DOI: 10.23977/infse.2024.050513
ISSN 2523-6407 Vol. 5 Num. 5

Combinatorial Optimisation Model for E-Commerce Retail Merchant Demand Forecasting Based on ARIMA and LSTM

Shitian Li[1,a,*,#], Junzhe Zhang[2,b,#], Ziyu Zhang[3,c,#], Xv Chu[4,d], Lili Song[1,e], Xiaojun Wang[1,f]

[1]Medical Information Engineering, Shandong Traditional Chinese Medicine University, Jinan, China
[2]Health Management, Shandong Traditional Chinese Medicine University, Jinan, China
[3]Ophthalmology and Optometry, Shandong Traditional Chinese Medicine University, Jinan, China
[4]School of Health, Shandong Traditional Chinese Medicine University, Jinan, China
[a]13341882658@163.com, [b]zjz20241216@163.com, [c]du19511588247@163.com, [d]18306373568@163.com, [e]13355449516@163.com, [f]15275243371@163.com
[#]Co-first author
[*]Corresponding author

Keywords: Demand Forecasting, Combinatorial Optimisation Model, ARIMA, LSTM, Inventory Management

Abstract: With the rapid development of e-commerce, accurate demand forecasting in the e-commerce retail industry is crucial for inventory optimisation and supply chain management. In this paper, a combined optimisation model based on ARIMA and LSTM is proposed to improve the accuracy and robustness of demand forecasting. Firstly, the data are preprocessed and clustering analysis is performed to group similar categories into one class to simplify the data structure and reduce the modelling complexity; then the ARIMA and LSTM models are used to forecast the demand respectively, and the combined optimization model is constructed by combining the advantages of the two models through difference forecasting. The experimental results show that the model can significantly improve the prediction accuracy, optimise the inventory management, and provide a scientific basis for the resource planning of the e-commerce platform. Finally, this paper analyses the limitations of the model and looks forward to the future research direction.

1. Introduction

With the rapid development of e-commerce, the number of merchants on e-commerce platforms has reached thousands, and each merchant has its own merchandise goods, which are stored in e-commerce supporting warehouses and managed by e-commerce platforms [1]. In e-commerce retailing, accurate demand forecasting and inventory optimisation are crucial for merchants' operations and profits. Inaccurate demand forecasting may lead to an oversupply or undersupply situation, which can affect sales and customer satisfaction. At the same time, excessively high or low inventory levels can have a negative impact on capital flows and costs. At this stage, how to scientifically manage and make decisions about these goods in order to reduce inventory costs and ensure on-time performance of goods has become an urgent problem [2-3]. The supply chain optimisation problem involved is crucial for the operation of e-commerce platforms.

Firstly, this paper describes the importance and challenges of demand forecasting in e-commerce in the introduction section; then in the related work section, we review the research progress in the field of sales volume forecasting, including qualitative forecasting, quantitative forecasting, and machine learning methods; then we construct the basic volume forecasting model, and carry out the data preprocessing, clustering analysis, and demand forecasting by using the ARIMA and LSTM models in turn; Further, a combined optimisation forecasting model based on ARIMA and LSTM is proposed, and the superiority of the model is verified by comparing its forecasting effect; finally, the research results are summarised in the conclusion section and future research directions are discussed.

2. Related Work

The topic of sales forecasting appeared a long time ago, but the related researches have been enriched only in the last forty years, the reasons are mainly two: firstly, the development of supply chain technology, and sales forecasting plays an important role in supply chain management, the enterprises have more and more requirements on the accuracy of sales forecasting [4]. The second reason is the continuous development of computer computing power, so that people can process more data, sales forecasting methods have been further developed [5].

After years of development, the technical methods of sales forecasting can now be summarised into two directions: qualitative forecasting and quantitative forecasting. Qualitative forecasting refers to the forecaster to rely on personal ability and experience, to analyse the existing data, so as to make their own subjective speculation on the direction of things, at present, the main expert meeting method [6], Delphi method, brainstorming method [7] and so on. Quantitative prediction refers to the existing data, the use of certain mathematical methods to find out the connection between the variables, so as to predict the future value based on their own values. Early research in quantitative forecasting focused on time series analysis, by analysing historical sales to predict future sales, mainly divided into moving average method, exponential smoothing method, ARIMA model, etc. [8-9].

