Forecasting superstore fruit and vegetable sales based on ARIMA-LSTM

Yutong Wang*, Can Cui

School of Mathematical Sciences Soochow University, Soochow University, Suzhou, China, 215006

*Corresponding author: 2207404017@stu.suda.edu.cn

Abstract. In the context of a fresh food supermarket, managing perishable vegetables, which have a notably limited shelf life, maintaining optimal stock levels becomes imperative for safeguarding profitability. This research paper intricately explores the analysis of numerous datasets using Power Query and Power BI, with a pivotal focus on devising a succinct coding system. To ensure data quality, outliers are meticulously eliminated using advanced techniques like linear interpolation and Z-Score processing. The ARIMA time series model is employed to predict the sales volume for each category in July, having been trained on data from June 2022 to June 2023, adeptly capturing historical sales patterns and emerging trends. Additionally, a parameter-tuned LSTM recurrent neural network, optimized with Adam's algorithm, is utilized to refine predictions, considering its inherent memory capabilities. A thoughtfully designed weighted average approach combines outputs from both models, strategically minimizing overfitting and enhancing prediction accuracy. The resultant research fortifies inventory management optimization, bolstering the supermarket's profitability while effectively meeting customer demands.

Keywords: Fresh vegetables, Sales forecasting, ARIMA, LSTM.

1. Introduction

In fresh food superstores, general vegetable fresh goods are characterized by perishable, fragile and random short life cycle cycles. Most varieties cannot be re-sold the next day if they fail to be sold on that day. Therefore, supermarkets usually replenish goods on a daily basis based on the historical sales data and demand for each commodity. If the purchase quantity is significantly higher than the demand, too many discounted vegetables that are not sold and shipped will reduce the profitability of the superstore. Therefore, it is very important to forecast the amount of fruits and vegetables to be stocked on the same day.

The traditional method of predicting the sales of fruits and vegetables is to analyze the process and direction of the development of the object over time series through econometric models or statistical methods, so as to make analogical inferences and predict the future trend of change, and the most common prediction models include the differential integration moving average autoregressive model (ARIMA), grey model (GM), and the support vector machine (SVM) [1][2]. Compared to traditional statistical methods, AI has greater predictive power when dealing with nonlinear relationships, large-scale data, and complex features[3].

In this paper, Power Query and Power BI are used to correlate multiple tabular data to study the creation of a short code system while eliminating outliers, performing linear interpolation and Z-Score processing to ensure data quality and improve model accuracy[4]. To make category replenishment plans and pricing decisions for the coming week, sales are predicted by ARIMA time series[5] and parameterized Adam Optimized LSTM[6] recurrent neural network for each category and weighted average to achieve high accuracy and reduce overfitting[7].

2. Data analysis and processing

In this paper, the data used in the analysis is obtained by crawling data from MCM websites (https://cumcm.cnki.net/).Due to the large size of the data and the need to align it with data from other sources before analysis, a combination of Power Query and Power BI is chosen to merge all the data into a unified dataset.

As a result of the complex coding structure of individual products, it is not convenient for subsequent data analysis and statistical conclusions. Therefore, all the single products will be recoded based on their respective categories. The categories are assigned numeric values from 1 to 6, representing the thousands digit of the code. Additionally, a three-digit code will be assigned to each product based on an index, resulting in a new four-digit code for the category.

Furthermore, to enhance data comparability and address outliers in the analysis, the selling prices of individual products will undergo z-score normalization[8]. This normalization procedure transforms the selling prices into a standard normal distribution with zero mean and unit variance. By doing so, the values can be easily compared across different features, facilitating meaningful data analysis.

3. ARIMA-LSTM model sales forecasting

3.1. ARIMA model

3.1.1. Modeling

Simple Season (Simple Season), a special model in the SARIMA model, is the basic model used to deal with time series data with significant seasonal variations[9].

The system of equations to construct the simple seasonal model is as follows:

Horizontal smoothing equation:

$$l_t = \alpha\left(x_t - s_{t-m}\right) + \left(1-\alpha\right) l_{t-1} \tag{1}$$

Seasonal smoothing equations:

$$s_t = \gamma\left(x_t - l_{t-1}\right) + \left(1-\gamma\right) s_{t-m} \tag{2}$$

Predictive equations:

$$\hat{x}_{t+h} = l_t + s_{t+h-m(k+1)}, k = \left[\frac{h-1}{m}\right] \tag{3}$$

For the dataset that has been preprocessed, the total number of sales per month for each category is extracted as the independent variable x [10], which is brought into the solution.

