Research on E-Commerce Commodity Demand Forecasting and Inventory Optimization Based on Time Series Forecasting Model

Jing Li [1,*,†], Wen Li [2,†], Ruixuan Hao [1,†], Rui Li [3]

[1] School of Materials Science and Engineering, Anhui University of Science and Technology, Huainan, China

[2] School of Safety Science and Engineering, Anhui University of Science and Technology, Huainan, China

[3] School of Wildlife and Nature Reserves, Northeast Forest University, Harbin, China

* Corresponding Author Email: 2645982388@qq.com

† These authors contributed equally to this work

Abstract. In this study, 1996 sets of time series data are extracted from the three dimensions of merchant, warehouse and commodity, and the outliers are removed using the 3σ criterion. Through visual analysis, it was found that home improvement building materials and food and beverages were popular commodities. Then ARIMA and LSTM models were used to forecast the demand. Through 1-wmape index evaluation, LSTM model was found to perform better. Finally, the K-Means cluster analysis is carried out based on the categories of goods, merchants and warehouses, and the cluster number is selected as 70. The cosine similarity method is used to match the newly added dimensions to the category of the cluster in problem 1, and the ARIMA model is used for prediction. The results show that the cosine similarity method can accurately predict the commodity demand, the 1-wmape index is 0.801, and the residual plot presents randomness. The data of November 2022 is added to the original data, and the missing data is processed by linear interpolation. The LSTM model was used to forecast the increase of the total demand during Double 11. By setting a periodic time series with a step size of 30, the model achieves a good accuracy in predicting the demand in June, with a 1-wmape index of 0.829. To sum up, this study comprehensively considers three dimensions of merchants, warehouses and commodities, and uses different models to make accurate predictions. The ARIMA and LSTM models are adopted, while the promotion characteristics are considered and the appropriate LSTM parameters are selected. The research results provide practical forecasting tools and methods for supply chain management.

Keywords: Three sigma criterion, seasonality, cosine similarity, K-means clustering, time series prediction.

1. Introduction

On an e-commerce platform, managing the inventory of thousands of merchants is a complex and critical task. Through scientific management methods and intelligent decision-making, inventory costs can be reduced, while ensuring timely performance. In supply chain optimization, demand forecasting and inventory optimization are two important aspects. Demand forecasting is the basis of intelligent supply chain decision-making, by forecasting demand in advance to understand the demand situation in each region, and to place inventory in warehouses close to demand. The task of demand forecasting is to predict the future demand of each commodity in each warehouse based on the demand of a historical period of time. Prediction accuracy is an important index to evaluate the quality of prediction model. However, in the actual e-commerce supply chain forecasting task, we often encounter various problems. For example, some goods are sold for too short a time, resulting in less historical data; The increase in sales during large promotions brings irregular demand. Therefore, it is necessary to find general rules in historical data through algorithms to achieve accurate demand forecasting. Inventory optimization is to effectively manage the inventory level, enterprises usually develop a variety of inventory strategies to control the frequency and quantity of replenishment.

Among them, the regular inventory strategy is a common strategy, through the regular inventory level to decide whether to carry out replenishment. The goal of inventory optimization is to reduce the total cost of holding inventory, meet service levels, and reduce inventory turnover days.

2. Model building and solving

2.1. Data analysis and visualization

Due to the large amount of data, data analysis and visualization are carried out to explore the potential rules in the data. Perform an initial exploratory analysis of the data, using statistical and visual methods to identify patterns, trends and anomalies in the data. This helps to establish the basis for subsequent analysis to discover the characteristics and patterns of the data. Firstly, the data is preprocessed, and a certain commodity in the same warehouse of the same merchant is defined as one data dimension, and a total of 1996 data dimensions are obtained. Then, the outlier value processing is carried out using the 3σ criterion to facilitate the subsequent analysis of the data.

