Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 6, No. 2, 2023

Research on Establishing Inbound Strategies for Supermarkets based on LSTM and Gaussian Process Regression Modeling

Wei Weng †, Yifu Lin † and Jiawei Wu †

School of Internet Economics and Business, Fujian University of Technology, Fuzhou, China

† These authors also contributed equally to this work

Abstract: This paper provides an in-depth study on the challenges of vegetable merchandising in fresh produce supermarkets, aiming to provide a comprehensive set of management strategies to optimize supermarket operations. First, the sales volume and sales of six types of vegetables were analyzed by descriptive statistics and the cyclical trend was explored by time series processing; second, good correlations between edibles and aquatic roots and tubers as well as edibles and eggplants were found by plotting correlation matrices and heat maps of Spearman's coefficients. Next, this paper analyzed the relationship between cost-plus pricing and total sales and predicted the total replenishment and pricing of vegetables in the coming week using an LSTM time series forecasting model and evaluated the model performance using root mean square error (RMSE). Finally, a Gaussian regression model was used to predict a small sample of data to develop an optimal replenishment volume and pricing strategy for the superstore, which maximized the superstore's revenue. The results of the study show that the inventory management efficiency of fresh supermarkets can be effectively improved by these methods.

Keywords: Vegetable Commodity Management; Time Series Forecasting; LSTM; Gaussian Regression Modeling.

1. Introduction

Fresh food supermarkets play a key role in the daily lives of residents by providing them with fresh vegetables, fruits, and other ingredients, ensuring convenient, dependable, and diversified food choices. Reasonable replenishment and pricing strategies are crucial in the operation of these supermarkets, especially for vegetable items with a short shelf life. This paper investigates automatic pricing and replenishment strategies based on data on wholesale prices, wastage rates and sales volumes of vegetables, aiming to improve the operational efficiency and profitability of fresh food supermarkets. The study first analyzes the sales distribution patterns and interrelationships between different vegetable categories and individual products, and then proposes a pricing strategy and replenishment volume planning for vegetables during the week to maximize benefits. This paper will also consider the market demand and product diversity and formulate the replenishment volume and pricing strategy for the vegetable varieties that can be sold in a specific time, to ensure the diversity of products and at the same time, maximize the interests of the superstore.

2. Relevance Analysis

2.1. Data Preprocessing and Statistical Description

The descriptive statistics section of this paper focuses on analyzing the distribution of sales volume of six vegetable categories. Through the presentation of Figure 1, it is learned that the sales volume is, in descending order, the flower and leaf category, the pepper category, the edible mushrooms, the aquatic root and tuber category, and the cauliflower category. In addition, by analyzing the ranking of sales volume of individual items, this paper identifies the vegetable items with the highest sales volume, in which Xixia mushrooms, broccoli, Wuhu green peppers, and millet peppers are ranked among the top four, which indicates the degree of hot sales of these categories in the market. This analysis provides an important basis for further pricing strategies and replenishment plans.

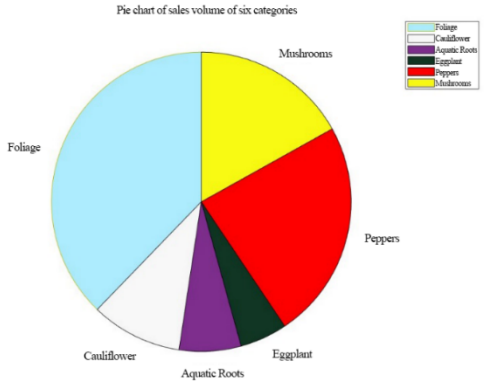

Figure 1. Pie Chart of Sales Volume of Six Categories

After that, in this paper, the major categories of vegetable sales are processed in time series to check whether the data has a significant periodicity or trend [1], and the six categories are sorted in chronological order through matlab output as shown in Figure. 2.