In recent years, with the rise of big data, enterprises have more and more data for decision support, and machine learning method has gradually become a popular forecasting method. Compared with the time series method, the machine learning method considers more causal links between sales and related factors, and uses a large amount of data for repeated learning to adjust the parameters, so as to get a more satisfactory result [10].

Combination prediction has also been a popular research direction for sales volume prediction, and Bates and Granger were the first to propose that combining different prediction methods can improve the accuracy of prediction, because combination prediction takes into account the influence of multiple factors and has strong robustness [11]. Zou Haofei et al. introduced genetic algorithm in neural network model, which made the convergence ability and prediction accuracy of the model improved significantly.

3. Base Cargo Forecasting Model

Accurate warehousing volume prediction is an important basis for category binning planning, for accurate prediction results can foreseeably determine the future use of warehousing resources decision-making, in order to plan ahead for warehousing resources and reduce the investment of redundant sites. Generally speaking, the scenario needs to predict two objectives, which are the inventory volume and sales volume. Among them, the inventory is the total inventory of the category to be stored in all warehouses, which is limited by the warehouse capacity in the warehouse results; the sales volume is the total amount of the category to be packed out of all warehouses, which is limited by the capacity in the warehouse results.

Our core objective is to choose a suitable clustering method to classify very large data sets on time series based on some similar features, and to combine the same category so that the same category has the most similar features in terms of demand. According to different categories, the appropriate prediction model is selected for prediction, which can greatly reduce the workload of prediction. Because of the large dataset, it is not practical to use individual data to predict the solution step by step.

3.1 Data Pre-processing

In order to better analyse the data, we combined and processed the data in the annexes and finally obtained the results as shown in Table 1 below.

Table 1: Integration Result Data Part Display

Seller no	19	19	19	19	19	19	19
Product no	448	448	448	448	448	448	448
Warehouse no	30	30	30	30	30	30	30
Date	20230509	20230417	20230109	20230120	20230213	20230215	20230311
Qty	10	14	2	1	22	21	14
Category1	mobile communications	mobile communications	mobile communications	mobile communications	mobile communications	mobile communications	mobile communications
Category2	Mobile Phone Accessories	Mobile Phone Accessories	Mobile Phone Accessories	Mobile Phone Accessories	Mobile Phone Accessories	Mobile Phone Accessories	Mobile Phone Accessories
Category3	Mobile Phone Accessories_12	Mobile Phone Accessories_12	Mobile Phone Accessories_12	Mobile Phone Accessories_12	Mobile Phone Accessories_12	Mobile Phone Accessories_12	Mobile Phone Accessories_12
Category	numerals	numerals	numerals	numerals	numerals	numerals	numerals
Inventory category	C	C	C	C	C	C	C
Level	Large	Large	Large	Large	Large	Large	Large
Warehouse _category	central warehouse	central warehouse	central warehouse	central warehouse	central warehouse	central warehouse	central warehouse
Warehouse _region	Southern China	Southern China	Southern China	Southern China	Southern China	Southern China	South China

For raw data, there are a large number of data indicators that cannot be modelled directly. Subsequent modelling usually requires numerical inputs, and the data contains textual data (e.g., product categories, warehouse areas, etc.) that need to be converted to numerical data. Therefore, data transcoding is required to fix inconsistencies. We used SPSS Pro for transcoding and some of the data transcoding results are shown in Table 2 below.

Table 2: Partial Data Transcoding Results

Inventory Classification Name	**Transcoding results**	**Merchant Size Name**	**Transcoding Results**	**Warehouse Area Name**	**Transcoding Results**	**Warehouse Area Name**	**Transcoding Results**
A	1	Large	1	north-eastern	1	central China	5
B	2	Medium	2	Northern China	2	northwestern	6
D	3	Small	3	Eastern China	3	southwestern	7
C	4	Special	4	Southern China	4		

3.2 Cluster Analysis

Prior to inventory and sales forecasting, cluster analysis allows similar categories to be grouped together based on characteristics such as historical inventory levels, sales, piece types, and relatedness of the categories. This helps simplify the data structure, reduce modelling complexity, and build separate forecasting models for each cluster to capture their common patterns more accurately. In addition, the clustering results can also reveal the correlation between categories, providing a scientific basis for subsequent category binning planning and ultimately improving the overall efficiency of forecasting and warehouse optimisation.

