The supply chain demand forecasting model based on LSTM and multiple clustering techniques

Hongfeng Xun[#], Wenhui Li[#, *]

College of intelligence and information Engineering, Shandong University of Traditional Chinese Medicine, Jinan, China, 273500

*Corresponding author: 15020792683@163.com

Abstract. This study addresses the optimization of supply chain management in e-commerce platforms through the analysis of historical data related to e-commerce activities and product demand. By processing data and conducting anomaly detection, a combination of linear regression, ARIMA, and LSTM models is employed to analyze time series features, with LSTM selected for predicting the demand of various products across different warehouses for each merchant. K-means clustering is utilized to categorize time series data, identifying distinct demand patterns for different products. For new time series data, DBSCAN density clustering and ARIMA models are applied for prediction. Additionally, considering the impact of promotional events such as Singles' Day on demand, ARIMA models are employed to analyze periodic time series data.

Keywords: ARIMA, LSTM, K-means, DBSCAN.

1. Introduction

This paper explores how e-commerce platforms can optimize their supply chains through predictive modeling based on historical data analysis. Prior studies by Wang and Zhu (2022) focused on improving supply chains using information technology and profit distribution methods, with GM(1,1) models used for forecasting vegetable production trends[1]. Others, such as Mao (2023), utilized Python and ARIMA models for strategic recommendations in the industry. Chietal[2]. (2022) investigated loss scenarios in the vegetable supply chain in Shandong province and proposed improvement strategies. Jiang[3] (2023) developed a resilience assessment model for e-commerce vegetable supply chains, ensuring their smooth operation[4]. Lietal. (2023) proposed a combined model for effectively predicting vegetable price fluctuations, emphasizing the importance of supply chain management[5]. Que (2024) presented a solution for fresh vegetable supply chains based on "blockchain+", supporting technological advancement in agriculture and rural revitalization[6].

The data for this study is sourced from http://www.mathorcup.org/detail/2433.Data preprocessing involves integration and cleaning. Based on data characteristics and patterns, a mixed random forest model is employed to predict product demand for merchants across various warehouses. K-means density clustering is applied for time series segmentation. Seasonal patterns are incorporated using ARIMA for predicting demand during events like Singles' Day.

2. ARIMA time series model development

The 3sigma method is used for detecting missing values, revealing no anomalies in the sample data. Hence, demand data distribution for product categories is directly analyzed[7].

$$\begin{aligned} &\left(1-\varphi_t B^i\right)\left(1-B\right)^d Y_t = \left(1+\theta B^t\right)\partial_t \\ &\Leftrightarrow \left(1-B\right)^d Y_t - \sum_{i=1}^{p}\varphi_i \left(1-B\right)^d Y_{t-i} = \partial_t + \sum_{t=1}^{q}\theta_t \partial_{t-i}, \end{aligned} \tag{1}$$

The ARIMA formula is applied for model training and validation to forecast demand for merchant products in different warehouses from May 1, 2023, onwards,as shown in Fig 1.

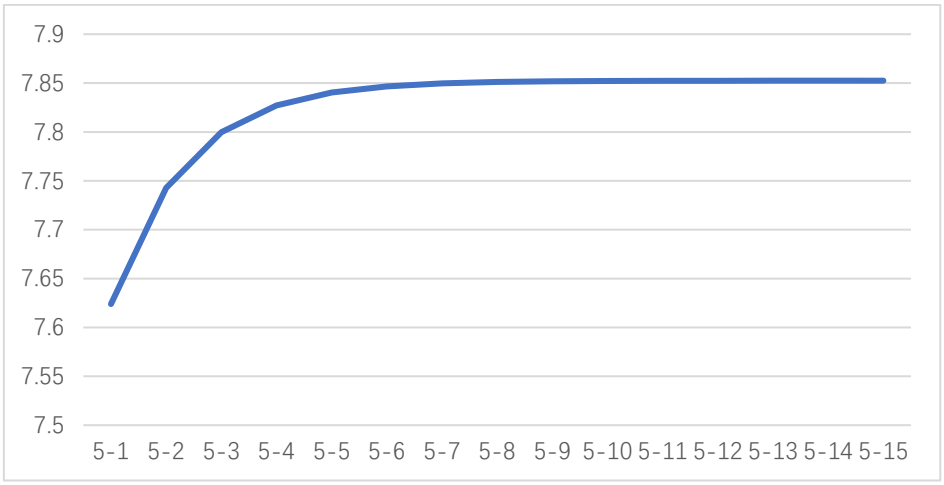

Fig 1. Partial product forecast results

The 1-WMAPE value for ARIMA is found to be 0.9721475920178584, indicating that only 97% of the data aligns with actual conditions.

