

Data driven Comprehensive Replenishment and Pricing Model

Hao Huang^{1,*}, Chuyu Zhao^{1,#}, Xiangyi Liu^{2,#}

¹School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, 100091

²School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China, 100091

*Corresponding author: H1un0ter1@163.com

#These authors contributed equally.

Abstract. Pricing and replenishment of vegetable commodities is a key step in solving the practical problem of short shelf life in vegetable markets. Our research introduces a nuanced, data-driven approach to optimize replenishment and pricing decisions. At the core of our methodology is the development of a comprehensive model that leverages an enhanced ARIMA, meticulously tailored for distinct vegetable categories. This model is designed to capture the intricate dynamics of market demand and supply fluctuations, enabling precise predictions that guide effective pricing and replenishment strategies. To further refine our model's predictive capabilities, we integrate advanced machine learning techniques, including an optimized XGBoost algorithm, facilitated by a genetic algorithm. This integration not only augments the model's accuracy but also provides a robust framework for analyzing and interpreting complex market data. Through a detailed example analysis, we demonstrate the model's efficacy in forecasting demand and optimizing inventory levels, thereby ensuring that pricing and replenishment actions are both strategic and data-informed.

Keywords: Vegetable Pricing and Replenishment, Data Driven, CRPM, ARIMA-GXGBoost.

1. Introduction

The field of fresh produce has received a lot of attention in the study of agricultural products [1]. Accurate forecasting of fresh commodity prices is essential to guide farmers' production adjustments, reduce market risks, and ensure the stable operation of agricultural markets [2]. However, most of the existing literature focuses on macro-level analysis, with fewer studies on the classification of specific vegetable products [3]. Traditional forecasting models have certain shortcomings, such as the single model approach based on ARIMA model [4] and dynamic model averaging [5], which can only forecast the price problem from a certain perspective and cannot accurately capture the complex characteristics of price fluctuations [6].

In this paper, new ideas and methodological innovations are proposed to address the price problems of six major vegetable categories such as edible mushrooms and foliage. In order to further improve the accuracy of agricultural price trend prediction, this paper first optimizes the XGBoost model and introduces a genetic algorithm to construct the GXGBoost model, which yields better prediction results than XGBoost. ARIMA and GXGBoost are further considered as the basis to get the ARIMA-GXGBoost combined model based on ARIMA-GXGBoost, and a data-driven integrated replenishment and pricing model is further proposed to predict the price and replenishment demand of fresh vegetables more effectively.

2. Data driven Comprehensive Replenishment and Pricing Model (CRPM)

2.1. ARIMA Model

The basic principle of the ARIMA model [7] is to predict the future period of data with the help of its own original data set after difference processing and then through the characteristics of its own data. The ARIMA model is used if the time series is smoothed by Nth order differencing. The following formula shows the calculation of ARIMA model:

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^p \gamma_i y_{t-1} - \sum_{i=1}^q \theta_i \varepsilon_{t-1} \quad (1)$$

Where: μ represents the constant term; ε_t is the random error value; γ_i is the autocorrelation coefficient; θ_i is the moving average coefficient; p is the autoregressive term, q is the moving average term, and d is the number of times required for the time series to be smooth. If the predicted data set is a non-smooth series, it is generally transformed into a smooth series by differential processing, and then fitted with AR-MA model; if it is a smooth non-white noise series, it is necessary to require the autocorrelation coefficient and partial correlation coefficient combined with correlation graph analysis to obtain p and q values.

2.2. XGBoost Model

Extreme Gradient Boosting [8] (XGBoost) is an algorithm that integrates multiple base learners to build a strong learner through the idea of Boosting. It is improved by GBDT algorithm, and its learner can be a CART decision tree (GBTree) or a linear classifier (GBLinear). Generally speaking, XGBoost, GBDT algorithm and Random Forest are in a progressive relationship. Among them, the GBDT algorithm incorporates the XGBoost idea on the basis of random forests, so that the trees in the forest are connected rather than existing alone, forming a kind of overall ordered decision measurement system. Similarly, the XGBoost algorithm is based on decision trees, and introduces a second-order Taylor expansion and a regular term, which can effectively control the complexity of the model (the model variance is greatly reduced), and the trained model is simpler and more stable. XGBoost is able to set the default direction of branching for missing values generated during the computation process. In addition, XGBoost also supports parallel computation at the feature level, and the added regular term prevents overfitting and underfitting phenomena. The integrated model of the tree is shown in the following equation: Where f_k is the k th base model and $\hat{y}_i$ is the predicted value for the i th sample.

$$\hat{y}_i = \sum_{t=1}^k f_t(x_i) \quad (2)$$

Where f_k is the k th base model and $\hat{y}_i$ is the predicted value for the i th sample.