The 3σ criterion assumes that a set of detection data contains only random errors, and the standard deviation is obtained by calculating and processing it, and an interval is determined according to a certain probability. It is considered that any error exceeding this interval is not a random error but a gross error, and the data containing this error should be removed. The maximum error range is 3σ. Given a set of measured values $x_i(i = 1,2,\cdots,n)$ whose sample average is $\bar{x}$, the deviation is $\Delta x_i = x_i - \bar{x}$ the standard deviation calculated as follows. If there is $\Delta x_i > 3s$ a deviation of a measured value $x_i(1 \le i \le n)$, it is considered to be a bad value containing a gross error, and it is eliminated.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\Delta x_i)^2} \tag{1}$$

In addition, in order to more accurately grasp the demand for goods and inventory optimization of e-commerce retailers, the characteristics of time series data should be fully mined, such as the periodicity of goods, seasonality and distribution of merchants.

According to the preliminary statistics, the distribution of first-level product categories is shown in Figure 1. It can be seen that the number of merchants is 35, the number of products is 1212, the number of warehouses is 54, and the total demand in the past six months is 4758809. Through the analysis of product categories, it is found that home improvement, building materials, food and beverage, makeup and skin care are favored by the public. Therefore, its forecast demand will inevitably be greater than other products, including the follow-up replenishment process should also focus on the purchase of these types of products.

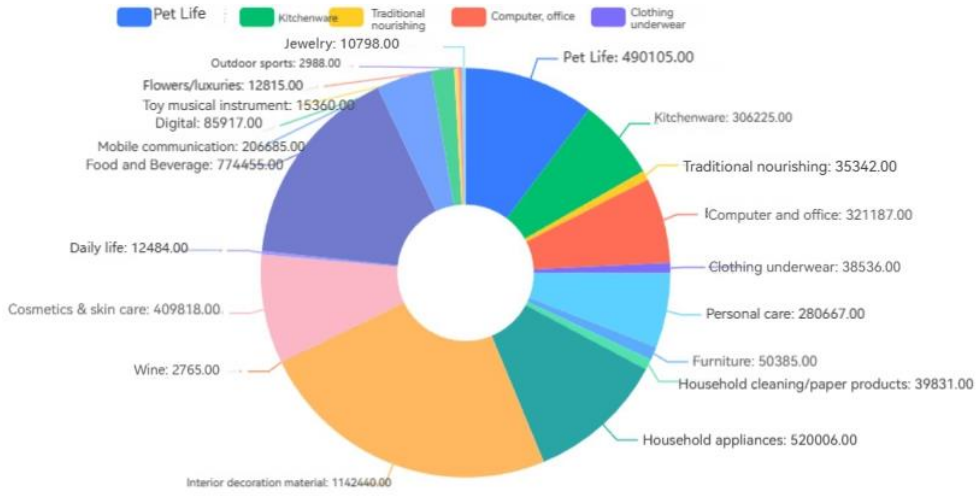


Figure 1. Product category distribution

In order to further explore the seasonal or periodic characteristics of products in the historical 6-month data, a product product_no of 311 was randomly selected for research. As shown in Figure 2, it is easy to know that the demand of such products presents a certain periodicity, and the demand increases sharply on some special days, showing a seasonal trend. It is speculated that this may have something to do with a specific promotion or holiday.

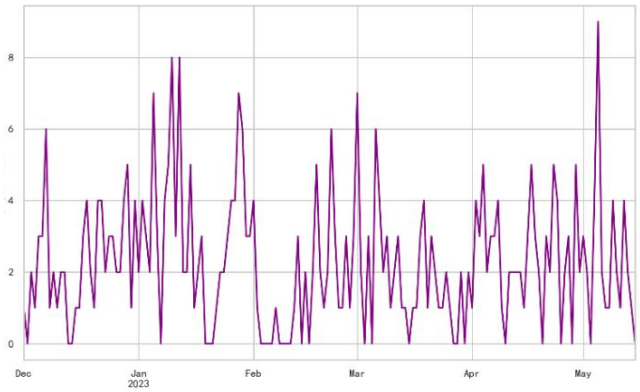


Figure 2. The sales trend of product no. 11 in a warehouse

Secondly, the warehouse demand with warehouse_no as 12 is randomly selected for research, and the demand is regarded as its actual purchase quantity. The demand of the warehouse presents a certain periodicality, as shown in Figure 3. The demand of a warehouse, that is, the actual purchase quantity, also reflects the storage capacity of the warehouse, and the size or storage capacity of the warehouse will not fluctuate greatly in the short term.

