Marketing Strategy Optimization: A Case Study Based on SARIMA And Genetic Algorithm

Kaiyuan Li [1, #, *], Jingyang Bao [2, #], Minyue Yang [3, #]

[1] Department of Polymer Materials and Engineering, East China University of Science and Technology, 200237, Shanghai, China

[2] Department of Automation, East China University of Science and Technology, 200237, Shanghai, China

[3] Department of Computer Science and Technology, East China University of Science and Technology 200237, Shanghai, China

* Corresponding Author Email: kyli0301@163.com

[#]These authors contributed equally.

Abstract. As societal living standard elevates, a heightened demand for enhanced quality of life, food quality, and freshness is emerging. In the fiercely competitive market economy, businesses are motivated to devise sensible pricing strategies and optimize their replenishment methods to maximize profits. In pursuit of profit maximization, this paper proposes a novel fusion optimizing-forecasting model based on SARIMA and a Genetic Algorithm for determining the optimal daily replenishment volume and pricing strategy in two steps: considering vegetable categories and considering vegetable items. In the case of vegetable sales volume forecasting and sales strategies, the applied model produces a 7-day prediction of sales volumes and the recommended pricing strategy selecting 27 products, resulting in a maximum profit of 1243.9 yuan. The outcome holds considerable significance for supermarkets in formulating profit-driven pricing strategies.

Keywords: SARIMA, Genetic Algorithm, Big Data.

1. Introduction

In contemporary society, there is a growing demand for food, and the perishable nature of most vegetables poses a challenge as they tend to lose their pristine appearance and flavor rapidly. Only some selected specialized vegetable categories have a longer shelf life, which sustains food marketability to some extent. Since consumers have taken food freshness as a pivotal factor influencing purchasing decisions, both businesses and researchers have been working on prioritizing the management of the sales duration for all vegetable products.

The price prediction problem, mainly used on a time-series dataset for analysis, is commonly solved by the following methods: Ariyo A A and others (2014) used the ARIMA Model in stock price prediction and concluded that ARIMA models can compete reasonably well with emerging forecasting techniques in short-term prediction [1]. The deep natural networks-based prediction model (Hu et al. 2021) displayed a higher prediction capacity in the context of stock indices than the alternatives [2]. Advanced machine learning algorithms such as gradient boosting algorithm (Truong Q, Nguyen M, Dang H, et al 2020) displayed the highest accuracy regarding house price predictions [3]; and Gegic E, Isakovic B, Keco D (et al, 2019) discovered that on car price prediction task, novel machinery learning techniques could elevate the accuracy to 87.38%, reaching a new height in the goodness of fit [4]. Since the ARIMA model was built for time serial data and predicting future trends, the fusion research method was given birth (Poongodi M and others 2020) to plumb deeper into the technology lying underneath Bitcoin's network and the various machine learning predictive algorithms. [5]

Genetic algorithm can be implemented in optimization, feature decisions, and path-planning problems. The official IEEE conference essay [6] overviewed the past evolution (Lambora A and others 2019) and proposed future development [7] (Katoch S, Chauhan S S, Kumar V 2021) of GA,

as conventional GAs have deduced way of choosing appropriate crossover and mutation operators while upcoming challenges fall to simulation of more natural evolution process such as human immune system. Gen M et al. (1997) applied genetic algorithm to partitioning problems [8]and set a precedent for the variation of research fields in upcoming research. To be more specific, in the area of optimization, D'Angelo G. and Palmieri F.(2021) [9] effectively applied modified genetic algorithms with gradient-based local search, which later beat its rivals with fewer iterations and higher scores of precision.

However, the traditional ARIMA model performs poorly in handling complex non-linear patterns inherent in financial time series data; also, the application result of genetic algorithms in optimization tasks can sometimes be undesirable, mainly problem-specific, for its performance may even worse than a random solution.

