Application of Long Short-Term Memory for prediction of monthly tobacco sales volume

Ruochen Feng [1, #], Ziyi Wu [2, *, #], Jiayu Gong [2, #]

[1] School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan, China, 430074

[2] School Of Economics Hust, Huazhong University of Science and Technology, Wuhan, China, 430074

* Corresponding Author Email: u202216427@hust.edu.cn

[#]These authors contributed equally.

Abstract. The prediction of tobacco sales is crucial, for it informs businesses about future demand trends, facilitating production optimization and minimizing stockouts or overstocking. The tobacco market's complexity, with numerous influencing factors and rapid fluctuations, poses significant challenges for accurate sales forecasting. Responding to this challenge, this paper introduces the application of the LSTM (Long Short-Term Memory) network approach to develop a predictive model to accurately forecast future sales volumes. This paper initiated by conducting data processing on open-source tobacco sales datasets, utilizing box plots to identify and eliminate outliers, and subsequently filling in missing values with Hermite interpolation, thereby completing the preprocessing stage. Following this, based on the distinct sales data of four different brands, this paper independently established an ARIMA (Autoregressive Integrated Moving Average) time-series prediction model and a LSTM (Long Short-Term Memory) network prediction model seperately. Subsequently, the performance of these two prediction models was examined. Upon comparison, the ARIMA model exhibited suboptimal performance in predicting nonlinear and complex time series, thereby leading to the conclusion that the LSTM model outperforms ARIMA in terms of predictive capability. Compared with the ARIMA time series prediction model, the LSTM model has better performance in terms of robustness to outliers and long-term prediction ability model fitting, which can provide analysis and support for decision-making in the tobacco market.

Keywords: Tobacco Sales Volume, ARIMA Time-Series, LSTM Memory, Box Plot, Hermite Interpolation.

1. Introduction

Amid socioeconomic development and rising living standards, China's tobacco retail market exhibits steady growth, with consumers now keenly focused on cigarette safety and prices [1]. The complex and dynamic nature of the industry, shaped by emerging tobacco products, technological advancements, and stringent government policies, underscores the importance of accurate sales forecasting [2] .Leveraging deep learning and ensemble learning models for time-series prediction is vital to anticipate future sales, analyze consumption patterns, and grasp market trends, ultimately enhancing both economic and social benefits within the tobacco sector [3].

Zhao Chenzhi et al. [4] introduced a prediction model for cigarette sales, leveraging the ARIMA time series analysis. Initially, they collected weekly sales data of cigarettes from Guizhou Provincial Tobacco Company spanning from May 2019 to March 2021 and utilized SPSS software to conduct a thorough analysis. Based on this analysis, they established an ARIMA model to forecast future weekly cigarette sales, achieving an R-squared value of 0.898, indicating a good level of fit. Meanwhile, Deng Chao et al. [5] proposed a sales prediction model that incorporates LSTM technology. To start with, LSTM was employed to extract temporal features from the cigarette sales data, and subsequently, expert-extracted features were integrated into the model to enhance the prediction accuracy of product sales. The model was validated using historical sales data for 189 different cigarette types from Shandong Qingdao Tobacco Co., Ltd, demonstrating a prediction accuracy of 95.67% for these cigarette types.

In this paper, for the data, the aforementioned LSTM and ARIMA were respectively used to construct prediction models. This paper conducted a comparative analysis of the predictive performance between these two models. Among them, the LSTM model performed relatively better. It had a better fitting effect for complex patterns such as nonlinear trends and periodic changes, could automatically extract time series features, performed better when facing outliers, and had a better effect for long-term predictions. This paper conclusively demonstrates that, when applied to our datasets, the LSTM prediction model exhibits a significantly enhanced predictive power compared to other methods.

The marginal contribution of this paper is that this paper employs both the LSTM network approach and the ARIMA time series forecasting model, which have been applied separately before, to solve the problem of tobacco sales forecasting. And a detailed comparison of the two methods is made in this article, which demonstrates that the LSTM model is superior in predicting nonlinear and complex time series, providing more accurate analytical support for tobacco market forecasting.

2. Exploration of Tobacco Sales Data

2.1. Data Sources

The data comes from: http://www.nmmcm.org.cn/match_detail/33.

The data selected in this paper are the monthly sales volume (boxes) and monthly sales amount (yuan) of tobacco for Brand One from January 2011 to September 2020; and the monthly sales volume (boxes) and monthly sales amount (yuan) of tobacco for Brand Two from January 2011 to June 2019. Due to space limitations, only some are presented here, as shown in Table.1. And Table.2. Below:

Table 1. Presentation of Data Sources,Data of band one.

