Research on Optimization of Taobao Product Sales Forecast Based on ARIMA model

Chengyang Li*

School of Information Management, Shanghai Lixin University of Accounting and Finance, Shanghai, 201209, China

*lichengyang2023@163.com

Abstract. Accurate sales forecast is of great guiding significance for online e-commerce. First of all, merchants can make corresponding marketing strategies and inventory plans in advance according to the predicted sales volume of goods, so as to meet the needs of consumers and achieve profits, while avoiding the stock shortage or backlog of goods, so as to improve operational efficiency and customer satisfaction, and help merchants better understand consumer demand and market trends. By analyzing historical sales data and consumer reviews, merchants can tap into consumers' shopping habits and preferences, as well as hot items and trends in the market. This information is very important for merchants to formulate marketing strategies and adjust product positioning, which can help merchants better meet market needs and consumer expectations.

Keywords: E-commerce, sales, ARIMA, machine learning algorithms, forecast

1 Introduction

With the rapid development of the Internet, e-commerce has become an indispensable part of People's Daily life. As one of the largest e-commerce platforms in China, Taobao attracts hundreds of millions of consumers and sellers, covering a wide range of product categories and brands. In this huge market, how to accurately predict the sales volume of Taobao e-commerce products has become the focus of many researchers. This paper will optimize the sales forecasting method of Taobao e-commerce products based on ARIMA model, and provide useful reference for the research in related fields.

2 Related Work

From the perspective of influencing factors, existing literatures focus on specific feature analysis. For example, Zhang Xinchao et al. (2021) [1] focused on the attributes and features of commodities in their research. In their study, Sun Ming et al. (2022) [2] proposed to establish a hierarchy for time series according to the importance of each feature. Jiang Wenwu et al. (2020) [3] focused on the time factor in their research. Zhou Yuduan, Yong Rui et al. (2021) [4] mainly analyzed user behavior characteristics in

J. Liao et al. (eds.), *Proceedings of the 2024 2nd International Conference on Digital Economy and Management Science (CDEMS 2024)*, Advances in Economics, Business and Management Research 292,
https://doi.org/10.2991/978-94-6463-488-4_4

their research. Baiyun (2020) [5] considers the characteristics of market environment in the study; Li Jie et al. (2018) [6] focused on natural factors including weather, seasons and holidays in their research. These factors may have an impact on the purchasing power of consumers, thus affecting the sales volume of Taobao goods.

From the perspective of prediction methods, many current papers focus on obtaining the most suitable model for calculation by comparing various models. For example, in his research, Pu Jiapeng (2018) [7] used machine learning models such as Xgboost and LSTM to discuss the prediction effect of single time series numbers. Wang Yuxia (2019) [8] proposed a sales prediction model based on XGBoost algorithm, Granger thought and feature processing method of sliding window sampling for comparative analysis. But only a small amount of research has focused on optimizing one model in particular.

These few papers are not satisfied with the existing research methods, and most of them make specific improvements based on the original forecasting methods and models. For example, Huo Jiazhen et al. (2023) [9] proposed a sales forecasting model based on ensemble empirical mode decomposition (EEMD), Holt-Winters and gradient Lift Tree (GBDT). Chen Qiang et al. (2021) [10] combined Word2Vec model to conduct correlation clustering of commodities, and discussed the feasibility of sales forecasting model of associated commodities based on Conv LSTM network.

After reading a lot of literature on this kind of improved algorithm, it is found that arima has strong modifiability and high prediction accuracy compared with other algorithms in predicting commodity sales abroad. The main reason can be seen in the review by Wen Hu et al. (2020) [11], which measures the degree of similarity between one subsequence of the residual sequence and another by improving the similarity distance function of the comparison layer in the adaptive resonance theory neural network. The experimental comparison shows that the optimized residual model can reflect the sales law more accurately and comprehensively, and improve the accuracy of product sales forecast.

