Research and application of commodity demand forecasting algorithm based on improved ARIMA-LSTM

Jingqi Zhang*

Department of Aviation Engineering school, Air Force Engineering University, Xi'an, China, 710038

*Corresponding author: 18144757088@163.com

Abstract. Through thorough investigation and research, it has been determined that commodities possess a certain degree of timeliness, with long production cycles and short storage times. Accurately predicting customer demand can effectively reduce costs. This paper proposes an improved model based on ARIMA-LSTM, incorporating data cross-correlation analysis for effective data classification. By leveraging the unique characteristics of both ARIMA and LSTM prediction time series models, the data is divided into stationary and non-stationary models for accurate forecasting of future demand for goods. After rigorous testing on the test set, our proposed model achieved impressive results with an accuracy rate of 0.855, precision rate of 0.834, and recall rate of 0.854 respectively. Therefore, the model proposed in this paper has a good performance in forecasting the sales volume of general commodities with time series.

Keywords: ARIMA, LSTM, ADF inspection, Demand forecasting.

1. Introduction

With the development of economic globalization, the commodity retail industry is facing great competition pressure. As the most sensitive market signal, the sales volume of goods has always been the focus of merchants [1]. At the same time, it is also a category of big data and artificial intelligence application to obtain future sales forecast by analyzing sales volume. Especially for fresh vegetables with short storage time, it is necessary to know the purchase quantity of the previous day or even a week in advance to reduce the cost.

The asymmetry of commodity supply and demand information has become a pain point in its industry. In many cases, merchants only rely on their own experience and sales experience of the previous day to estimate today's demand and pricing, which leads to the situation that the supply and demand of goods do not correspond to. Through literature review [1], it is evident that the utilization of bp neural network in commodity sales analysis exhibits certain limitations. For periodic sequence data like commodity sales, this algorithm tends to generate overfitting and local solutions easily, while being susceptible to abnormal sales volume. Additionally, it incurs significant computational time and space complexity [2]. With the development of science and technology information, big data has been widely used in the field of commodity demand forecasting, among which ARIMA and LSTM are two commonly used algorithm models [3]. ARIMA is a classical model based on time series. Through the establishment of autoregressive model and moving average model to predict the future price trend, it is mainly applied to the prediction of stationary time series [4]. LSTM is a deep learning model suitable for sequence data, and its internal memory unit can effectively capture long-term dependencies in time series. Therefore, forecasting commodity demand has become an urgent need in commodity economy.

This paper first preprocesses the data, combines the advantages of the two, introduces the classification based on ADF test. [4] And adjusts the selectivity of the model, in order to better deal with the time series data of commodity sales, help supermarkets and even general retail workers to make the right purchase decision, which is particularly important for today's increasingly fierce industry.

2. Correlation theory

2.1. Concept of ARIMA

The $ARIMA$ model includes three parts: autoregressive AR, difference I, and moving average MA. Firstly, the AR model establishes the relationship between the current value and the past value through the autoregressive term of historical data, and the model must be stationary and only suitable for predicting the phenomenon related to its own previous period. The difference part is then used to smooth the data [4]. Finally, the MA model models random errors by a moving average term. ARIMA model Features:

Autoregressive model, p-order autoregressive process formula:

$$y_t = \mu + \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t \tag{1}$$

Where y_t is the current value, μ is the constant term, P is the order, α is the autocorrelation coefficient, ϵ is the error [5]. The average moving model, which focuses on error accumulation in autoregressive models, is a Q-order autoregressive process model.

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{2}$$

Where ϵ is the error function and θ is the autocorrelation coefficient [5]. The purpose of difference model is to transform non-stationary series into stationary series, reduce the influence of seasonality on time series, and improve the prediction accuracy [5]:

$$\Delta y_i = y_i - y_{i-1} \tag{3}$$

2.2. Concept of LSTM

$LSTM$ is a special RNN structure that is used to deal with the modeling and prediction of sequential data and time series data. Compared with the traditional recurrent neural network, $LSTM$ introduces a memory unit and a gating mechanism, which can effectively solve the long-term dependence problem. The cell state of the $LSTM$ is the key, namely the horizontal line $c^{(t-1)} \rightarrow c^{(t)}$ that runs through the top of the chart. The cell state is somewhat like a conveyor belt, which moves down the flow relation with only some slight linear relation. LSTMs are able to delete or add information to the cell state, a structure called a gate is carefully regulated, and the way information passes through the gate is determined by the "gate". They consist of σ neural network layers and a point-wise multiplication operation that outputs a number between 0 and 1 describing how much each component should pass, with a value of 0 meaning "nothing passes" and a value of 1 meaning "everything passes". The $LSTM$ has three such gates for protecting and controlling the cell state. Figure 1 is the appearance of the three gates in the unit cell [6].