Month	Sample Code	Name	Sales Volume (Boxes)	Amount (Yuan)
201101	43010103	A1	265.836	4785048
201102	43010103	A1	72.268	1300824
202008	43010103	A1	34.788	663755.04
202009	43010103	A1	59.816	1141289.28

Table 1. Presentation of Data Sources,Data of brand two.

Month	Sample Code	Name	Sales Volume (Boxes)	Amount (Yuan)
201101	43010104	A2	235.256	2646630
201102	43010104	A2	0.188	2115
201905	43010104	A2	51.436	681527
201906	43010104	A2	41.828	554221

2.2. Data Preprocessing

2.2.1. Processing of Outliers

Before establishing the time series prediction model, it is necessary to detect the out layers in the datasets and eliminate them, which is to ensure the accuracy of the model. Boxplot is a type of chart often used in explanatory data analysis. It’s named for its books-like shape and is widely used in highlighting outlier.

Sales of four brands is shown in Figure 1. Among them, the data represented by dots are abnormal data (excessive data), and these data are eliminated.

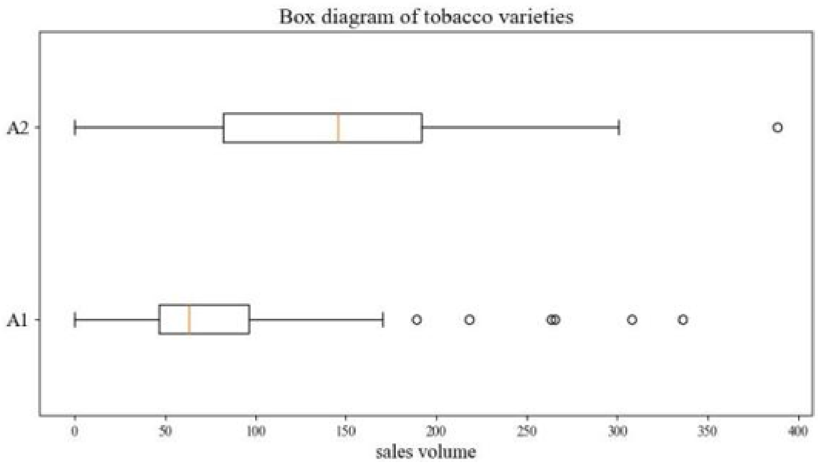

Figure 1. Boxplot.

2.2.2. Missing Value Imputation

Given the continuous nature of the time series and the correlation between consecutive data points, missing values were not directly omitted. Instead, the Hermite interpolation polynomial method was employed to impute missing values, ensuring data continuity and integrity [6].

Hermite interpolation is also called interpolation with derivative values. It requires not only to take known function values at a given node, but also to take known derivative values of the interpolating function at a given node. The calculation formula for n nodes is as follows:

$$h_i(x)=(1-\frac{w_n''(x_i)}{w_n'(x_i)}(x-x_i))l_i^2(x)=(1+2\sum_{j\neq i}\frac{x-x_i}{x_j-x_i})l_i^2(x) \tag{1}$$

$$H_i(x)=l_i^2(x)(x-x_i), i=0,1,...,n \tag{2}$$

$$l_i(x)=\frac{(x-x_0)...(x-x_{i-1})(x-x_{i+1}...(x-x_n))}{(x_i-x_0)...(x-x_{i-1})(x-x_{i+1})...(x_i-x_n)}, i=0,1,2,...,n \tag{3}$$

Use Python to write a program to read the data in the excel table. After Hermite interpolation, the sales of the two brands is shown in Figure 2 and Figure 3 below.

3. Establishment of Tobacco Sales Forecasting Models

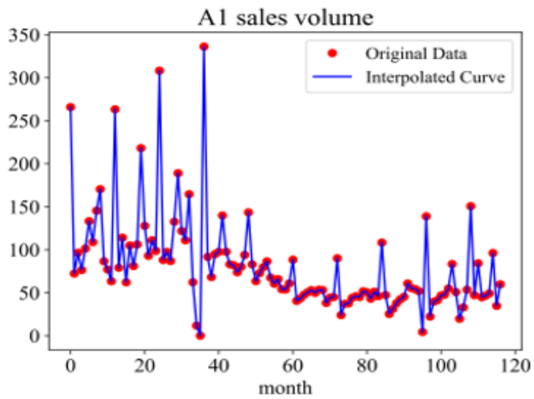

Figure 2. Hermite Interpolation of A1.