Cluster analysis was used to classify the pre-processed dataset, and the total number of cases to be classified was 1996, and the statistical results are shown in Table 3.

Table 3: The Required Classification Data Statistics

Case-by-Case					
Efficiently		Hiatus		(Grand) Total	
Number of Cases	Per Cent	Number of Cases	Per Cent	Number of Cases	Per Cent
1996	100.0	0	.0	1996	100.0

The systematic cluster analysis dendrogram shows the clustering process and the classification results of the cases belonging to different classes when they are shown in Fig. 1. Observation of the dendrogram reveals that the samples can be classified into 2, 3, 4, 5 classes and so on, using the difference between the breast-waist and the waist as the basis for classification. By approximating the classification with the average monthly demand and the size of the business, it is now only necessary to predict the classification results when there are only 12 classifications, which greatly reduces the workload and improves the prediction efficiency. Finally, the results of the 12 partial classifications are shown in Table 4.

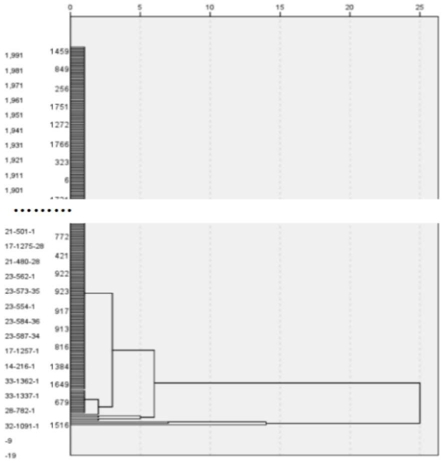


Figure 1: Systematic Cluster Analysis Dendrogram

Table 4: Part of the Data Classification Results

Member of a Clustering		
Case-by-case	12 Clusters	Number of Cases in each Cluster
1:10-1664-13	1	173.000
201:11-178-16	2	1.000
209:11-178-9	3	6.000
295:12-323-1	4	1.000
426:14-1590-1	5	18.000
447:14-194-1	6	2.000
740:18-1505-1	7	1.000
786:21-474-1	8	1783.000
924:23-559-1	9	8.000
1504:32-1085-1	10	1.000
1516:32-1091-1	11	1.000
1650:33-1375-54	12	1.000

3.3 ARIMA Forecasting Model

Indicators such as inventory level and sales volume have obvious trends and fluctuations, so we consider using ARIMA model for forecasting. Through the historical data, ARIMA model can capture the long-term trend of the inventory level and help to predict the average monthly inventory level in the coming months, which can provide reference for the planning of warehouse resources. At the same time, it can also make a fine prediction of daily sales, accurately estimate the demand for outgoing stock, and provide a basis for warehouse capacity allocation.

From the classification of data in Table 3, we can see that the number of cases in the 8th cluster is the largest, containing 1783 cases, which is the most representative, so we take the data in the 8th cluster as an example, and the final prediction results are shown in Fig. 2.