The objective function consists of a loss function L of the model with a regular term Ω that suppresses the complexity of the model, and thus has:

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{t=1}^k \Omega(f_t) \quad (3)$$

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (4)$$

Where $\sum_{i=1}^n l(y_i, \hat{y}_i)$ is the error between the true value and the predicted value of the model.

Eq. (4) is the regular term used to control the complexity of the model, where T is the number of leaf nodes, and w is the value of the leaf nodes, and γ and λ are used to control the number of leaf nodes and the value of the leaf nodes, respectively, and the model can be avoided from overfitting by regularization.

Finding the minimizing objective function is equivalent to solving $f_t(x_i)$. For function $l(y_i, \hat{y}_i^{t-1} + f_t(x_i))$, a Taylor second-order expansion at $\hat{y}_i^{t-1}$, $f_t(x_i)$ is taken to be $\square x$, and the objective function is approximated by

$$Obj^{(t)} \approx \sum_{i=1}^n \left[l(y_i, y_i^{t-1}) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right] + \sum_{i=1}^t \Omega(f_i) \quad (5)$$

Since $l(y_i, y_i^{t-1})$ is a constant and has no effect on the optimization of the objective variable function, the objective function can be written as:

$$\begin{aligned} Obj^{(t)} &\approx \sum_{i=1}^n \left[g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right] + \sum_{t=1}^t \Omega(f_i) = \\ &\sum_{i=1}^n \left[g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2 \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 = \\ &\sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T = \\ &\sum_{j=1}^T \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \end{aligned} \quad (6)$$

$$\text{Where } G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i$$

From the above equation, it can be seen that the objective function is transformed into a one-dimensional quadratic function, and its solution yields the optimal w and the simplified objective function, as shown in equations (7) and (8):

$$w_j^* = -\frac{G_j}{H_j + \lambda} \quad (7)$$

$$Obj = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (8)$$

2.3. Genetic Algorithm (GA)

Genetic algorithm is a computational model that simulates the biological evolution process of Darwin's theory of biological evolution by natural selection and genetics [9], and it is a kind of method that searches for the optimal solution adaptively by simulating the natural evolution process. The implementation process of the genetic algorithm is as follows:

Randomly generate n chromosomes of definite length, $pop_i(t), i=1, 2, 3, \dots, n$, as the initial population. Assuming that the fitness function is $f(i)$, its corresponding value is calculated for all individuals as: $f(i) = f(pop_i(t))$. Determine whether the convergence state is reached according to the set termination condition in order to select the next operation. Select the individuals with large fitness values for the crossover operation, then the selection probability of each individual is P_i :

$$P_i = \frac{f(i)}{\sum_{i=1}^n f(i)}, i=1, 2, 3, \dots, n \quad (9)$$

Select individuals from the current population to reorganize the new population of individuals with selection probability P_i as the new probability distribution:

$$newpop(t+1) = \{pop_j(t) | j=1, 2, 3, \dots, n\} \quad (10)$$

Two different individuals are randomly selected from the population and genes are exchanged with probability P_c to obtain two new individuals, which is performed $n / 2$ times to obtain a new population $crosspop(t+1)$.

Individuals are randomly selected from the population and mutated with mutation probability P_m to obtain a new population $mutpop(t+1)$, which serves as a sub-population that completes a genetic operation, i.e., $pop(t) = mutpop(t+1)$, which is then passed on to the fitness calculation.

2.4. ARIMA-GXGBoost Combined Model

2.4.1 Model based on improved XGBoost

The steps of GXGBoost, a fusion algorithm based on GA and XGBoost, are as follows:

(1) The XGBoost prediction model is first constructed and the input dataset is divided into a training set and a test set.

(2) The parameters in the XGBoost prediction model are initialized using the GA algorithm, and the initial parameter values are input into the XGBoost prediction model

(3) XGBoost is trained according to the parameter values and the training set to obtain the prediction results, and the prediction error between the prediction results and the test set is calculated.

