot of time series prediction of cauliflower species

Table 3. Time series forecasting table

Order (time)	Predicted results
1	21.424553324384338
2	16.516327587038127
3	16.136902383268886
4	14.51993559008259
5	14.862638145939183
6	14.023394088676426
7	14.289844261542276

The sales volume of cauliflower from July 1 to July 7, 2023, can be obtained from the prediction table Table 3. According to the original assumption of production-sales equilibrium (sales volume = replenishment), the predicted sales volume can be used to make the daily replenishment volume of cauliflower in the coming week, which means that the replenishment strategy can be obtained. Then the predicted sales volume was brought into the logistic regression fitting equation to estimate the daily selling price of cauliflower, i.e., the pricing strategy from July 1 to July 7, 2023, could be obtained. The remaining five categories of vegetables are analyzed as described above to obtain the replenishment strategy and pricing strategy. The results obtained are shown below in Table 4 Forecasted Replenishment and Table 5 Pricing Strategy table.

Table 4. Projected replenishment

	July 1	July 2	July 3	July 4	July 5	July 6	July 7
edible mushrooms	45.668	43.996	44.419	43.406	43.224	42.561	42.176
cabbage	21.425	16.516	16.137	14.520	14.863	14.023	14.290
capsicum	84.552	82.031	81.889	81.439	81.436	81.166	80.811
philodendron	184.595	184.595	184.595	184.595	184.595	184.595	184.595
eggplant	25.748	22.980	23.043	23.106	23.170	23.233	23.296
aquatic rhizomes	15.354	14.524	14.705	14.654	14.374	14.066	13.787

Note: The unit is kg and the value is retained to three decimal places.

Table 5. Pricing Strategy Sheet

kind	July 1	July 2	July 3	July 4	July 5	July 6	July 7
edible mushrooms	11.50	11.52	11.52	11.53	11.53	11.54	11.54
cabbage	9.11	9.10	9.10	9.09	9.09	9.09	9.09
capsicum	13.61	13.64	13.64	13.65	13.65	13.65	13.66
philodendron	11.09	11.09	11.09	11.09	11.09	11.09	11.09
eggplant	13.40	13.43	13.43	13.43	13.43	13.43	13.43
aquatic rhizomes	10.34	10.35	10.35	10.35	10.35	10.35	10.35

Note: The unit is Yuan/kg, and the value retains two decimal places.

4. Conclusions and outlooks

This paper addresses the problem of replenishment pricing decisions for vegetable commodities in supermarkets. The processed data is first used to fit the data to derive a linear equation for sales volume and cost-plus pricing using a logistic regression model and Adam's optimized gradient descent algorithm. Finally, the daily replenishment and specific pricing for the coming week are given based on the previous daily sales data using the time series AMIRA forecasting model prediction. The results show that the relationship between total category sales and cost-plus pricing has good prediction results, and the replenishment volume of edible mushrooms from July 1-7 is 45.668kg, 43.996kg, 44.419kg, 43.406kg, 43.224kg, 42.561kg, 42.176kg, and the pricing is 11.50 Yuan/kg, 11.52 Yuan/kg, 11.52 Yuan/kg, 11.53 Yuan/kg, 11.53 Yuan/kg, 11.54 Yuan/kg, 11.54 Yuan/kg, and the rest of the rest of the categories of the replenishment and ordering program are shown in Table 4 and Table 5 of the main text, respectively.

Based on the multiple models established in this paper, we analyze them and find the following advantages and disadvantages:

1. In the process of regression analysis of the six categories based on Logistic regression model, the gradient descent algorithm optimized by Adam is used with 10,000 iterations, which makes the algorithm able to find the optimal solution faster;

2. Forecasting using ARIMA time series model, while choosing to automatically find the optimal parameters based on the AIC information criterion, led to better forecasting results;

3. In this paper, when considering seasonal influences in time series forecasting, the factor considerations are not too comprehensive, which makes the time series forecasting model fitting effect general.

References

[1] Liu Baozheng, Liu Depo, Gao Liqun. Price control and production and marketing decision modeling for seasonal commodities in short supply [J]. Journal of Northeastern University, 2005, (11): 2 3 - 26.

[2] Lu Jing. Research on inventory control and dynamic pricing of fresh agricultural products [D]. Tianjin University, 2019. DOI: 10. 27356/d. cnki. gtjdu. 2019. 000182.

[3] Nie Yuxuan. 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. DOI: 10.19995/j.cnki.CN10-1617/F7. 2024. 05. 019.

[4] Poongodi M, Vijayakumar V, Chilamkurti N. Bitcoin price prediction using ARIMA model [J]. International Journal of Internet Technology and Secured Transactions, 2020, 10 (4): 396 - 406.

[5] Kumar Mahto A, Biswas R, Alam M A. Short term forecasting of agriculture commodity price by using ARIMA: based on Indian market [C]//Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, April 12 – 13, 2019, Revised Selected Papers, Part I 3. Springer Singapore, 2019: 452 - 461.

[6] Dharavath R, Khosla E. Seasonal ARIMA to forecast fruits and vegetable agricultural prices [C]//2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS). IEEE, 2019: 47 - 52.

[7] Jamuna C J, Patil C, Kumar R A. Forecasting the Price of Potato Using Time Series ARIMA Model[C]//Proceedings of International Conference on Communication and Computational Technologies: ICCCT 2021. Singapore: Springer Singapore, 2021: 493 - 518.

[8] Quartey-Papafio T K, Javed S A, Liu S. Forecasting cocoa production of six major producers through ARIMA and grey models [J]. Grey Systems: Theory and Application, 2021, 11 (3): 434 - 462.

[9] Madziwa L, Pillalamarry M, Chatterjee S. Gold price forecasting using multivariate stochastic model [J]. Resources Policy, 2022, 76: 102544.

[10] Mohanty M K, Thakurta P K G, Kar S. Agricultural commodity price prediction model: a machine learning framework [J]. Neural Computing and Applications, 2023, 35 (20): 15109 - 15128.