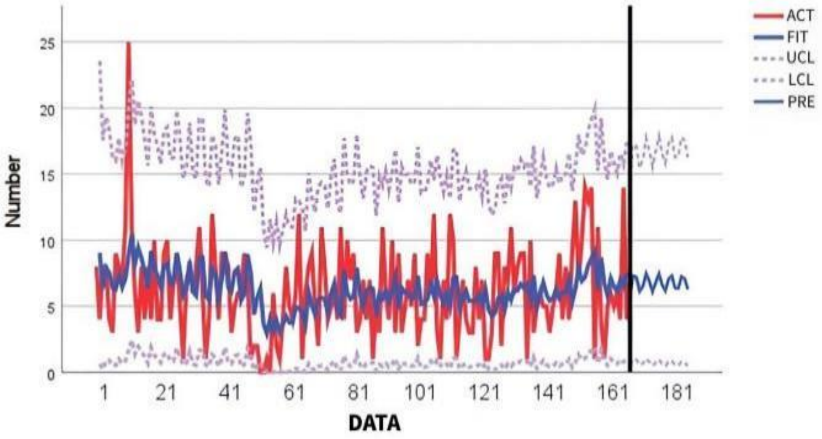


Figure 2: Model Prediction Result Diagram

3.4. LSTM Predictive Modelling

The LSTM model is able to effectively handle time-series data with non-linear and long-term dependencies such as inventory and sales, and the memory unit of LSTM can capture the long-term effects of historical data while ignoring short-term noise to generate more accurate forecasts. It is also able to handle multi-dimensional inputs, incorporating factors such as category characteristics and historical sales into the model to further improve prediction accuracy. With the strong modelling capability of LSTM, we can better predict future inventory and sales, providing a scientific basis for warehouse resource planning and warehouse allocation decisions.

In order to reflect the contrast effect, enhance the persuasiveness and universality, we use the same dataset as the ARIMA model in the eighth cluster, for the prediction of the LSTM model, through the repeated training of the LSTM model, the encapsulation function, and constantly modify the corresponding parameters, such as the number of iterations, the learning rate, the descent factor, the validation step, etc., so as to make the prediction results more close to the real value, the final model prediction results are shown in Figure 3. The results are shown in Figure 3, and the analysis of the relative error of the model training set shows that the relative error of the LSTM model is large and the fitting effect is not ideal.

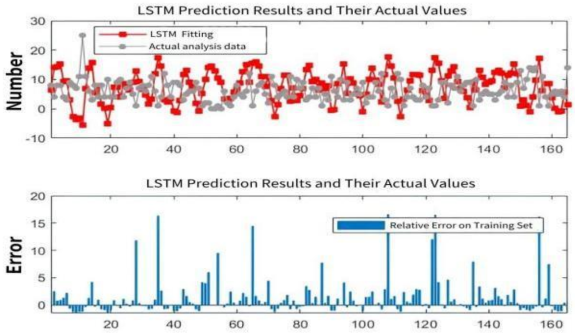


Figure 3: LSTM Model Prediction Results and Relative Error Plots

We therefore consider combining LSTM with ARIM to improve the model fit.

3.5 Combined Volume Optimisation Forecasting Model

We used ARIMA model and LSTM model for demand forecasting respectively, but the actual forecasting results were not satisfactory. In order to improve the accuracy of demand forecasting, we combine the ARIMA model with the LSTM model. We first use the ARIMA model to predict, then subtract the data set corresponding to the real demand data set from the data set obtained by using the ARIMA prediction model, and then use the difference data set to predict the LSTM model. The final actual predicted value is equal to the value predicted by the ARIMA model plus the difference predicted by the LSTM model. The flowchart of the combinatorial optimisation prediction model is shown in Figure 4.

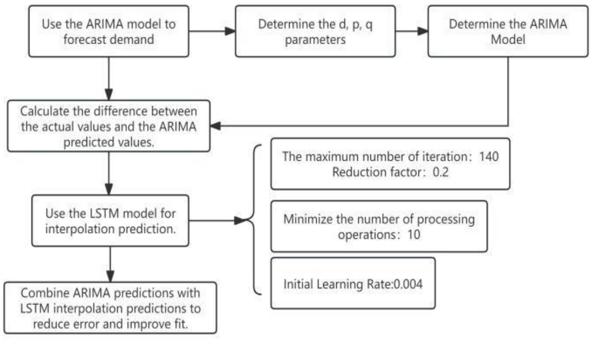

Use the ARIMA model to forecast demand
Determine the d, p, q parameters
Determine the ARIMA Model
Calculate the difference between the actual values and the ARIMA predicted values.
The maximum number of iteration: 140 Reduction factor: 0.2
Use the LSTM model for interpolation prediction.
Minimize the number of processing operations: 10
Initial Learning Rate:0.004
Combine ARIMA predictions with LSTM interpolation predictions to reduce error and improve fit.