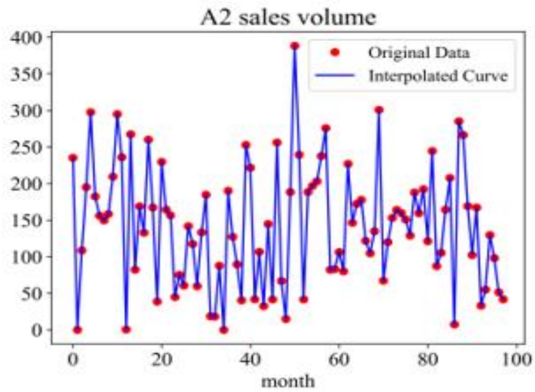

Figure 3. Hermite Interpolation of A2.

3.1. Time Series ARIMA Model

The ARIMA (Autoregressive Integrated Moving Average) model was chosen due to its capability in transforming non-stationary time series into stationary ones, thereby facilitating accurate forecasting [7].

3.1.1. Determination of Differencing Order (d)

The sales and turnover of Brand 1 from January 2011 to September 2020 and the sales and turnover of Brand 2 from January 2011 to June 2019 are shown in Figure 4 and Figure 5 below, respectively.

Figure 4. The sales of Brand 1.

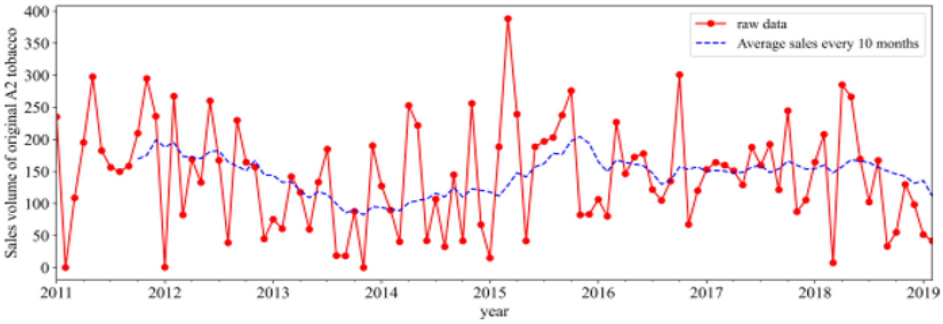

Figure 5. The sales Brand 2.

Obviously, the initial series is unstable, which does not meet the requirement of stationarity of ARIMA model. The sequence needs to be differenced until it is transformed into a stationary sequence. ARIMA (p, d, Q) model can be used if the model meets the model requirements after d times of difference. The third-order difference operation is performed on the sequence and the best difference resultis selected, the differencing coefficient d is chosen as 1, as shown in Figure 6. For brand 2, do the same, as shown in Figure 7. The differencing coefficient d is also 1:

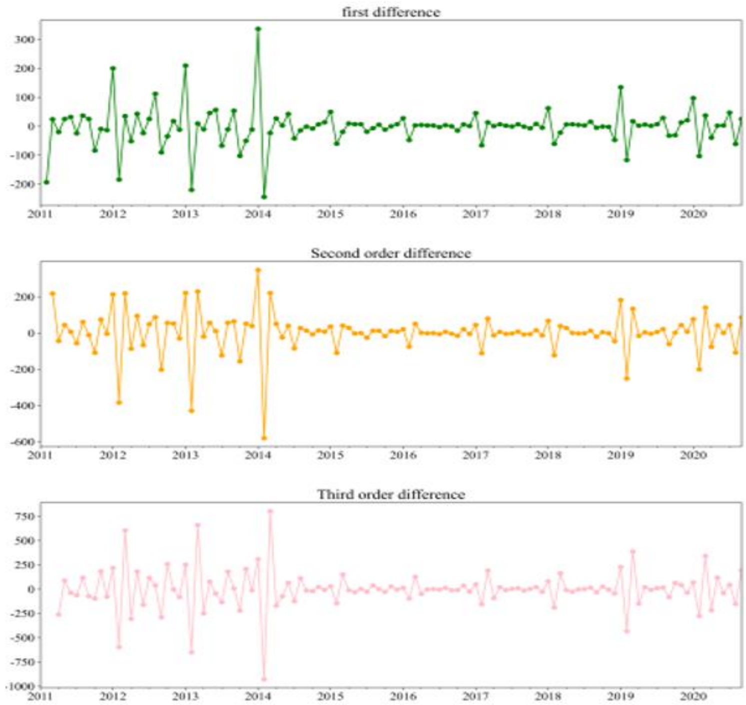

Figure 6. Illustrative Diagrams of A1.