(4) If the error fails to satisfy the termination condition, the GA algorithm calculates the next set of parameter values based on the error, inputs this set of parameters into the XGBoost prediction model, and repeats step (3). If this error satisfies the termination condition, output the prediction result.

2.4.2 Model based on ARIMA-GXGBoost

In this system, we have designed an ARIMA-GXGBoost algorithmic model based on wavelet transform [10], which is used to decompose the data into low frequency data and high frequency data using wavelet transform. After that these two parts of data are predicted using ARIMA algorithmic model and GXGBoost model respectively and finally the future prediction of pricing and replenishment of goods is obtained.

The process of model construction based on ARIMA-GXGBoost algorithm is as follows:

(1) Obtain vegetable data data of 6 categories and perform data preprocessing and related feature extraction

(2) Use the discrete wavelet transform to do a layer of decomposition of the selected sequence to obtain high-frequency nonlinear sequences and low-frequency trend sequences, and the length of the two sequences is halved.

(3) Construct an ARIMA model for low-frequency sequences, the main steps of which are smoothness judgment, model identification, parameter estimation and model testing, and ultimately get the ARIMA model most suitable for low-frequency sequences and predict the target number of steps forward for low-frequency sequences with this model.

(4) Utilize the GXGBoost model for high-frequency sequences and predict the target number of steps forward.

(5) Reconstruct the predicted results of the high-frequency and low-frequency sequences by wavelet reconstruction to obtain the actual predicted values.

(6) Compare the difference between the predicted and real values using evaluation metrics.

2.4.3 Assessment of indicators

After the prediction results, they need to be evaluated to get the evaluation of the model fitting effect, which is usually judged by the gap between the predicted and actual values, and the smaller the gap is, the better the model fitting effect is. In this section, MAE (Mean Absolute Error) and R^2 (Coefficient of Determination) are considered as the evaluation indexes, and its mathematical expressions are as follows, $\hat{y}_i$ is the predicted value, y_i is the actual value, $\bar{y}_i$ is the average of the actual values, and n is the size of the data to be calculated.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (11)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (12)$$

3. Results

3.1. The establishment of CRPM

In this paper, we choose the dataset provided by the website: <http://www.mcm.edu.cn>, which is divided into annexes 1-4, which includes the commodity information of 6 vegetable categories distributed by a superstore, 3-year sales flow details and the data related to the wholesale price, and the data of the recent attrition rate of each commodity, etc. We used the data to predict the sales volume of various vegetable categories in the coming week. We preprocessed the data and constructed an algorithm centered on the ARIMA-GXGBoost algorithm to present a comprehensive forecast of the sales volume of major vegetable categories in the coming week. We also proposed pricing and replenishment strategies for vegetable categories in the coming week by combining the joint order pricing method [11].

3.2. Analysis of experimental results

We carried out wavelet reconstruction on the predicted values of the high-frequency part and the low-frequency part. Finally, the forecast results of edible mushroom sales next week are shown in Figure 1:

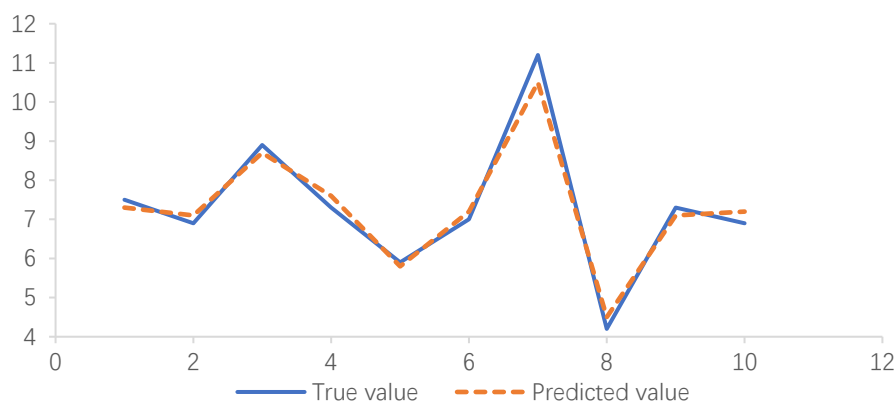


Figure 1. Predictive effects of CRPM

It can be seen that the model largely predicts the wholesale prices for the coming week after July 1, 2023 onwards. It can be seen that the predicted values of the ARIMA-PGXGBoost model based on the wavelet transform broadly match the actual values. The evaluation indexes are compared with BPNN, linear regression, random forest and ARMI-XGBoost based on wavelet transform, and the results are shown in Table 1.