Figure 4: Flowchart of Combinatorial Optimisation Forecasting Model

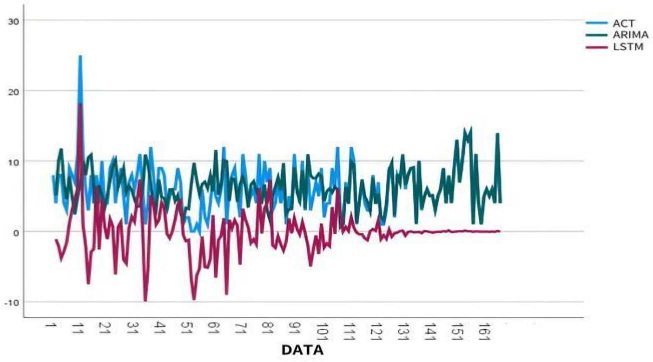


Figure 5: Plot of True Values, ARIM Predictions and LSTM Interpolation

We still use the dataset in the 8th clustering for prediction, firstly we use ARIMA model to make preliminary prediction on the dataset, then we use LSTM model to go to the difference prediction, and the difference prediction results are shown in Figure 5 below.

3.6 Analysis of Model Evaluations

In order to quantify the advantages and disadvantages of each model's prediction results, we give the prediction accuracy index as: $1 - \mathrm{wmape}$, which is given by the following formula:

$$1 - \mathrm{wmape} = 1 - \frac{\sum |y_i - \widehat{y_i}|}{\sum y_i}$$

wherey_i is the number ofi the true demand for the sequence (the daily quantity of various goods stored by the merchant in each warehouse), and$\widehat{y_i}$ is the predicted demand for the ith sequence.

The data predicted using the combinatorial optimisation prediction model was brought into the formula for the prediction accuracy metrics, and from the results Table 5, it can be seen that the model predicted a higher accuracy and a better model.

Table 5: Tests of Prediction Results by Category

Form	Wmape(be) worth	Form	Wmape(be) worth	Form	Wmape(be) worth
1	0.8900	5	0.8700	9	0.9100
2	0.9400	6	0.8130	10	0.9000
3	0.9520	7	0.9975	11	0.8900
4	0.9600	8	0.8100	12	0.8500

4. One Product, One Warehouse Planning Model

4.1 Second-order Clustering Classification Model

Since the product category and warehouse category are fixed variables, we decided to use the second-order clustering algorithm for comparative analysis. We find out the average value of all goods in April and May, as well as the four values of product category, merchant level and warehouse category. Using them as eigenvalues for second-order clustering, we finally clustered 4 classes, and some of the clustering results are shown in Table 6.

Table 6: Clustering Results for Selected Commodities

Seller	Category1	Warehouse_category	Seller_level	April average	Average May	TSC_9844
1-1-4	daily necessities	central warehouse	New	0.6153	7.4000	4
1-1-2073	daily necessities	central warehouse	New	0.3334	0.6000	4
1-1-2084	daily necessities	central warehouse	New	0.7778	4.0000	4
1-1-2262	daily necessities	central warehouse	New	0.8235	0.3334	4
1-1-2350	daily necessities	central warehouse	New	0.9615	0.9334	4
1-2-9	daily necessities	Regional warehouse	New	0.8667	1.5334	4

The quality of the clustering results was evaluated and analysing Figure 6 shows that the clustering into 4 classes is of good quality.

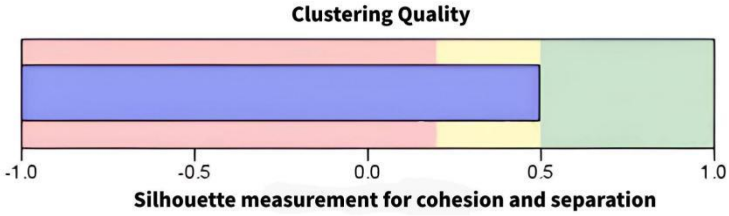
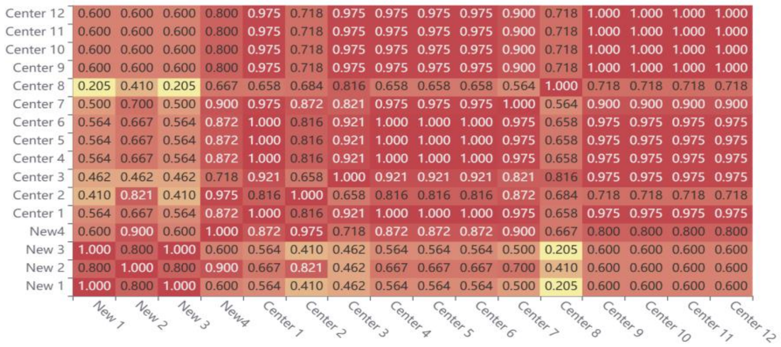