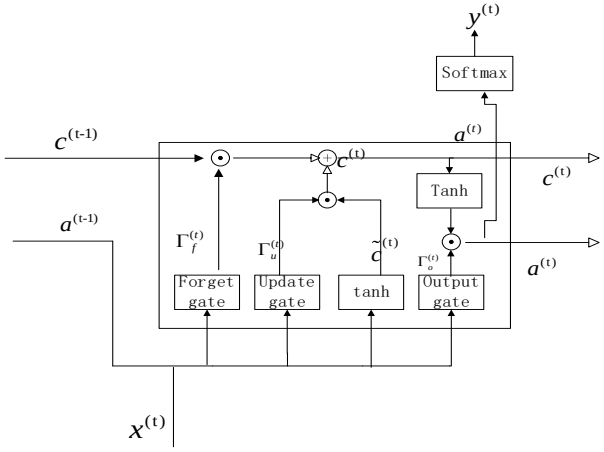


Figure 1. LSTM cell unit

The first step of the $LSTM$ is to go through the Forget Gate, which determines what information from the previous memory state needs to be forgotten; The next step is to go through the Input Gate, which determines what information in the current input data needs to be updated into the memory cell. Finally, it goes through the Output Gate, which determines what information needs to be output to the next layer or as a final prediction based on the current input and memory state. Combining the two output functions $a(t), c(t)$, we can obtain the equation of the $LSTM$ model as follows [7].

$$\begin{cases} \Gamma_f = \sigma(w_f\left[a^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_f) \\ \Gamma_i = \sigma(w_i[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_i) \\ \tilde{c}^{\langle t\rangle} = \tanh(w_c[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_c) \\ \Gamma_o = \sigma(w_o[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_o) \\ c^{\langle t\rangle} = \Gamma_u \cdot \tilde{c}^{\langle t\rangle} + \Gamma_f \cdot c^{\langle t-1\rangle} \\ a^{\langle t\rangle} = \Gamma_o \cdot \tanh(c^{\langle t\rangle}) \end{cases} \tag{4}$$

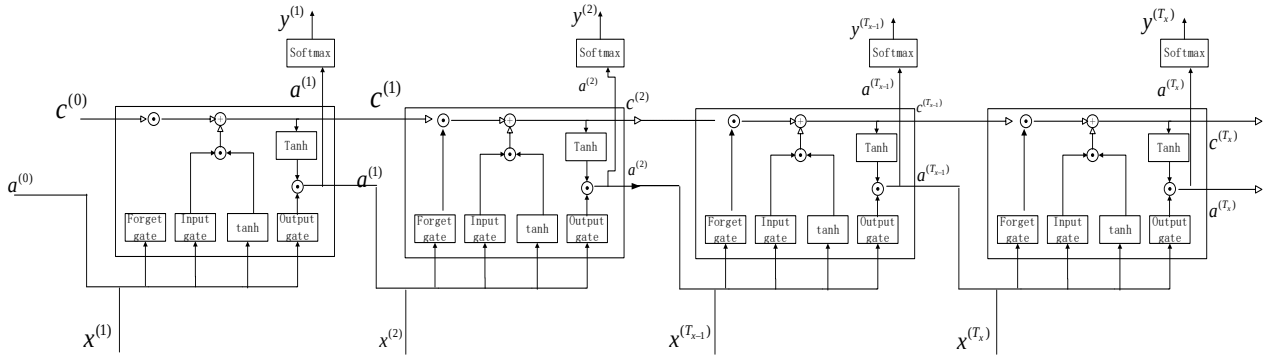


Figure 2. The Forward propagation model for LSTM

$LSTM$ is a special recurrent neural network for sequence data modeling. By introducing memory unit and gating mechanism, it solves the long-term dependence problem in traditional recurrent neural network, has strong memory and learning ability, Figure.2 is a reification of the model and has excellent effect on the sales prediction of this paper [7].