3.1.2. Model solving

A simple seasonal model is applied to the data for each of the 6 categories and the model is solved as follows from SPSS.The result is shown in the figure 1~6, where, the red line in all charts is the actual sales volume and the blue line is the predicted sales volume.

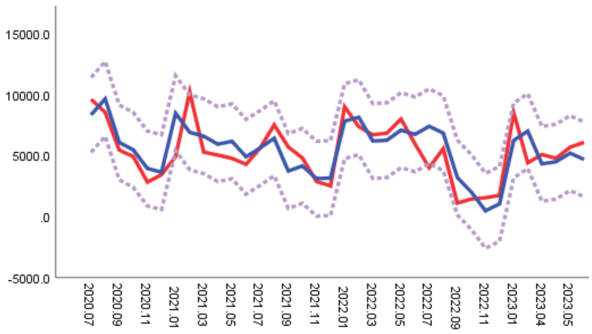

Figure 1 Trends in cauliflower sales

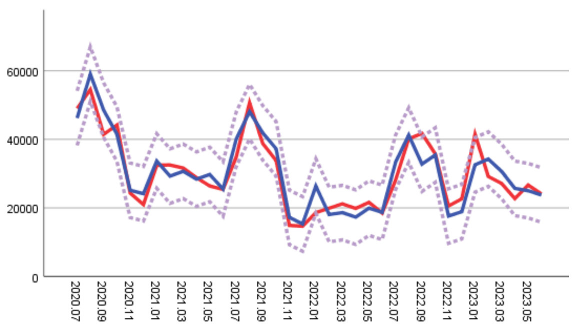

Figure 2 Trends in sales of foliage

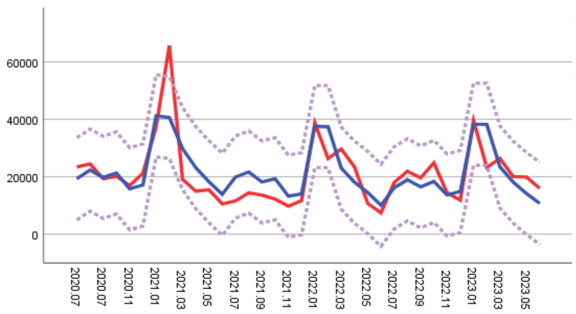

Figure 3 Trends in sales of chili peppers

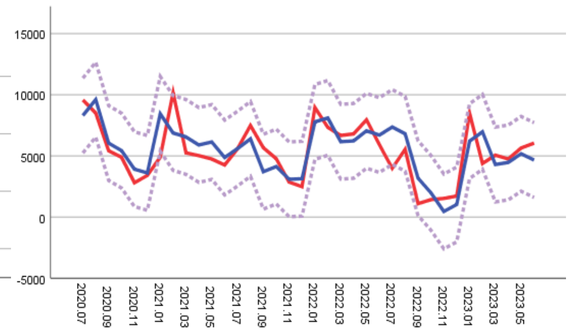

Figure 4 Trend in sales of eggplant

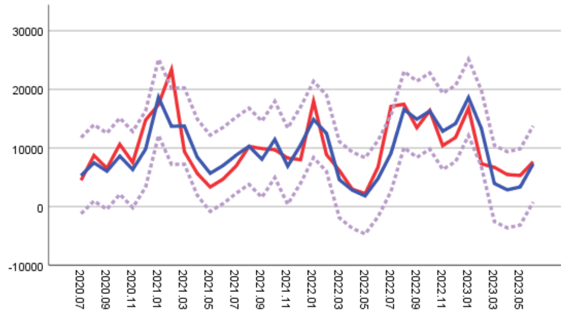

Figure 5 Trends in sales of aquatic rhizomes

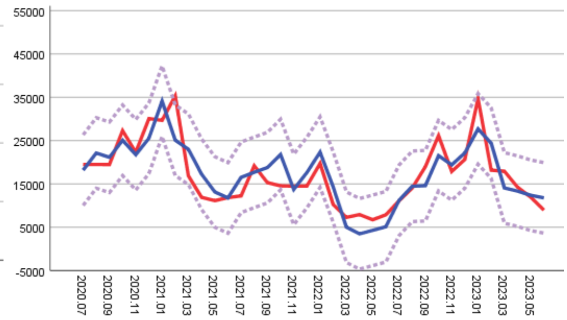

Figure 6 Trend in sales of edible mushrooms

The fitted metrics which comes from SPSS are also obtained is shown in Table1.