Figure 6: Clustering Quality Evaluation Diagram

4.2 Solving with the Optimised Prediction Model

We would like to refer to the data prior to the new product's April date with the data from the clustering centres, and apply the Combinatorial Optimisation Prediction Model to solve it. To do this, a spearman correlation analysis is performed between the four categories and the 12 cluster centres identified in the previous section to identify the five cluster centres that are closest to each other. We select the merchant category, warehouse category, and the average daily demand in April and May for analysis, and since it contains more fixed class variables, Spearman correlation analysis is adopted. From the correlation heat coefficient diagram in Fig. 7, it can be seen that the correlation coefficient between new product four and the second cluster centre is 0.975, and the correlation coefficient between new product two and the second cluster centre is 0.821, and both clusters of These two clusters have strong correlation with the 12 centres of Question 1 cluster, so we decided to replace the missing demand information of this new product before April with its corresponding cluster centre data, while for the other two new product clustering centres, we screened the time series from the original time series and selected the more similar time series to be used in place of them, so as to finally obtain a set of relatively complete data sets conducive to the continuation of the prediction.



	New 1	New 2	New 3	New4	Center 1	Center 2	Center 3	Center 4	Center 5	Center 6	Center 7	Center 8	Center 9	Center 10	Center 11	Center 12
Center 12	0.600	0.600	0.600	0.800	0.975	0.718	0.975	0.975	0.975	0.975	0.900	0.718	1.000	1.000	1.000	1.000
Center 11	0.600	0.600	0.600	0.800	0.975	0.718	0.975	0.975	0.975	0.975	0.900	0.718	1.000	1.000	1.000	1.000
Center 10	0.600	0.600	0.600	0.800	0.975	0.718	0.975	0.975	0.975	0.975	0.900	0.718	1.000	1.000	1.000	1.000
Center 9	0.600	0.600	0.600	0.800	0.975	0.718	0.975	0.975	0.975	0.975	0.900	0.718	1.000	1.000	1.000	1.000
Center 8	0.205	0.410	0.205	0.667	0.658	0.684	0.816	0.658	0.658	0.658	0.564	1.000	0.718	0.718	0.718	0.718
Center 7	0.500	0.700	0.500	0.900	0.975	0.872	0.821	0.975	0.975	0.975	1.000	0.564	0.900	0.900	0.900	0.900
Center 6	0.564	0.667	0.564	0.872	1.000	0.816	0.921	1.000	1.000	1.000	0.975	0.658	0.975	0.975	0.975	0.975
Center 5	0.564	0.667	0.564	0.872	1.000	0.816	0.921	1.000	1.000	1.000	0.975	0.658	0.975	0.975	0.975	0.975
Center 4	0.564	0.667	0.564	0.872	1.000	0.816	0.921	1.000	1.000	1.000	0.975	0.658	0.975	0.975	0.975	0.975
Center 3	0.462	0.462	0.462	0.718	0.921	0.658	1.000	0.921	0.921	0.921	0.821	0.816	0.975	0.975	0.975	0.975
Center 2	0.410	0.821	0.410	0.975	0.816	1.000	0.658	0.816	0.816	0.816	0.872	0.684	0.718	0.718	0.718	0.718
Center 1	0.564	0.667	0.564	0.872	1.000	0.816	0.921	1.000	1.000	1.000	0.975	0.658	0.975	0.975	0.975	0.975
New4	0.600	0.900	0.600	1.000	0.872	0.975	0.718	0.872	0.872	0.872	0.900	0.667	0.800	0.800	0.800	0.800
New 3	1.000	0.800	1.000	0.600	0.564	0.410	0.462	0.564	0.564	0.564	0.500	0.205	0.600	0.600	0.600	0.600
New 2	0.800	1.000	0.800	0.900	0.667	0.821	0.462	0.667	0.667	0.667	0.700	0.410	0.600	0.600	0.600	0.600
New 1	1.000	0.800	1.000	0.600	0.564	0.410	0.462	0.564	0.564	0.564	0.500	0.205	0.600	0.600	0.600	0.600

