Research on the Stock prediction based on LSTM Algorithm

Joleen Mai*

Department of Digital Discovery, Bellevue, USA

*Corresponding author: s-maij@bsd405.org

Abstract. Predicting the value of a stock in the future is a widely recognized challenge. However, the Long Short-Term Memory (LSTM) network provides a distinctive solution as a specialized type of recurrent neural network proficient at tackling long-term dependency problems. Each time a new input is fed, the LSTM will first decide which previous memories to forget based on the new input and the output of the previous moment. This paper analyzes the LSTM algorithm for time series prediction in volume of stock. The dataset is initially split in a 7:3 ratio, with 70% allocated for the training set and 30% for the test set. This paper conducts data processing first, and then plots time series graphs to visualize the characteristics of volume of stock. Additionally, this paper conducts smoothness test together with white noise test, and makes time series predictions using LSTM. Values obtained from the experiments show that the training MSE of the LSTM algorithm is 28333951.7500 and the RMSE is 4722325.29. The final prediction comparison graph proves the potential of the LSTM algorithm to be used in market prediction.

Keywords: Long term short memory network; time series; stock prediction.

1. Introduction

Anticipating the future stock value is an infamously challenging task. In line with the well-established economic theory known as the Efficient Market Hypothesis, the current stock price is deemed to have already absorbed all publicly available information [1]. As a result, predicting future prices without supplementary data should be deemed unattainable. Nevertheless, financial markets are not entirely flawless in their efficiency, opening up the possibility of uncovering concealed patterns through the application of deep learning techniques [2]. The enduring debate regarding the predictability of financial markets has spawned numerous research endeavors, yet a unanimous agreement remains elusive [3]. The advent and progression of effective machine learning algorithms alongside robust computing power have rekindled this debate in recent years.

Although predicting financial time series is widely recognized as a formidable challenge, there exists a multitude of widely-recognized anomalies in capital markets that starkly contradict the concept [4]. Jacobs [5] and Zhang [6] presented surveys which effectively depend on return predictive signals to surpass market performance. Nonetheless, the models are typically straightforward in design and incapable of capturing intricate non-linear relationships.

As far as the securities market is concerned, in the short term, stock prices show random fluctuations due to disorganized operations arising from the lack of expertise of many investors. In the long run, due to different channels and ways of obtaining information, there are large differences in market participants, which in turn affects the stock market price trend. Considering the fact that the price of stocks fluctuates and has a certain trend, how to grasp the stock trend has become an important lesson that investors must master.

Although traditional analytical methods can more accurately predict the trend of stocks in the coming period of time, the traditional analytical methods have become increasingly unable to meet people's needs with the continuous increase of market information and the improvement of people's knowledge level. The traditional mode of analysis is demanding in terms of human and material resources, and requires investors to fully understand the stock market and master certain professional knowledge and skills. In today's information surge, the possibility of error has increased accordingly, which is simply a difficult task to many investors.

In recent years, initial findings have indicated that machine learning techniques have the capacity to detect non-linear financial data, as demonstrated in the work of Huck, Takeuchi et al. Although

differing perspectives exist on the effectiveness of technical analysis, Menkhoff and Taylor's survey predominantly revealed evidence of superior returns when employing technical analysis. Notwithstanding the truth that a majority of the studies examine the utilization of various tools concerning technical analysis, they are notably deficient in four aspects [7]. To begin with, certain studies, such as Sweeney and Willett, encompass only a limited span of approximately three years of data. Consequently, the impact of significant global events like the 2008 global financial crisis might not have been adequately captured in their application of technical analysis. Furthermore, some studies employed a wide range of technical analysis methods without conducting thorough testing. In the case of studies that depend on the Relative Strength Index (RSI) model, there appears to be an absence of a clear review of the underlying assumptions of the model.

Stock price forecasting is typically carried out through either price prediction with numerical data or price prediction with direction. The first task is regarded as a formidable challenge [8]. The second task seems more attainable [9]. While the price change direction doesn't provide the complete picture, it remains highly valuable and can yield significant profits. Also, it is crucial for researchers to identify outliers prior to conducting any analysis [10].

The focus of this paper is to forecast the direction of price changes. While minor alterations in price direction can occur frequently, substantial price shifts are infrequent and influenced by distinct underlying factors. Certainly, some of the noteworthy price fluctuations stem from unforeseen news events that are challenging to predict. Conversely, it appears feasible for individuals to discern instances where a stock is excessively bought or sold, potentially triggering a price reversal. This paper employs an LSTM network for this purpose.

2. Methodology

2.1. Data Source

The data used is the stock data of "Ping An of China" from 2016 to 2018, and the background is Ping An Insurance Group. The data variables are date, open, high, low, close, and volume.

2.2. Metrics Selection and Description

The metrics selected for evaluating the LSTM model are Mean Square Error (MSE) and RMSE. MSE (Mean Squared Error) is computed by taking the square of the difference between the true and predicted values, and then summing and averaging these squared differences. The smaller the value of MSE is, the better the model is. Root Mean Square Error (RMS) is the deviation of the observed value from the actual value. The smaller the value of RMSE is, the better the model is.

2.3. Methodology Introduction

2.3.1 Data processing

Data pre-processing includes the following five steps. (i) The training set and the test set are divided. (ii) Null values and duplicates are deleted, and then column names are modified. (iii) The date format is restored. (iv) Descriptive analysis is conducted. (v) Outliers are processed.

2.3.2 Plotting time series graphs

The time series plot is used to visualize the characteristics of the time series data, such as the observation of the smoothness of the data before the ARIMA model, or the smoothness of the data before the analysis of the VAR model. If the data fluctuates along a certain mean value and there is no obvious trend, then it indicates that the data has smoothness.

