Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 13, No. 2, 2024

Research on Vegetable Sales Strategies in Supermarkets Based on MLP Optimisation Algorithm and Random Forests

Zehao Qu *, Yibo Wang

School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China
* **Corresponding author**: Zehao Qu (Email: 2717782356@qq.com)

Abstract: The aim of this study is to propose a comprehensive solution to the challenges in vegetable merchandising in superstores. For the sales volume pattern study, data processing and correlation analyses reveal the association between vegetable categories and explore possible merchandising strategies for merchandise mix. For the pricing problem, a cost-plus pricing method and a machine learning model are introduced to develop a price strategy for superstores to maximise sales profit, and a prediction model for future pricing is proposed. For the replenishment plan, the MLP optimisation algorithm and regression model are used to design the replenishment plan to maximise the revenue within the limited sales space. Finally, the importance of data is emphasised and multifaceted data collection and analysis methods are proposed to help superstores better understand market demand, optimise inventory and pricing strategies, and ultimately maximise revenue and customer satisfaction. In summary, this study provides a systematic set of vegetable merchandising strategies that provide useful reference and guidance for superstores to cope with market challenges.

Keywords: MLP; Random forest; Cost-plus pricing.

1. Introduction

This study aims to provide a comprehensive set of vegetable merchandising strategies for superstores to help them better cope with market challenges. In this paper, we will address several key issues. First, we will explore the distribution pattern of vegetable sales volume and the association between categories to reveal possible merchandising strategies for merchandise mix. Second, we will develop a reasonable pricing strategy to maximise the sales profit of the superstore through cost-plus pricing method and machine learning model. Further, we will design a replenishment plan to control the total number of saleable items within a certain range to maximise the revenue of the superstore. Finally, we emphasise the importance of data and propose multi-dimensional data collection and analysis methods to help hypermarkets better respond to market demands and optimise their business strategies [1].

2. Sales Volume Analysis of Vegetables by Category

2.1. Data processing

In this paper, the data of vegetable sales in a superstore for three years were collected and combined.

Firstly, Shapiro-Wilk or Kolmogorov-Smirnov test was performed on the data to check the significance. If it does not show significance (P>0.05), it means that it conforms to the normal distribution, and vice versa, it means that it does not conform to the normal distribution [2]. The normality test was based on the S-W test or K-S test to get the output overall descriptive results, the overall descriptive results are as follows (Table 1).

Table.1 General descriptive results

variable name	Sample size	Median	Mean	Standard deviation	Skewness	Kurtosis	S-W test	K-S test
Foliage	1048	8.655	9.629	3.235	0.735	0.62	0.959(0.000***)	0.121(0.000***)
Cauliflower	1048	9.158	9.898	4.087	1.82	4.738	0.853(0.000***)	0.102(0.000***)
Aquatic Roots	1048	9.158	9.898	4.087	1.82	4.738	0.853(0.000***)	0.102(0.000***)
Eggplant	1048	8	8.214	3.242	0.72	0.529	0.964(0.000***)	0.051(0.009***)
Chilli	1048	9.736	10.269	4.027	0.965	1.126	0.941(0.000***)	0.086(0.000***)
Mushrooms	1048	8.899	9.224	2.815	0.738	0.529	0.963(0.000***)	0.053(0.005***)

For this reason, through the above table, it is known that there are outliers in only some of the data. After careful analysis of the data set, the data values that do not conform to normal distribution are normalised in the merged data.xlsx. On the basis of standard deviation, variance, mean, identify and deal with outliers, use the unique function to remove duplicate data points and stubborn data. In the above process, string data types such as leibi_num and product names are converted to numerical types using the str2double function. The eigenvalues are extracted using the statistical indicators spectrum analysis and wavelet transform, then the data are smoothed and de-trended using smoothdata, and finally the data are standardised and normalised using zscore and written into Excel tables using writable or writematrix [3].

