Prediction Of Vegetable Commodity Pricing Based on Machine Learning Algorithms

Xiaoyu Yang *, Zixuan Liu

North Alabama International College of Engineering and Technology, Guizhou University, Guizhou, China, 550025

* Corresponding Author Email: yxy040224@163.com

Abstract. This paper is dedicated to improving the profit of vegetable category goods in superstores and optimizing how to make efficient pricing strategy for vegetable category so as to maximize the revenue of superstores. Therefore, a pricing strategy model based on genetic algorithm for vegetable category goods is proposed. First, we analyze and calculate the Pearson correlation of sales volume between categories, and obtain the positive correlation of leafy and cauliflower categories and the negative correlation of aquatic root and eggplant categories. Then, the cost, price and profit margin of each vegetable category are calculated in a weighted way, and Spearman's correlation coefficient is applied to reveal the relationship between total sales volume and cost. At the same time, we find out the linear relationship between the total sales volume and the weighted sales price of vegetable commodities, and finally, with the total profit as the objective function and the linear regression equation between the total sales volume and the weighted sales price of vegetable commodities as the constraints, we use genetic algorithms to search for the optimum, find out the sales price and sales volume under the maximal profit, and then realize the formulation of the pricing strategy for vegetable commodities.

Keywords: Spearman Correlation Coefficient, Pearson Correlation, Linear Regression, Genetic Algorithm.

1. Introduction

The fresh produce market occupies a unique position in the contemporary retail industry, with vegetable items presenting retailers with a series of complex management challenges due to their short shelf life and variable nature [1]. Most vegetable varieties cannot be sold the next day if they are not sold on that day, forcing retailers to accurately predict market demand and implement timely and effective replenishment decisions to reduce wastage and increase sales revenue. However, retailers are faced with a wide range of vegetable varieties from different origins, and the transaction time for stocking is usually between 3:00am and 4:00am, when access to information is relatively limited [2-4]. Under such circumstances, retailers need to make replenishment decisions for various vegetable categories with incomplete information in order to maximise profits in this volatile market environment. The complexity of this task is self-evident, and it requires retailers to have strong data analytics capabilities to be able to conduct in-depth research from multiple perspectives, such as historical sales data, market price data, and consumer behaviour data, and to make quick and accurate decisions accordingly. In addition, because the freshness and quality of vegetables have a huge impact on consumers' purchasing decisions, retailers also need to take into account the display management of vegetables while making replenishment decisions to ensure the freshness of the products and enhance consumers' purchasing experience [5].

Therefore, for vegetable retailers, how to achieve effective management and profit maximisation of vegetable commodities through accurate data analysis and scientific decision-making in the face of these challenges is a notable and pressing issue in the industry today.

2. Descriptive statistical analysis of data

2.1. Data Presentation

The data sources in this article are related to http://www.mcm.edu.cn/html_cn/node/c74d72127066f510a5723a94b5323a26.html. Table. 1 below (only partially shown due to space reasons) shows the data of vegetable items categorised into categories, which contains the date of sale, item code, sales volume, sales unit price, and category information. Table 1 can be seen from the vegetable category sales volume and sales price are affected by the type of single product.

Table. 1 Pre-processing of data for item classification into categories

Sales date	code	Sales volume (kg)	Sales unit price (¥/kg)	Category
2020-07-01	102900005117056	0.396	7.60	Chilli
2020-07-01	102900005115960	0.849	3.20	Foliage
2020-07-01	102900005117056	0.409	7.60	Chilli
...	...	...	...	...
2023-06-30	102900005115250	0.125	24.00	Mushrooms

Through these labels, the sales volume and sales of each category are aggregated. This provides a comprehensive understanding of the sales of each category and enables a clear view of the differences in sales between categories. In order to deepen the understanding of the relationship between categories, this paper further processes the data and calculates the sales volume of each category over time. Sales volume serves as an intermediate variable that can reveal the dynamic relationship between different categories. In this paper, quarter is used as the independent variable, and the sales volume of different categories under each quarter is used as the dependent variable, according to which a line graph is plotted, as shown in Figure 1.

