Merchandise Pricing and Replenishment Decision Based on ARIMA Model and BP Neural Network

Pengjie Wang
Alibaba Cloud College of Big Data Application,
Zhuhai College of Science and Technology
Zhuhai, China
1297276320@stu.zcst.edu.cn

Kaihua Che
Alibaba Cloud College of Big Data Application,
Zhuhai College of Science and Technology
Zhuhai, China
openflower1202@stu.zcst.edu.cn

Chenxin Fan
Alibaba Cloud College of Big Data Application,
Zhuhai College of Science and Technology
Zhuhai, China
chenxin@stu.zcst.edu.cn

Xiaolin Zhu*
Alibaba Cloud College of Big Data Application,
Zhuhai College of Science and Technology
Zhuhai, China
fdsfsdxno@foxmail.com

2023 IEEE International Conference on Electrical, Automation and Computer Engineering (ICEACE) | 979-8-3503-0961-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEACE60673.2023.10442357

Abstract—**With the development of economy and society, vegetable goods occupy an important position in the retail industry. In order to improve the sales efficiency and profit of the superstore, this paper addresses the problem of automatic pricing and replenishment decision-making for vegetables, analyses a large amount of sales data of vegetable goods, and takes into account the cost, promotional discounts and attrition rate and other factors. Based on the idea of realizing the maximization of the revenue of the superstore, and by determining the total amount of daily replenishment and pricing strategy, we build a model based on the historical sales data and demand forecasting, to determine the optimal pricing strategy and replenishment plan. We also use machine learning and optimization algorithms to solve the model.**

Keywords—Data Mining, Cluster Analysis, Time Series Analysis, Regression Models, Neural Networks

I. Introduction

In fresh food supermarkets, because of the short shelf life of vegetable products and the deterioration of the quality of vegetable products with the increase of sales time, which leads to the low unit price of their sales, it is particularly important for supermarkets to develop reasonable pricing and replenishment strategies for supermarkets to make a profit. Supermarkets usually make replenishment and pricing according to the historical sales and market demand of each commodity.

In this paper, we firstly study the correlation between the sales volume of different vegetable categories and individual products, which starts from two main aspects: firstly, the distribution pattern and interrelationship of the sales volume between vegetable categories, and secondly, the distribution pattern and correlation of the sales volume within vegetable categories. In order to achieve this goal, the attached data are used for data processing, visualisation and correlation analysis, and during data processing we have to take into account linear and non-linear factors, use suitable correlation analysis algorithms, and finally solve these problems with the help of histograms, scatter plots, and Pearson correlation coefficients.

Secondly, in this paper, we find out whether there is a linear relationship between sales volume and cost-plus pricing for each category of vegetables by looking at the scatterplot and try to create a linear regression model to derive an expression for the functional relationship between sales volume and cost-plus pricing. Analyse the correlation based on correlation coefficients such as intercept and slope. For the total daily replenishment and pricing strategy for the vegetable category in the coming week, we consider first analysing whether the data is stable or not, and using appropriate algorithms, such as using time series to predict the total daily replenishment of each vegetable category, then solving the problem by taking the total daily replenishment as the variable and the superstore's revenue as the objective function, or attempting to use Bayesian algorithms in the application of strategic decision-making, and determining the function based on the distribution to establish a final decision-making.

II. Modelling and Solving

A. Analytical model based on Pearson's correlation coefficient for vegetable categories and sales of individual products

For the problem of analysing the distribution pattern and interrelationships between the sales volume of vegetable categories, we have chosen the Pearson correlation coefficient model. Pearson correlation coefficient is a statistical measure of the strength and direction of the linear relationship between two variables [1]. It measures the degree of linear correlation between two variables and takes values ranging from -1 to 1. The Pearson correlation coefficient calculation formula is presented as equation (1):

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (1)$$

When the correlation coefficient is 1, it means that the two variables are perfectly positively correlated, i.e. one variable increases and the other increases. When the correlation coefficient is -1, it means that the two variables are completely negatively correlated, i.e. one variable increases and the other decreases. And a correlation coefficient close to 0 indicates that there is no linear relationship between the two variables.