2.2. Model solution and analysis

In order to better analyse the distribution patterns and interrelationships of the sales volume of each category and

single product of vegetables, we carried out the following operations through matlab coding:We read the data in two Excel files and read them into the tables t1 and t2 respectively.The data in the columns of danpin_num (item number) and leibie_num (category number) were extracted from t1 and assigned to the variables item and leibie1 respectively. And assigned to the variables item and leibie1 respectively. At the same time, extracted the data of danpin_num column from t2, and assigned to the variable leibie2. Iterate through each element in leibie2 and find the index of the element in item. If a matching index is found, the corresponding leibie_num column in the t2 table is updated to the same value as the corresponding index in item. In each loop, the current loop index is output i. Finally, the updated table t2 is written to a new Excel file.

In order to make the presentation of data more intuitive, we show the sales volume share and specific sales volume of the six categories through visualisation, using a combination of pie charts and bar charts (as shown in Figure 1).

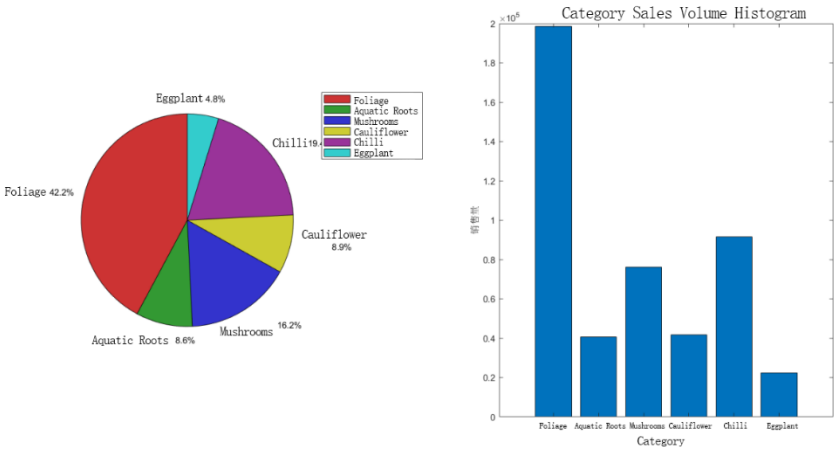

Figure 1. Pie and bar charts of sales volume share and specific sales volume for 6 categories

As can be seen from the graph above, by category, the 'Foliage' category has the highest sales volume, followed by 'Chilli' and 'Edible Mushrooms'.

Next, we will analyse the change in sales volume over time to check for any cycles or trends. We will first focus on category sales as a whole, and then break this down into individual product sales. By using MATLAB to plot a "Line graph of sales of each product on day x from 1 July 2020" (Figure 2), the background colour of the graph is set to white and several curves are plotted. The colour and line width of these curves were set separately. The x-axis represents the number of days from 1 July 2020 and the y-axis represents the sales of each product".

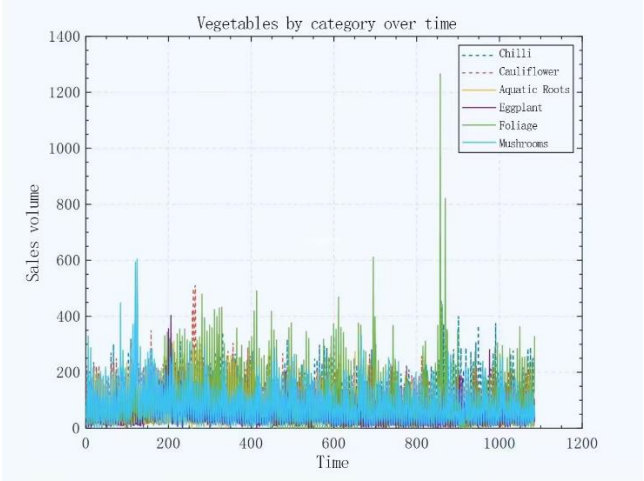

Figure 2. Sales by product on day x from 1 July 2020

From the above graph, we can learn that.