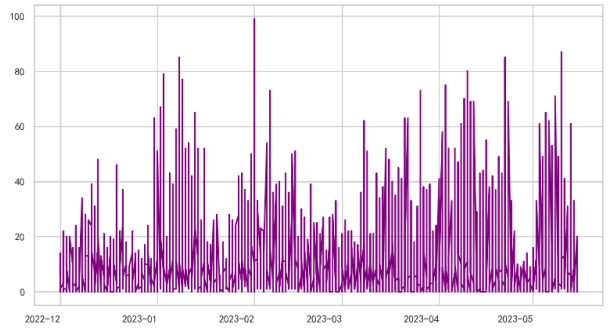


Figure 3. The sales trend of a warehouse no 12 products

Finally, the data dimension product_no is 300 and warehouse_no is 1 is randomly selected for analysis. As shown in Figure 4, it can be seen that the demand of this product in this warehouse is the highest from January to February, while the weather in this period is relatively cold, and the product demand decreases significantly in other months. This may be related to the characteristics of the product itself. Therefore, the timeliness of the product should be emphasized in the follow-up research of forecast replenishment, so as to determine the peak season and off-season of the product.

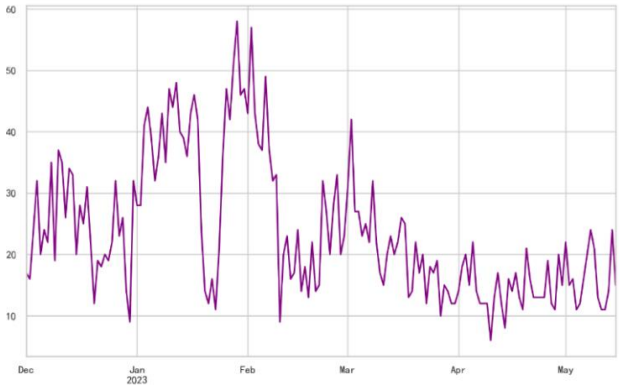


Figure 4. A warehouse - product sales trends

2.2. Similar demand categories based on ARIMA and LSTM forecasts

Through the above data analysis, it can be seen that the demand of the product presents a certain seasonality, and the ARIMA model is initially selected. This model is a commonly used time series prediction model, which first transforms non-stationary time series into stationary data through differential processing, then models the autocorrelation and moving average of time series, and performs parameter estimation and prediction based on historical data. In order to characterize the performance of the prediction model, indexes 1-wmape, root mean square error (RMSE) and mean absolute error (MAE) were selected to comprehensively evaluate the effect of the model. The calculation method of index 1-wmape was shown in Equation (2).

$$1-wmape=1-\frac{\sum\left|y_i-y_i\right|}{\sum y_i} \tag{2}$$

However, by solving the ARIMA model with its index value of 1_wmape 0.585, the fitting effect is relatively ordinary. At the end of the analysis of the 15-day warehouse sales forecast, the value of the model will tend to stabilize. Therefore, the LSTM model of machine learning is considered, and the classical evaluation indicators such as 1-wmape are combined to evaluate the effect. Through calculation, LSTM model 1_wmape index value 0.711, it can be found that the effect of this model is much improved compared with the ARIMA model. Therefore, considering the prediction results of the two models comprehensively, the LSTM prediction model is selected to predict the commodity demand of the e-commerce platform.