In this paper, a novel fusion-optimizing forecasting model based on Seasonal Autoregressive Integrated Moving Average Model (SARIMA) and Genetic Algorithm is proposed. The prices and sales volumes of some vegetables from a local supermarket are taken as the dataset for the model application. The core idea of the model is to combine the seasonal-ARIMA model with an optimization task using a Genetic Algorithm to solve a nonlinear prediction problem. By controlling the fitness function in GA, the forecasting model could avoid overlooking the goal of maximizing profit, overcoming its possible deviation from reality.

2. Methodology

2.1. Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

A seasonal autoregressive integrated moving average (SARIMA) model is an advanced time series forecasting technique that extends the traditional ARIMA model to account for seasonality in data.

A typical $SARIMA(p,d,q)(P,D,Q)_S$ model is characterized by 7 main parameters: (1) autoregressive order (p), representing the number of lag observations included in the model; (2) integrated order (d), indicating the number of times the raw observations are differenced to achieve stationarity; (3)moving average order (q), denoting the size of moving average window; (4) seasonal autoregressive order (P), adding lag observations at seasonal intervals; (5) seasonal integrated order (D), seasonal differences required for stationarity; (6) seasonal moving average order (Q), the size of the seasonal moving average window; (7) seasonal period (s), specifying the number of observations per season.

The SARIMA model is expressed in formula (1):

$$\Phi_p(B)\phi_P(B^s)(1-B)^d(1-B^s)^D X_t = \Theta_q(B)\theta_Q(B^s)Z_t \quad (1)$$

where $\Phi_p(B)$ and $\phi_P(B^s)$ are respectively the autoregressive and seasonal autoregressive polynomials. $(1-B)^d$ and $(1-B^s)^D$ are the non-seasonal and seasonal differencing operators.$\Theta_q(B)$ and $\theta_Q(B^s)$ are the moving average and seasonal moving average polynomials.X_t is the observed time series data. Z_t is a white noise error term.

The methodology of a typical SARIMA Model can be summarized into four steps, which is:

- Stationarity and Seasonality identification

Check for stationarity by examining the mean and variance over time. If the data is not stationary, apply differencing to remove trends or seasonality. Then utilize autocorrelation function (ACF) and partial autocorrelation function (PACF) plots to identify the seasonal pattern.

- Parameter Selection

Based on the ACF and PACF plots, identify potential values for the non-seasonal parameters (p,d,q) and seasonal parameters $(P, D, Q)_s$ .Typically, this procedure includes a nonlinear optimization method to minimize errors.

- Model Estimation

Apply a desirable estimation method, such as maximum likelihood estimation (MLE) to estimate the parameters selected. If model performance not satisfactory, adjust the seasonal components and parameters iteratively.

- Forecasting

Select the parameters that performs well on training datasets and deploy the model. Monitor its performance over time through diagnostic statistics and plots of residuals.

2.2. The Genetic Algorithm

The genetic algorithm (GA) is an intelligent optimization algorithm based on the natural selection of Darwin's theory. It tries to find the optimal solution by modeling the process of natural selection and reproduction. It has a wide range of applications and is also used in different areas for solving a wide range of problems [10].

In the Darwinian evolution theory, the individual traits of a population are preserved, while in the GA, the set of candidate solutions to the given problem is preserved. After that, it is further evaluated to create the next generation of solutions. Heritability and cross-mutation rates are also introduced so that the process of natural selection is better modeled.

The genetic algorithm consists of five key steps.

- Construction of individuals:

Individuals are usually identified by parameter vectors and initialized populations are randomly generated.

$$x = [x_1, x_2, x_3..., x_n] \tag{2}$$

- Construction of fitness function:

The fitness function f(x) is used to evaluate the survival quality of an individual in a genetic algorithm. In this question, it can be written as follows:

$$W = \sum_i x_i m_i - \sum_i r_i c_i \tag{3}$$

where m_i represents the sales volume of items, x_i represents the unit price, r_i represents replenishment quantities and c_i represents the unit price of purchase.