2.3. Hybrid LSTM and ARIMA model

The data is classified and the commodity data is studied. It is found that the sales volume of goods is affected by seasons and periodicity over time. When the goods are affected by seasonality, due to the different seasonal temperature demand, the fluctuation of sales volume has a great influence, and the $LSTM$ model is suitable for forecasting at this time. When the commodity is affected by periodicity and the demand is almost unchanged in a week, $ARIMA$ model is suitable for forecasting [9]. $ARIMA$ can capture some linear trends and seasonality, while LSTM can handle more complex patterns. The $ARIMA-LSTM$ hybrid model combines ARIMA and LSTM, which can provide more comprehensive series modeling capabilities and is suitable for time series data with both linear trends and complex nonlinear relationships. The ADF test is used to analyze the commodity sales data. Through the analysis, it can be obtained whether the commodity sales is not stationary time series, if it is stationary time series, it can be predicted by ARIMA model. If it is a non-stationary time series, it can be predicted by LSTM model. The specific model structure of this paper is shown in Figure 3, which combines ARIMA and LSTM models and ADF test to optimize the forecasting process of time series. Since the ARIMA model has lower space and time complexity than the LSTM model, this model can further improve the operation speed and accuracy.

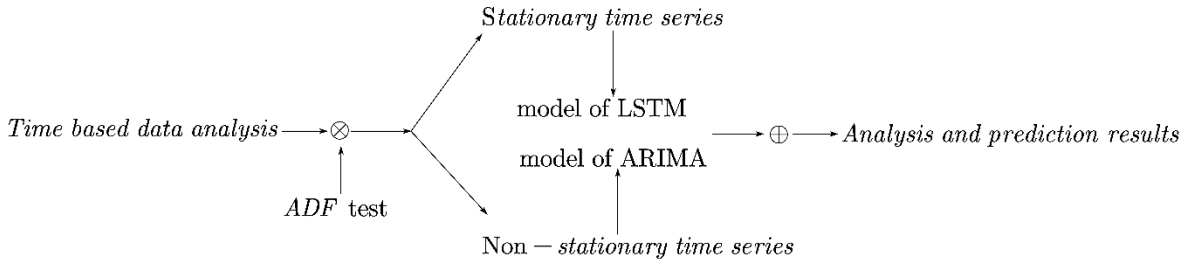


Figure 3. Model Structure of this Paper

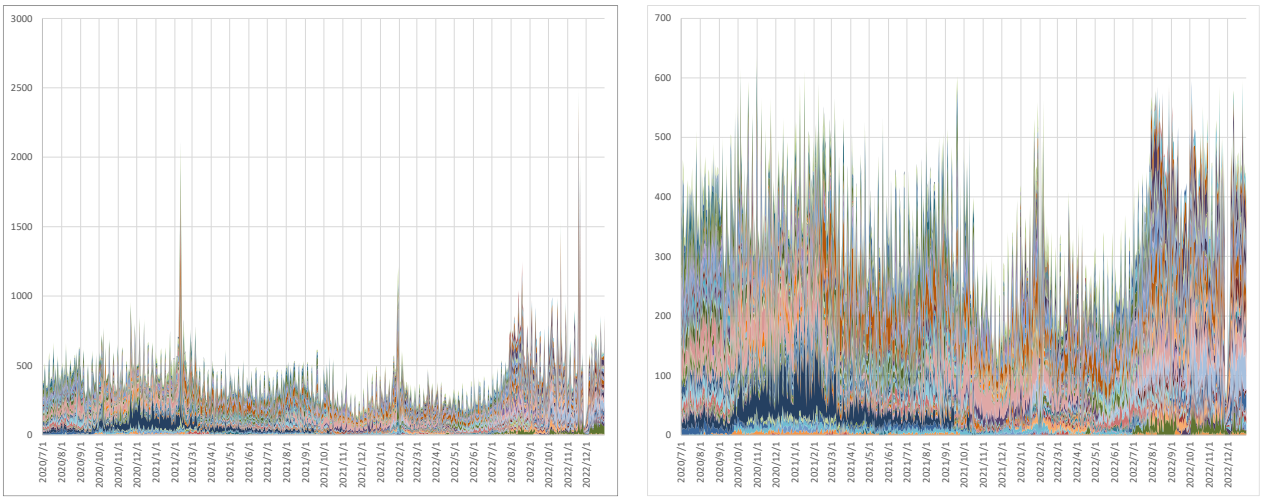
Through searching the literature [9], it can be seen that ARIMA algorithm is suitable for predicting stationary time series, that is, when the data fluctuation is relatively small, the prediction accuracy of ARIMA algorithm is better than that of LSTM. When the original data fluctuates violently, the LSTM algorithm can predict more similar results through its memory nerve (forget gate, input gate, output gate). Now, the data is classified and predicted, and the advantages of different algorithms are combined to improve the prediction accuracy compared with separate prediction.