For the problem of analysing the distribution patterns and interrelationships of the sales volume of individual products, we chose the cluster analysis model.

Cluster analysis is an exploratory data analysis method used to group similar objects or points together to form "clusters" or "agglomerations". The aim is to ensure that data points in the same cluster are as similar to each other as possible and data points in different clusters are as different as possible. K-Means clustering [2] formula is shown as equation (2).

$$J = \text{minimize} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{2}$$

Where I(x_i belongs to cluster k) is an indicator function that is 1 if x_i belongs to cluster k and 0 otherwise.

We've finished calculating the total sales volume for six different categories and presented them as Fig. 1 shows below.

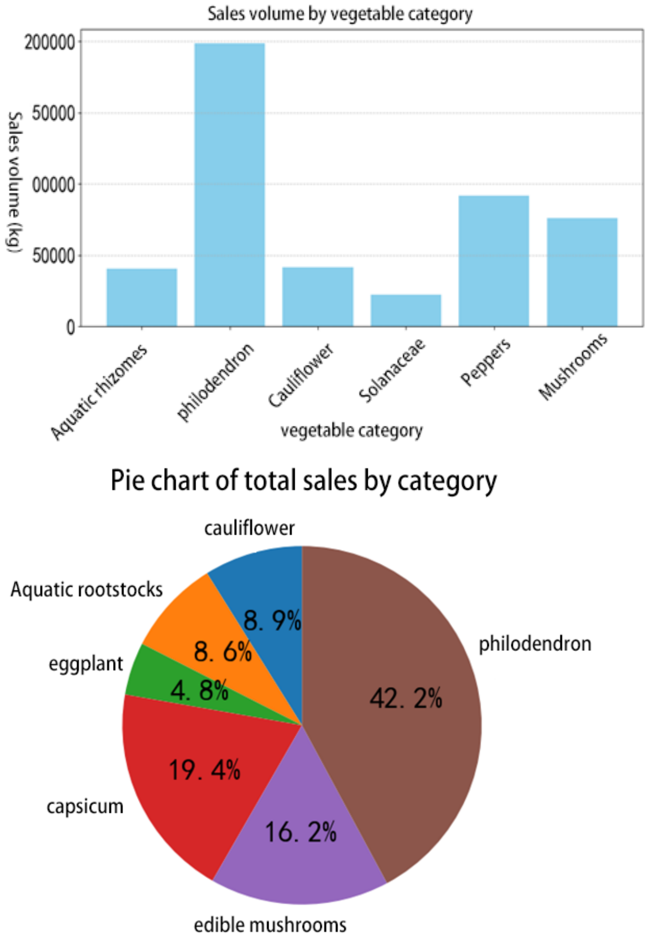

Fig. 1. Chart of total sales by vegetable category

From Fig. 1, it can be seen that the sales volume of foliage category is the highest, far exceeding the other categories, followed by the sales volume of edible mushrooms and chilli, but still lower compared to foliage, and the sales volume of eggplant and cauliflower is relatively low.

Considering that there is often a certain correlation between the sales volume of vegetable products and time, we take days as the unit, and the sales flow detail data recorded a total of 1,094 days from 1 July 2020 to 30 June 2023, calculated the sales volume of the day of all categories, and plotted the daily sales volume of each category with the time change curve. From Fig. 2, it can be observed that each vegetable category shows a seasonal pattern, with the flower and leaf category being particularly significant.

Due to the number of individual products, so the single product distribution pattern of the investigation we will use the clustering algorithm, using the "elbow rule [3]" to determine the optimal number of clusters for the 5 or 6, we choose 5 clusters.