2.1. Development and solution of LSTM neural network model

Data preprocessing involves standardization, merging, and scaling to normalize all features between 0 and 1. The LSTM model is constructed and trained using Keras. The dataset is created, and the model is built, compiled, and validated for accuracy. Visualization of training and testing scores is presented[8]as shown in Fig 2.

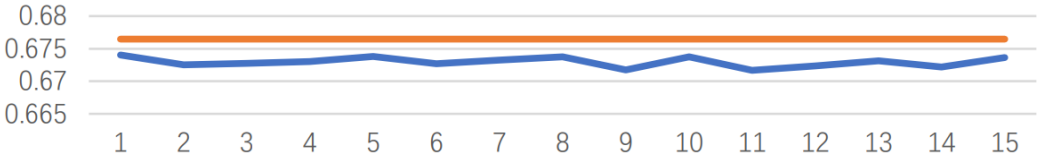

Fig 2. Visualization of training and testing scores

The close proximity of training and testing scores suggests that the model neither overfits nor underfits. The 1-WMAPE indicator for this model is found to be 0.992573402226743, higher than the accuracy of the ARIMA model analysis, leading to the adoption of LSTM model predictions.

2.2. Analysis of data correlation

One-hot encoding is applied to the dataset to independently process all data, followed by the selection of the optimal k-value using the elbow method. Clustering analysis is performed, dividing time series data of merchants, products, and warehouses into categories with either high or low demand, enhancing the similarity of features within the same category[9]as shown in Fig 3.

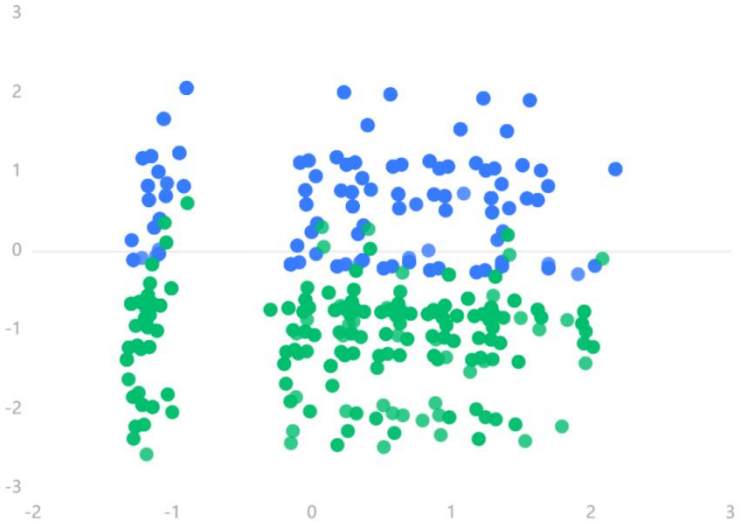

Fig 3. The cluster diagram for k=2

2.3. Selection and preprocessing of new data features

New dimensions of merchants + warehouses + products are selected, and similar feature data are encoded and averaged to calculate time series similarity. Data are merged, retaining only feature data, and DBSCAN density clustering algorithm is utilized to identify time series similar to the original data[10].

2.4. DBSCAN density clustering algorithm

Label encoding is applied to seller_no, warehouse_no, and product_no, and the distance matrix for each unique combination is computed. The optimal parameters are selected based on the Silhouette score, and DBSCAN's neighborhood radius ε is determined through the k-distance graph, with ε set to 0.5. DBSCAN algorithm is applied for clustering, with MinPts set to 5, yielding basic grouping results, which are then combined with ARIMA model for time series data processing,as shown in Fig 4.