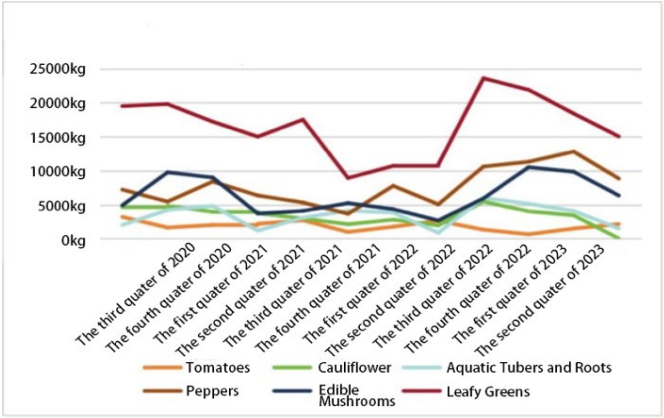

Fig. 1 Change in sales volume by category over the quarter.

The preliminary analysis reveals that there is a relationship between the sales volume of each category of vegetables, for example, the discount in the table from the second quarter of 2022 to the third quarter of 2022 are showing an upward trend (except for eggplant), while the first quarter of 2023 to the second quarter of 2023 are showing a downward trend (except for eggplant), so the next step continues to the analysis of the interactions between the different categories.

2.2. Pearson correlation analysis

Pearson's correlation coefficient is a statistical measure of the strength and direction of the linear correlation between two variables. It has a value between -1 and 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear relationship [6].

Equation (1) is the mathematical formula for Pearson correlation coefficient, and the Pearson correlation coefficient between the sales volume of each category is calculated by Equation (1), where P denotes the Pearson correlation coefficient, X, Y denotes two sets of vectors, and E(X), E(Y) denote the mean values of X and Y vectors, respectively.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} \tag{1}$$

The analysis of Pearson correlation coefficient gives a clear picture of the degree of correlation between the different categories and whether the correlation between them is positive or negative. According to Pearson correlation coefficient the correlation of each vegetable category was obtained as shown in Fig. 2.

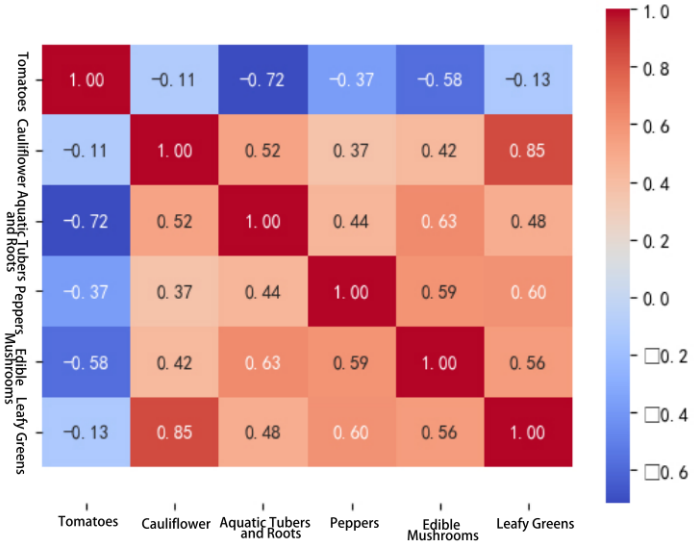

Fig. 2 Pearson's coefficient correlation analysis for each category of vegetables

Based on the analysis in Figure 2, the following conclusions are drawn: the Pearson coefficient on the main diagonal is 1. This is because this part of the data represents the sales volume of each category compared to itself, and naturally the results will show a perfect correlation. For the other non-diagonal elements, the values reveal the correlation between the different categories. Of all the category combinations, Cauliflower and Foliage have the highest Pearson's coefficient of 0.85. This value indicates that there is a significant positive correlation between Cauliflower and Foliage sales, i.e., an increase in Cauliflower sales tends to be accompanied by a simultaneous increase in Foliage sales. Comparatively, the lowest Pearson's coefficient of -0.72 was found for aquatic rootstocks and eggplant, which indicates a significant negative correlation between the sales volume of these two categories. Specifically, an increase in sales volume of aquatic rootstocks may lead to a decrease in sales volume of eggplant and vice versa. Finally, the Pearson coefficient for the eggplant and cauliflower categories is closest to zero, which indicates that the correlation between eggplant and cauliflower sales volumes is the weakest. In other words, changes in sales volume of eggplant are unlikely to have an impact on sales volume of cauliflower and vice versa.