Sales volume trends over time: each category has its peaks and troughs during the year, showing a clear seasonality. Foliage: sales in this category peak at certain times of the year, probably related to certain festivals or seasonal events.

Next, we will use Pearson's correlation coefficient to investigate the link between the different vegetable categories [4]. By calculating the correlation coefficients between the sales volumes of different categories, we can see which sales trends are correlated. This means that when the sales volume of one category increases, the sales volume of another category may also increase. By using this method, we can better understand the relationship between different vegetable categories so that we can better forecast and manage sales.

Then, a correlation heat map was drawn (Figure 3) which shows the values of the correlation coefficient matrix. This heatmap uses a parula colour map with the colour range restricted to -1 to 1, the grid lines are not visible and the title is 'Correlation Heatmap'. Overall, this code is mainly used to plot multiple curves and show the correlation between the sales of each product.

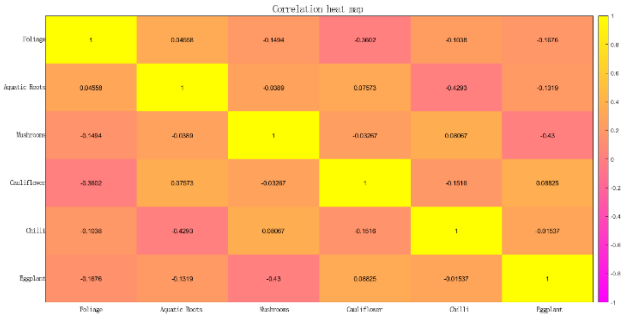

Figure 3. Heat map of Pearson's coefficient for vegetable category

1. High positive correlation: For example, in this case, for example, there may be a high positive correlation between "fruits" and "roots". This means that when the sales of "Fruit" increase, the sales of "Roots and Tubers" may also increase, and vice versa.

2. low or no correlation: for example, the correlation between 'foliage' and most of the other categories is low, which means that their sales patterns are likely to be independent.

A heat map of the correlation coefficients between the sales of individual items is also obtained, partly as shown in Figure 4.

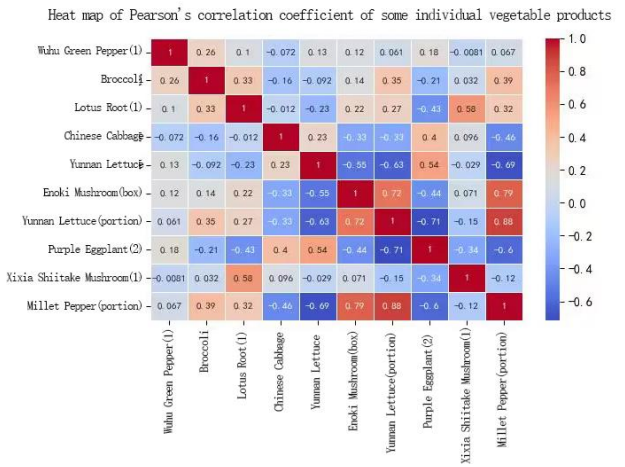

Figure 4. Heat map of Pearson's coefficient for selected vegetable items

We then identified pairs of items with correlations greater than 0.7, and the results are shown in Table 2.

Table.2 Individual items with correlation greater than 0.7

Item1	Item2	Correlation
102900011033562	102900011033531	1.0
102900011033531	102900011033562	1.0
102900011033586	102900011033562	1.0
102900011033562	102900011033586	1.0
102900011033586	102900011033531	1.0

Based on the results of the analyses, it can be surmised that these groups of items were purchased at the same time and in the same quantities as a combination or package. The results of these correlation analyses provide insights for superstores to develop combination sales strategies and promotions to maximise sales and revenue.