3 Related Theories Based on ARIMA Model

3.1 AR(P) Model

Autoregressive Model is a widely used statistical model for analysis of time series data. For a time series data, the first order form of the autoregressive model can be expressed as:

$$Y_t = c + \varphi Y_{t-1} + \xi_t \tag{1}$$

Seeing equation (1) as an example, where Y_t represents the time series value at time t, that is, the observed value at time t. c represents the constant term of the model, also known as the intercept term. In some cases, this value may be set to 0 so that the model does not contain constant terms. φ is the autoregressive coefficient, Y_{t-1} is the observed value at time $t-1$, and ξ_t is the error term, also known as the white noise term.

In the standard AR model, in the strictest case, only the mean is 0, the variance is a specific binary $б^2$, and the normal distribution can be called white noise.

When the order of the autoregressive model is p, the model can be expressed as:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \xi_t \tag{2}$$

As shown in equation (2), where, φ_1、φ_2 and φ_p are autoregressive coefficients. And $\varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p}$ said the items of the past time points of time series and p values for the current point in time. Where, φ_p is the coefficient of the PTH lag term, representing the relative influence of the PTH lag term on the current time point. Y_{t-p} is the time series value at the time point t-p. We can specify the specific time interval between t and $t-1$ according to different scenarios, but in the same time series, the interval between t and $t-1$ must be the same as the interval between $t-p$ and $t-(p-1)$.

3.2 MA(Q) Model

Moving Average Model (MA(Q) Model) is a kind of model in time series analysis, which describes the relationship between the current time point data and the past noise. The MA model is defined as:

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{3}$$

This formula in equation (3) shows the moving average model of order q, where Y_t is the observed value at time point t of the time series we are interested in. μ is the mean or expected value of a time series, and this value is the same for all time points. In many practical time series analyses, we assume that the time series has somehow been converted to a series with a mean of zero. But in the full MA model, this mean term μ exists. ε_t、ε_{t-1}、ε_{t-2}、ε_{t-q} are so-called white noise terms, and the values at each point in time are independently and equally distributed, usually assumed to be normally distributed. The mean of these terms is 0 and the variance is σ^2. ε_t is white noise at the current time, ε_{t-1} is white noise at the previous time, and so on, ε_{t-q} is white noise before time q. These are the parameters of the MA model, and each parameter θ corresponds to a white noise term. They measure how much the corresponding white noise affects the current point in time. q is the order of how many past white noise terms are incorporated into the model, referring to the number of past white noise terms included in the model.

The implication of this formula is that at point t, the observed value Y_t is determined by a linear combination of the white noise at the past point q plus a constant (i.e. the mean μ) and the white noise at the current point in time. The weight of each white noise term is determined by the parameter θ.

4 Case study and Analysis

The data of this experiment comes from the sales data of a certain product on Taobao published on Kaggle from March 2019 to June 2020 and other relevant factors, including: time, sales, reviews, price change, weekend and day.

4.1 Data Preprocessing Based on Correlation

This study first extracts the data set from Kaggle, which can reflect the changes of other statistical factors when the sales volume of a certain product is different every day on Taobao. This study selected sales data from March 2019 to June 2020 and other relevant factors.

Table 1. Taobao commodity sales data set

time	sales	reviews	price change	weekend	day
2019/3/1	151	18	10	2	1

As shown in Table 1, the basic statistical characteristics of the data set are introduced. The specific data information is shown in Table 1. time represents the time, that is, the date. sales represents the sales volume, that is, the sales number of specific days; reviews indicates the number of comments added on the day; price change represents the amount of price change, that is, the change in price on that day compared with the original price. weekend stands for days off. There are several days off during the week. day indicates the day in the month.

