Electronics, Communications and Networks
A.J. Tallón-Ballesteros et al. (Eds.)

doi:10.3233/FAIA231179

Automated Pricing and Replenishment Decision for Vegetable Products Based on Hybrid Machine Learning Models

Yujia ZHENG [a] Tianhao LI [a] Weizhi MA [a] Zhengping LI [a] Lijun WANG [a] and Ying LI [a,1]

[a] *School of Information, North China University of Technology, Beijing, China*

Abstract. Fresh produce supermarkets play a vital role in modern cities, but their management is challenging due to the perishable nature of vegetables. This research proposes and implements an automated pricing and replenishment strategy based on hybrid ML models and massive historical sales data. The combination of Seasonal ARIMA, Linear Regression, and Gradient Boosting Decision Tree (GBDT) results in an average R^2 for different categories of vegetable products of 0.993, indicating high model accuracy and fitting, which could guide pricing and replenishment strategies for the merchant who sells time-sensitive products.

Keywords. Pricing, Replenishment, Seasonal ARIMA, Linear Regression, GBDT

1. Introduction

Fresh produce supermarkets, a critical commercial entity [1] in daily life, are pivotal in urban landscapes. They specialize in distributing highly perishable vegetables, managing complex challenges due to short shelf life and quality vulnerability. Daily restocking is based on past sales and current demand from various sources.

Procurement typically occurs in the early morning, between 3:00 and 4:00 AM, with merchants making replenishment decisions on the same day without complete product details or purchase prices. To remain competitive, these supermarkets often employ cost-plus pricing strategies [2] to maximize revenue.

This research conducts a thorough statistical analysis of supermarket data covering vegetable sales, wholesale prices, and spoilage rates from July 2020 to June 2023. The dataset includes 878,503 vegetables categorized into six groups: **Foliage (331,968), Cauliflower (86,570), Peppers (207,996), Eggplant (44,898), Edible Mushrooms (148,424), and Aquatic Rhizomes (58,647)**. **Table 1** shows the data information of some vegetable categories.

This research is motivated by the critical role of vegetables in people's daily lives. The primary goal is to address the issue of waste in supermarket replenishment processes, which has negative ecological impacts and can result in significant financial losses for

[1] Corresponding Author: Ying Li, School of Information, North China University of Technology, Shijingshan, 100144, Beijing, China; E-mail: liying4590@gmail.com.

supermarkets. This research focuses on analyzing data related to different vegetable categories and introduces an already deployed automated pricing and replenishment module. The key contributions of this study are as follows:

- Proposed and implemented a quantitative approach to pricing strategy
- Proposed and implemented a replenishment model for predicting future replenishment of individual vegetable categories
- Proposed and implemented an automated pricing model for predicting the future of each vegetable category.

Table 1. Sample Data Information For The Vegetable Categories From July 1, 2020 To July 30, 2023.

Category Name	Sales Date	Sales Volume (kg)	Sales Unit Price (RMB/kg)	Wholesale Price (RMB/kg)	Wastage Rate (%)
Cauliflower	2022-05-27	1.809	8	6.82	9.26
Cauliflower	2022-11-05	0.417	5.8	2.4	9.43
Aquatic Rhizomes	2022-08-25	0.39	18	14.24	9.43
Eggplant	2022-08-25	0.455	11	4.04	9.43
Foliage	2022-08-25	1	4.5	3.11	9.43
Edible Mushroooms	2023-05-16	0.388	7	4.67	6.07
Foliage	2021-08-19	0.67	16	9.98	18.51
Peppers	2023-05-16	0.321	7.2	4.71	5.7

2. Related Works

Banerjee et al. [3] proposed an LSTM network model for vegetable price forecasting, capable of capturing nonlinear and seasonal patterns while considering various price-affecting factors. However, it lacks comparisons with other neural network models.