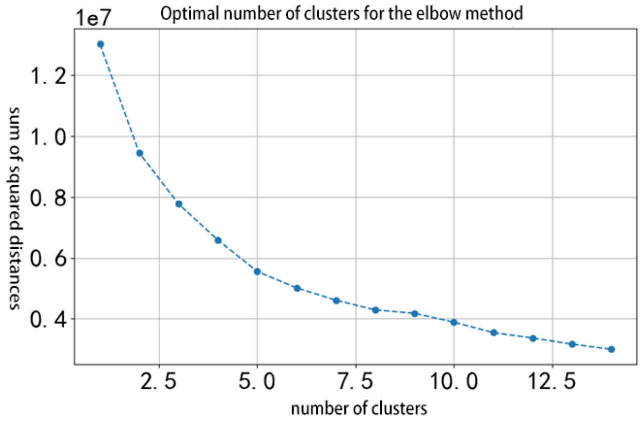

Fig. 2. Number of clusters

This is visualised by a scatterplot of the individual items, where the x-axis represents the average sales volume, the y-axis represents the standard deviation of the sales volume, and each point represents a cluster. The size of the points will be determined based on the number of vegetable singles in each cluster. The following information is obtained as Fig. 3 presented.

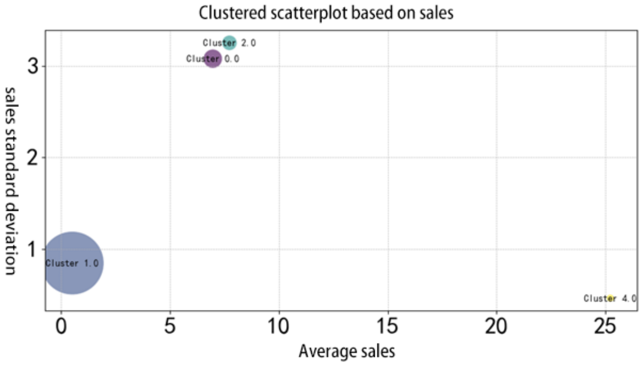

Fig. 3. Clustered scatterplot

Most of the clusters have mean sales volume and standard deviation that are relatively close to each other. However, there are some clusters, such as Cluster 3 and Cluster 4, whose mean sales volume and standard deviation are significantly different from the other clusters. Through the cluster analysis we get the following conclusions.

Cluster 0 mainly includes some common vegetables, such as amaranth, Yunnan lettuce, choy sum, Yunnan oleander, spinach and so on. These may be common vegetables in daily life and have medium sales volume.

Cluster 1 had a wider variety of vegetables, including Niushou lettuce, Sichuan red parsnip, local small hairy cabbage, cabbage moss, and so on. This may be a category of vegetables with low sales volume but a wide variety of vegetables.

Cluster 2 included only three vegetables: broccoli, net root (1) and turnip peppers (1). These vegetables may have similar sales patterns or be seasonal vegetables.

Cluster 3 There is only one vegetable, Chinese cabbage, which implies that Chinese cabbage has a very different sales pattern from the other vegetables and may be a vegetable with very high sales volume.

Cluster 4 Includes baby lettuce, Yunnan lettuce (portion), Yunnan oilseed rape (portion), etc. This may be a class of vegetables in specific packages or in specific sizes, with a different sales pattern than regular vegetables.

We used a scatterplot matrix as Fig. 4 demonstrated, to analyse the sales relationship between different vegetable categories. Positive linear relationships indicate that sales usually increase or decrease together, whereas negative linear relationships imply that when sales increase in one category, they may decrease in another. For example, different individual items within the leafy and flowering vegetables usually complement each other, showing an overall negative linear correlation, e.g., if there is a shortage of Chinese cabbage (item code 102900005115793) in the inventory, consumers may choose to purchase Chinese cabbage (item code 102900005115960) as a substitute rather than choosing other vegetable categories such as chillies or eggplant as their commodities.

