Commodity Sales Forecast Based on Cluster Analysis and Time Series

Wanying Lu[a], Dongyao Ren[b*]

[a]1816728182@qq.com, [*b]3032135352@qq.com

University of Information Engineering, Zhengzhou, Henan, 450001, China

Abstract. The price of vegetables in fresh supermarkets is affected by many factors such as time and sales. Reliable market demand analysis is especially important for replenishment and pricing decisions. This article delves into related issues in vegetable pricing. We select the sales data of the latest 2023 Mathematical Modeling Competition Question C as the data in this paper, first preprocess the data, summarize the data according to vegetable categories and single products every day, and observe the sales trend of categories and single products over time by drawing a line chart. Second, the distribution of the data is tested, and the P-P plot of each variable is drawn to determine whether it follows a normal distribution; Pearson correlation analysis and Spearman correlation analysis were used to calculate the correlation coefficient between categories and items, respectively. According to the correlation coefficient test, the model is not significant. Therefore, cluster analysis is used to explore the correlation between each vegetable category and a single product, and the cluster analysis lineage map is drawn to visually represent the correlation between each category and a single product. The data is then aggregated to plot a scatter plot between sales and cost plus price to observe possible functional relationships between them. Linear regression, loglinear regression and polynomial regression were used to train the data, plot the function curve and data scatter plot, and calculate the evaluation indicators of the regression model. The time series model is used to predict the sales volume of each category in the next week, and the relationship between sales volume and cost plus price is trained into the time series model as a penalty function. Finally, get the best replenishment volumes and pricing for each category next week so that merchants can get the highest profits.

Keywords: Pearson correlation analysis, Time series model, Prediction accuracy.

1. Introduction

In the fresh supermarket, the shelf life of general vegetable goods is relatively short, and the product deteriorates with the increase of sales time, most varieties such as unsold on the same day, can not be sold again the next day. Therefore, the supermarket usually replenishes the stock daily based on the historical sales and demand of each product.

In view of the nonlinear and linear influence of vegetables in sales forecasting [1]. Diets rich in fruit and vegetables (FV) are associated with favorable public health outcomes [2]. In recent years, affected by the new crown pneumonia, the fresh e-commerce industry has developed rapidly, and online buying fresh vegetables has gradually become people's first choice. However, fresh vegetables are easily affected by seasonal fluctuations, weather changes, holidays, prices and other factors, and sales often show a non-linear change trend [3]. With the advent of the era

ICEMBDA 2023, October 27-29, Tianjin, People's Republic of China

DOI 10.4108/eai.27-10-2023.2341974

of artificial intelligence, enterprises are facing various cost challenges, although the modern management of the catering industry has become increasingly mature, but also due to a large number of dishes backlog, resulting in rotten and wasteful dishes, thereby reducing profits, or insufficient supply of dishes, unable to meet consumer demand, resulting in a decline in turnover and a decline in customer satisfaction [4]. With the development of Internet technology and the wide application of big data, online and offline shopping methods have brought consumers a more convenient consumption experience [5]. But at the same time, for brick-and-mortar supermarkets, the competition is more intense [6].

With the influence of the Internet, the real economy has entered a new era of digitalization, in fresh supermarkets, the shelf life of general vegetable commodities is relatively short, and the product deteriorates with the increase of sales time, most varieties such as the same day is not sold, the next day can not be sold [7]. Therefore, supermarkets usually replenish daily according to the historical sales and demand of each product [8]. In the past, replenishment was done through the experience of store associates [9]. Now we can introduce clustering and time series based methods to forecast future replenishment volumes based on historical data to achieve the goal of maximizing revenue [10].

Based on the above research results, there are various results of using time series for sales forecasting, but due to the limitation of datasets and no precedent for large-scale dataset training, this paper selects the latest and most complete sales volume dataset to use time series models for prediction and discusses the results.

2. Introduction to datasets

The dataset contains two important information, one is time information and one is sales information. The dataset selects the sales data of six categories of vegetables in supermarkets, and counts all sales data from July 1, 2020 to June 30, 2023 by day.

3. Correlation analysis between different product sales

3.1. Pearson correlation analysis

Pearson correlation analysis is a method used to measure the degree of linear correlation between two variables. Its mathematical principles are based on the concepts of covariance and standard deviation. Covariance indicates the degree of overall joint variation between two variables and is calculated as:

$$Cov(X,Y)=E[(X-E(X))(Y-E(Y))] \tag{1}$$

where X and Y represent two variables, respectively, and E(X) and E(Y) represent their expected values, respectively.The standard deviation represents the degree of dispersion of a variable and is calculated as:

$$SD(X)=\sqrt{E[X-E(X)]^2} \tag{2}$$

The Pearson correlation coefficient r represents the degree of linear correlation between the two variables and is calculated as:

$$r=\text{cov}(X,Y)/(SD(X)*SD(Y)) \quad (3)$$

The Pearson correlation coefficient r ranges from -1 to 1. When r equals 1, it means that the two variables are completely positively correlated