- Probability of selection:

The selection operation is based on the fitness function to decide whether the individual can be selected or not, usually, the selection probability can be solved by the following equation:

$$p(x) = \frac{f(x)}{\sum_{i=1}^{N} f(x_i)} \tag{4}$$

- Crossover operation:

Crossover operation is based on the parent individuals, which generate the offspring individuals. This is similar to the process of natural mating, in which each natural gene locus is selected from the parents.

$$x_{child}[i] = p_1 x_1[i] + p_2 x_2[i] \tag{5}$$

where p_1 and p_2 represent the probability of the selection of parent individuals.

- Mutation operation:

To bring out diversity, the mutation operation is usually given a probability of mutation, which makes the individuals of the offspring more diversified.

The specific flowchart is shown in Figure 1.

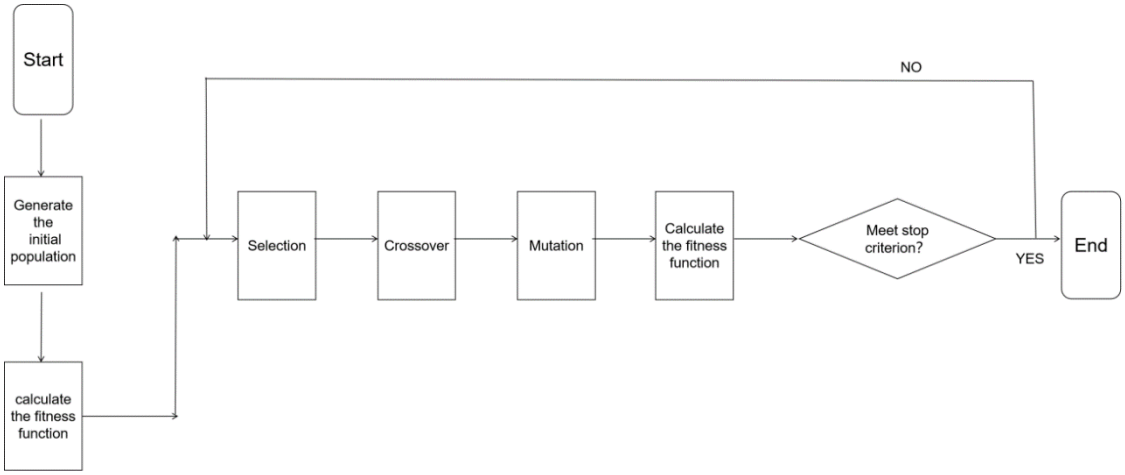

Figure 1: Scheme of the genetic optimization algorithm

3. Results

3.1. Description of data

Our study uses daily sales data of diversified vegetables from a supermarket in China, as a case study from July 1st, 2020, to June 30th, 2023. The data has 87,8504 lines, depicting the sales volume, the sales price of every individual vegetable item in different sales times, and the unit price of purchase of every individual vegetable each day. Additionally, the data is divided into six vegetable categories, cauliflower, foliar, chili, eggplant, edible mushrooms, and aquatics rootstocks categories, and each category has plenty of individual items. For example, the foliar category has 100 different individual items which are subordinate to the foliar category.

We are committed to using optimized algorithms to make the best sales strategy for the supermarket, not just predicting the profit trend using the predicted method. Firstly, we analyze the data of vegetable categories, and formulate a strategy of replenishment quantities and unit prices of different categories in the coming week. Secondly, we focus on the individual vegetable items, developing a sales and replenishment strategy to make the most profits the next day.

3.2. Sales prediction for vegetable categories in one week

To simplify the elaboration for each category, here foliar category is taken as the model case below. For the remaining categories, the same procedures are done, and the final predictions and analysis results are given at the end of this part.

Firstly, we visualize the tendency, seasonal indicator, and real sales volumes (Figure 2). Judging from the outcomes, time sequence data for the foliar category shows high seasonality, which means further analysis using the SARIMA model should be effective.

Subsequently, we conducted ADF tests (to assess whether a time series has a unit root, thereby determining the stationarity of the series) on the data, both on the original dataset, first-order difference dataset, and first-order difference-first order seasonal difference dataset. The results, as presented in Table 1, indicate that after performing first-order differencing, the data satisfies the 1% significance level, affirming the stationarity of the time series.