Wang et al. [4] developed a Markov dynamic programming model that addressed product quantity and quality losses, along with the impact of freshness investment on product quality and demand. This model provided a comprehensive description of loss and demand characteristics for cold chain products. It was solved using two methods: strategy iteration and value iteration. However, the model had limitations. It didn't account for the influence of replenishment cycle and cost on decision-making, neglected the effects of various product types and sales environments on freshness investment, and lacked real-world case studies or numerical simulations to validate its effectiveness and applicability.

Shi et al. [5] created a dynamic pricing and freshness strategy for perishable goods, factoring in price, freshness, and inventory's impact on demand. They employed the maximum principle to determine the optimal strategy. This innovative approach introduced freshness inputs to slow freshness decay, analyzing their cost and effects on strategy optimization for perishable product management. However, the model has limitations. It assumes known demand functions, which may not be practical or consistent in practice. Additionally, it doesn't consider competition's influence on pricing and freshness strategies. Empirical analyses and case studies are also absent, leaving room to validate the model's applicability and effectiveness.

3. Methods

3.1. Pricing Strategy Quantification

Given that 'pricing strategy' is inherently abstract and challenging to represent with specific numerical values or expressions, this section conducts an extensive analysis to make it more concrete and quantifiable. As depicted in **Figure 1**, pertaining to supermarkets adopting the 'cost-plus pricing'[6] methodology, the process from wholesale procurement to final sale depends on three key numerical parameters: 1) Wholesale Price (WP); 2) Increasing Rate (IR); 3) Selling Price (SP), calculated using Eq.1.[7] Thus, our primary focus is on two numerical values: Wholesale Price and Increasing Rate. It's important to note that, as per our model's assumptions, businesses have no control over wholesale prices, making the determination of the increasing rate the essence of the pricing strategy.

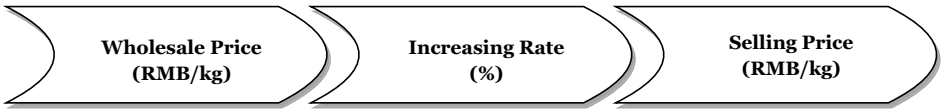

Figure 1. Pricing Flowchart Centered On Increasing Rate In The Context of Cost-plus Pricing.[7]

$$SP = WP + WP \times IR \tag{1}$$

3.2. Analysis Process

A supermarket's daily revenue relies primarily on two factors: 1) Daily sales volume, measured in kilograms (kg), representing the total quantity of vegetables sold during the day. 2) Daily profit margin, denoting the profit earned per kilogram (kg) of vegetables sold, measured in Chinese Renminbi (RMB).

To optimize revenue for the upcoming week (July 1-7, 2023), the supermarket will follow this methodology:

- Perform data preprocessing, including quartile-based outlier detection and data calculation.
- Utilize historical sales time series data to predict daily sales (kg) for each category in the coming week using the **Seasonal ARIMA Model** [8].
- Forecast the pricing increasing rate for each category for each day in the coming week based on the daily sales predictions from step 2, using **Linear Regression** [9] [10].
- Combine daily sales data with pricing additives for each category. Employ the **Gradient Boosted Decision Tree (GBDT) Model** [11]to make inferences about daily revenue for each category. Fine-tune the pricing increasing rate for optimal daily revenue.

Figure 2 illustrates the data and model used to address this issue. To account for losses during transportation and natural deterioration, the predicted future daily sales values are divided by the wastage rate for each product category, resulting in the recommended daily replenishment quantity. **Algorithm 1** outlines this recommended replenishment process.