3. Cost-plus pricing for each vegetable category

3.1. WSO_BiLSTM model

By processing the data in the previous section, the data have been made to have a certain degree of completeness and accuracy. Next, we will analyse the relationship between total sales volume and cost-plus pricing for each vegetable category and determine the total daily replenishment volume and pricing strategy for the coming week in order to maximise the revenue of the superstore [5].

Predicting Sales and Pricing Using the WSO_BiLSTM Model:

A WSO_BiLSTM model was trained for predicting sales volume and future pricing of vegetable categories based on historical sales data and category characteristics. The model was used to predict the daily sales volume and corresponding pricing for each category from 1-7 July 2023.

Calculate the total minimum replenishment for each category for each day from 1-7 July 2023 based on projected sales volumes and category wear rates. In conjunction with the cost-plus pricing relationship, determine the pricing strategy for each category for each day from 1-7 July 2023 to maximise the supermarket's revenue. A linear regression model can be used to analyse the relationship between total sales and cost-plus pricing for each vegetable category, and based on this model, the total daily replenishment and pricing strategy for 1-7 July 2023 can be given to maximise the supermarket's revenue.

The linear regression model is given by:

$$Y = \beta_0 + \beta_1 * X^1 + \beta_2 * X_2 + \cdots + \beta_n * X_n \quad (1)$$

Where, Y denotes the total sales volume, $X1, X2, \ldots, Xn$ are the factors affecting the total sales, $\beta 0, \beta 1, \beta 2, \ldots \beta n$ are the parameters of the regression model.

3.2. Model solving

Firstly, the necessary libraries are loaded, including pandas, numpy and Sklearn's StandardScaler module. These libraries are used for data processing and model training.

Next, load the data from an Excel file using the read_excel method of pandas. Then, save the feature names of the data in a variable named ata_biao (by converting the column names of the data into a list).

Create a variable named A_data1 to hold the data feature values and label values. Then, store the index of the data feature in a variable named feature_need_last (using numpy's range function to generate a sequence of integers from 0 to the number of features minus 1). Next, divide the data into x_feature_label and y_feature_label by using the slice operation.

Create an index variable called index_label1 (a sequence of integers from 0 to the number of rows of x_feature_label minus 1, generated by numpy's arange function). The data index is then stored in a variable called index_label (which may have been defined elsewhere). If index_label is empty, set it to index_label1. Then, divide the data by counting the number of training sets, the number of validation sets, and the number of test sets (by multiplying the division ratio with the total number of data).

Z-score normalisation of x_feature_label and y_feature_label was performed using StandardScaler objects. The features and labels of the training, validation and test sets were normalised separately and the results were saved in objects named train_x_feature_label_norm, vaild_x_feature_label_nor, test_x_feature_label_norm, train_y_feature_ label_norm, vaild_y_feature_label_norm and test_y_feature_label_norm.

Next, set some parameters of the model, such as the number of populations, the number of population iterations, the optimisation method, the number of Bayesian iterations, the batch size, the maximum number of iterations and the size of the hidden layer. The optimize_fitrBiLSTM function is called to train the training, validation and test sets to obtain the model Mdl and fitness fitness. The trained model Mdl is then used to predict the training, validation and test sets, and the prediction results are stored in a file named y_train_predict_norm, y_vaild_ predict_norm and y_taild_ predict_norm. predict_norm, y_vaild_predict_norm and y_test_predict_norm.

Then, the predictions are inverted by using StandardScaler object to get the actual predicted values, and the results are stored in the variables named y_train_predict, y_vaild_predict and y_test_predict. Next, various evaluation metrics such as Mean Absolute Error (MAE), Mean Relative Error (MAPE), Root Mean Square Error (RMSE) and R2 are computed for the training, validation and test sets.

Then, the results of the training, validation and test sets are plotted according to the given parameters. The time of the

whole process is obtained by calculating the difference between t2 and t1 and stored in a variable named Time.