Following this, to ascertain the parameters in the model, we computed the autocorrelation function (ACF) and partial autocorrelation function (PACF) separately (Figure 3). The determination of parameters p, q, P, and Q was based on the decay or truncation patterns observed in the autocorrelation and partial autocorrelation coefficients.

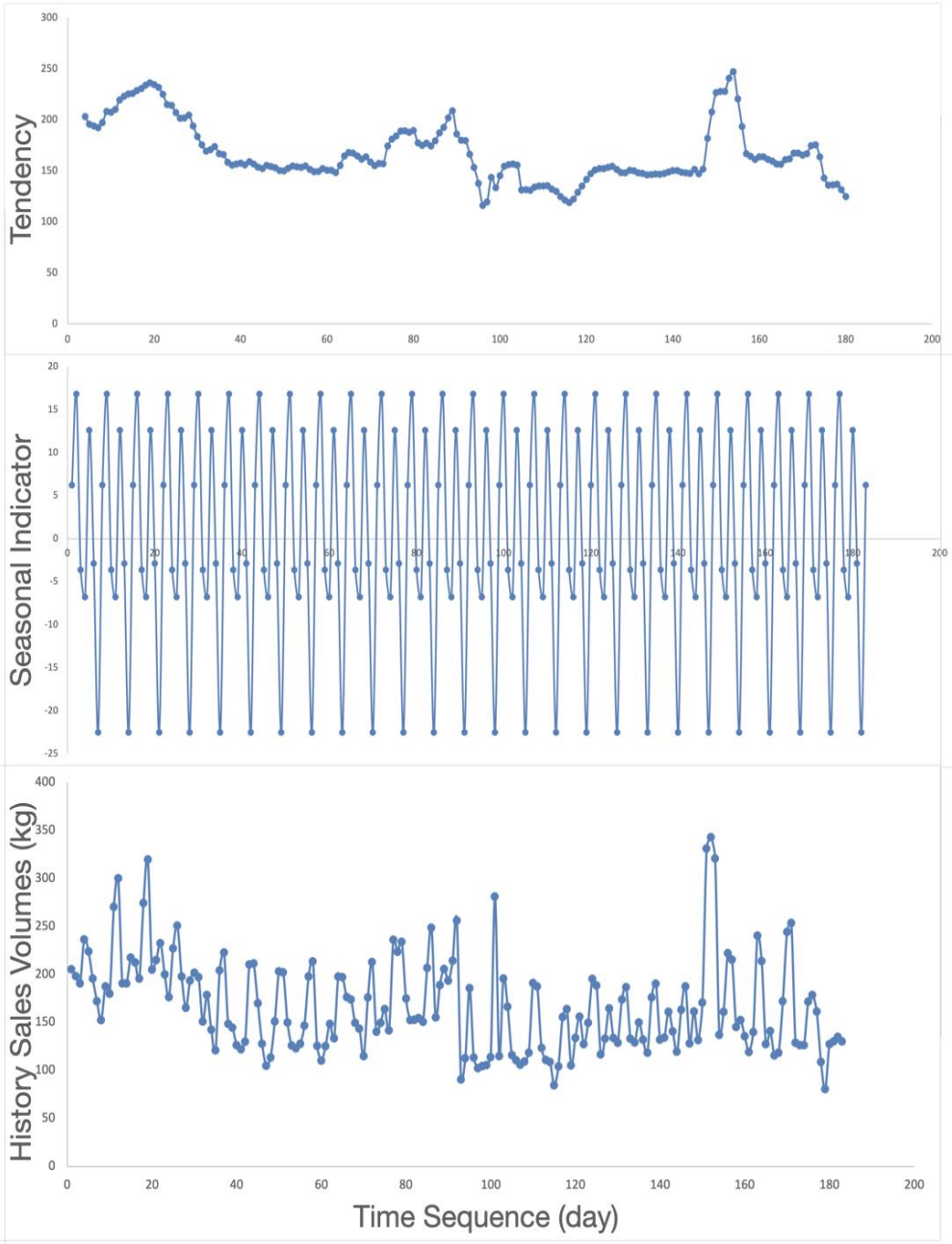

Figure 2: Tendency(top), seasonal indicator(middle) and real sales volumes(bottom) of foliar