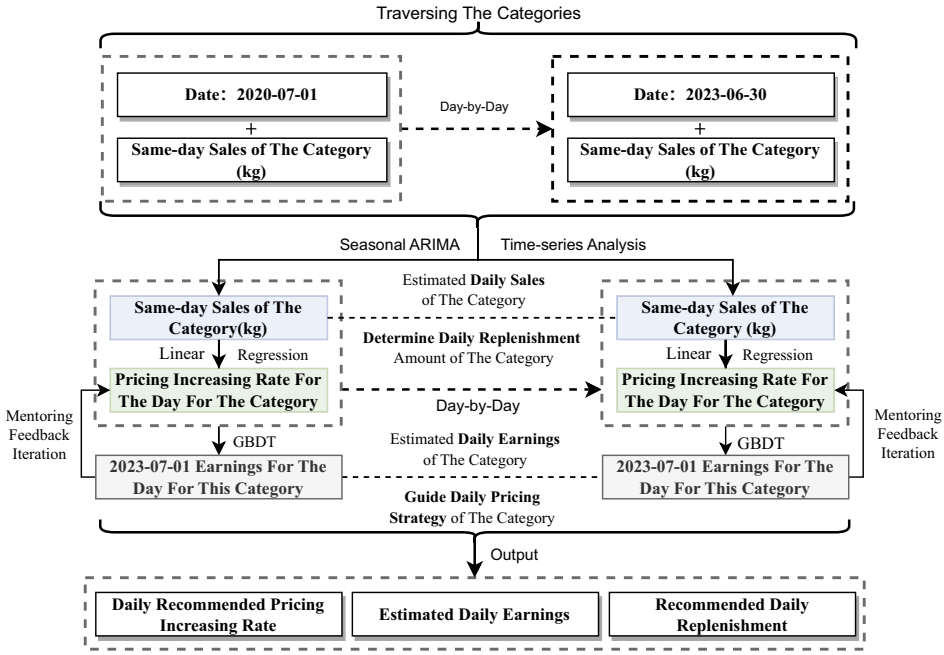

Figure 2. Model Design Solutions For Pricing and Replenishment Strategies For The Coming Week.

Algorithm 1 Daily Recommended Replenishment Algorithm

Require: Sales forecast for each product category $P[1...n](n = 6)$, Loss rate for each product category $R[1...n](n = 6)$

Ensure: Recommended daily replenishment quantity for each product category $Res[1...n]$ $(n = 6)$

$Res \Leftarrow [1...n]$

for $i = 1$ to n **do**

$Res[i] \Leftarrow P[i]/R[i]$

end for

return Res

4. Experiments

4.1. Data Pre-processing

In the realm of big data analysis and modeling, data quality holds paramount importance[12], forming the foundation for subsequent analyses. [13] Given the extensive volume and complexity of supermarket data, effective pre-processing is essential to ensure data reliability and usability.

(a) Outlier Detection:

Outliers can significantly impact analysis accuracy. The Box Plot method [14], is employed for their identification, offering a visual representation of data distribution. Outliers, defined as data points beyond the interquartile range, are handled by replacing

them with mean values to enhance data accuracy and stability, ensuring reliable results in subsequent analyses.

(b) Daily Sales Calculation

After data cleaning, the next step involves data integration, including the calculation of daily sales for each category.

(c) Dataset Segmentation

To evaluate the model, the dataset was split into training set [15] (70%) and test set [16] (30%) using a random approach. Each of the six vegetable categories followed this same division, with 'Foliage', for example, having 232,378 samples in the training set and 99,590 samples in the test set.

4.2. Predicting Daily Sales and Daily Replenishment Based on Seasonal ARIMA Model

In this research, the steps for predicting weekly replenishment using the Seasonal ARIMA model are as follows:

(a) Decomposition of Time Series: Time series data for various categories were decomposed into four components: the original series, trend series, seasonal series, and random series. Initial analysis revealed a discernible seasonal effect in each category's data, consistent with real-world observations. Furthermore, it was evident that vegetable sales closely correlated with changing seasons.

(b) Verification of Smoothness: Augmented Dickey-Fuller [17] (ADF) tests were conducted to confirm the smoothness of the time series data. The results yielded P-values below 0.05 for all six categories, rejecting the null hypothesis and affirming that all categories represented smooth time series data.

(c) Time-series Ordering: Graphical analysis, utilizing autocorrelation function [18] (ACF) and partial autocorrelation function [19] (PACF), along with estimation based on truncated tails, determined the seasonal order of the model and seasonal difference order.