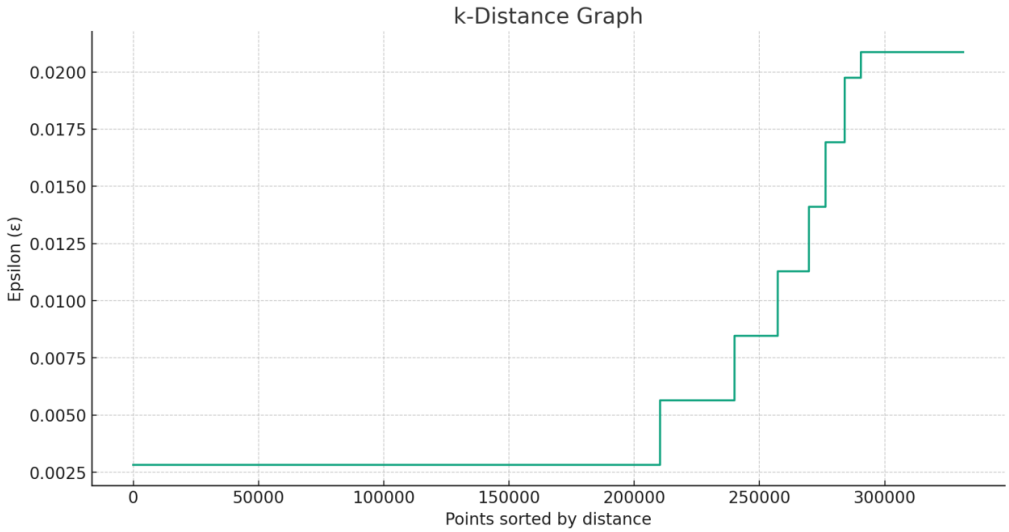

Fig 4. K-distance plot

2.5. Basic data analysis and preprocessing

ARIMA time model is employed for data analysis of large time series with a seasonal cycle of six months.

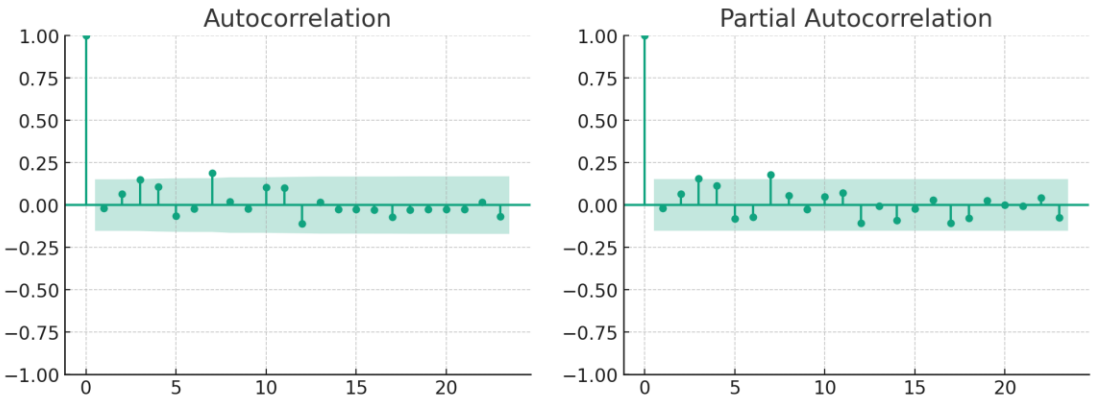

Fig 5. Autocorrelation graph of real demand values and predicted values

The close alignment of the autocorrelation graph in Fig 5 indicates that the demand forecast values obtained using the ARIMA model closely match the actual values, leading to the selection of the subsequent data for time series prediction, as depicted in the conclusion.