Figure 1 is shown below

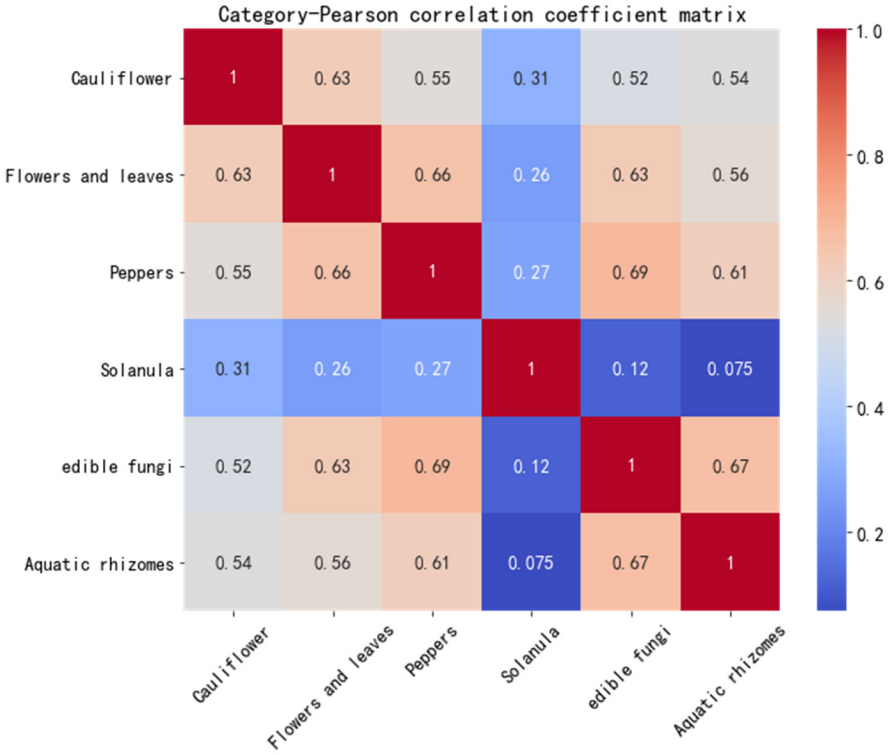


Figure 1. Pearson correlation analysis.
(Photo credit : Original)

3.2. logistic regression model

Spearman correlation analysis is a nonparametric statistical method used to measure the degree of correlation between two variables. Its mathematical principle is based on the sorting of hierarchical data. Suppose there are two variables X and Y, which take values x1, x2,..., xn and y1, y2,...,yn, respectively, we can sort them by size to get the rankings r1, r2,..., rn and s1, s2,...,sn. We can then calculate the rank difference d=r-s for each observation and calculate the sum of squares Sd of the rank difference.

The Spearman correlation coefficient rho can be calculated using the following formula:

$$rho=1-(6*Sd)/(n*(n^2-1) \quad (4)$$

where n represents the sample size. When the rho is 1, it means that the two variables are completely positively correlated; When the rho is -1, it means that the two variables are

completely negatively correlated; When the rho is 0, it means that there is no linear correlation between the two variables.

The advantage of Spearman correlation analysis is that it can handle non-normally distributed data and is not sensitive to outliers. But it also has some limitations, such as not being able to detect nonlinear relationships.