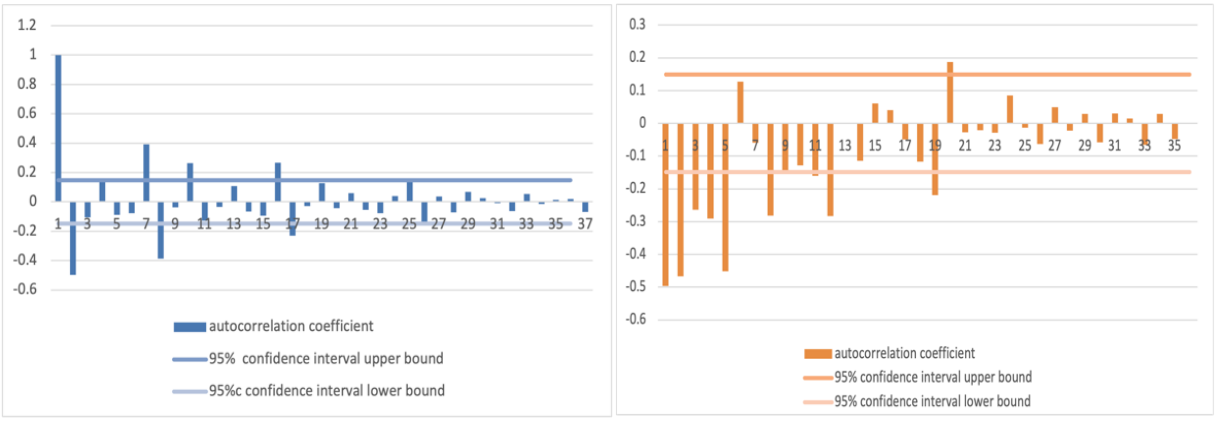

Figure 3: ACF plot (left) and PACF plot (right)

Combining the results from the aforementioned calculations and the automatic parameter selection based on the Akaike Information Criterion (AIC) approach, we employed SARIMA (0, 1, 2) (1, 0, 0)$_7$ as the model parameters.

Table 1: ADF Test table

Sequence	t	P value
Original	-2.858	0.051*
First order difference	-7.21	0.000***
First order difference- First order seasonal difference	-5.358	0.000***

Table 2: Model evaluation

Statistic	Item	Value
Q-Statistic for Residuals	Q_6	0.744
	Q_{12}	0.904
	Q_{18}	0.427
	Q_{24}	0.567
	Q_{30}	0.574
Information Criteria	AIC	1865.837
	BIC	1878.653

Analysis of the residual Q-statistics (Table 2) indicates that Q_6 does not exhibit significance at a certain level, failing to reject the hypothesis that the residuals of the model follow a white noise sequence. Thus, the model is deemed to be generally satisfactory. As depicted in Figure 4, predicted sales volumes have deviations from real sales volumes. Whereas, from a macroscopic view, the predictive values exhibit a certain degree of representation of the true data at each point, indicating that model possesses reference significance. The goodness-of-fit for the model, as measured by the R-squared value, is 0.264, indicating less-well performance. However, the model has achieved the most desirable results from this model.