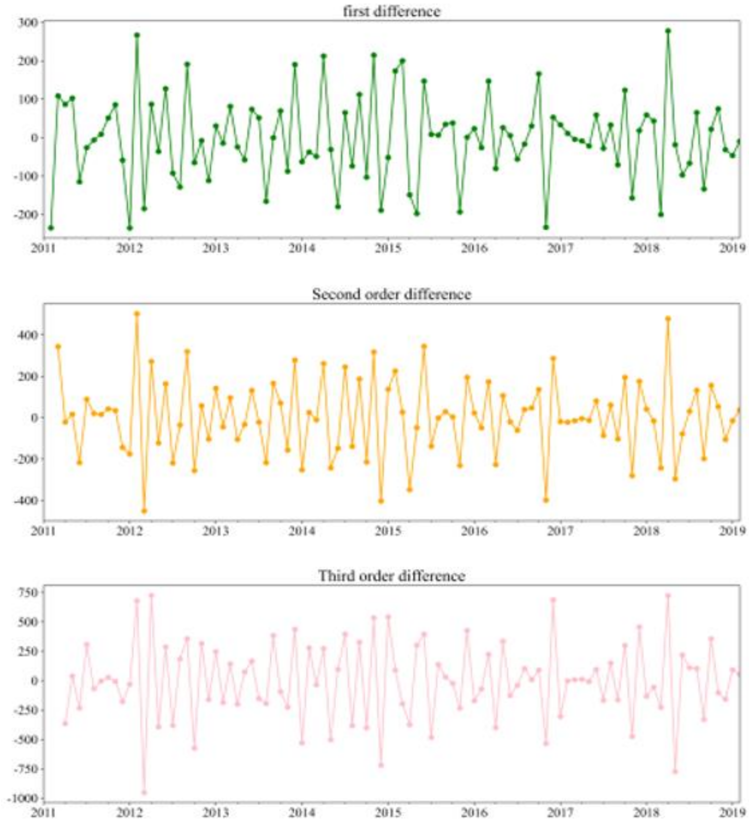

Figure 7. Illustrative Diagrams of A2.

3.1.2. Determination of Autoregressive Order (p) and Moving Average Order (q)

After first-order differencing, the sales sequences of Brand one and Brand two exhibit a relatively stationary state. Therefore, seasonal ARIMA **(**p, 1, q) (p, 1, q) models can be established separately for them. By plotting the ACF and PACF graphs, the autoregressive coefficient p and the moving average order q can be further determined when the ACF or PACF graph approaches 0 or equals 0.

Employing the aforementioned method, the ARIMA model parameters for A1 is chosen to be (1,1,1). The same to A2 sales data, this process identified the optimal model parameters as (3,1,1). The PACF and ACF graphs for Brand one and two are shown in Figure 8. And Figure 9.

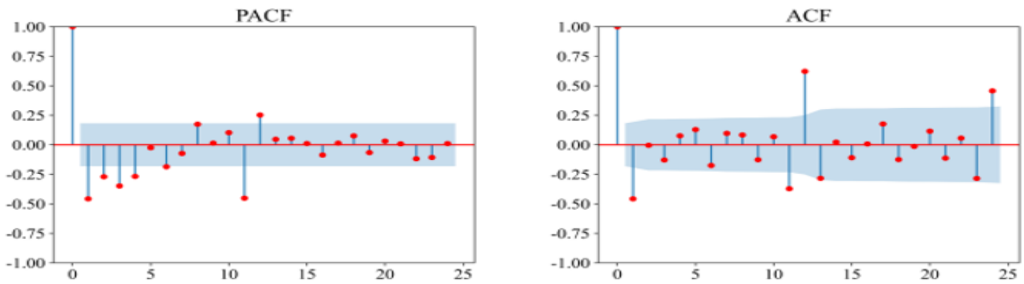

Figure 8. PACF and ACF graphs for A1.