Table.1. Comparison of power load forecasting of 403 line

Model	MAE	R^2
BPNN	6154	0.096
Linear Regression	5956	0.12
Random Forest	2403	0.86
ARIMA-GXGBoost	1325	0.91

For MAE , the smaller the value, the better the prediction effect; for R^2 , the larger the value, the better the prediction effect. The above table shows that ARIMA-GXGBoost is better than BPNN, Linear Regression and Random Forest in predicting commodity prices in the coming week, i.e., an improved XGBoost model with excellent prediction performance is constructed. From the accuracy of the prediction results in the above table, it can be seen that compared with other models, the improved model has a higher degree of fitting, which reduces the prediction error of the model and gives better prediction results. Therefore, it is reasonable to use this model to predict future vegetable commodity prices.

From the comparison between prediction data and actual data, the BP neural network has better prediction performance and relatively small error, which can meet the demand completely, and has fast prediction speed and convenient operation.

In the end, we give the pricing and replenishment strategies for the major categories for the coming week based on the above analysis as shown in the table 2 below:

Table.2. Pricing and replenishment strategies for selected categories

The coming week	Edible mushroom	
	Replenishment (kg)	Pricing (\$/kg)
Day 1	237.810	12.43
Day 2	110.557	13.77
Day 3	72.602	16.73
Day 4	218.974	12.43
Day 5	99.062	16.57
Day 6	164.291	11.92
Day 7	383.227	11.29

4. Conclusions

The XGBoost model has been widely used in various types of forecasting work in various studies, however, the prediction accuracy of a single model is limited to a certain extent. The dataset used in this paper has a variety of vegetable categories and a large amount of data, so the XGBoost is optimized by using genetic algorithm, and the ARIMA-GXGBoost combined model is proposed by combining with ARIMA, and at the same time, the CRPM based on ARIMA-GXGBoost is proposed by combining with the joint order pricing method for vegetable pricing and replenishment prediction. The ARIMA-GXGBoost model is experimentally compared with other models, and the results show that the ARIMA-GXGBoost model improves the prediction accuracy over a single model, and by optimizing the defects of the model, the prediction error of pricing and replenishment can be reduced. The CRPM model based on ARIMA-GXGBoost established in this paper provides a reference way for vegetable pricing and replenishment, which is of some significance.

References

- [1] FENG Ying, WANG Yuanfang, ZHANG Yanzhi et al. Impact of business model on supply chain operation of fresh agricultural products under stochastic output [J]. Systems Engineering Theory and Practice, 2020, 40(10): 2631-2647.
- [2] XU Guangshu, JIANG Yanli. Research on pricing and preservation strategy of fresh produce supply chain based on quality grading [J]. Mathematical Practice and Understanding, 2021, 51(03): 321-328.
- [3] Zhao H. Futures price prediction of agricultural products based on machine learning [J]. Neural Computing and Applications, 2020: 1-14.

- [4] Yan Haowei, Mu Yueying. Research on the construction of vegetable industry monitoring and early warning system based on vegetable market price monitoring and early warning [J]. Vegetable, 2023, (05): 1-10.
- [5] Xiong Tao, Bao Yukun. Research on soybean futures price prediction based on dynamic model averaging [J]. China Management Science, 2020, 28(05): 79-88.
- [6] Liza Mao. Research on pricing strategy and production and marketing model of vegetable wholesale market under the perspective of supply chain [D]. Central South Forestry University of Science and Technology, 2022.
- [7] Ma XB. Forecasting the price index of fresh glutinous corn in northern Hebei based on time series reconstruction[D]. Hebei North College,2023.
- [8] R. Cao,B. Liao,M. Li et al. An XGBoost-based price prediction and feature analysis model for online short-term rental market [J]. Data Analysis and Knowledge Discovery, 2021, 5(06): 51-65.
- [9] ZHANG Ge,QIAO Min,LIU Xiaohui. Analysis of influencing factors of total retail sales of consumer goods based on GA-XGBoost model [J]. China Business Journal,2023,(24):11-14.
- [10] Bonnie F. Li. Epidemiological characterization and prediction system based on ARIMA-PGXGBoost model with wavelet transform [D]. Huazhong University of Science and Technology, 2021.
- [11] Qiao Xue. Joint replenishment pricing strategy for fresh products considering sales loss [D]. Southeast University, 2021.