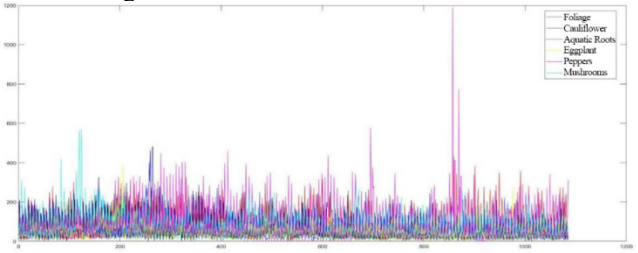

Figure 2. Timeline diagram of sales change of six major vegetable categories

From Figure 2, it is not difficult to find the following in this paper:

Firstly, sales change over time with obvious peaks and troughs, and there are obvious patterns of change and cyclical fluctuations throughout the year.

Secondly, data that fluctuate significantly from the above data are discussed separately in this paper, and examples of foliage and chili peppers are given here as shown in Figures 3 and 4.

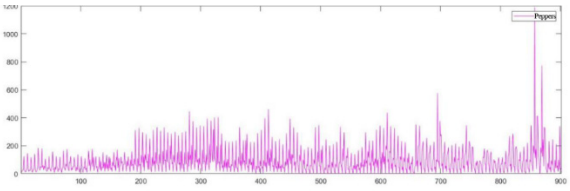

Figure 3. Line graph of the time evolution of sales in the chili pepper category

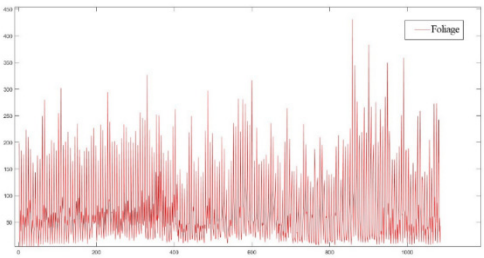

Figure 4. Time series of changes in sales of flowers and foliages

The analysis in this paper reveals the prominence of the leafy and flowering vegetable category in terms of sales, especially when sales peak at specific times of the year. This pattern may be closely related to consumer preferences, geographic location, or seasonal events. Meanwhile, analysis of the data for the chili pepper vegetable category revealed varying degrees of fluctuation at different time periods (190 days ago, 190 to 860 days, and 860 days to the end of the data record). These fluctuations remained consistent in the overall pattern, but the magnitude of fluctuations increased over time on the vertical axis. This phenomenon may be related to the scarcity of chili pepper supply or the continued decrease in production.

2.2. Correlation Analysis

The correlation coefficient matrix is calculated as shown in Figure 5:

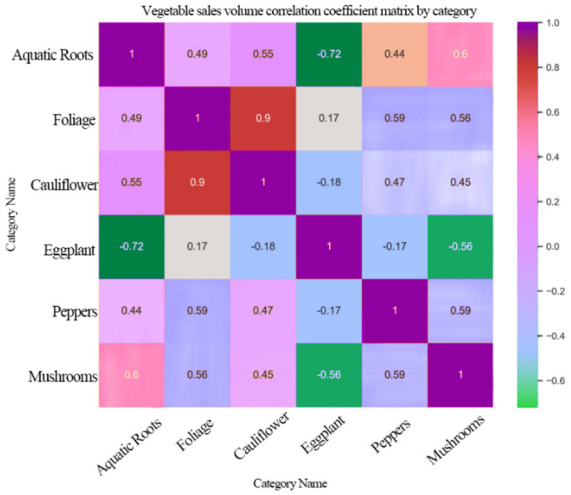

Figure 5. Matrix of correlation coefficients of sales volume of vegetables by category

This paper continues to explore in depth the correlations between different vegetable categories, particularly between their sales volumes. To this end, a two-by-two correlation was calculated using the Spearman coefficient for the six vegetable categories to identify which categories have correlations between their sales volumes [2]. For example, the analysis revealed an extremely high correlation between the cauliflower category and the foliage category, suggesting that sales trends for these two categories are intricately linked. In contrast, sales volumes between the eggplant category and the other vegetable categories showed a significant negative correlation, suggesting that these categories may be selling independently of each other. In addition, to more fully understand the impact of the time factor on sales, this paper combines the time variable with the sales data to generate a heat map of Spearman correlation coefficients between vegetable categories as shown in Figure 6.