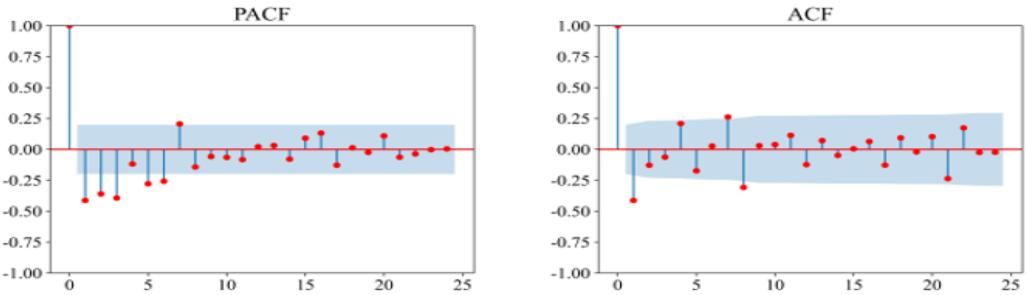

Figure 9. PACF and ACF graphs for A2.

3.1.3. ARIMA Model Prediction Results

After configuring the model parameters, the ARIMA model was applied to predict the monthly sales of A1 from October 2020 to July 2022, and the monthly sales of A2 from July 2019 to May 2020. A partial demonstration of the A1 and A2 data is shown below. The predictive outcomes of are presented in Figures 10 and Table 3.

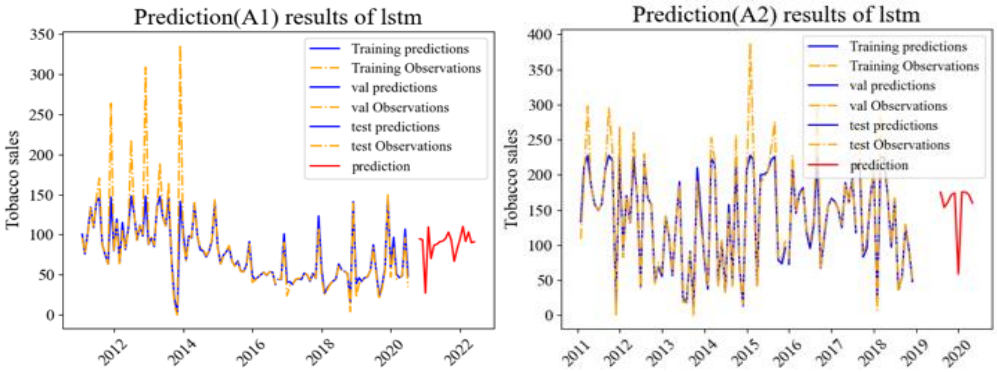

Figure 10. Predictive Outcomes of A1 and A2.

Table 3. Predictive Outcomes of A1

Month	Sales Volume (Boxes)
2020-10	56.0297
2020-11	57.3265
....	....
2022-07	57.4357
2022-07	57.4357

3.1.4. Validation of ARIMA Model

To ensure the model's robustness, residual analysis was conducted. The residual plots indicated that the residuals approximately followed a normal distribution, thus validating the model's stationarity and predictive accuracy. The residual test results for the ARIMA (1, 1, and 1) and ARIMA (3, 1, and 1) time series prediction model are shown in Figure 11.and Figure 12.

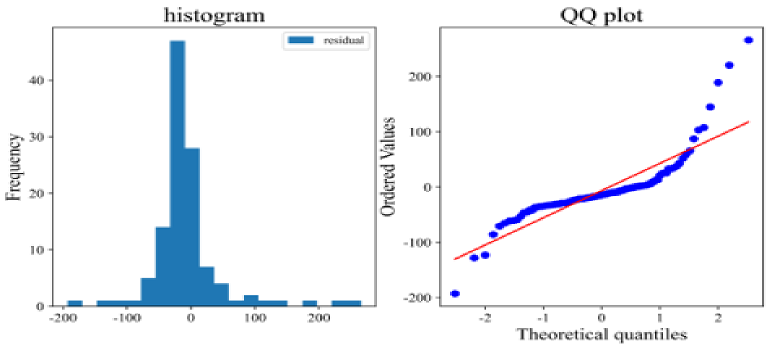

Figure 11. Model Validation for A1.

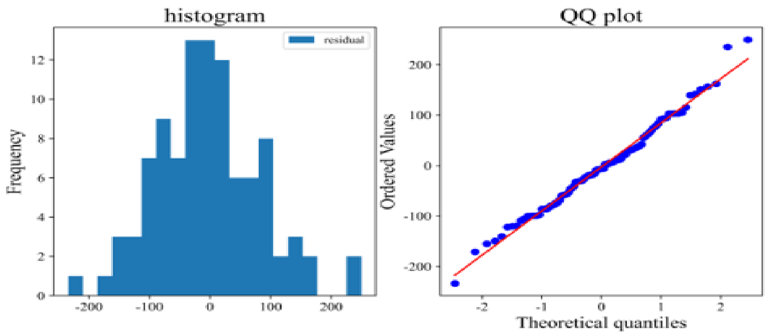

Figure 12. Model Validation for A2.