Figure 2 is shown below

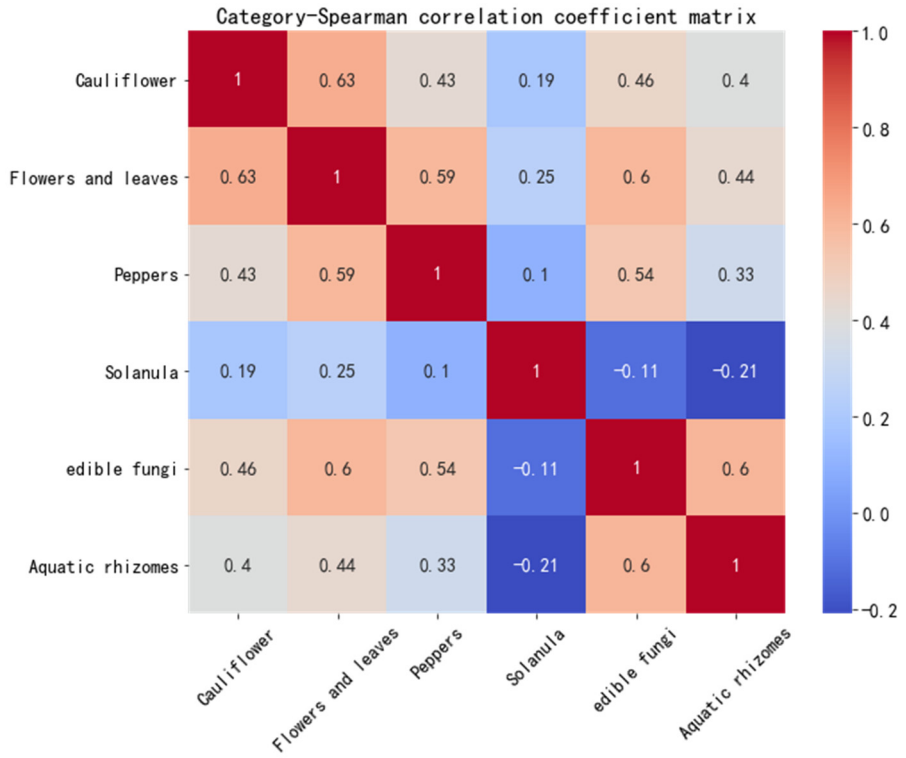
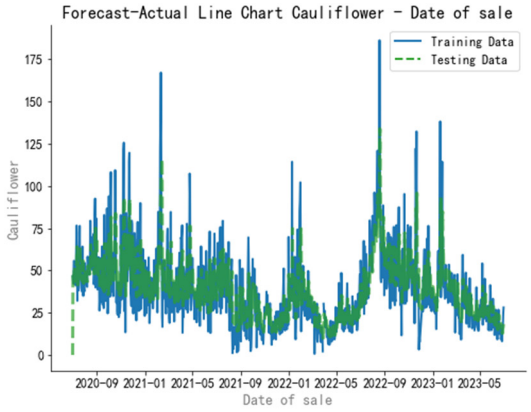


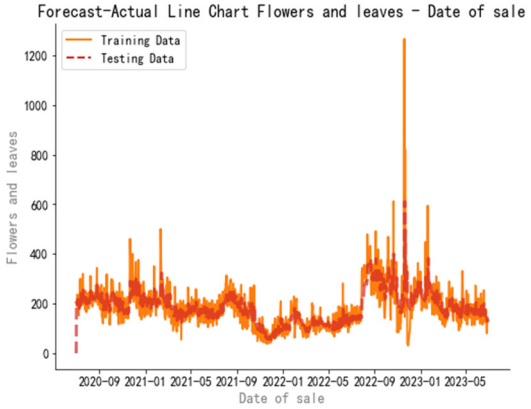
Figure 2. Spearman correlation analysis
(Photo credit : Original)

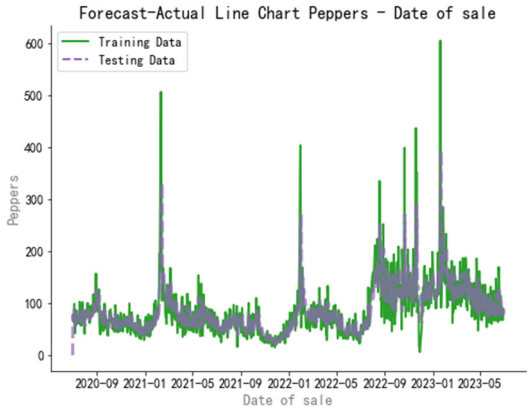
4. Comparison of individual machine learning classifiers

The ARMA model is a commonly used time series model that consists of two parts: autoregressive (AR) and moving average (MA). The AR part indicates that the observations at the current moment are correlated with the observations at several moments in the past, and the MA part indicates that the observations at the current moment are related to the noise of the past moments. For product categories, enter historical sales data to predict the sales volume of this category in the coming week by time series. First, a line chart of the predicted and true values of the time series model is given.

Figure 3 is shown below


Forecast-Actual Line Chart Cauliflower - Date of sale
Training Data
Testing Data
Cauliflower
0
25
50
75
100
125
150
175
2020-09
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
2023-05
Date of sale


Forecast-Actual Line Chart Flowers and leaves - Date of sale
Training Data
Testing Data
Flowers and leaves
0
200
400
600
800
1000
1200
2020-09
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
2023-05
Date of sale


Forecast-Actual Line Chart Peppers - Date of sale
Training Data
Testing Data
Peppers
0
100
200
300
400
500
600
2020-09
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
2023-05
Date of sale