Figure 7: Heat Map of Correlation Coefficients

4.3 Solving with the Optimised Prediction Model

We again used the combinatorial optimisation prediction model from the previous paper to solve the problem and some of the results are shown in Table 6 below.

Table 6: Some Commodity Prediction Results

Seller_no	Product_no	Warehouse_no	Date	Forecast_qty
3	31	2074	2023-5-16	17.32903953
3	31	2074	2023-5-17	12.81330274
3	31	2074	2023-5-18	10.14539891
3	31	2074	2023-5-19	8.265889635

5. Conclusion

In this study, a combined optimal forecasting model based on ARIMA and LSTM is proposed for the demand forecasting problem in the e-commerce retail industry, which combines the advantages of the two models and significantly improves the accuracy and robustness of the forecast. Through the steps of data preprocessing, cluster analysis and model evaluation, the results of the study show that the method has good potential for application in optimising inventory management and improving the efficiency of supply chain decision-making.

However, this study still has some limitations. The model is more dependent on the quality of historical data, which makes it difficult to effectively respond to sudden market changes or anomalies. In addition, the training and tuning of the LSTM model is computationally expensive for large-scale datasets. Future research could further introduce external factors, such as real-time market trends or consumer behaviour data, to improve prediction accuracy. Meanwhile, the development of generic prediction frameworks that can cope with dynamic and uncertain scenarios is also a direction worth exploring.

References

[1] Alsharif M H, Younes M K, Kim J. Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea [J]. Symmetry, 2019, 11(2): 240.

[2] Alqatawna A, Abu-Salih B, Obeid N, et al. Incorporating Time-Series Forecasting Techniques to Predict Logistics Companies' Staffing Needs and Order Volume[J]. Computation, 2023, 11(7): 141.

[3] Tang Y M, Ho G T S, Lau Y Y, et al. Integrated smart warehouse and manufacturing management with demand forecasting in small-scale cyclical industries[J]. Machines, 2022, 10(6): 472.

[4] Ribeiro A M N C, do Carmo P R X, Endo P T, et al. Short- and very short-term firm-level load forecasting for warehouses: a comparison of machine learning and deep learning models[J]. Energies, 2022, 15(3): 750.

[5] Ho G T S, Choy S K, Tong P H, et al. A forecasting analytics model for assessing forecast error in e-fulfilment performance[J]. Industrial Management & Data Systems, 2022, 122(11): 2583-2608.

[6] Tufano A, Accorsi R, Manzini R. A machine learning approach for predictive warehouse design[J]. The International Journal of Advanced Manufacturing Technology, 2022, 119(3): 2369-2392.

[7] Alaziz S N, Albayati B, El-Bagoury A A H, et al. Clustering of COVID-19 multi-time series-based K-means and PCA with forecasting[J]. International Journal of Data Warehousing and Mining (IJDWM), 2023, 19(3): 1-25.

[8] Chen J, Xu S, Liu K, et al. Intelligent transportation logistics optimal warehouse location method based on internet of things and blockchain technology [J]. Sensors, 2022, 22(4): 1544.

[9] Wan Y, Wang S, Hu Y, et al. Multiobjective Optimization of the Storage Location Allocation of a Retail E-commerce Picking Zone in a Picker-to-parts Warehouse[J]. Engineering Letters, 2023, 31(2).

[10] Tang Z, Ge Y. CNN model optimisation and intelligent balance model for material demand forecast[J]. International Journal of System Assurance Engineering and Management, 2022, 13(Suppl 3): 978-986.

[11] Martins E, Galegale N V. Sales forecasting using machine learning algorithms[J]. Revista de Gestão e Secretariado, 2023, 14(7): 11294-11308.