4. The relationship between sales volume and cost-plus pricing for each individual product

4.1. MLP-RF regression models

MLP-RF regression model is a regression model based on Multilayer Perceptron (MLP) and Random Forest (RF) [6].

Multilayer Perceptron (MLP) is a forward feedback neural network model, which consists of an input layer, a hidden layer and an output layer. Each neuron has connection weights with the neurons in the previous layer and nonlinearly transforms the input through an activation function. The multilayer perceptron model uses a back propagation algorithm to adjust the connection weights to minimise the prediction error. It is capable of learning complex nonlinear mapping relationships between inputs and outputs, and is suitable for dealing with regression problems with highly nonlinear characteristics.

Random forest is an integrated learning method that consists of multiple decision trees. Each decision tree is constructed by random subsampling of data and random selection of features. Random forests improve prediction performance and reduce the risk of overfitting by integrating the predictions of the decision trees. It can handle high-dimensional data and data with complex interaction effects and is suitable for regression problems.

The MLP-RF regression model combines the MLP and RF models. First, the MLP model is used to train and predict the data, and the prediction results of the MLP model are obtained. Then, the prediction results of the MLP model together with the original features are used as inputs, and the RF model is used to further train and predict the data to obtain the final regression results. By combining the advantages of the two models, the MLP-RF regression model can better cope with the nonlinearity and complex interaction effects of the features and improve the prediction performance.

4.2. Model solution and analysis

An MLP-RF regression model is used in this study. The model combines both multilayer perceptron (MLP) and random forest (RF) methods to improve the accuracy and robustness of prediction.

In the prediction process, the data are first preprocessed, including data cleaning, feature selection and missing value processing to ensure the quality and consistency of the input data. Then, the MLP model is used to train and predict the sales data. The MLP model is able to model the complex relationship of sales data through multiple levels of neurons and nonlinear activation functions. Subsequently, the prediction results of the MLP model are used as features, which are provided to the RF model together with the original features. The RF model consists of multiple decision trees, and the final sales volume prediction results are obtained by integrating the prediction results of each decision tree. The advantages of the RF model are that it is able to capture nonlinear relationships and interactions in the data, avoid overfitting, and improve the stability and reliability of the prediction. The MLP-RF regression algorithm was used for model training and model evaluation. Combination and weighting of models. Predictions are made based on the models and weights, and the predictions are obtained. Evaluate the prediction results for the training, validation and test sets and print out the results. Plot the results of the training, validation and test sets. Perform real prediction, using the features of the data to be predicted, and save the prediction results in the data_out table.

Finally, the prediction results are output to the table for subsequent decision making and planning. The processing and forecasting of sales data can help superstores to have a more accurate prediction of future sales trends and thus make better business decisions and market strategies.

5. TOPSIS analysis

5.1. Ideas for solving the problem

Identify evaluation indicators: Based on the requirements of the problem, identify evaluation indicators to measure replenishment and pricing strategies for individual items. Possible indicators include superstore revenues, sales growth rates, inventory fulfilment rates, cost control, etc. These indicators should reflect the overall performance of the product.

By using the TOPSIS method, multiple evaluation indicators can be considered and the optimal replenishment and pricing strategy can be found in terms of overall performance. When using the TOPSIS method, attention needs to be paid to the weighting of the evaluation indicators, the choice of standardisation methods and the sensitivity analysis of the results to ensure the reliability and validity of the results.

5.2. Modelling

This algorithm is TOPSIS, in which the relevant variables are divided into positive and negative indicators, positive indicators and good weather, customer purchases, customer feedback, daily supply, negative indicators have competitors, the index term is the name of the category, the parameter variable weighting entropy weighting method.

Step 1: Distinguish the indicators according to their categories (high or low merit) and use appropriate formulas to normalise them according to the different types of indicators. Construct a matrix X with n rows and m columns, where X_{ij} denotes the value of the j indicator for the i object.