Table 1 Data table of fitted indicators for the six categories

	cauliflower	philodendron	capsicum	eggplant	Aquatic rhizomes	edible mushrooms
Smooth R2	0.646	0.583	0.848	0.695	0.727	0.746
R2	0.705	0.861	0.602	0.586	0.608	0.717
BIC	15.571	16.744	17.917	14.832	16.341	16.791

As can be seen from the data in the table, the simple seasonal model for each category is smoothed and processed with small values of the normalized Bayesian information criterion, which fits the data better, and it can be seen that the sales volume data for each category show seasonal cyclical changes, which verifies the previous data conjecture. Observing the annual change pattern of sales volume of each category, except for eggplant, the sales volume of the other five categories reached the maximum value of the year in February and then declined rapidly; there will be a small and steady increase in sales volume from July to September until December when the growth rate began to accelerate significantly until it reached the maximum value again in February of the following year. However, the trend of the eggplant category was observed to be different from that of all other categories, so it was hypothesized that the eggplant category was negatively correlated with the sales of the other categories, and it was hoped that this conclusion would be verified in the next correlation analysis[11]. The total sales forecast for July 2023 for each vegetable category which comes from SPSS is shown in Table2.

Table 2 Forecasted sales by category in July 2023 (in kg)

	philodendron	cauliflower	Aquatic rhizomes	eggplant	capsicum	edible mushrooms
July	6020.0751	1257.0558	1013.0007	695.3563	2760.2729	2069.3051

3.2. LSTM

When choosing a pricing decision model, it is also necessary to consider that the price of vegetable commodities may show cyclical changes over time, based on which Long Short-Term Memory (LSTM) is chosen as the decision model. It is a variant of Recurrent Neural Network (RNN) for processing and modeling time series data and sequence information. Compared to the traditional RNN structure, LSTM has stronger memory and is better able to capture and process long-term dependencies[12].

The structure of LSTM is formulated as follows:

Oblivion Gate:

$$f(t_i)=\sigma\left(w_f * x(t_i)+w_{hf} * h(t_i-1)+b_f\right) \tag{4}$$

Input Gate:

$$a(t_i)=\sigma\left(w_a * x(t_i)+w_{ha} * h(t_i-1)+b_a\right) \tag{5}$$

Candidate Memory:

$$c(t_i)=f_t * c(t_i-1)+a_t * \tanh\left(w_c * x(t_i)+w_{hc} * h(t_i-1)+b_c\right) \tag{6}$$

$$c(t_i)=f_t^* * c(t_i-1)+o(t_i)^* * c(t_i) \tag{7}$$

Memory Updates:

$$o(t_i)=\sigma\left(w_o * x(t_i)+w_{ho} * h(t_i-1)+b_o\right) \tag{8}$$

Hidden status:

$$h(t_i)=o(t_i) * \tanh c(t_i) \tag{9}$$

In order to find the optimal parameters of the model, this paper sets the objective function to be the MSE, which is a measure of prediction error of the time series on the test set, and sets the set of learning rates to be {0.001, 0.005, 0.01}, and the set of the number of hidden units to be {100, 500, 800}.

From the tuning matrix, the optimal parameter is the learning rate of 0.005 and the MSE is minimized when the number of hidden units is 100.

Time series charts for six categories from MATLAB is shown in Figure 7~12, and Table3 shows Forecasted sales by category in July 2023.