3. Conclusions

For the analysis of the original data, the forecasted values from May 16th to May 30th for items categorized under inventory class A, with a product grade of large items, first product category of food and beverages, warehouse class of regional warehouses, and items sourced from East China, are as follows: 1.24, 1.22, 1.20, 1.19, 1.19, 1.18, 1.18, 1.18, 1.18, 1.17, 1.17, 1.17, 1.17, 1.17, 1.17. For the analysis of new data, items categorized under inventory class B, with a product grade of goods, first product category of computers and office supplies, warehouse class of central warehouses, and items sourced from East China, have the following forecasted values from May 16th to May 30th: 2.54, 2.51, 2.48, 2.45, 2.44, 2.43, 2.43, 2.44, 2.44, 2.4, 2.45, 2.46, 2.44, 2.46, 2.46. For the analysis of data related to Singles' Day, items categorized under inventory class A, with a product grade of special items, first product category of food and beverages, warehouse class of regional warehouses, and items sourced from East China, have the following forecasted values from June 1st to June 15th: 5.35, 5.33, 5.20, 5.15, 5.12, 5.08, 5.05, 5.03, 5.01, 5.00, 4.98, 4.97, 4.97, 4.96, 4.96.

The article predicts the demand for each store's products in each warehouse from May 16, 2023, to May 30, 2023, and evaluates the accuracy of the model's performance in predicting demand. It categorizes businesses, warehouses, and products that are similar in demand characteristics within the time category. Additionally, to validate the model's applicability, it adds a product dimension, finding sequences similar to it in the data and predicting the forecast values for these additional dimensions from May 16, 2023, to May 30, 2023. Furthermore, it incorporates demand volatility as a feature and data among businesses, warehouses, and products showing significant numerical spikes. Based on the previous predictions of each store's products in each warehouse, it forecasts their values from June 1, 2023, to June 20, 2023.

The research problem of this article is to explore the importance of goods management and supply chain optimization in e-commerce platforms. To reduce inventory costs and ensure supply and demand balance, it establishes a predictive model through historical data analysis, forecasts future demand, and thus achieves reasonable planning and management of the supply chain to improve inventory allocation efficiency.

References

[1] Wang Jianhua, Zhu Fangxiao. Analysis of Agricultural Product Supply Chain Integration—Discussion on the Necessity of Vegetable Production Forecasting with GM(1,1) Model. Logistics Technology, 2022, 45(04): 123-129.

[2] Mao Lisha. Research on Pricing Strategy and Production-Sales Model of Vegetable Wholesale Market from the Perspective of Supply Chain. [Dissertation]. Central South University of Forestry and Technology, 2023.

[3] Chi Xiaojun, Song Yizhao, Li Yu, et al. Investigation and Analysis of Vegetable Supply Chain Loss in Shandong Province. China Fruits and Vegetables, 2022, 42(05): 78-84.

[4] Jiang Yi. Research on Resilience Evaluation of E-commerce Vegetable Supply Chain under Demand Fluctuations. [Dissertation]. Shijiazhuang Tiedao University, 2023.

[5] Li Hong, Wu Yanjie, Yang Li. Research on Price Prediction Method of Vegetable Supply Chain—Analysis Based on the Perspective of Supply Chain Scenario. Price: Theory and Practice, 2023(06): 77-82.

[6] Que Lijuan. Exploration of Fresh Vegetable Supply Chain Based on "Blockchain." Modern Business, 2024(02): 3-6.

[7] Cao Xinyue, He Chunlin, Cui Mengtian. Urban Vegetable Price Fluctuation Patterns and Forecasting Based on X12-ARIMA and LSTM Composite Model. Journal of Southwest University for Nationalities (Natural Science Edition), 2021, 47(04): 418-425.

[8] Zhang Yao, Sang Bohan, Wang Chen, et al. Coherent Transmission System of 64QAM Optical Frequency Division Multiplexing Based on Nonlinear Equilibrium of LSTM. Chinese Journal of Lasers, 1-15

[9] Zhao Wenjuan, Cheng Yuhan, Li Mei. Optimization of Drainage Pipe Network Monitoring Points Based on Improved K-means Algorithm. Environmental Monitoring Management and Technology, 2024, 36(01): 79-83.

[10] Li Zhicong, Sun Xuyang. Three-Branch DBSCAN Algorithm Based on Outlier Detection and Adaptive Parameters. Computer Application Research, 1-7