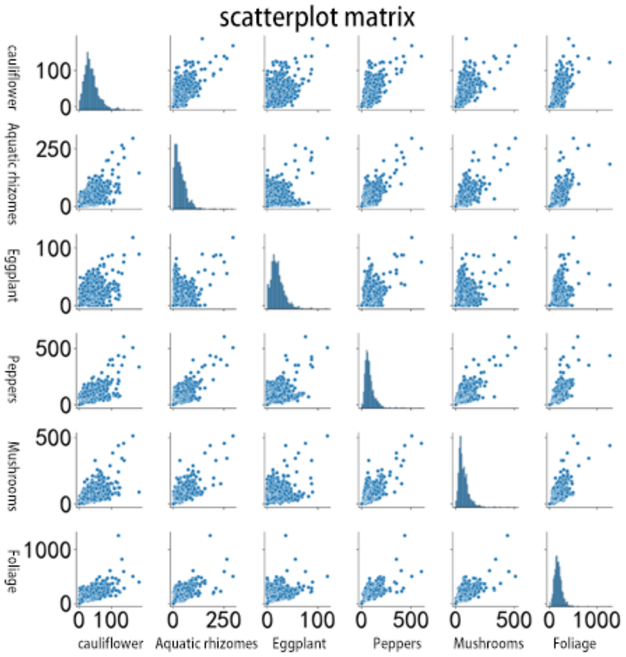

Fig. 4. Scatterplot matrix

Subsequently, we calculated the Pearson correlation coefficients between the categories and plotted heat maps to clearly show the relationships between them. The results showed that there was a strong correlation between the foliar, chilli, aquatic root, edible fungi and aquatic plant categories, while the eggplant category had a weak correlation with these categories. This led us to initially classify the six categories into two groups, where the first five categories were strongly correlated with each other, while the eggplant category was weakly correlated with them, a finding that was also confirmed in the scattering matrix, as Fig. 5 presented below.

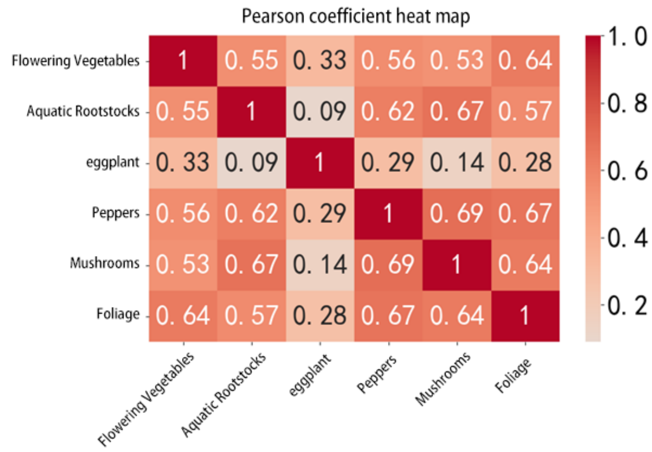

Fig. 5. Heat map of Pearson's correlation coefficient

B. Analytical model of the relationship between sales volume and cost-plus pricing for the vegtable category based on a linear regression model

In order to better understand the relationship between sales volume, cost plus, and pricing across categories, we will plot scatter plots. Fig. 6 presented below demonstrates the relationship.

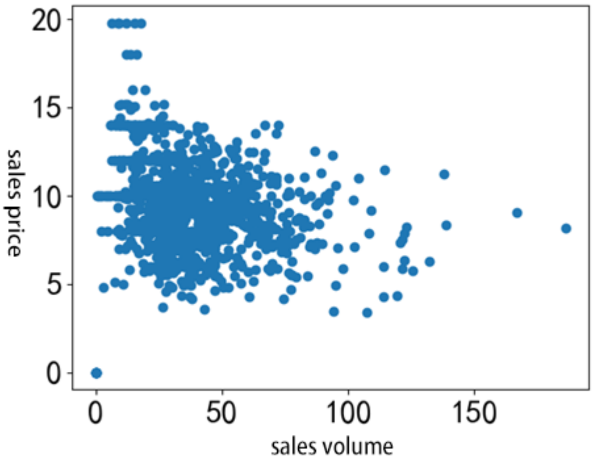

Fig. 6. Scatterplot of cauliflower sales and sales prices

Through the scatter plot we can roughly see that there is a linear relationship between the two, so a linear regression model is established for correlation analysis. It is shown as equation (3):

$$y_j = q_j + p_j w_j \tag{3}$$

Where y_j denotes the pricing of cauliflower on day j, p_j denotes the coefficients of the independent variables, q_j denotes the coefficients of the constant term, and w_j denotes the total sales volume of cauliflower on day j. We used a LinearRegression model (LRM) to fit the relationship between sales volume and cost and obtained a mathematical expression for the relationship between total cauliflower sales volume and cost. This is shown as Fig. 7.