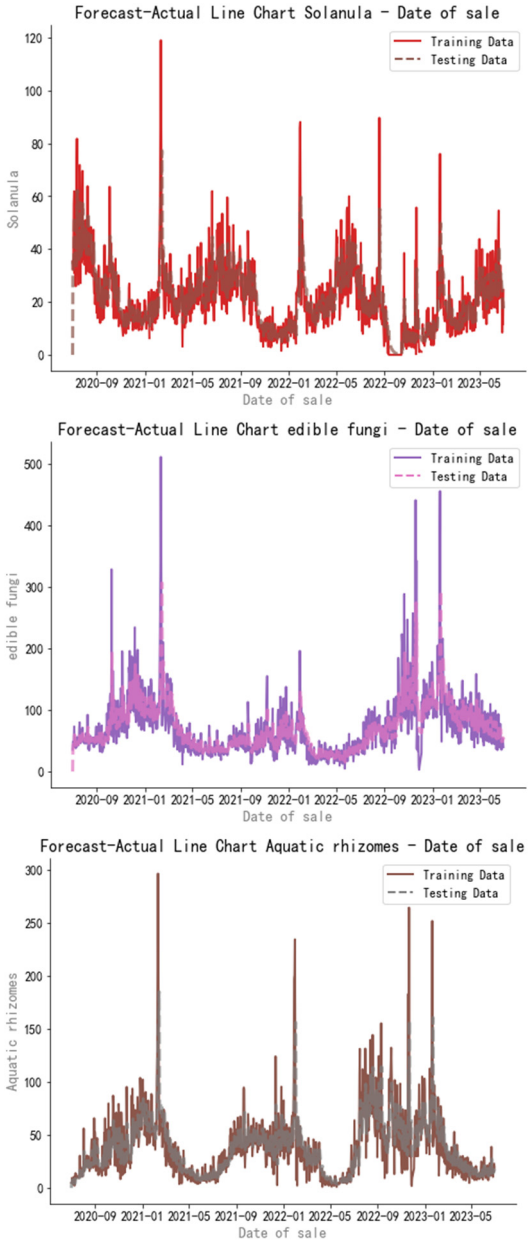


Figure 3. Time series forecast-actual line chart.
(Photo credit: Original)

Table 1. Future availability of time series forecasts.

data	Cauliflower	Flowers and leaves	Peppers	Solanula	edible fungi	Aquatic rhizomes	Cauliflower
Monday	23	140	85	23	48	18	23
Tuesday	21	143	86	23	51	18	21
Wednesday	20	144	87	23	52	18	20
Thursday	19	144	87	23	53	18	19
Friday	19	144	87	23	53	18	19
Saturday	19	144	87	23	53	18	19
Sunday	19	144	87	23	53	18	19
Monday	23	140	85	23	48	18	23

The table 1 shows the sales volume of each type of time series forecast for the coming week, and stores can restock sales according to this forecast to achieve maximum revenue.

5. Conclusion

In this paper, the sales data of various categories of goods are explored by using correlation analysis and time series forecasting, and Pearson correlation analysis and Spearman correlation analysis are used to calculate the correlation coefficient between each commodity category, and the correlation relationship between each category and single product is visually represented according to the correlation heat map.

The time series model is used to predict the sales volume of each category in the coming week, and the relationship between the sales volume and the cost-plus price is added to the time series model as a penalty function training, and finally the optimal replenishment volume and pricing of each category in the coming week are obtained, so that the merchant has the highest revenue.

References

[1] Daniyal I,Xuan T,Vipul N, et al. Prediction of Quality Food Sale in Mart Using the AI-Based TOR Method[J]. Journal of Food Quality,2022,2022.

[2] Zhu H D,Chang P Y. Robot with humanoid hands cooks food better?[J]. International Journal of Contemporary Hospitality Management,2020,32(3).

[3] Zhang H,Sun Y. Application of LightGBM and LSTM combined model in vegetable sales forecast[J]. Journal of Physics: Conference Series,2020,1693(1).

[4] Erica M,E K P,Colleen R, et al. Predictors of college-student food security and fruit and vegetable intake differ by housing type.[J]. Journal of American college health : J of ACH,2016,64(7).

[5] Dunstan J,Aguirre M,Bastías M, et al. Predicting nationwide obesity from food sales using machine learning[J]. Health Informatics Journal,2020,26(1).

[6] Kayyali Y E,Sebastian M,W. R S. A machine learning-based framework for forecasting sales of new products with short life cycles using deep neural networks[J]. International Journal of Forecasting,2023,39(4).
[7] XinXin Y,MengQi L,XiaoYan L, et al. Qualitative and quantitative prediction of food allergen epitopes based on machine learning combined with in vitro experimental validation.[J]. Food chemistry,2022,405(Pt A).
[8] Daniyal I,Xuan T,Vipul N, et al. Prediction of Quality Food Sale in Mart Using the AI-Based TOR Method[J]. Journal of Food Quality,2022,2022.
[9] Deng Y,Xiao H,Xu J, et al. Prediction model of PSO-BP neural network on coliform amount in special food[J]. Saudi Journal of Biological Sciences,2019,26(6).
[10] Truett R. Mini forecasts sales rebound with new product[J]. Automotive News,2021,95(6973).