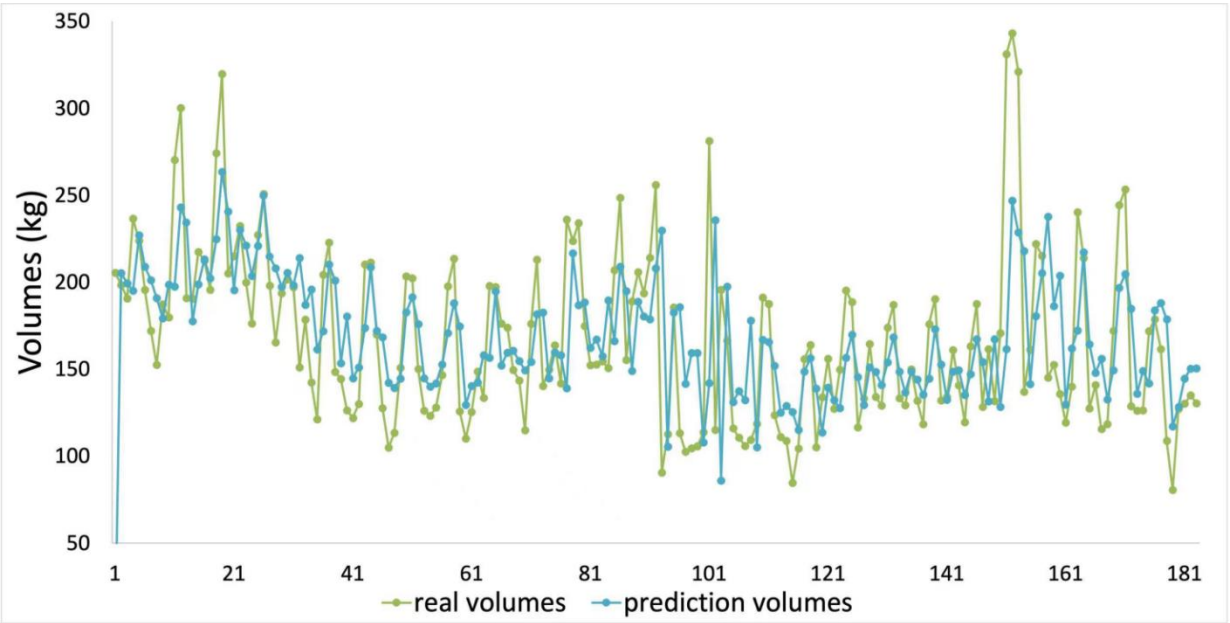

Figure 4: Model Fit Graph for Foliar

Table 3: Sales prediction for 6 vegetable categories in one week

Prediction date	Cauliflower	Foliar	Chili	Eggplant	Edible mushroom	Aquatics rootstocks
2023/7/1	21.61	144.66	92.14	28.61	51.31	19.99
2023/7/2	19.11	135.76	85.50	28.07	46.50	20.13
2023/7/3	16.98	127.64	74.90	21.92	44.60	18.28
2023/7/4	18.04	141.03	74.15	17.74	43.90	17.14
2023/7/5	18.73	141.86	75.53	20.53	50.94	20.89
2023/7/6	20.01	143.27	86.84	20.75	50.91	24.92
2023/7/7	19.05	141.94	80.71	23.66	43.93	19.32

Finally, apply the methodology to each vegetable category. Complete prediction results are shown in Table 3.

In light of the aforementioned results, it is observed that the forecasting sales volumes for the majority of vegetable categories are higher on Saturdays and Sundays, compared to the following days, aligning well with our daily experience. (Aquatics rootstocks here are an exception to this trend.) This observation directly correlates with our choice of setting the parameter S to 7 in the SARIMA model.

In terms of the absolute values of the forecasting sales volumes, they all fall within reasonable magnitude ranges for their respective categories, with no apparent deviations. Therefore, it can be considered that the utilization of this model is effective.

3.3. The optimal selling strategy for individual vegetable items

After developing an optimal strategy for each vegetable category, we are starting to investigate the optimal selling strategy for individual vegetable products under each category. The selling strategy contains two parts, the quantity of imported items and the price of the sale items.

We denote that W represents the maximum profits of the next day, and we need to develop an optimal model for x_i and r_i of i individual vegetable products.

$$W = \sum_i x_i m_i - \sum_i r_i c_i \tag{6}$$

where m_i represents the sales volume of items, x_i represents the unit price, r_i represents replenishment quantities and c_i represents the unit price of purchase.