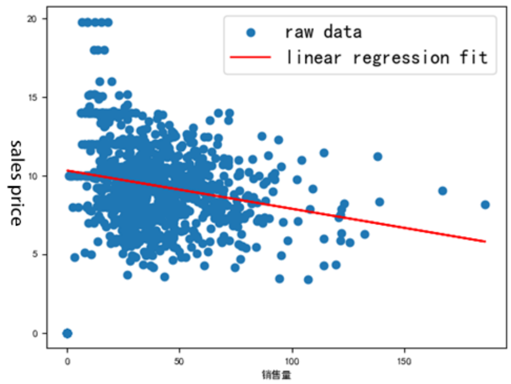

Fig. 7. Scatterplot fit of cauliflower sales and sales price

Looking at the linear regression plot, we can clearly see that the model fits the data linearly very well, which indicates that the linear regression model we chose basically meets the requirements of the problem and captures the relationship between sales volume and cost very well. In the end, we obtained a concise but powerful mathematical expression to describe the relationship between sales volume and cost. It is demonstrated as equation (4):

$$y_j = 10.3156195 - 0.024288\, w_j \tag{4}$$

The individual correlation coefficients are shown as table I and table II.

TABLE I. CORRELATION SLOPE COEFFICIENT

Foliage	Cauliflower	Aquatic Roots	Eggplant	Peppers	Mushrooms
-0.002	-0.024	-0.033	-0.024	-0.007	-0.016

TABLE II. RELEVANT INTERCEPT COEFFICIENT

Foliage	Cauliflower	Aquatic Roots	Eggplant	Peppers	Mushrooms
6.623	10.316	11.485	9.313	10.702	13.117

With the help of the pre-processed data, we conducted an exhaustive cyclical analysis and judgement of the time series of total sales. Through this process, we were able to gain a clearer insight into the underlying cyclical patterns and trends in the sales data, providing more profound insights and guidance for the replenishment planning and pricing strategies of the superstore. The results of these analyses on cyclicality will help us better understand sales behaviour and provide a solid basis for future decisions. Take the cauliflower category as an example, as Fig. 8 and Fig. 9 shown below.