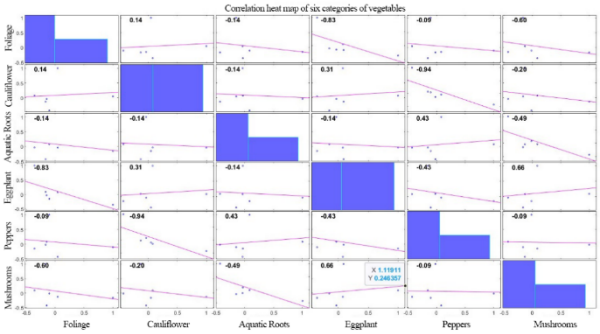

Figure 6. Heatmap of correlation of six vegetable groups

From the above figure, there is a significant correlation between edibles and aquatic rhizomes with a correlation coefficient of 0.66 and between eggplant and edibles with a correlation coefficient of 0.66.

Therefore, it can be found that edible mushrooms have a good correlation with both aquatic roots and eggplants, and therefore there is a possibility to consider that edible mushrooms are linked to the purchase of these two vegetable commodities in terms of routes, purposes, and composition.

3. Category Pricing and Replenishment Strategies

3.1. Modeling

In this paper, a simple linear regression method was first tried in predicting the restocking quantity and pricing strategy of vegetable items [3]. Although the prediction error of this method is within an acceptable range, a Long Short-Term Memory (LSTM) temporal recurrent neural network is introduced in order to further improve the accuracy of the prediction. the LSTM model is able to learn richer and more generalized patterns due to its advantage in dealing with large-sample data, which is especially important for predicting sales data with complex time-series characteristics [4]. Large sample data not only contains more information and diversity, but also promotes a more stable training process. In the optimization process, LSTM is not easy to fall into local minima due to its network structure, which increases the possibility of finding the global optimal solution. One of the linear regression results is shown in Figure 7:

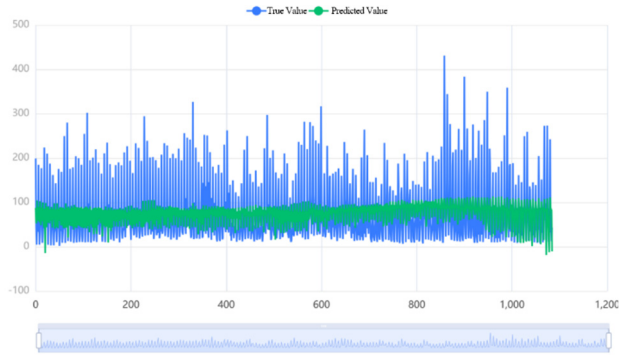

Figure 7. Linear regression of sales volume and cost-plus pricing for foliage and flowers

In the study of this paper, an LSTM model is used to predict the total daily replenishment volume and pricing strategy from July 1 to 7, 2023, especially for the large-scale time series data from July 2020 to June 30, 2023. This method aims to optimize the daily operations of the superstore, especially in predicting the replenishment volume and pricing to maximize the revenue of the superstore. With this advanced machine learning approach, this paper shows how historical data and advanced analytics can be used to optimize the operational decisions of fresh supermarkets.

Where in LSTM, each time step has a hidden state and a cell state. These states are updated and passed between time steps. Three of the gates are responsible for controlling the flow of information and updating of the cell state, which are Forget Gate, Input Gate and Output Gate.

Oblivion Gate: This gate serves to determine how much previous information the network should retain at the current time step. Its output ranges from 0 to 1, where 0 means completely forgotten, and one means completely retained. This means that the forgetting gate can decide what information is retained or forgotten from the cellular state of the previous moment.

Input Gate: The function of the input gate is to control which added information will be added to the cell state. It also has an output between 0 and 1, which is used to determine which added information should be included when updating the cell state at the current time step.

Output Gate: The output gate is responsible for controlling which cell state information will be output to the hidden state of the current time step. It serves to selectively output cell state information for the current time step, which is crucial for prediction and next time step computation.

By precisely controlling the operation of these gates, LSTM can efficiently capture long-term dependencies in time series data.