Due to the significant positive correlation between sales of cauliflower and foliage, merchants may consider implementing joint or bundled sales strategies, such as promoting them together or offering some discounts for purchasing packages. This may further increase their sales volume because when consumers' likelihood of purchasing one of the items increases, their likelihood of purchasing the other increases accordingly. Because of the significant negative correlation between sales of aquatic roots and eggplants, merchants may need to avoid placing these two categories next to each other or promoting them at the same time. For example, merchants may choose to reduce eggplant inventory during peak sales periods for aquatic roots and tubers, and vice versa. This will avoid stock build-up and reduce the risk of obsolescence.

3. Establishment of pricing model for vegetable commodities

In this paper, the data are first pre-processed with the aim of removing returned goods and discounted goods in order to obtain valid data for each individual product, including wholesale price, selling price, total sales and cost margin, and some of the results of the data processing are shown in Table 2 (detailed data are shown in the supporting materials).

Table. 2 Results of removing returned goods and discounted goods

	date	Total sales/kg	wholesale price/¥	sale price/¥	cost margin	category
1	2020/9/18	0.195	14	23.6	0.685714	mushrooms
2	2020/9/27	1.308	17.1	27.6	0.614035	mushrooms
3	2020/9/28	0.163	17.1	27.6	0.614035	mushrooms
...	...	...	...	...	...	...
22694	2023/6/28	6	1.96	3	0.530612	mushrooms

From Table 2, it is clear that the impact of different individual items on the category is different and disparate, so a weighted approach was used to calculate the weighted price, weighted cost and weighted margin for each category. The purpose of weighting is to calculate overall metrics for the category based on the sales volume and importance of each individual item. The idea of weighting allows for the averaging of unit costs calculated when market prices rise or fall, which is more eclectic for the apportionment of inventory costs. This gives a more accurate picture of the price level, cost level and profit level of the category.

3.1. Spearman's rank correlation coefficient analysis

In order to explore in depth, the relationship between total sales and cost-plus pricing, a Spearman's correlation coefficient matrix was used to perform an aggregation analysis for each category and individual product [8].

The results shown in Figure 4 were obtained. As an example, a high positive correlation can be found between total sales and total profit for both the foliage category and the eggplant category, while a very high negative correlation exists between sales price and wholesale price. This finding may suggest a potential correlation that may exist between total sales and profitability. By looking at Figure 4 and Figure 5, both total sales and total profit for foliage and eggplant showed high positive correlations, implying that these two variables usually increase or decrease simultaneously. In other words, when total sales increase, total profit also increases and vice versa. This could be attributed to the increase in sales revenue due to the increase in total sales, which in turn increases total profit. On the other hand, there is a very high negative correlation between sales price and wholesale price, which means that these two variables usually decrease when one increases and the other decreases. This may be due to the fact that an increase in sales price may lead to a decrease in sales volume, which in turn leads to a decrease in wholesale prices, given that costs remain constant.