First, visualization of sales data is performed in this experiment, and the results are shown in the figure 1 below:

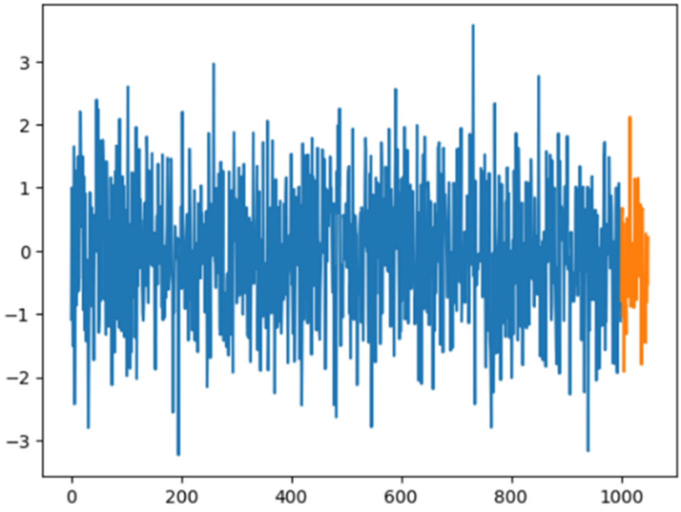


Fig. 1. sales price fluctuation chart

As can be seen from the Figure 1, it is of little significance to analyze the data only based on the timeline, but it is very messy to look at the sales data. Therefore, the experiment first conducts correlation analysis on sales volume and other factors according to correlation coefficient. The correlation coefficient is a statistic used to measure the degree of correlation between two variables, and its value is between -1 and 1. If the correlation coefficient of two variables is close to 1, it means that they have a strong positive correlation. If the correlation coefficient is close to -1, it indicates that there is a strong negative correlation between them. If the correlation coefficient is close to 0, there is little correlation between them. The following is the formula for the correlation coefficient:

$$r = \frac{\Sigma[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{[\sum(x_i - \vec{x})^2 \Sigma(y_{i.} - y)^2]}} \tag{4}$$

As shown in equation (4), where x_i and y_i represent the values of each observation point in the two variables respectively, x̄ and ȳ represent the average of the observed values of the two variables respectively, and Σ represents the sum.

The data is calculated by correlation coefficient, as shown in the figure 2 below:

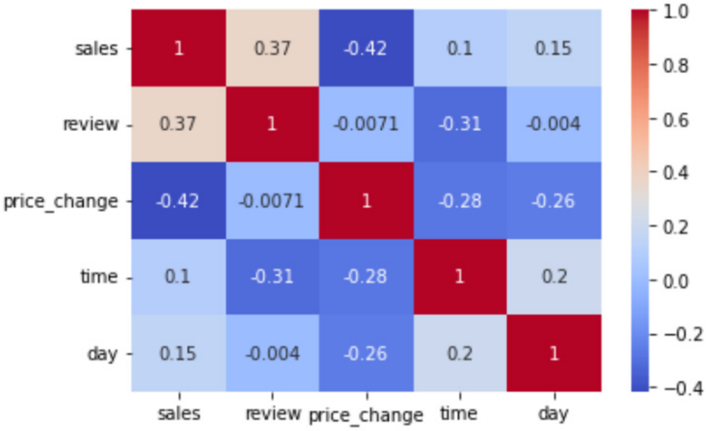
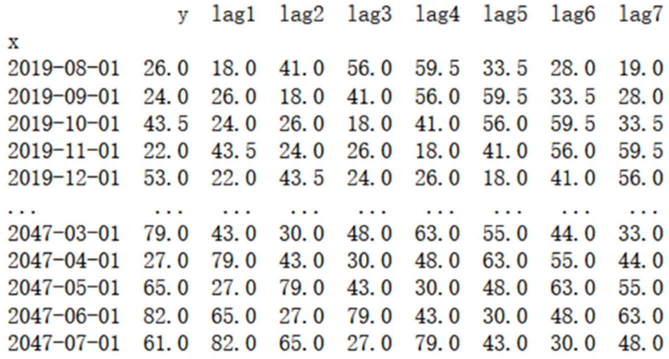


Fig. 2. Correlation coefficient diagram

Through Figure 2, we can draw the following conclusions:

The correlation coefficient of Sales and review is the highest, and the correlation degree between them is the highest, while the relationship between sales and several other factors is negative, indicating that the change of sales has little relationship with other factors. Therefore, we can establish a model between variables based on review and sales to help us predict the changing trend of future data.