3. Results

The data source is the sales information of a supermarket in the past three years. The first data set is the sales volume of each vegetable, and the second data set is the unit price of vegetable sales. Data for a total of 251 vegetables 257 weeks of sales data. In order to further facilitate the prediction of ARIMA model and LSTM model, the data is preprocessed and classified, and the characteristics of vegetables sold in supermarkets in the past three years are analyzed. It is preliminarily found that the supermarket selling data has certain seasonality, that is, it meets certain timing. Compared with the ordinary nonlinear prediction model and ARIMA model, the LSTM model has better learning ability for the purchase time series of customers, and can better capture the long-term dependence of customer demand for goods. LSTM model has the ability to process multiple variables at the same time. Compared with RNN model, LSTM model solves the problem of gradient disappearance and gradient explosion, which makes the model function converge better. The design of this model uses the fusion of ARIMA model and LSTM model, and uses a variety of optimization methods, which makes the model have higher generalization ability and robustness, and can be used to predict the sale of goods or the demand for goods [9].

3.1. Data preprocessing analysis

Commodity sales volume is a kind of time series data, and there is no strong linear correlation between variables. First, the data is preprocessed. Preprocessing helps to further explore the relationship between data and enhance the learning ability of the model on data [9]. Figure.4 is the comparison diagram before and after data preprocessing, after preprocessing, the model is used to check the calculation on the learning set and the test set, which can further enhance the persuasive power [9]. Based on the data, we have to understand that we have to understand that there is a certain amount of chance in the sales rate. In addition, we also need to analyze the impact of holidays on the sales rate of goods. First of all, we analyze the quantity of goods sold under general conditions. We can clearly see that after data preprocessing, the sharp part of the curve in Figure.4 is significantly reduced, and the prediction accuracy can be significantly improved after eliminating the outliers and introducing them into the model.



(1)Before pretreatment (2)After pretreatment

Figure. 4. Comparison before and after outlier processing

Now, the daily sales volume of each commodity for nearly three years is accumulated to obtain the daily sales volume of each commodity. $SPSS$ software was used to preprocess the data, and the outliers were screened out by $3-\sigma$ criterion, and the average value was used instead. The$3-\sigma$ criterion can effectively identify the outliers in the data. In general, if an observation is outside the interval $(\mu-3\sigma,\mu+3\sigma)$, we can identify it as a possible outlier. The preprocessed data were brought into the $ARIMA-LSTM$ model for prediction. The steps are as follows:

1) Determine whether each observation in the dataset is within the upper and lower bounds.

2) For the observation value beyond the interval, it is regarded as an outlier, and the outlier value is changed to the average sales volume of the item, which is helpful to further solve the model accurately.

3) The data is brought into the model according to the 5:1 relationship between the training set and the test set.

3.2. Analysis of results

By setting stationary time series and non-stationary time series: let T=7, $\sigma<1$ is a stationary time series, $\sigma>1$ is a non-stationary time series. This condition was brought into the data for cross-correlation analysis. Using commodity 1 as data analysis, the data set is obtained in Table.1.

Table 1. Experimental data distribution

Dataset (time series)	stationary time series	nonstationary time series	Sum up
Training set	100	106	206
Test set	19	22	41
All set	119	128	257

3.3. Evaluation index

(1) Accuracy(A), the probability that the prediction is correct across all samples [10]:

$$A=\frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

(2) Precision (P), the probability that the sample is correct [10]:

$$P=\frac{TP}{TP+FP} \tag{6}$$

(3) Recall (R), the actual probability of being correct in a correct sample [10]:

$$R = \frac{TP}{TP + FN} \tag{7}$$

(4) Score rate (F1), comparing the comprehensive ability of the model [10]:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{8}$$

TP: predicted true, actually true FN: predicted false, actual false FP: predicted false, actually true TN: Predicted true, actual false; The predicted object is the sales volume of the product. When the predicted sales volume is greater than the real sales volume and less than 10%, it is considered as true, and the rest is false. Now, the sales volume of product Choy sum which in Figure.5 in the store in the past three years is used as the learning object to learn the model [10].