However, there are two other variables we need to forecast based on previous data. For sales volume m_i, we use the weighted-average method to forecast the future m_i. After data analysis, we found that the sales volume of products is significantly influenced by time. Sales tend to be higher at the end of each week and lower in the middle of each week.

We denote the weight factor $\omega = 0.7$.

$$m_i = 0.7 \times (\frac{m_{sat} + m_{sun}}{2}) + 0.3 \times (\frac{m_{mon} + m_{tue} + m_{wen} + m_{thr} + m_{fri}}{5}) \tag{7}$$

Combined with the historical data analysis, the price of replenishment doesn't fluctuate highly in the short-term. So, for the unit price of purchase c_i, we average the price of last week to obtain the predicted values of the coming day.

As for the number of individual vegetable products, we count the number of vegetables sold on each day of the previous week, and we find that the number is between 27 to 33. Therefore, we can define the range of variable i to be 27 to 33.

Until now, we have completed the prediction of two variables in Eq. 6. The other two sets of variables will be solved and analyzed in the next step. However, considering the need to find the

optimal solution to this problem rather than simply predicting it, we use the genetic algorithm (GA) for an iterative solution.

The objective function sees Eq. 6:

The restrictive condition: (i = 27 for example)

$$\begin{cases} r_i \geq \max\{2.5, m_i\} \\ x_i^{\min} \leq x_i \leq x_i^{\max} \\ \sum_a r_a = 22.61 \\ \sum_b r_b = 141.66 \\ \sum_c r_c = 92.14 \\ \sum_d r_d = 28.61 \\ \sum_e r_e = 53.31 \\ \sum_f r_f = 19.99 \end{cases} \quad \begin{array}{l} i = 1,2,3...,27 \\ a,b,c,d,e,f < i \end{array} \tag{8}$$

Where a,b,c,d,e,f denote the quantity of imported items in the cauliflower, foliar, chili, eggplant, edible mushrooms, and aquatics rootstocks categories, which data are derived from 3.2. As for $x_i^{\min}$ and $x_i^{\max}$, we determine the value based on the previous week's maximum x and minimum x for each individual item.

Table 4: Decision variables table

DV	value	DV	value	DV	value	DV	value	DV	value	DV	value
x_1	7.77	x_{10}	9.20	x_{19}	5.89	r_1	41.11	r_{10}	24.07	r_{19}	7.58
x_2	14.00	x_{11}	5.83	x_{20}	3.55	r_2	18.99	r_{11}	9.02	r_{20}	2.50
x_3	5.88	x_{12}	25.80	x_{21}	4.97	r_3	4.98	r_{12}	12.86	r_{21}	28.73
x_4	6.00	x_{13}	4.66	x_{22}	20.58	r_4	19.73	r_{13}	10.15	r_{22}	10.77
x_5	3.70	x_{14}	1.96	x_{23}	5.56	r_5	3.86	r_{14}	14.43	r_{23}	4.62
x_6	15.99	x_{15}	15.09	x_{24}	6.00	r_6	21.99	r_{15}	4.54	r_{24}	3.51
x_7	17.96	x_{16}	14.00	x_{25}	4.74	r_7	13.30	r_{16}	3.52	r_{25}	12.57
x_8	2.77	x_{17}	11.91	x_{26}	8.00	r_8	7.12	r_{17}	2.83	r_{26}	36.18
x_9	19.81	x_{18}	5.15	x_{27}	5.06	r_9	29.77	r_{18}	2.62	r_{27}	3.95

*DV is short for decision variables.

By using the Genetic Algorithm, we get the r_i and x_i of 27 individual items, details are in Table 4. The final total profit is 1243.9 yuan.

After analyzing the case where i = 27, the following analyzes the profitability of the superstore in other cases.