(d) Forecasting: The Seasonal ARIMA Model, based on the determined order, was employed to make predictions of future daily sales. Daily replenishment was calculated using the formula from **Section 3.1**.

4.3. Predicting Increasing Rates Based on Linear Regression

In this research, the steps for predicting each category's increasing rate for the upcoming week using a gradient descent model with linear regression are as follows:

(a) Linear Regression Modeling: Separate Linear Regression Models were constructed for each category to predict the pricing increasing rate. These models utilized historical daily sales data as input features, expressed as follows:

$$IR = \beta_0 + \beta_1 \times SP \tag{2}$$

In the Eq.2, β_0 and β_1 are the coefficients of the linear regression model, determined by training the model to minimize the loss function.

(b) Linear Regression Model Training: Training set was used to train a Linear Regression Model. Find the best coefficients, β_0 and β_1, to minimize the loss function, ensuring a good fit to historical data's increasing rate.

(c) Predicted Increasing Rate: The daily sales predictions for each category for the upcoming week, as determined in **Section4.2**, were combined with the previously developed linear regression model. This combined approach was used to predict the increasing rate for each category in the coming week.

4.4. Predicting Overall Revenue Based on GBDT Model

In this research, the steps for predicting each category's earnings for the upcoming week using the GBDT Model are as follows:

(a) GBDT Modeling: The analysis involved characteristics such as daily sales and pricing increasing rates for each category, with the target variable being the revenue generated on the same day.

(b) GBDT Model Training: The GBDT model was fitted using the training set, and decision trees were iteratively generated. The model was progressively improved using gradient boosting. After each training round, model performance was assessed with the test set to determine if more trees or hyperparameter adjustments were necessary.

(c) Final Predictions: Leveraging the established GBDT Model along with previously forecasted daily sales and increasing rates, predictions were made for each category's future revenue, aiming to maximize overall revenue.

4.5. Results

Through these steps, a comprehensive forecasting and decision-making system has been created to help superstores optimize their pricing and replenishment strategies, ultimately maximizing revenue. This approach combines three composite models: Seasonal ARIMA, Linear Regression, and GBDT, leading to improved forecasting accuracy and reliability. The models' accuracy is demonstrated in Table 2, where it's evident that the the Goodness of Fit [20] R^2 values are all close to 1 and the MSE values are all close to 0, indicate high model accuracy and improved fitting.

Table 2. Model Performance Evaluation Results for Test Set Data.

	Cauliflower	Peppers	Edible Mushrooms	Eggplant	Foliage	Aquatic Rhizomes
R^2	0.993	0.994	0.991	0.995	0.987	0.997
MSE	0.065	0.015	0.072	0.084	0.056	0.047

5. Conclusion

This research adeptly tackles the challenges faced by urban fresh produce supermarkets, particularly with perishable vegetables. It utilizes machine learning and extensive sales data to create precise pricing and replenishment strategies. By blending SARIMA, Linear Regression, and GBDT, it attains an impressive average R^2 of 0.993 across diverse vegetable categories, showcasing exceptional model accuracy.

Future work will refine our approach to meet evolving market demands. It will include a detailed analysis of inventory, competition, weather, and other factors to customize replenishment and pricing strategies, optimize inventory management, and reduce

waste. These efforts will enhance agricultural product supermarkets' sustainability and provide valuable support to decision-makers in dynamic markets.

This research exemplifies the integration of data science and management practices, providing valuable insights for future research and practical applications.