After determining the factors for analysis, the next step is to preprocess the data, and the first step is to add the hysteresis features to the data. In time series analysis, data is usually time-related, that is, data at a point in time may be affected by data at a previous point in time. In this case, the hysteresis feature can be used to explain the trend and periodicity of such time series data.



```
               y  lag1  lag2  lag3  lag4  lag5  lag6  lag7
x
2019-08-01  26.0  18.0  41.0  56.0  59.5  33.5  28.0  19.0
2019-09-01  24.0  26.0  18.0  41.0  56.0  59.5  33.5  28.0
2019-10-01  43.5  24.0  26.0  18.0  41.0  56.0  59.5  33.5
2019-11-01  22.0  43.5  24.0  26.0  18.0  41.0  56.0  59.5
2019-12-01  53.0  22.0  43.5  24.0  26.0  18.0  41.0  56.0
...          ...   ...   ...   ...   ...   ...   ...   ...
2047-03-01  79.0  43.0  30.0  48.0  63.0  55.0  44.0  33.0
2047-04-01  27.0  79.0  43.0  30.0  48.0  63.0  55.0  44.0
2047-05-01  65.0  27.0  79.0  43.0  30.0  48.0  63.0  55.0
2047-06-01  82.0  65.0  27.0  79.0  43.0  30.0  48.0  63.0
2047-07-01  61.0  82.0  65.0  27.0  79.0  43.0  30.0  48.0
```

Fig. 3. Hysteresis feature addition

As shown in Figure 3, by adding 6 lag quantities to a specific time, a time plane consisting of 7 time points is formed. This time plane is used to explain the trend and periodicity of time series data on a weekly basis. Adding hysteresis features can enhance the explanatory and predictive ability of the model and improve the performance of the model.

The next step is to smooth the lagging data. The exponential smoothing model is a weighted average based method for forecasting time series data. It is based on data from past points in time and averages them using different weights to predict future values. Its main advantage is the ability to assign weights to future forecasts based on data at different points in time, thus better reflecting fluctuations in demand.