3.2. Solving the Model

Using the above model, univariate time prediction was first done with aquatic rhizomes as an example, while RMSE, as an evaluation metric, was used to observe the iteration of the LSTM model, as exemplified in Figures 8 and 9:

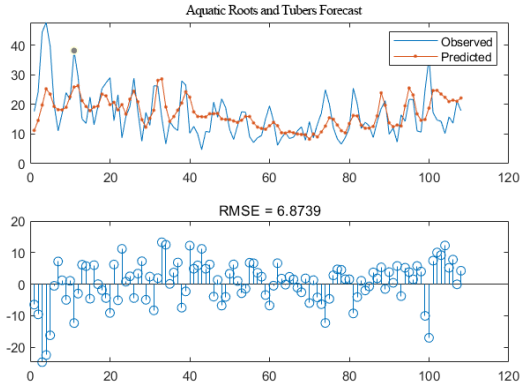

Figure 8. Map of prediction errors for aquatic rhizomes

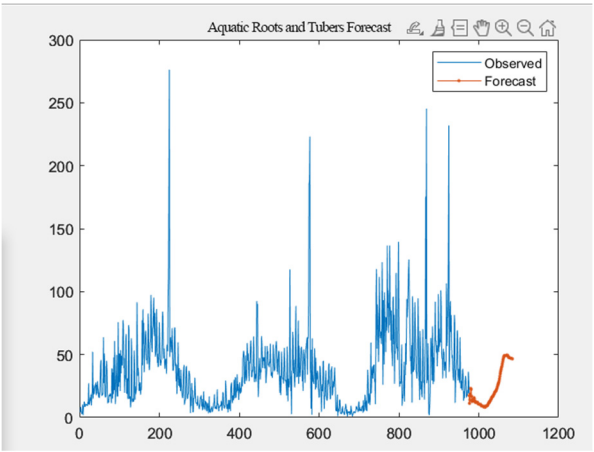

Figure 9. Forecasted sales of aquatic roots and tubers

Predictions were made for the remaining five categories using LSTM fitted to the above data, and the results are shown in Tables 1 and 2:

Table 1. Six Vegetables 2023701-2023707 Forecast Pricing

Dates	Philodendron	Cauliflower	Aquatic rhizomes	Eggplant	Capsicum	Edible mushroom
2023701	12.3866	6.6802	5.8978	20.1239	4.9662	5.4987
2023702	4.0191	14.4091	14.0623	5.831	7.5081	13.8197
2023703	6.9726	5.4397	6.1032	5.8089	7.349	15.954
2023704	6.9672	9.2294	9.9114	11.9924	3.6533	8.6223
2023705	7.9709	16.3965	13.9264	4.9914	8.4725	7.7846
2023706	15.8386	5.5117	5.2457	9.3137	3.8634	9.7836
2023707	12.1566	6.2549	5.339	20.3168	4.9379	5.7845

Table 2. Forecast Sales of Six Vegetables 2023701-2023707

Dates	Philodendron	Cauliflower	Aquatic rhizomes	Eggplant	Capsicum	Edible mushroom
2023701	134.0364	7.3361	14.193	9.389	115.37	50.183
2023702	127.1269	8.8237	16.1751	10.6	199.2	44.853
2023703	136.2311	13.0762	22.5194	14.31	418.77	46.09
2023704	201.233	22.828	33.1267	16.35	235.34	63.16
2023705	194.4921	29.9806	30.7682	14.18	291.78	69.756
2023706	149.4206	22.7928	21.413	11.26	79.603	42.336
2023707	158.4612	26.7922	19.9221	9.704	90.706	52.038

Conclusion, based on the LSTM time series prediction model, the stocking and pricing of the six vegetable categories from 2023701-2023707 were obtained, i.e., the pricing strategy and stocking according to the above data can be the most profitable for the superstores.

4. Individual Product Pricing and Replenishment Strategies

To accurately predict the optimal replenishment and pricing strategy for each individual item, this paper introduces Gaussian Regression (GRP), a machine learning model based on Bayesian theory and kernel functions, which is particularly suitable for prediction of small sample data [5]. Gaussian regression can play a key role in the prediction of continuous type target variables by using Gaussian distribution (normal distribution) to model the data and make reliable predictions about the uncertainty in the data. In this paper, Gaussian regression is used to analyze the relationship between input and output variables with the aim of providing more accurate replenishment and pricing strategies for fresh produce supermarkets, optimizing their operational efficiency, and improving profits. After the model training is completed, the Gaussian regression model is used to predict the sales volume of the remaining thirty-three individual items on July 1 [6]. Details are shown in Table 3:

Table 3. Top20 forecast sales table.