As depicted in the frequency distribution histograms and QQ plots, the residuals of the ARIMA(1, 1, 1) and ARIMA(3, 1, 1) models approximate a normal distribution with a constant mean, with their scatter points closely aligning on a straight line, indicating a strong fit to normality and thus suggesting high predictive accuracy for both models.

3.2. Establishment of LSTM Time Series Prediction Model

LSTM, a special type of Recurrent Neural Network (RNN), is renowned for its capability in capturing long-term dependencies in sequential data [8]. Its memory units, governed by forget, input, and output gates, enable effective learning and prediction of time series patterns. Figure 13 below shows how it works.

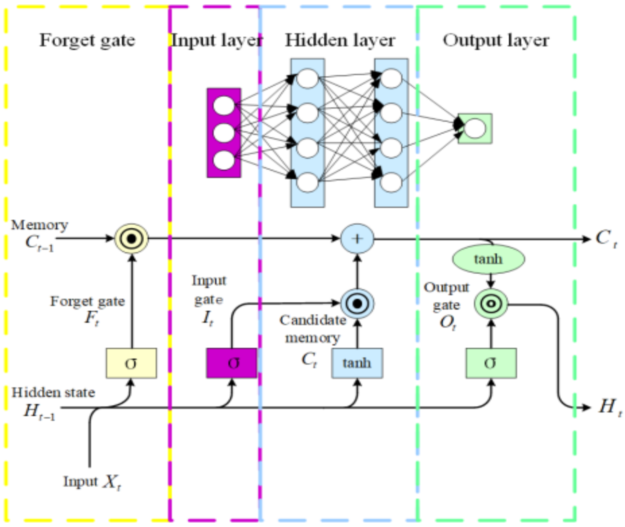

Figure 13. Schematic diagram of LSTM neural network.

3.2.1. LSTM Model Development and Solution

To predict the monthly sales with the LSTM model, the data needs to be processed. The specific steps are as follows:

(1) Step 1. Data Preparation: The datasets were temporally ordered and split into training (60%), validation (20%), and testing (20%) sets.The division results of A1 and A2 are shown in Figure 14:

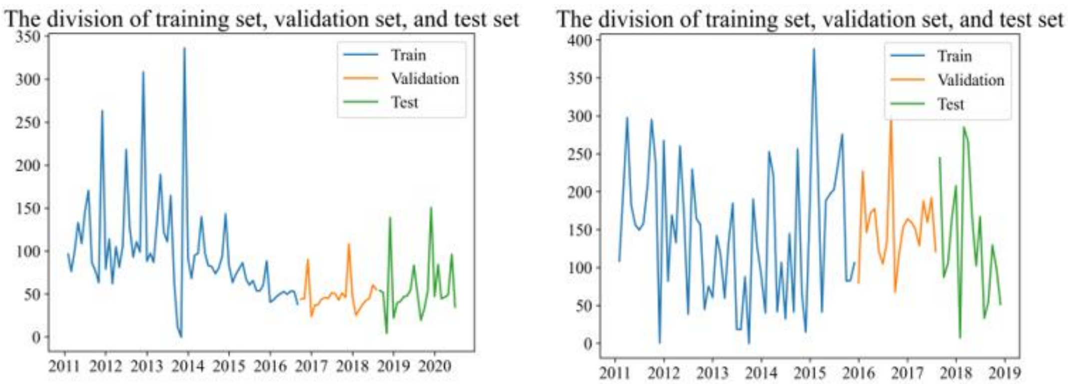

Figure 14. Division of A1 (left), A2 (right), training set, validation set and test set.

(2) Step 2. Data Standardization: To enhance model training efficiency and prediction accuracy, the data were standardized.

(3) Step 3. Formatting for LSTM Input: The standardized data were reshaped into a three-dimensional array compatible with LSTM's input requirements, with dimensions representing samples, time steps, and features, respectively.The time step represents the sales of A1 and A2 in the past period of time and the number of features is 1.

(4) Step 4. Optimizer Selection: Optimizer selection.The LSTM time series prediction model adopted in the paper uses the Adam optimizer [9].The Adam optimizer is an adaptive learning rate

optimization algorithm that improves the convergence speed and stability of the model by dynamically adjusting the learning rate [10].