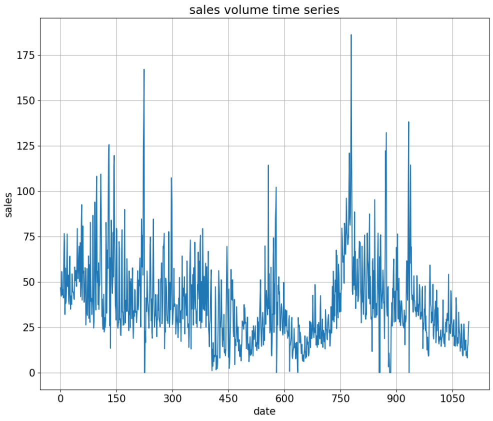

Fig. 8. Cauliflower Time Series Chart

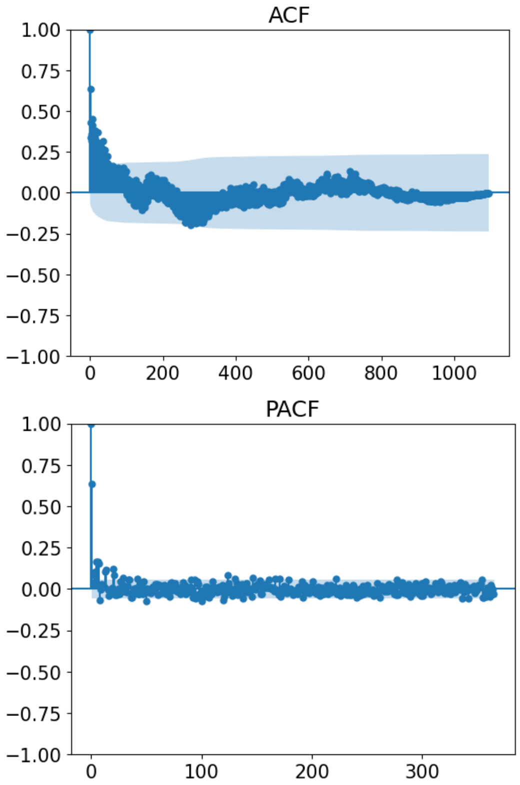

Fig. 9. Plot of autocorrelation function and partial autocorrelation function

By analysing the results of the time series plots, autocorrelation function plots and partial autocorrelation function plots, we can conclude that the sales volume data of the cauliflower category does not have any obvious strong cyclical changes in time. This means that the fluctuation of sales volume is relatively smooth and there is no obvious seasonal or cyclical trend.

Of course, we did not only rely on the results of the time series analysis, but also employed a periodicity test to further delve into the sales data, we used the Augmented Dickey-Fuller [4] test to test the smoothness of the time series, as well as the cyclical nature, which presented as Table III.

TABLE III. CAULIFLOWER PERIODICITY CHECKLIST

parameters	results
ADF statistic	-3.1003514436201547
P-value	0.026527191012485355
hysteresis order (math.)	20
Number of observation	1074
threshold value	{'1%': -3.4364533503600962, '5%': -2.864234857527328, '10%': -2.568204837482531}
significance level	9082.60095557394

After a series of cyclical tests, we find that only the Edible Mushroom category exhibits a clear cyclical trend, so we will discuss the Edible Mushroom category separately and in depth. Next, we use the cauliflower category as an example to explore the total daily replenishment and pricing strategy for each category in detail.

For the Cauliflower category, we used an ARIMA forecasting model to predict the total daily replenishment for the coming week (1-7 July 2023).

Autoregressive component (AR): denoted as AR(p), where p is the autoregressive order, which is used to represent the correlation between the time series and its past p points in time. the mathematical representation of the AR(p) component is presented as equation (5):

$$\phi_p(L)(Y_t - \mu) = \epsilon_t \qquad (5)$$

Y_t is the value of the time series, μ is the mean, and ϵ_t is the white noise error.

The difference part (I, denoting the integral): denoted as I(d), where d is the difference order, which is used to smooth the time series. The mathematical representation of the I(d) part is shown as equation (6):

$$(1 - L)^d Y_t = \epsilon_t \qquad (6)$$

where L is the lag operator, Y_t is the differenced time series, and ϵ_t is the white noise error.

Moving Average Part (MA): denoted as MA(q), where q is the sliding average order, which is used to represent the correlation between the time series and its past q error terms. the MA(q) part is mathematically represented as equation (7):

$$\left(1 - \theta_q(L)\right)\epsilon_t = \eta_t \qquad (7)$$

where $\theta_q(L)$ is the sliding average polynomial, ϵ_t is the white noise error, and η_t is the sliding average term.

Combining these three components yields a mathematical representation of the ARIMA model as equation (8) presented:

$$\left(1 - \phi_p(L)\right)(1 - L)^d Y_t = \left(1 + \theta_q(L)\right)\epsilon_t \qquad (8)$$

where $\phi_p(L)$, $\theta_q(L)$ are polynomial functions, Y_t is the observed value, d is the difference order, p is the autoregressive order, q is the sliding average order, and ϵ_t is the white noise error.

In contrast, edible mushrooms were predicted using a BP (back-propagation) neural network [5] model, where we pass the preprocessed data through a hierarchy of multiple neurons, each of which performs a nonlinear transformation and a weighted summation operation. With the back-propagation algorithm, the model can automatically adjust the connection weights between neurons to minimise the prediction error and perform accurate prediction.

Based on the nature of the time series data and the effectiveness of the model fit, we set p, d, and q to 1 to fit the ARIMA model is shown as Fig. 10.