Item Name	Sales on July 1	Item Name	Sales on July 1
Yunnan lettuce (portion)	35	Sweet potato tip	4.2521
Yunnan oilseed rape (portion)	21	Long term eggplant	3.7629
Peppers (portions)	20	Kogua (1)	2.5953
Broccoli	16.1261	Water caltrop or water chestnut (genus Trapa)	2.7957
Brussels sprouts (Brassica oleracea var. Botrytis)	13.0084	Cordyceps flowers (portions)	3
Small wrinkled skin (portions)	10.8571	Green & Red Pepper Combo (Servings)	2
Xixia Mushroom (1)	9.5195	Red Pepper (2)	1.915
Baby Chinese cabbage (mini-sized variety)	9.4312	Yunnan lettuce	1.6331
Screw peppers (portions)	8	Chinese flowering cabbage	1.0884
Agaricus bisporus(box)	8	White Mushroom (Bag)	1

Afterwards, through the data after the prediction of the July 1 stocking volume, and as a basis for drawing the distribution of vegetable stocking volume, it is not difficult to find thirty-three single products to meet the needs of the supermarket. Finally, the thirty-three specific product July 1 net income summary, get the maximum gain of 397.3364. Therefore, on July 1 as far as possible to meet the supermarket on the market supply meets the market on the premise of several types of vegetables, through the sales volume of 2.5 kg divided into benchmarks, on the two types of data pure income summation, to get the super July 1 income maximization data for 397.3364.

5. Conclusion

In this paper, an efficient forecasting model is successfully constructed, which is used to optimize the inventory management and pricing strategy of fresh food supermarkets. The significant advantages of the model include making full use of time series data and applying statistical analysis and visualization tools such as line plots, correlation matrices, Spearman coefficients, and heat maps to intuitively understand the data structure. The introduction of LSTM time series forecasting efficiently handles the problem of long-run dependence and multivariate sequences, and the Gaussian regression forecasting model helps to maximize the revenue of the superstore in the face of the changing market demand the model is not suitable for the market demand. However, the models have limitations in terms of applicability to certain time series data types as well as data integration and updating issues in practical applications. To improve the model, it is recommended to adjust the kernel function in Gaussian regression or to use integrated learning methods such as random forests or gradient boosting trees. In addition, this model is also applicable to sales forecasting of other commodity types such as meat, beverages, and cold goods, demonstrating its broad potential for time series purchase problem processing. With these improvements, the model will be more accurately adapted to different market demands, providing dedicated support for operational optimization of superstores and other retail businesses.

References

[1] Jing Wang, Miaomiao He,Jian Ding et al. Spatio-temporal graph convolutional network for multidimensional time series anomaly detection[J/OL]. Journal of Xi'an University of Electronic Science and Technology,1-11[2023-11-29]https://doi.org/ 10.19665/j.issn1001-2400.20230804.

[2] 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,2023,43(14): 5372-5381.DOI:10.13334/j.0258-8013.pcsee.220035.

[3] YANG Wei-Lun, GAO Yu-Xuan, CAO Lei. Linear regression method combined with MLP to predict the comprehensive water quality index of Lijiahe Reservoir[J]. Shaanxi Water Resources,2023,(06):19-21+25.DOI:10.16747/j.cnki.cn61-1109/ tv.2023.06.061.

[4] Chen ZY, Yang B, Ruan WJ et al. Short-term electrical energy load forecasting based on LSTM neural network[J]. Power Big Data,2021,24(04):8-15.DOI:10.19317/j.cnki.1008-083x. 2021.04.002.

[5] Xiong W.L., He D.F.,Wang X.L. et al. Scene-optimized robust predictive control based on Gaussian regression learning[J]. Journal of Zhejiang University (Engineering Edition),2023, 57(04): 693-701.

[6] ZHANG Zhongqiu, ZHANG Yufeng. Research on image relationship of ecological restoration attribute mapping in national land space based on Bayesian theory[J/OL]. Resource Development and Market,1-15[2023-11-29]http://kns.cnki.net/ kcms/detail/51.1448.N.20231107.1718.008.html.