In order to conduct subsequent cluster analysis, it is first necessary to select appropriate cluster indicators. According to the process of data analysis and modeling, the time series formed by merchants, warehouses and commodities can be divided according to the following classification methods, and the mean value, variance, warehouse area and first-level classification of products of each time series can be selected for cluster analysis of 1996 product dimensions.

After predicting 1996 time series data, in order to make the characteristics of the same category most similar in demand, K-Means clustering is used to explore the time series classification problem formed by merchants, warehouses and commodities. Since the number of categories is not clear, in order to calculate the categories more accurately, the elbow chart is used to evaluate. With the increase of cluster number k, sample division will be more refined, and the degree of aggregation of each cluster will gradually increase, so the error square and SSE will naturally become smaller. In addition, when k is less than the real cluster number, SSE will decrease greatly because the increase of k will greatly increase the degree of aggregation of each cluster, and when k reaches the real cluster number, the return on the degree of aggregation obtained by increasing k will rapidly decrease, so SSE will decrease sharply, and then tend to be flat with the continuous increase of k value. That is to say, the relationship between SSE and k is the shape of an elbow, and the corresponding k value of this elbow is the real cluster number of data.

As shown in FIG. 5, it can be seen from the elbow chart that the square error and SSE of the cluster category number between 60 and 100 are both small. As a large cluster number will also affect the subsequent prediction accuracy, cluster=70 is finally adopted to allocate the original time data series.

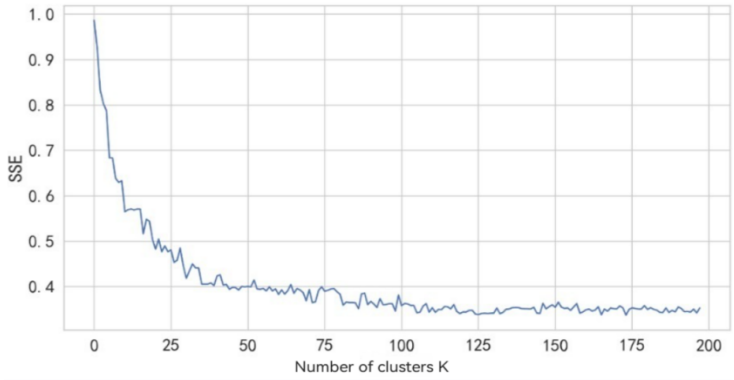


Figure 5. The elbow chart determines the k value

2.3. Category matching based on cosine similarity

For the data of these new dimensions (merchant - warehouse - goods), if there is a large amount of historical sales data, the time prediction model can be trained directly on the basis of the new data, but if the data is insufficient, the prediction must be based on the historical data. According to the data analysis, the sales days of the new product are far lower than that of the normal product, which is only 1/4 of that of the normal product, as shown in Figure 6. Therefore, the effect of using the short-term sales series of the new product to predict the sales situation in the next 15 days will be much worse than that of using the normal product to predict the effect. Secondly, in these data dimensions, there are completely new products, that is, new products on the market, which has almost no historical data for reference, that is, "cold start prediction"; The other is that the previous product only changes the storage warehouse or sales business, which is much easier to deal with than the former, that is, refer to the past demand for the same product.

Finally, this paper decides to make a prediction for the future sales of the new product by referring to the similar series of new sales data series. That is, according to the K-Means clustering model trained above, the new dimension is classified into a certain category sequence, and then the product demand in the next 15 days is predicted according to the data of this category.