Step 2: Construct the standardisation matrix. For matrix X, calculate the mean and standard deviation of each indicator. Then, X_{ij} is standardised using the following formula. Through this standardisation process, the differences in the scales of the indicators are eliminated so that the standard deviation of the indicators is 1.

$$D_i^+ = \sqrt{\sum_{j=1}^{m} \omega_j \left(Z_j^+ - z_{ij}\right)^2} \tag{2}$$

$$D_i^- = \sqrt{\sum_{j=1}^{m} \omega_j \left(Z_j^- - z_{ij}\right)^2} \tag{3}$$

$$Z_{\mathrm{ij}} = \frac{X_{ij}}{\sqrt{\sum_{k=1}^{n} \left(X_{ij}\right)^2}} \tag{4}$$

Step 3: Calculate the gap between each evaluation index and the optimal vector and the worst vector. Where, w_j denotes the weight (importance) of the j attribute. By calculating the gap between the evaluation indexes and the optimal vector and the worst vector, we can measure the degree of superiority or inferiority of the evaluation object in each index.

Step 4: Use the measure C_i to evaluate how close the evaluation object is to the optimal solution. The larger the value of C_i, the better the evaluation object. This measure can be used to help determine the proximity of the subject to the optimal solution, which can assist in decision-making or in selecting the best solution.

$$C_{\mathrm{i}} = \frac{D_i^-}{D_i^+ + D_i^-} \tag{5}$$

6. Conclusions

This study proposes a comprehensive set of solutions by conducting in-depth research on several key issues in vegetable merchandise management in superstores, aiming to help superstores better manage vegetable merchandise and improve sales efficiency and customer satisfaction.

First, in the study of sales volume patterns, we revealed the association between different vegetable categories through data processing and correlation analysis, providing possible merchandising strategies for superstores to combine goods. Second, in terms of pricing issues, we introduced the cost-plus pricing method and machine learning model to effectively formulate a pricing strategy to ensure that the superstore achieves the maximisation of sales profit. In addition, we established a prediction model for future pricing to provide a more reliable pricing reference for hypermarkets.

In terms of replenishment planning, we designed a replenishment plan using MLP optimisation algorithms and regression models, which enabled the hypermarket to control the total number of saleable items and maximise revenue within a limited sales space. Finally, we emphasise the importance of data and propose a multifaceted data collection and analysis method, which provides the superstore with more accurate market demand forecasts and operational decision support.

References

[1] Huang Yujia. Research on the sales strategy of agricultural products in LD organic mall based on AIDA model[D]. Dalian University of Technology, 2019.

DOI:10.26992/d.cnki.gdlqc.2019.000240.

[2] Y. Xia,Y. Liu,D. Liu,X. Huang,et al. Application of normality test in the evaluation of tobacco structure quality[J]. Light Industry Science and Technology,2022,38(06):14-16.

[3] Liang YT. Anomalous vibration monitoring method of centrifugal compressor return line based on wavelet transform[J]. Mechanical Management Development, 2024, 39 (04): 37-39. DOI: 10. 16525/j.cnki.cn14-1134/th.2024.04.012.

[4] Zhao L, Wang SG, Wang N, et al. Fully generalised spatial modulation visible optical communication system based on Pearson correlation coefficient selection[J]. Journal of Optics, 2024, 44(04):116-124.

[5] An Qi. Research on the pricing of scientific and technological research services based on cost-plus pricing method[J]. Library Research and Work,2021, (10):25-31+24.

[6] STRICTLY QI, ZHAO WANYING, YU ZHENWEI, et al. Prediction of respiration rate of dairy cows based on random forest model under hyperparametric optimisation algorithm[J/OL]. Journal of Agricultural Engineering,1-9[2024-06-01]. http: // kns. cnki. net/ kcms/ detail/ 11. 2047. S. 20240528. 1333.030.html.