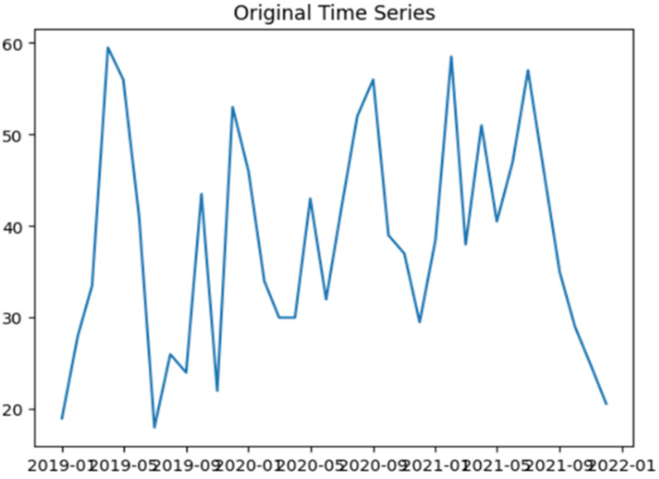


Fig. 4. The smoothness of the original data

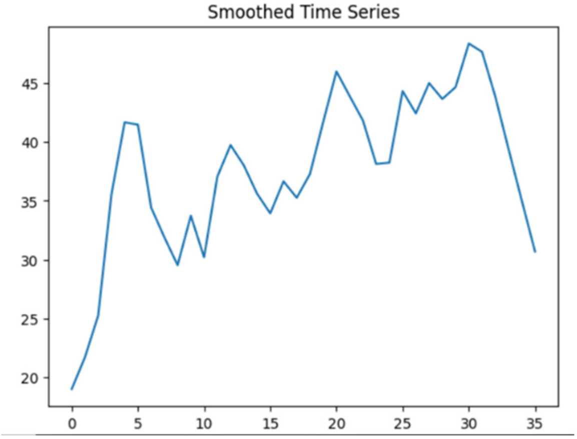


Fig. 5. Smoothing Indicates the degree of data smoothness after processing

As can be seen from the Figure 4 and Figure 5, the smoothed data can better reflect the actual changes in the market. By giving more weight to recent observations, exponential smoothing allows the forecast to quickly catch up with actual market changes, thus better meeting actual demand. At the same time, the data is more reliable.

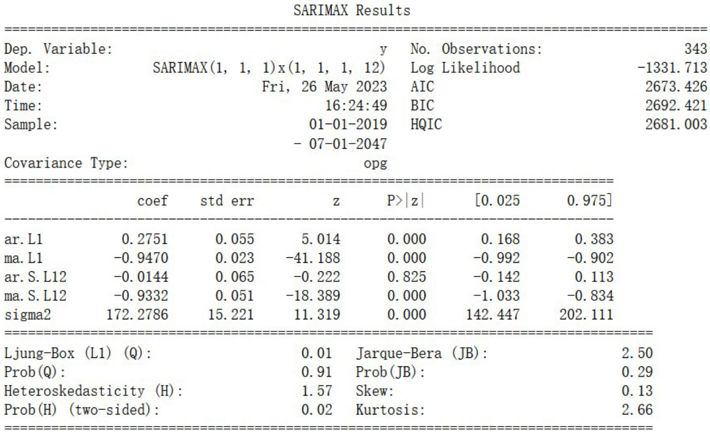
At this point, the data preprocessing part of the experiment is over.

4.2 Establish the ARIMA Model and Obtain the Results

Next, the ARIMA model is established, before which the pdq parameters of the model need to be determined. Among them, p represents lags of the time series data used in the prediction model (lags), also known as AR/Auto-Regressive term. d means that the time series data needs to be differentiated by several orders of difference, which is stable, also called Integrated term. q represents the lags (lags) of the prediction error used in the prediction model, also known as the MA/Moving Average term.

Since the selection of parameters in the ARIMA model greatly affects the accuracy of the final result, the selection of parameters is a very important matter. The traditional method is to continuously adjust the three parameters manually until the final result is optimal. However, a method to automatically select the optimal parameters is designed in this experiment.

Subsequently, Ljung-Box test, Jarque-Bera test and heteroscedasticity test were carried out on the model. The main reason was that AIC value alone could only explain the good fitting effect of the model theoretically, but ARIMA model with AIC value might not exist in practical application, so further testing of the model was needed. It is proved that the model exists under this AIC value.



```
                                   SARIMAX Results
==========================================================================================
Dep. Variable:                                y   No. Observations:                   343
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 12)   Log Likelihood               -1331.713
Date:                          Fri, 26 May 2023   AIC                           2673.426
Time:                                  16:24:49   BIC                           2692.421
Sample:                              01-01-2019   HQIC                          2681.003
                                   - 07-01-2047
Covariance Type:                            opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2751      0.055      5.014      0.000       0.168       0.383
ma.L1         -0.9470      0.023    -41.188      0.000      -0.992      -0.902
ar.S.L12      -0.0144      0.065     -0.222      0.825      -0.142       0.113
ma.S.L12      -0.9332      0.051    -18.389      0.000      -1.033      -0.834
sigma2       172.2786     15.221     11.319      0.000     142.447     202.111
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 2.50
Prob(Q):                              0.91   Prob(JB):                         0.29
Heteroskedasticity (H):               1.57   Skew:                             0.13
Prob(H) (two-sided):                  0.02   Kurtosis:                         2.66
===================================================================================
```