(5) Step 5. Epoch Selection: To prevent overfitting and underfitting, the number of iterations Epoch can not be too large or too small.For A1, the loss rate is almost zero when the number of iterations is 90, and for A2, the loss rate is almost zero when the number of iterations is 60.So for A1 and A2, the optimal epoch counts were identified as 90 and 60 respectively, shown in Figure 15 below.

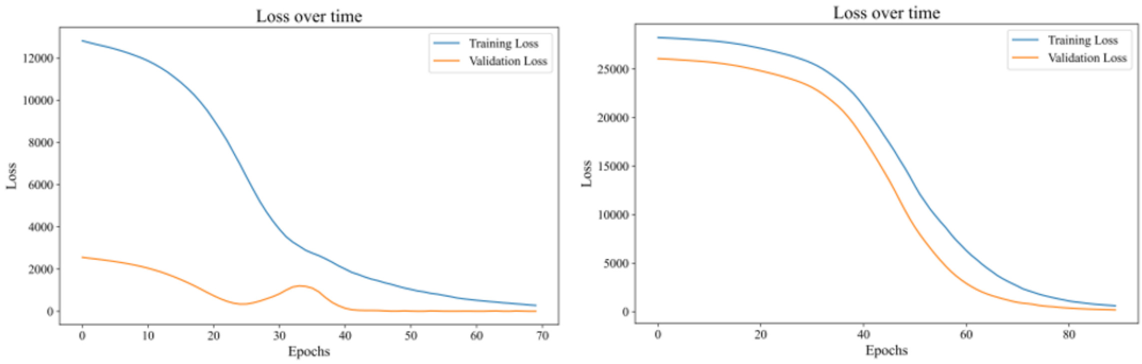

Figure 15. Loss rate variation with iterations for A1 (left) and A2 (right).

3.2.2. LSTM model prediction results

Based on the analysis of the model, a program is written in Python to first obtain the visualization plots of the predicted values versus the actual values for the training set, validation set and test set, and the visualization results are shown in Figures16 below:

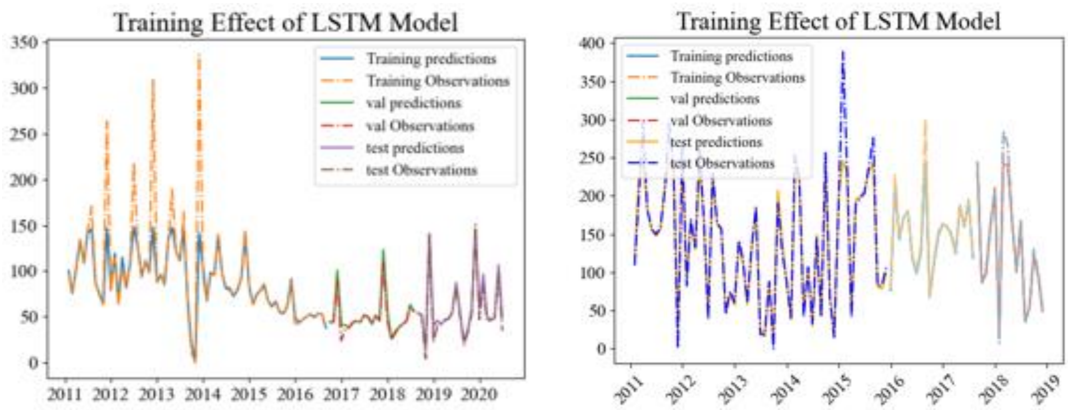

Figure 16. A1 (left), A2 (right) LSTM model corresponding to the training results.

The trained LSTM time series prediction model is used to predict the future sales of A1 and A2 and the prediction results are visualized. The visualization results are shown in Figure 17 below:

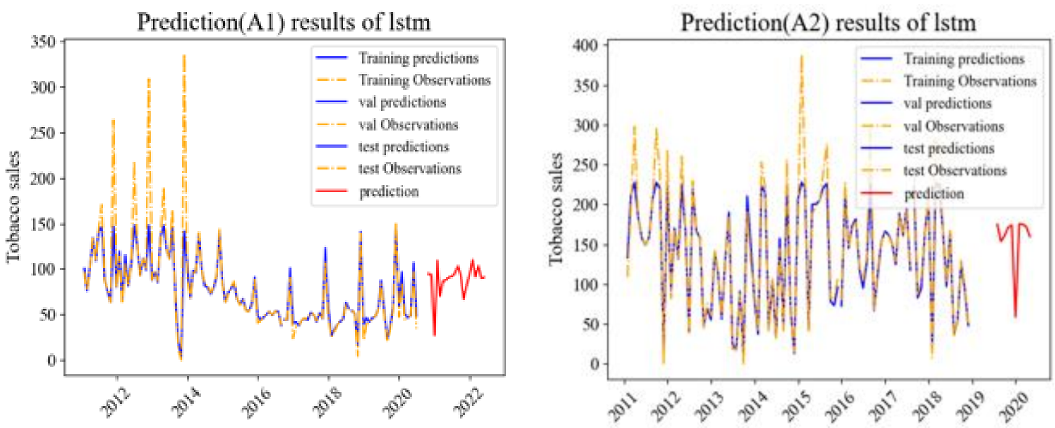

Figure 17. A1 (left), A2 (right) LSTM model prediction results.

3.3. Comparison of the two models

The predictive results obtained from the above have been verified and indicate that the ARIMA time series forecasting model has relatively poor fitting effects for complex models with nonlinear trends and periodic changes, and can only reflect future sales trends and periodic variations to a certain extent. In contrast, the LSTM (Long Short-Term Memory) network, as a method based on neural networks, has better effects in fitting complex patterns such as nonlinear trends and periodic changes. It can perform long-term predictions by automatically extracting time series features, reflecting the trends and periodic variations in sales, and the resulting predictions are more accurate.

4. Conclusions

This paper concludes that the Long Short-Term Memory (LSTM) model is more superior in predicting the future sales of cigarette varieties compared to traditional Autoregressive Integrated Moving Average (ARIMA) time series models. The ARIMA model has limited predictive capabilities when faced with complex and volatile sales data, whereas the LSTM model excels in fitting complex patterns, such as nonlinear trends and cyclical variations. Therefore, the LSTM model is more suitable for predicting the future sales of cigarette varieties.

Of course, due to not very sufficient training datasets, there may still be errors between the predicted values and the actual values. However, it's believed that as the sample size continues to increase and after rigorous training, the accuracy of tobacco market sales forecasts based on LSTM model could become increasingly high, enabling tobacco enterprises to make more reasonable and scientific decisions.

References

[1] Liang Shangjian, Wang Ying. Research on the Impact of the Novel Coronavirus Epidemic on China's Stock and Bond Market: An Empirical Analysis Based on ARIMA Model [J].China Securities and Futures, 2024, (04): 59-66+89.

[2] Chen Hong, Wang Yi, Zhou Qian. The Influence and Prediction of Macroeconomic Indicators on the Economic Operation Indicators of Tobacco Firms: A Case Study of Tobacco Company in N City [J]. Modern Enterprise, 2024, (07): 65-67.

[3] Guo Yuanbo. Reform and Innovation of Cigarette Marketing in Tobacco Industry Enterprises under the New Retail Model [J]. Modern Enterprise Culture, 2024, (13): 38-40.

[4] Zhao Chenzhi, Wang Xueliang, Chen Z, et al. Research on the construction of ARIMA weekly sales prediction model for dynamic allocation of cigarette sorting line tasks[J].China Storage and Transportation, 2022, (08): 203-204.

[5] Deng Chao, Liu Song, Wang Ludi, et al. Construction method of cigarette intelligent delivery model based on deep neural network[J].Tobacco Science and Technology, 2021, 54(02):78-83.

[6] Liang Qi, Gong Xunqiang, Lu Tieding, et al. Improved three-time Hermitian interpolation data interpolation method for high-speed railway pier settlement[J].Science of Surveying and Mapping, 2024,49(07):29-36.

[7] Wang Bin, Yin Lijun, Liu Chunlan, ET al. Comparison of the prediction effect of ARIMA model and exponential smoothing method on the reported incidence of pulmonary tuberculosis in Wushan County, Chongqing [J/OL]. Journal of Preventive Medicine Information, 1-10 [2024-10-20].

[8] Su Xiangjing, Cheng Zifan, Nie Liangzhao, et al. Power probabilistic prediction of offshore wind farm based on AGCN-LSTM model[J/OL]. Power System Automation, 1-13[2024-10-20].

[9] Wu Qingyun, GAO Jinghui, Li Zhao, et al. Testing and application of early warning model of big data platform for thermal power plant based on Adam optimized convolutional neural network [J]. Science Technology and Engineering, 2023, 23(35):15075-15083.

[10] Qian Zhang. Optimization method of LSTM network in speech-to-text application [J]. Electroacoustic Technology, 2024, 48(09):85-87.