References

[1] Stuber, J., Lakerveld, J., Kievitsbosch, L., Mackenbach, J. & Beulens, J. Nudging customers towards healthier food and beverage purchases in a real-life online supermarket: a multi-arm randomized controlled trial. *BMC Medicine*. **20**, 1-13 (2022)

[2] Feng, L., Wang, W., Teng, J. & Cárdenas-Barrón, L. Pricing and lot-sizing decision for fresh goods when demand depends on unit price, displaying stocks and product age under generalized payments. *European Journal Of Operational Research*. **296**, 940-952 (2022)

[3] Banerjee, T., Sinha, S. & Choudhury, P. Long term and short term forecasting of horticultural produce based on the LSTM network model. *Applied Intelligence*. pp. 1-31 (2022)

[4] Wang, X. Joint Decision Making of Replenishment, Pricing, and Fresh Keeping Input in Fruit and Vegetable Cold Chain: Based on Markov Process. *Mobile Information Systems*. **2022** (2022)

[5] Shi, R. & You, C. Joint dynamic pricing and freshness-keeping effort strategy for perishable products with price-, freshness-, and stock-dependent demand. *Journal Of Industrial And Management Optimization*. **19**, 6572-6592 (2023)

[6] Guilding, C., Drury, C. & Tayles, M. An empirical investigation of the importance of cost-plus pricing. *Managerial Auditing Journal*. **20**, 125-137 (2005)

[7] https://en.wikipedia.org/wiki/Cost-plus_pricing

[8] Tseng, F. & Tzeng, G. A fuzzy seasonal ARIMA model for forecasting. *Fuzzy Sets And Systems*. **126**, 367-376 (2002)

[9] James, G., Witten, D., Hastie, T., Tibshirani, R. & Taylor, J. Linear regression. *An Introduction To Statistical Learning: With Applications In Python*. pp. 69-134 (2023)

[10] Maulud, D. & Abdulazeez, A. A review on linear regression comprehensive in machine learning. *Journal Of Applied Science And Technology Trends*. **1**, 140-147 (2020)

[11] Zhao, X., Li, N., Wang, W. & Yang, X. Research on mobile service satisfaction prediction model based on GBDT regression algorithm. *2023 5th International Conference On Communications, Information System And Computer Engineering (CISCE)*. pp. 26-29 (2023)

[12] Ghasemaghaei, M., & Calic, G. (2019). Can big data improve firm decision quality? The role of data quality and data diagnosticity. Decision Support Systems, 120, 38–49.

[13] Zha, D., Bhat, Z., Lai, K., Yang, F. & Hu, X. Data-centric AI: Perspectives and Challenges. *Proceedings Of The 2023 SIAM International Conference On Data Mining (SDM)*. pp. 945-948, https://epubs.siam.org/doi/abs/10.1137/1.9781611977653.ch106

[14] Moeini, B., Haack, H., Fairley, N., Fernandez, V., Gengenbach, T., Easton, C. & Linford, M. Box plots: A simple graphical tool for visualizing overfitting in peak fitting as demonstrated with x-ray photoelectron spectroscopy data. *Journal Of Electron Spectroscopy And Related Phenomena*. **250** pp. 147094 (2021)

[15] Cheng, J., Xu, Z., Wu, W., Zhao, L., Li, X., Liu, Y. & Tao, S. Training set selection for the prediction of essential genes. *PloS One*. **9**, e86805 (2014)

[16] Berrar, D. & Others Cross-Validation.. (2019)

[17] Roza, A., Violita, E. & Aktivani, S. Study of Inflation using Stationary Test with Augmented Dickey Fuller & Phillips-Peron Unit Root Test (Case in Bukittinggi City Inflation for 2014-2019). *EKSAKTA: Berkala Ilmiah Bidang MIPA*. **23**, 106-116 (2022)

[18] Basri, G., Streichenberger, T., McWard, C., Edmond IV, L., Tan, J., Lee, M. & Melton, T. A New Method for Estimating Starspot Lifetimes Based on Autocorrelation Functions. *The Astrophysical Journal*. **924**, 31 (2022)

[19] Yakubu, U. & Saputra, M. Time series model analysis using autocorrelation function (acf) and partial autocorrelation function (pacf) for e-wallet transactions during a pandemic. *International Journal Of Global Operations Research*. **3**, 80-85 (2022)

[20] Liu, D., Zhu, X., Greenwell, B. & Lin, Z. A new goodness-of-fit measure for probit models: Surrogate R 2. *British Journal Of Mathematical And Statistical Psychology*. **76**, 192-210 (2023)