Fig. 6. Ljung-Box inspect, Jarque-Bera inspect

As shown in Figure 6, after the Ljung-Box test and Jarque-Bera test, it is found that the Q value here is 0.01, indicating that the autocorrelation of the residual sequence is very small, and the residual can be considered independent. Here the JB value is 2.50 and the probability is 0.29, indicating that the residual sequence is close to a normal distribution. Here the H value is 1.57 and the probability is 0.02, indicating slight heteroscedasticity in the residual sequence. Here the Skew value is 0.13, indicating that the residual sequence is slightly skewed to the right. Here the Kurtosis value is 2.66, indicating that the residual sequence is slightly flatter than the normal distribution. In summary, the residual series of this time series model performs well, but there are slight heteroscedasticity and slight right skew, and the overall performance is excellent, through detection.

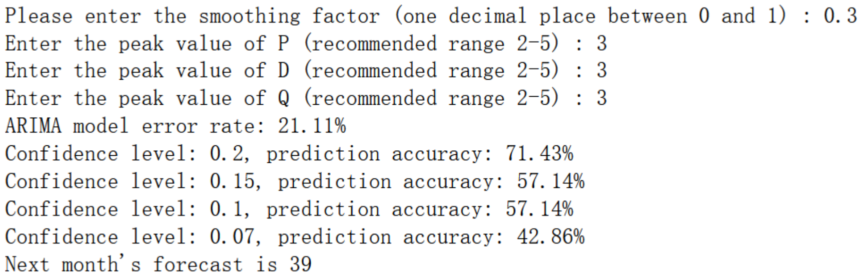
Finally, the ARIMA model is synthesized from the data of re-detection, and the prediction results are obtained, and the accuracy of the prediction results is analyzed. The accuracy calculation formula of this experiment is:

$$ER = 1 - \frac{rmse}{\bar{x}} \tag{5}$$

As shown in equation (5), where ER is the accuracy, $\bar{x}$ is the Mean of the data, and RMSE, the full name of Root Mean Square Error, is a statistic that measures the difference between the predicted value and the true value. The theoretical formula is as follows:

$$rmse = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n}} \tag{6}$$

Finally, as seen in equation (6), the prediction accuracy was calculated at confidence levels 0.2, 0.15, 0.1 and 0.07 respectively. The confidence level refers to the probability measure, which represents the degree of confidence for the estimate in the estimation of the population parameter. Prediction accuracy within the confidence level refers to the proportion of estimates within the confidence level over multiple samples. Within the confidence level, the higher the prediction accuracy, the more accurate our prediction model is, and vice versa. The reason for using confidence level is that it is impossible to achieve completely accurate prediction on the current basis, so it is necessary to set some errors to better evaluate the reliability of the model. In practical application, we may need to choose different confidence levels according to specific situations. In commodity sales, if the seller has a higher risk tolerance, a higher confidence level can be selected to obtain a higher prediction accuracy. Conversely, if the seller has a low tolerance for risk, a lower confidence level can be selected to reduce the risk of prediction errors. By calculating the prediction accuracy at different confidence levels, the appropriate confidence level can be better selected.



```
Please enter the smoothing factor (one decimal place between 0 and 1) : 0.3
Enter the peak value of P (recommended range 2-5) : 3
Enter the peak value of D (recommended range 2-5) : 3
Enter the peak value of Q (recommended range 2-5) : 3
ARIMA model error rate: 21.11%
Confidence level: 0.2, prediction accuracy: 71.43%
Confidence level: 0.15, prediction accuracy: 57.14%
Confidence level: 0.1, prediction accuracy: 57.14%
Confidence level: 0.07, prediction accuracy: 42.86%
Next month’s forecast is 39
```

Fig. 7. ARIMA model prediction results

As shown in Figure 7, the final error rate of ARIMA model is 21.11%, the prediction accuracy of confidence level 0.2 is 71.43%, the prediction accuracy of confidence level 0.15 is 57.14%, the prediction accuracy of confidence level 0.1 is 57.14%, and the prediction accuracy of confidence level 0.07 is 42.86%. The model predicts that next month's sales will be 39.