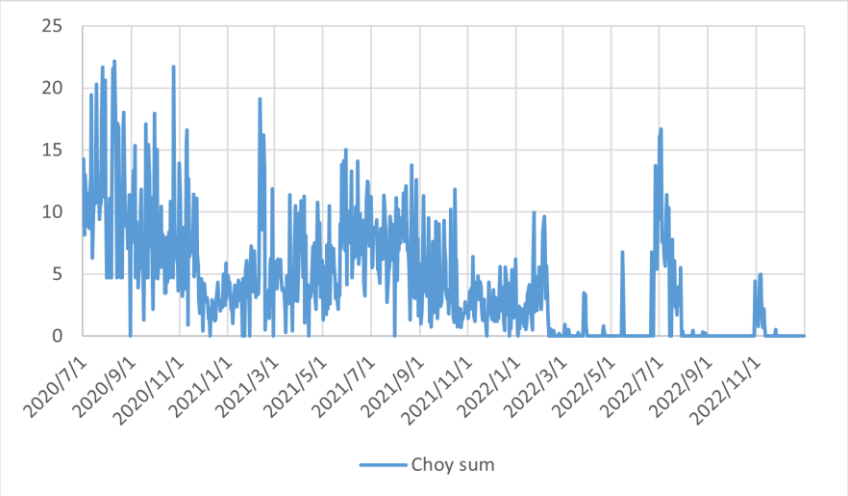


Figure 5. Daily demand change of commodity A

Table 2. Comparison of model results

Model	Acc	P	R	F1
ARIMA	0.756	0.743	0.621	0.677
LSTM	0.801	0.741	0.752	0.746
Algorithm in this paper	0.855	0.834	0.854	0.844

The proposed model demonstrates superior precision, accuracy, and recall rate in predicting the sales volume of goods compared to ARIMA or LSTM models, as shown in Table.2. It is preliminarily inferred that leveraging three years of historical sales data enables more effective and reliable prediction of next week's sales volume. But, the model of the paper is a little difficult for computation, LSTM time series prediction model and ADF test model will increase the time and space complexity.

In short, integrating the ARIMA and LSTM models while incorporating data correlation classification surpasses standalone prediction models when accompanied by appropriate data preprocessing. Based on this, it is believed that applying the model in this paper to the prediction analysis of commodity sales volume can further calculate the demand of commodities more accurately, help commodity sellers to control the budget, and can further reduce the price and promote the circulation of commodities in the process of commodity circulation.

4. Conclusions

In this paper, an algorithm for predicting the sales volume of commodities is given. By using the feature that the sales volume of commodities is a time series, the data preprocessing step is added to obtain the sales data without outlier data. After that, the data is classified by using the cross-correlation coefficient algorithm, and the sales volume of goods is divided into two types: stationary time series and non-stationary time series. The ARIMA model and LSTM model are used to predict

the sales volume respectively. It is believed that the model can meet the daily sales unit to estimate the sales volume of goods. However, there are still some defects, such as the inability to predict the sales volume demand during holidays and sudden times, at the same time, the correlation between commodities is not considered. Nevertheless, the model in this paper has a good score in verification and is still convincing. At the same time, it can reduce the imaginary high commodity economy, reduce the production of inflation, and promote social stability.

References

[1] Zhou Shang. Research on Supermarket Commodity Pricing Based on Data Mining [D]. Fuzhou University, 2013.

[2] Zhang Jianhua. On the Formation of Market Commodity Demand [J]. Farm Staff, 2020, (16): 281.

[3] Zhao Juanhe. Research on Short-Term Demand Forecasting Method of E-commerce Goods Based on Multi-layer Hybrid Deep Neural Network [D]. Chang'an University, 2019.

[4] Lu Yitong, Liu Zhiquan, Mo Qiao pin et al. Feasibility analysis of ARIMA model in predicting clinical demand of apheresis platelets in primary blood stations [J]. Chinese Journal of Applied Medicine,2023, 18(23): 144-148.

[5] Yan Xun, Tie Chengcheng, Yan Wei et al. Global Temperature Prediction Analysis Based on ARIMA Model and CNN-LSTM Combination Model [J]. Science, Technology and Innovation, 2024: 19-22.

[6] Shang Xueyi, Chen Yong, Chen Jie et al. Noise reduction method of mine microseismic signal Based on $Adaboost\_LSTM$ prediction and its application [J/OL]. Journal of China Coal Society, 2024: 1-9.

[7] Wang Rui, Zhou Zuojian, Li Can et al. Research on atrial fibrillation recognition Algorithm based on improved CNN and LSTM [J]. Computer Times, 2023, (08): 69-73.

[8] Wang Xu, Liu Bo, Chen Zhengchao et al. Winter wheat yield estimation at county level based on multi-source data and LSTM model [J/OL]. Research on Agricultural Modernization, 2024: 1-13.

[9] Wang Rui, Li Ruiyi, Cao Peigen et al. Prediction analysis of infectious diseases based on ARIMA-LSTM hybrid model [J]. Modern Information Technology, 2024, 8(01): 116-120.

[10] Wang Shiming, Zhang Shaotong, Lou Jiayi. Research on Ultra-short-term Wind Speed Prediction based on CNN-LSTM-ARIMA [J/OL]. Advances in New Energy, 2024: 1-9.