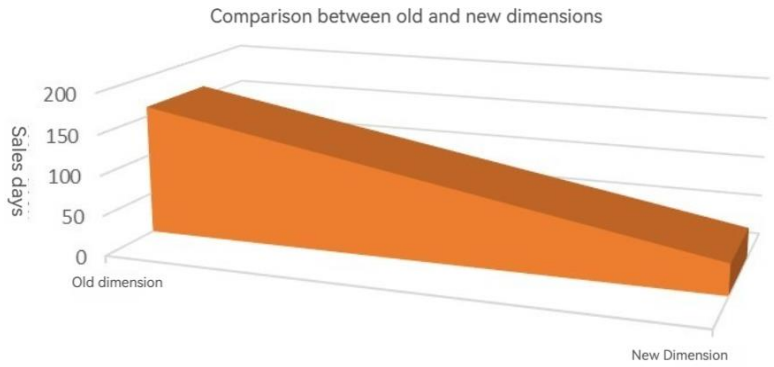


Figure 6. Comparison of days sold in old and new dimensions

In order to accurately match the time series data of the new dimension with 70 cluster categories, 210 groups of product dimensions were also selected from the candidate indicators to perform cluster analysis on the mean value, variance, warehouse area and product level classification of each time series. In this paper, cosine similarity is used to distinguish. Cosine similarity method is a common similarity measure used to measure the degree of similarity between two vectors. Cosine similarity method is often used in text similarity comparison, recommendation system, image similarity comparison and so on. By calculating cosine similarity between vectors, we can quickly judge their similarity degree, so as to make corresponding analysis and decision. It should be noted that the cosine similarity method may have some problems in dealing with sparse vectors and high-dimensional vectors, which can be improved and optimized by feature selection, dimensionality reduction and other methods.

The cosine similarity is defined as follows: For two vectors X and Y, their cosine similarity is defined as the product of their inner product divided by their modules by the formula shown below.

$$sim(X,Y) = X \square Y / (\|X\| \times \|Y\|) \tag{3}$$

Where $X \cdot Y$ represents the inner product of the vectors X and Y, and $||X||$ and $||Y||$ represent the modules of the vectors X and Y.

The above mentioned cosine similarity method has been used to match each new dimension, and some of them have a similarity of more than 0.9. Obviously, these new dimensions can make good use of their category data to obtain higher accuracy in the subsequent prediction.

ARIMA and LSTM time series models were used to predict 210 new dimension data. The residual diagram of the ARIMA model is intuitively analyzed, as shown in Figure 7. It is not difficult to find that its periodicity is not significant, and the residual is random and independent. Secondly, the 1-wmape index is 0.801, which indicates that the accuracy is satisfactory. At the same time, it indicates that the method of determining cluster categories by cosine similarity method is reasonable, so the LSTM time prediction model is no longer considered here.

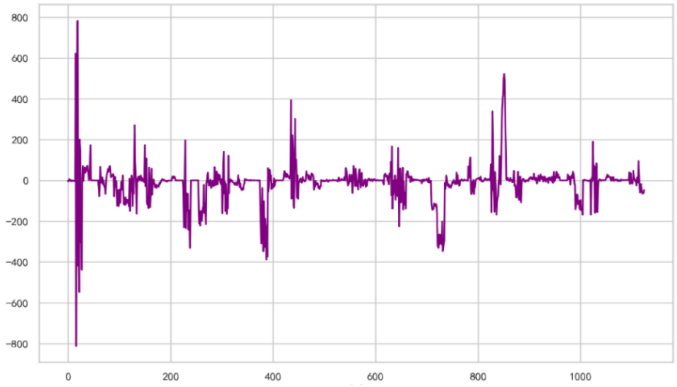


Figure 7. Residual plot of ARIMA model

2.4. Prediction based on LSTM time series model

Firstly, in order to ensure the continuity of the time series, linear interpolation is performed on the data. Roughly observing the data after fitting, it is found that the total demand increases significantly during the Double Eleven, as shown in Figure 8. Therefore, in order to better describe the sudden increase in commodity demand caused by a special period of time, the promotion feature marking is introduced, that is, the feature marking is carried out during this period, that is, the sudden increase of the commodity is marked as 0, otherwise it is marked as 1, so as to give it trend characteristics in advance when predicting the demand in June.