4.3 Experimental Summary

This experiment is based on ARIMA model optimization of Taobao product sales forecast research experiment. Based on the original arima model, the lag feature is added to the data preprocessing, and the data set is smoothed. AIC value was used as the criterion for parameter selection. Ljung-Box test and Jarque-Bera test are added to prove the existence of the model. Finally, the accuracy of the model was evaluated by RMSE and confidence level.

5 Conclusion

Accurate sales forecast is of great guiding significance for online e-commerce. Therefore, based on the ARIMA model, this paper explores and optimizes the Taobao product sales forecasting model to provide decision support and optimize merchants to formulate marketing strategies and adjust product positioning. The main achievements are as follows:

(1) In terms of model training time, the optimized ARIMA model in this paper has significant advantages over the traditional machine learning model. Due to the optimization of parameter adjustment steps, the efficiency of establishing a model with high precision has been significantly improved.

(2) In terms of model accuracy, due to the introduction of hysteresis features and smoothing processing in data pre-processing, the pre-processed data is more consistent with the data fitting of the ARIMA model, and finally the prediction accuracy of this experiment is significantly improved compared with the traditional machine learning model.

(3) In terms of model accuracy evaluation, by calculating the prediction accuracy under different confidence levels, the appropriate confidence level can be better selected. By understanding the prediction accuracy of the model at different confidence levels, we can better understand the limitations of the ARIMA model and thus better interpret and communicate the predictions.

References

1. Zhang Xin-Chao. Research on commodity sales forecast based on residual optimization of ARIMA model [D]. Harbin University of Commerce,2021.DOI:10.27787/d.cnki.ghrbs.2021.000355.
2. Sun Ming. Supermarket Commodity Sales Forecast based on LightGBM [D]. Dalian University of Technology,2022.DOI:10.26991/d.cnki.gdllu.2021.003534.
3. Jiang Wenwu. Research on product sales forecast based on WaveNet-LSTM network [D]. Guangdong University of Technology,2020.DOI:10.27029/d.cnki.ggdgu.2019.000614.
4. Zhou Yu, DUAN Yongrui. Retail Sales Forecast Based on Clustering and Machine Learning [J].Computer System Application,2021,30(11):188-194.DOI:10.15888/j.cnki.csa.008147.
5. Bai Yun. Research and Implementation of Commodity Sales Forecasting Model [D]. Xidian University,2021.DOI:10.27389/d.cnki.gxadu.2020.001827.
6. Li Jie, WANG Yuxia, ZHAO Xudong. Forecasting Method of commodity sales of e-commerce enterprises [J]. Statistics and Decision,2018,34(22):176-179.DOI:10.13546/j.cnki.tjyjc.2018.22.042.
7. Pu Jiapeng. Application of Machine Learning in Commodity Sales Forecasting [J]. Electronic Production,2018(22):81-82+70.DOI:10.16589/j.cnki.cn11-3571/tn.2018.22.035.
8. Wang Yuxia. Research on commodity sales forecasting method of e-commerce enterprises based on XGBoost algorithm [D]. Hebei University of Technology,2019.
9. Huo Jiazhen, XU Jun, Chen Mingzhou. Multi-step forecasting of retail sales based on EEMD-Holt-Winters-GBDT model [J/OL]. Industrial Engineering and Management:1-14[2023-09-10]http://kns.cnki.net/kcms/detail/31.1738.T.20230418.1652.006.html.
10. Chen Qiang. Research on sales forecasting model algorithm based on ConvLSTM network in new retail formats [D]. Guangdong University of Technology,2021.DOI:10.27029/d.cnki.ggdgu.2020.000877.
11. Hu W,Zhang X . Commodity sales forecast based on ARIMA model residual optimization[C]//Sichuan University.Proceedings of the 5th International Conference on Communication, Image and Signal Processing (CCISP 2020).IEEE Express Conference Publishing,2020:5.DOI:10.26914/c.cnkihy.2020.032038.