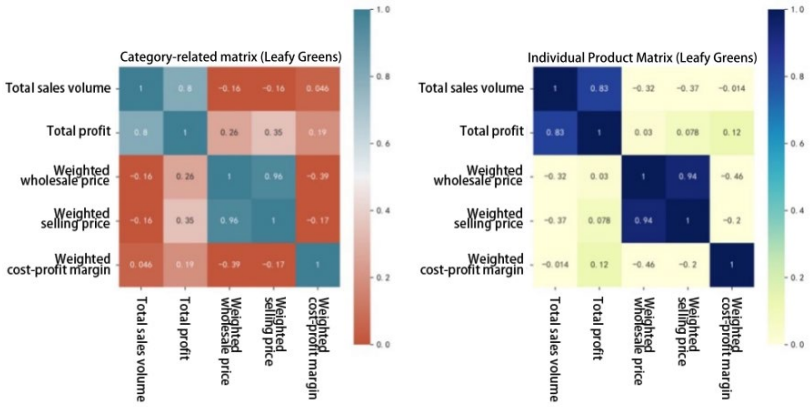

Fig. 3 Plot of Spearman rank correlation matrix for floral and foliar species

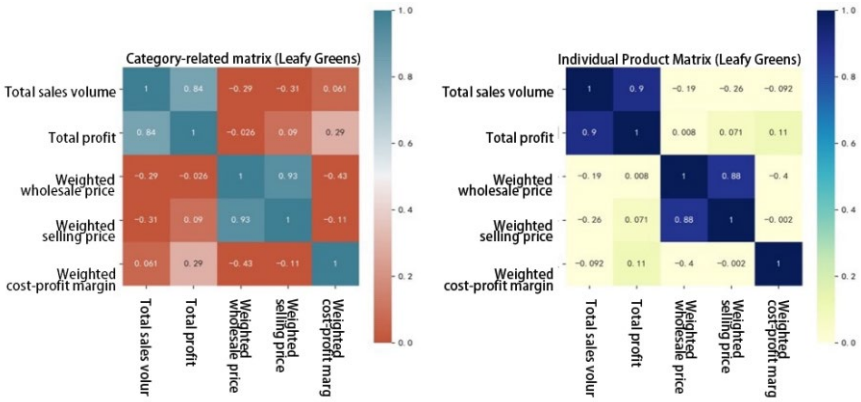

Fig. 4 Spearman rank correlation matrix for eggplant classes

These results hint at potential associations that may exist between total sales and profitability, and may provide important insights into product pricing, sales strategy, and inventory management. For example, managers can influence total sales volume and profitability by adjusting sales prices to achieve the goal of maximising profits. However, these correlations can only reveal associations between variables, not causality. Therefore, we cannot simply assume that changing one variable will change the other. A more in-depth study is needed in order to determine causality.

3.2. Pricing Model for Vegetable Products Based on Machine Learning Algorithm

To establish a linear regression equation (2) between total sales and weighted selling price of vegetable commodities.

$$Y = mx + b \tag{2}$$

Taking the flower and leaf category as an example, Figure 6 is obtained, in which the vertical coordinate of the left graph of the number of times each sales price occurs the proportion of the total number of times the station is recorded as the density, and the frequency of the occurrence of each sales price can be visualized from the left graph. The right figure is the linear regression equation obtained from the calculation, through the construction of the linear regression equation can be obtained by the sales volume of the flower and leafy vegetables category and weighted sales price is inversely proportional.

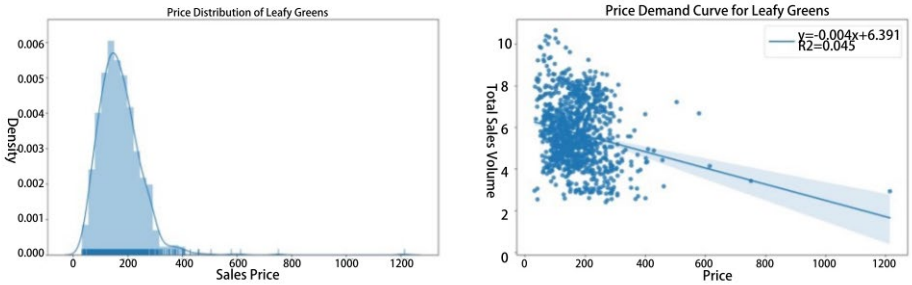

Fig. 5 Foliage sales density versus sales price

The price point and the corresponding replenishment volume that can maximize the expected return can be found according to the objective function of maximum return, which is shown in Equation (3), where W_1 denotes the total profit of the six categories, S_i denotes the sales volume of each category, P_i denotes the weighted sales price of each category, and B_i denotes the weighted wholesale price of each category.

$$W_1 = \sum_{i=1}^{6} (S_i P_i - S_i B_i) \tag{3}$$

st.

$$S_i = mP_i + b$$

Taking the total profit of W_1 as the objective function, and the linear regression equation of the total sales volume and the weighted sales price of vegetable products as the constraints, a genetic algorithm was used to find out the corresponding sales price and sales volume under the maximum profit.The results of profit-maximizing selling price and sales volume of cauliflower and leafy vegetable categories are shown in Tables 3, 4 as follows,Each row in the table 3 and 4 represents a possible strategy, including date of sale, weighted wholesale price, sales volume, and projected revenue:

Table. 3 Profit-maximizing sales price and volume of cauliflower category

Classification	Date	Sales Price	Sales Volume	Margins
cauliflower	2023-07-01	7.790023174	51.19999	260.8645423
cauliflower	2023-07-02	7.819433334	51.30726	257.0348356
cauliflower	2023-07-03	7.812731106	36.19845	184.8425181
cauliflower	2023-07-04	7.799150334	34.56317	176.2574834

Table 3 demonstrates is the price points and replenishment volumes that maximize returns for the cauliflower category. For the cauliflower category, the projected revenue is highest at 260.86 on July 1, 2023, with a volume of 51.20 and a weighted selling price of 7.79.

Table. 4 Profit-maximizing sales price and volume for foliage categories

Classification	Date	Sales Price	Sales Volume	Margins
foliage	2023-07-01	3.39856972	240.59914	793.8050999
foliage	2023-07-02	3.40178317	239.35398	814.0169364
foliage	2023-07-03	3.40219335	283.85173	965.4071751
foliage	2023-07-04	3.402245708	283.85173	965.414606

Table 4 then shows the price points and replenishment volumes that maximize returns for the Foliage category. For the floral and foliage category, July 1, 2023, has the highest projected revenue of 793.8, with a volume of 240.6 and a weighted sales price of 3.4.

4. Conclusions

In this paper, we first analyze and calculate the Pearson correlation of sales volume among vegetable categories, then calculate the cost, price and profit margin of each vegetable category by weighted method and use Spearman correlation coefficient to reveal the relationship between total sales volume and cost. At the same time, we find out the linear relationship between the total sales volume of vegetable commodities and the weighted sales price, and finally, with the total profit as the objective function and the linear regression equation between the total sales volume of vegetable commodities and the weighted sales price as the constraints, we use genetic algorithms to search for the optimum, find out the sales price and sales volume under the maximum profit, and then realize the formulation of the pricing strategy for vegetable commodities. The pricing strategy for vegetable products is an optimization problem, and since the sales volume and the weighted sales price are inversely proportional to each other, the objective of the strategy is to maximize the expected return. However, the effectiveness of this approach may be affected by many factors, such as changes in

market demand, competitors' strategies, etc. Therefore, managers may need to take these factors into account and flexibly adjust the replenishment volume and selling price in practice.

References

[1] Li Yuan. Research on Joint Optimisation of Pricing and Inventory for Dual-Channel Retailers Considering Reference Price Effect [D]. Yanshan University, 2022.

[2] Qiao Xue. Joint replenishment pricing strategy for fresh products considering sales loss [D]. Southeast University, 2021.

[3] Kang Sha. A study on joint decision-making of dual-channel sales pricing and inventory replenishment of fresh produce under different demand characteristics [D]. Chongqing Jiaotong University, 2022.

[4] Li Liang. Research on vegetable price formation based on industry chain perspective [J]. China Agricultural University, 2018.

[5] Ning Liyuan. Pricing strategies and techniques of green vegetables [J]. Guangdong Sericulture, 2020, 54(7):2.

[6] Yuan-Shang Zhao,Wei-Fang Lin. Study of typical scenarios based on Pearson correlation coefficient fusion of peak density and entropy weight method [J]. China Electric Power, 2023, 56(5):193-202.

[7] Carreteroayuso M J, Pinheiroalves M T, Bienvenidohuertas D. Sustainable building repair: A K-means approach to addressing fissures in ceramic brick partition walls [J].2023.

[8] YU Qun, HUO Xiaodong, HE Jian, et al. Trend prediction of power outages in China based on Spearman correlation coefficient and system inertia[J]. Chinese Journal of Electrical Engineering, 2023.

[9] ZHANG Ruiting, LIU Yu, HAN Aiqing, et al. Construction of hypoglycaemia prediction model for elderly type 2 diabetes patients based on random forest algorithm [J]. Chinese Journal of Practical Nursing, 2023, 39(23):1829-1835.

[10] Deng Y, Fan H, Wu S. A hybrid ARIMA-LSTM model optimized by BP in the forecast of outpatient visits [J]. Journal of ambient intelligence and humanized computing, 2023.