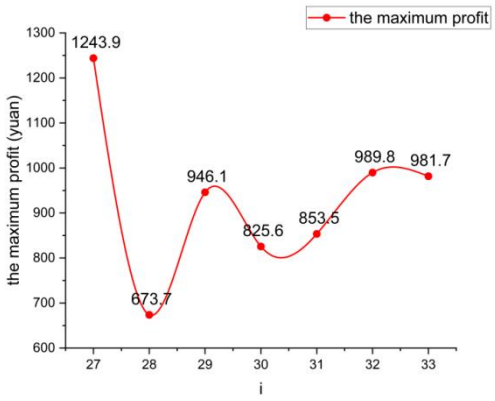

Figure 5: The maximum profit when different i

Figure 5 shows that the maximum profit fluctuates with the number of items in stock. With the selection possibilities of 27-33 items, the superstore makes the most profit when the item selection is 27, and the maximum profit is 1,243.9 yuan.

From the graph, it appears that the superstore earns the highest profit when it replenishes the least amount of individual items. The reason is that the superstore can meet the needs of the customers when its replenishment is 27. However, it doesn't exert too many appeals to customers when i = 28; in contrast, some sales volume of the individual items in 27-selections decrease slightly, which makes the total profits reduce highly. In the end, we can also draw the conclusion about which individual items we should replenish primarily to keep the total profit at a high level.

4. Conclusions

The issue of vegetable preservation in supermarkets has consistently garnered significant attention from consumers, and the implementation of an effective pricing strategy stands as a crucial determinant of a supermarket's competitive advantage. In this study, we employed a Seasonal ARIMA model and a Genetic Algorithm model to forecast future vegetable pricing based on available data, yielding relatively promising results.

Vegetable pricing occupies a pivotal role in a market economy. A comprehensive understanding of the intricate dynamics involving supply and demand relationships, production costs, and market competition enables the formulation of judicious pricing strategies. These strategies not only safeguard supermarket revenue but also cater to the evolving needs of consumers.

However, it is imperative to acknowledge the limitations of our approach. The utilization of genetic algorithms, while effective, introduced a significant challenge in terms of complexity, resulting in prolonged programming and execution times. Streamlining the objective function may enhance the efficiency of the model and expedite the computational process. Moreover, the pricing of vegetables is intrinsically linked to multifaceted factors that extend beyond the scope of our current models. Recognizing this, it becomes evident that a more nuanced and accurate prediction necessitates the incorporation of additional variables. For instance, the influence of weather factors on vegetable pricing should be considered to augment the precision of our conclusions.

References

[1] Ariyo A A, Adewumi A O, Ayo C K. Stock price prediction using the ARIMA model[C]//2014 UKSim-AMSS 16th international conference on computer modelling and simulation. IEEE, 2014: 106-112.

[2] Hu Z, Zhao Y, Khushi M. A survey of forex and stock price prediction using deep learning [J]. Applied System Innovation, 2021, 4(1): 9.

[3] Truong Q, Nguyen M, Dang H, et al. Housing price prediction via improved machine learning techniques[J]. Procedia Computer Science, 2020, 174: 433-442.

[4] Gegic E, Isakovic B, Keco D, et al. Car price prediction using machine learning techniques[J]. TEM Journal, 2019, 8(1): 113.

[5] 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.

[6] Lambora A, Gupta K, Chopra K. Genetic algorithm-A literature review[C]//2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE, 2019: 380-384.

[7] Katoch S, Chauhan S S, Kumar V. A review on genetic algorithm: past, present, and future[J]. Multimedia tools and applications, 2021, 80: 8091-8126.

[8] D'Angelo G, Palmieri F. GGA: A modified genetic algorithm with gradient-based local search for solving constrained optimization problems[J]. Information Sciences, 2021, 547: 136-162.

[9] Gen M, Cheng R, Wang D. Genetic algorithms for solving shortest path problems[C]//Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC'97). IEEE, 1997: 401-406.

[10] Behúnová A, Zemanová L, Behún M. Design of an Intelligent Application Using a Genetic Algorithm to Determine the Structure and Sales Volumes of Customized Products[J]. Mobile Networks and Applications, 2022: 1-9.