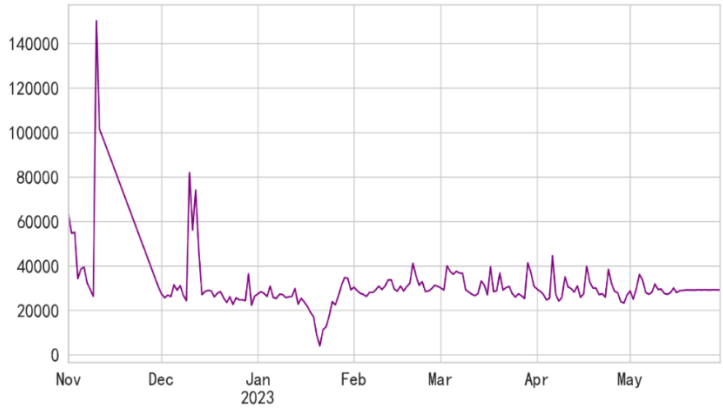


Figure 8. Total demand for all goods

LSTM is a kind of recurrent neural network, which supports the input of multiple external variables. Since the data has more trend features, it is equivalent to adding a new promotion feature in addition

to its own demand series value, which is obviously impossible for ARIMA model to train. Therefore, LSTM time series model is used for prediction. Since there will be corresponding promotion activities in each month, and based on the assumption that the promotion intensity of the activities is similar, a typical periodic time series is formed. The step length of the LSTM forecasting model is set to 30, that is, the prediction is made according to the historical data of the past 30 years, which not only ensures the rationality of the step length, but also ensures that the model forecasting time is not too long.

As shown in Figure 9, the intuitive analysis of the residual diagram of the LSTM prediction model shows that there is no obvious periodicity, and its residual is random and independent. The 1-wmape of the second model is as high as 0.829, and the accuracy of the model is better. Therefore, after supplementing the data with linear interpolation, it is reasonable to introduce the trend feature and specify the step size of 30 for LSTM prediction.

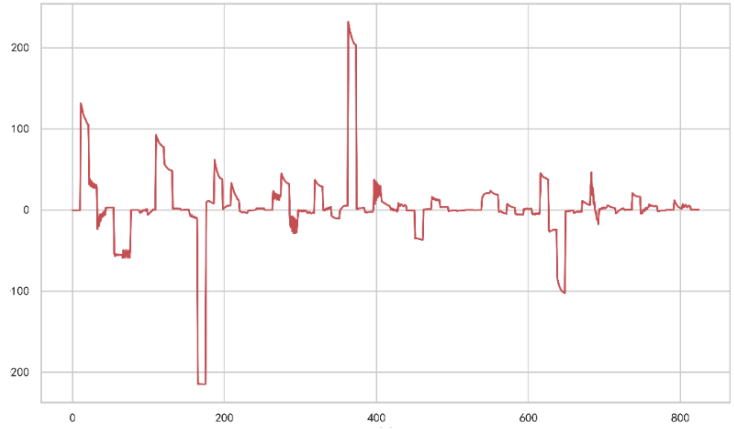


Figure 9. LSTM residual curve

3. Literature References

The literature review in this paper mainly covers the supply chain prediction algorithm, time series prediction method and the application of relevant data mining and machine learning technology. Literature [1] provides theoretical and practical experience in the field of intelligent supply chain. This paper expounds the application of prediction algorithm in supply chain management from two aspects of theory and practice. It has guiding significance for the research of demand forecasting and inventory management in supply chain. Reference [2] provides the results, findings and conclusions of the latest time series prediction algorithms. This is of great reference value for understanding the latest trends and effective methods in the current forecasting field. Literature [3] covers a large amount of time series data and a variety of forecasting methods. It is important to understand the challenges and solutions in the field of time series forecasting, and to evaluate the performance of different methods. Literature [4] provides the basic concepts and methods in the field of data mining. It may contain some data mining techniques used in the research, which provides a theoretical basis for the research data prediction and model building. Literature [5] focuses on the application of Bayesian methods in machine learning. In the research, Bayesian method may be used to optimize the model and improve the prediction accuracy, so it has certain reference value for understanding its application in supply chain prediction. In literature [6], it is mentioned that the algorithm using the Rajda criterion is used to identify abnormal data, which provides a way of thinking for the processing of outliers in the research. Reference [7] introduces the application of LSTM network in short-term power load forecasting. Due to the commonality of time series forecasting problems in different fields, this paper is valuable for understanding the potential application of LSTM in supply chain demand forecasting. Reference [8] describes the application of LSTM neural networks in language modeling. Considering the sequential nature of time series data, understanding the successful application of LSTM in other fields is instructive for model selection and optimization in research. Reference [9]