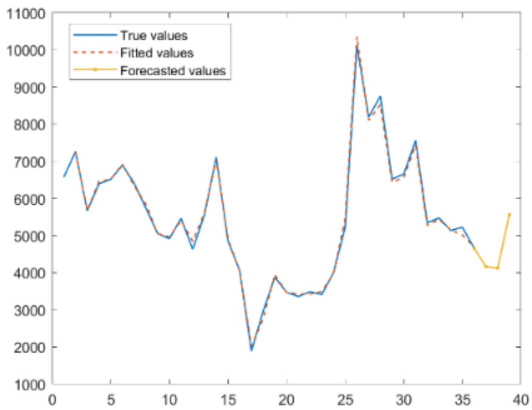

Figure 7 Foliage

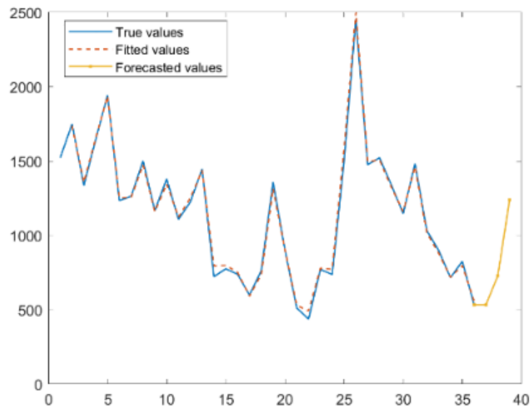

Figure 8 Cauliflower

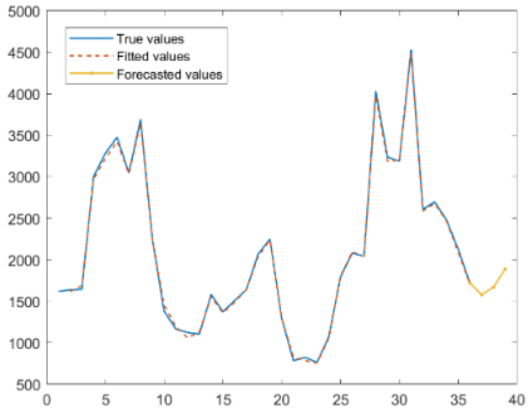

Figure 9 Edible Mushrooms

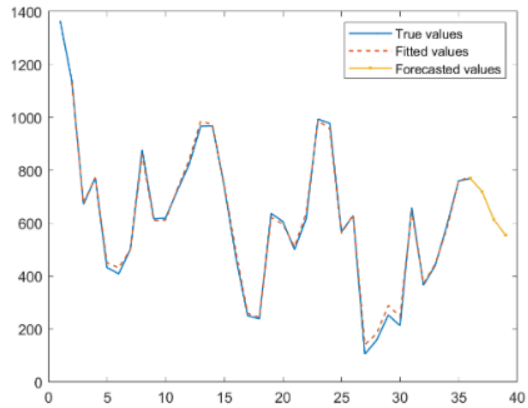

Figure 10 Solanaceae

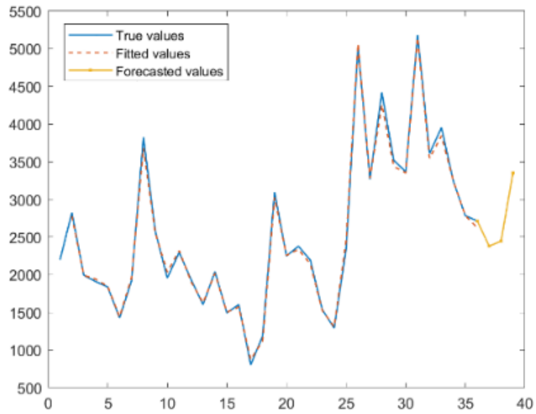

Figure 11 Chili Peppers

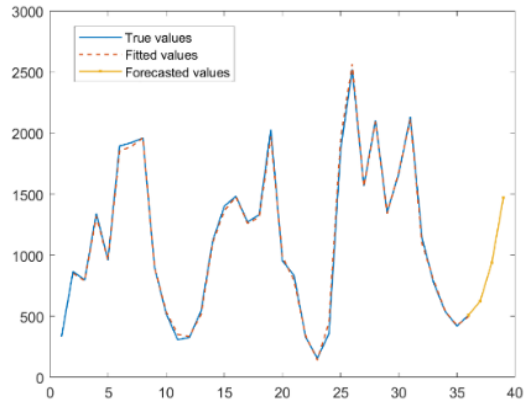

Figure 12 Aquatic Roots and Tubers

Table 3 Forecasted sales by category in July 2023 (in kg)

	philodendron	cauliflower	Aquatic rhizomes	eggplant	capsicum	edible mushrooms
July	4166.284	533.245	624.81	719.814	2380.11	1578.140

RMSE (Root Mean Square Error) is used to measure the prediction accuracy of the model, as can be seen from the training progress graph of LSTM, after 500 rounds of training, the RMSE is less than 0.5, which indicates that the prediction error of the model is relatively small, and it fits the data better; and as can be seen from the time series graphs of the LSTM model fitting for the above six categories, Fitted values and True values fit well, and the model prediction is more accurate. From the figure, except for eggplant, the sales of the other categories in the next month have increased, and aquatic roots and tubers, cauliflower, chili peppers rose more obviously, eggplant and other categories of sales were negatively correlated with the conclusions reached by the ARIMA model.[13]

Considering the high accuracy of both models, this paper weights and averages the forecasts of the two models to arrive at the final forecasted sales volume, as shown in the Table 4.