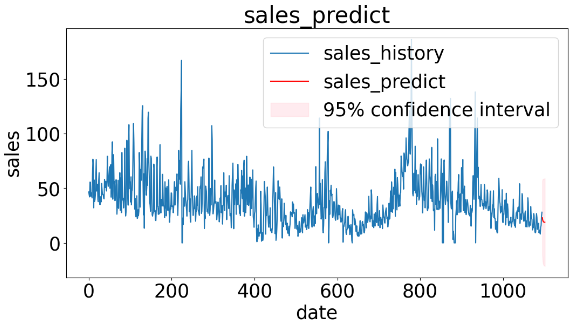

Fig. 10. Sales Volume Forecast Chart

Since the predicted value of sales volume lies exactly in the centre of the confidence interval, combined with the assessment of the model's performance and the characteristics of the historical data, it is reasonable to conclude that the model is very accurate. This means that the model performs well in predicting sales volume, the deviation between its predictions and actual observations is relatively small, and the model can be trusted.

We have used the BP (back propagation) neural network model for prediction for edible mushrooms category instead, after adjusting the hyperparameters and standardising the input data, the final Mean Squared Error (MSE) is 0.005454340422182085, and the predicted sales volume is as shown as Fig. 11.

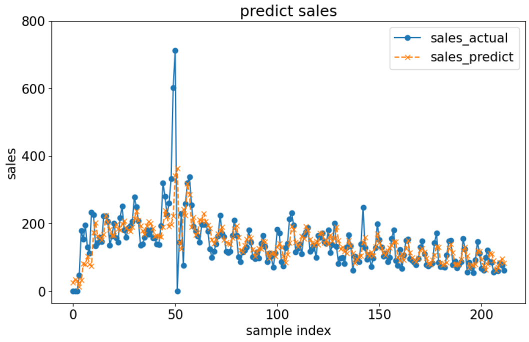

Fig. 11. Model Forecast Sales Chart

In the above Fig.11, a high degree of fit between the predicted and actual values can be clearly observed, which demonstrates that the BP neural network performs well in handling periodic data.

III. Conclusions

In this paper, we firstly consider the non-linear relationship between categories to get a reasonable Pearson model instead of Spearman model, and the Pearson model is close to the reality and has high application value. The clustering model is used in the single product analysis, the model can effectively divide the data into communities for multiple data volumes. It can be extended to such as supermarkets, food markets, aquatic markets and so on.

Secondly, BP neural network and ARIMA model are used to forecast the time series, BP neural network is a powerful nonlinear model that can capture nonlinear relationships in complex time series. The neural network is adaptive, automatically adjusting weights and biases to accommodate changes in the data, while easily handling multidimensional features, including lagged features and other external variables, which enhances the modelling capabilities.

However, in practice, the number of clusters was selected using only the elbow rule, followed by the solution division through K-means clustering, and no other clustering algorithms were used for multi-algorithm score evaluation, and the cluster division may be relatively homogeneous. At the same time, neural networks usually require a large amount of data for training, otherwise they are easy to overfit, and they also need to adjust many parameters, including the number of layers, the number of neurons, and the learning rate, leading to a more complex selection and adjustment of the model.

Acknowledgments

This work was supported by Zhuhai College of Science and Technology Doctoral Enhancement Program.

References

[1] Zhang JY, Gao Ran, Hu J et al. Comparison of the application of gray correlation degree and Pearson correlation coefficient[J]. Journal of Chifeng College(Natural Science Edition),2014,30(21):1-2. DOI:10.13398/j.cnki.issn1673-260x.2014.21.001.

[2] J.P. Zhang,X.Y. Liu. Research and application of K-means algorithm based on cluster analysis[J]. Computer Application Research, 2007(05):166-168.

[3] LONG Wenjia, ZHANG Xiaofeng, ZHANG Lian. A business process clustering method based on k-means and elbow rule[J]. Journal of Jianghan University(Natural Science Edition),2020,48(01):81-90. DOI:10.16389/j.cnki.cn42-1737/n.2020.01.011.

[4] Xia, N. X. A comparative study of DF, ADF test and PP test for unit root[J]. Research on Quantitative and Technical Economics, 2005(09):130-136.

[5] ZHOU Feiyan, JIN Linpeng, DONG Jun. A review of convolutional neural network research[J]. Journal of Computing,2017,40(06):1229-1251.