discusses the method of applying autoregressive moving average model in wind power generation capacity prediction. This paper provides a possible modeling approach to the seasonality and trend problems considered in the study. Based on the above literature, it can be seen that the research integrates the methods of supply chain management, prediction algorithm, data mining, machine learning and other fields. By drawing on the theory and practical experience of different fields, it provides various references for solving the problem of supply chain demand prediction.

4. Summary

In this paper, merchants, warehouses and products are considered comprehensively when selecting clustering indicators. Secondly, when the number of clusters is unknown, K-means clustering method is used to determine the K value with the help of elbow graph, aiming at the minimum SSE. In this paper, ARIMA and LSTM models are established, and the accuracy of prediction is compared comprehensively. In this paper, cosine similarity method is used to judge the similarity between different time series. The calculation method is simple, easy to understand and implement, and is not affected by dimension. In this paper, considering the large amount of data in the research problem, it is simple and convenient to select the criterion of removing error data. However, seasonality can cause data fluctuations, and the predictive performance of a process elimination model that does not take seasonal differences into account may be affected by data quality and model selection. In addition, the cosine similarity in this paper only considers the shape of the sequence, without considering the time order of the sequence. For problems that require consideration of time order, such as time series prediction, this can lead to poor cosine similarity. When evaluating the model, the selected index is too simple, and other more comprehensive similarity measurement methods need to be considered to improve the applicability of the model. Finally, when evaluating the model, it is pointed out that the selection index of the model is too simple. In order to evaluate the model performance more comprehensively, multiple indexes can be considered comprehensively to reflect the accuracy, stability and applicability of the model. To sum up, this paper has certain advantages in method selection and model evaluation, but there is still room for further improvement in data processing and model application.

References

[1] Intelligent Supply Chain: Prediction Algorithm Theory and Practice [M], Beijing: Publishing House of Electronics Industry, 2023.

[2] Makridakis S, Spiliotis E, Assimakopoulos V. The M5 Accuracy competition: Results, findings and conclusions [J]. International Journal of Forecasting, 2020, 36 (1): 224-227.

[3] Makridakis S, Spiliotis E, Assimakopoulos V, et al. The M4 Competition: 100,000time series and 61 forecasting methods [J]. International Journal of Forecasting, 2020, 36 (1): 54-74.

[4] P.N. Tan, M. Steinbach & V. Kumar, "Introduction to Data Mining", Addison-Wesley (2005).

[5] Ochiai A. Zoogeographical studies on the soleoid fishes found Japan and its neighboring regions. II // Bull. Jap. Soc. sci. Fish. 1957. V. 22. № 9. P. 526-530.

[6] Polykovskiy, D. and Novikov, A., Bayesian Methods for Machine Learning · Coursera and National Research University Higher School of Economics 2018.

[7] Wu Hao, Wang Jianguo. Anomaly data recognition algorithm of GNSS deformation monitoring based on Laida Criterion [J]. Science and Technology Innovation and Productivity, 2019 (01): 30-34.

[8] Chen Liang, WANG Zhen, Wang Gang. Application of LSTM Network in Short-term power load forecasting under Deep Learning framework [J]. Electric Power Information and Communication Technology, 2017 (5): 8-11.

[9] Dong Lei, WANG Lijie, Hao Ying, HU Guofei, Liao Xiaozhong. Wind power capacity prediction based on autoregressive moving average model [J]. Journal of Solar Energy, 2011, 05: 617-622.