Table 4 Forecasted sales by category in July 2023 (in kg)

	philodendron	cauliflower	Aquatic rhizomes	eggplant	capsicum	edible mushrooms
ARIMA	6020.0751	1257.0558	1013.0007	695.3563	2760.2729	2069.3051
LSTM	4166.284	533.245	624.81	719.814	2380.11	1578.140
weighted average	5093.180	895.1504	818.90535	707.5852	2570.1915	1823.72255

4. Conclusions

Sales volume data for each category showed seasonal cyclical changes, with total vegetable sales reaching a yearly maximum during the Chinese New Year period in February each year, and sales volumes for all categories significantly higher than in other months. Flowering and leafy vegetables had the most varieties and the highest sales volume; chili peppers and edible mushrooms had similar totals, and cauliflower and aquatic roots and tubers were close to each other; while eggplant had the fewest varieties and had a significantly lower sales volume than the other categories.

There were some correlations between categories, with the strongest positive correlations for foliar and cauliflower, followed by edible mushrooms and cauliflower; eggplant had weaker correlations with the other categories and was negatively correlated with edible mushrooms, aquatic rootstocks and cauliflower.

The prediction accuracy of the LSTM model is high, and the root mean square error is less than 0.5 after 500 rounds of training, and the model can fit the data better. Combined with the prediction results of the ARIMA model, except for eggplant, the sales of other categories in the next month showed an upward trend, and the rise of aquatic roots and tubers, cauliflower, and chili peppers was more obvious.

Taking into account the accuracy of the LSTM and ARIMA models, a weighted average can be used to arrive at the final sales forecast.

References

[1] Cui Yunhao. Research on fresh vegetables sales prediction based on CNN-PSO-LSTM combination model[D]. Anhui Agricultural University, 2022.

[2] Lv Wang. Research on trend prediction of fresh vegetables sales based on improved SVM[D]. Anhui Agricultural University, 2021.

[3] Guo, S.J., et al. "An artificial intelligence-based method for predicting daily shipments of fruit and vegetable fresh produce stores.", cn202111271616.9. 2022.

[4] Zhu J.-Q. Study on pre-seismic longwave radiation anomalies based on time series prediction model[D]. Institute of Earthquake Prediction, China Earthquake Administration, 2023.

[5] LU Zhiping, YU Xiaojing, LU Chengyu. A new energy vehicle sales prediction method integrating VMD and LSTM models[J]. Journal of Wuhan University of Technology (Information and Management Engineering Edition), 2023, 45(04): 546-551.

[6] HU Xiongwei. Research and application of ARIMA-LSTM based combined model in cylinder sales forecasting[D]. Shanghai Second University of Technology, 2022.

[7] Zhou Yubing. Research on inventory control and dynamic pricing of fresh products based on deep reinforcement learning[D]. Central South University, 2022.

[8] Lv Wang. Research on trend prediction of fresh vegetables sales based on improved SVM[D]. Anhui Agricultural University, 2021.

[9] Xu Qi, Zhang Yangkui. An analysis of dynamic sales forecasting model for retail apparel based on ARIMA model and random forest model[J]. China Management Informatization, 2022, 25(06): 100-104.

[10] Kangsha. Research on joint decision-making of pricing and inventory replenishment for dual-channel sales of fresh produce under different demand characteristics[D]. Chongqing Jiaotong University, 2022.

[11] Wu X. Research on synergistic optimization of inventory and distribution of urban fresh produce cold chain logistics [D]. Beijing Jiaotong University, 2019.

[12] CHEN Dehong, LI Changyun. Short-term clothing sales prediction based on LSTM neural network[J]. Modern Information Technology, 2021, 5(20): 106-108.

[13] GUI Sisi, SUN Wei, XU Xiaofeng. Analysis of automobile sales forecasting based on combined ARIMA and linear regression model[J]. Computer and Digital Engineering, 2021, 49(08): 1719-1723.