2.3.3 Smoothness test and white noise test

There are four types of time series patterns. (i) Trend, which means a long-term increase or decrease in the data, can be any function, such as linear or exponential. It can change direction over time. (ii) Seasonal pattern, which means a repetition in a series at a fixed frequency (hours of the day,

days of the week, months, years), exists for a fixed known period. (iii) Periodicity, which occurs when data go up or down, has no fixed frequency or duration caused by economic conditions. (iv) Noise, which means a random variation in a series.

For time series data, the most important tests are whether the time series data are white noise data and smooth. In the context of a time series, smoothness can be described as the absence of systematic changes in the mean (no trend), the absence of systematic changes in the variance, and the complete elimination of cyclical variations.

2.3.4 Time series prediction using LSTM

LSTM network is a special kind of recurrent neural network which can deal with long-term dependency problems. It can use past information to influence the current output. The LSTM network consists of a series of LSTM units, each of which has an internal state that stores and updates long-term memories. The LSTM units also have three gate structures: forgetting gates, input gates, and output gates. They can control the inflow, outflow, and retention of information. LSTM network has a wide range of applications in natural language processing, speech recognition, time series prediction, etc.

3. Results and Discussion

3.1. Research Design

3.1.1 Dividing the training set and test set

The random seed is set as 100, and the total data set is divided into a training set and a test set by random sampling in the ratio of 7:3. There are 511 samples in the training set and 220 samples in the test set after division.

3.1.2 Training set data after basic processing

By previewing the training set data (shown in Table 1), it can be seen that the data attributes of this experiment are date, open, high, low, close, and volume.

Table 1. Preview of some training set data

date	open	high	low	close	volume
2016-01-04	30.57	30.57	28.63	28.78	70997200
2016-01-05	28.41	29.54	28.23	29.23	87498504
2016-01-06	29.03	29.39	28.73	29.26	48012112
2016-01-07	28.73	29.25	27.73	28.50	23647604
2016-01-08	28.73	29.18	27.63	28.67	98239664
2016-01-11	27.73	28.06	26.73	26.76	99355696

3.1.3 The overall trend of the training set data

The goal of this paper is to predict the volume of a stock, so the focus is on exploratory analysis of the VOLUME column. A simple, overall analysis of the data is done first to understand the overall trend of the data, and then outliers are removed or replaced.

First, the paper analyzes the centralized trend of the data. It can be observed that the minimum value of stock volume is 12768281. The maximum value of stock volume is 260286256, the mean value is 63570456, and the median value is 56186120. Then the paper analyzes the degree of dispersion of the data. The standard deviation is 34853768, the extreme deviation is 247517975, and the quartile deviation is 41016350. The result shows that the degree of dispersion of the data is high (as shown in Figure 1).

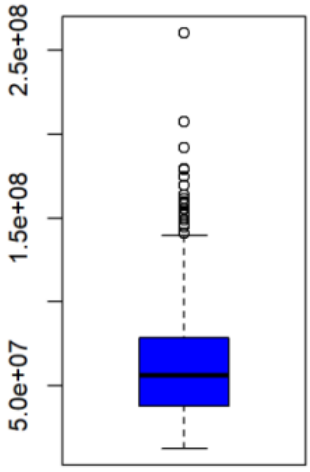

Fig. 1 Training set stock volume box plot

By observing the box plot, it is found that there are some anomalies in the training set stock volume data. The data is abnormal and has a slight impact on the next day's stock volume. To ensure the accuracy of the data, the anomaly data are removed and subsequently interpolated. Multiple imputation is a technique involving repeated simulation to create a full dataset by filling in missing data using the Monte Carlo method. The "mice()" function in the R package "mice" can be employed to implement multiple imputations for missing data.

3.2. Plotting the Training Set Time Series

The time series plot is plotted after multiple interpolations of the missing data of the training set to roughly observe the pattern and trend of stock volume over time. From the time series plot (Figure 2), it can be roughly assumed that this is a smooth time series plot due to the constant fluctuation of values around the mean. Of course, the training set data needs to be tested later.

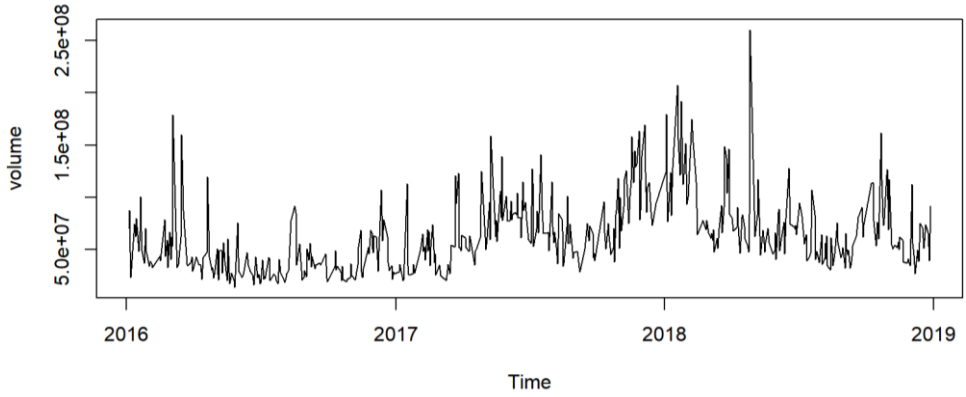

Fig. 2 Time series plot of volume of stock in training data

3.3. Smoothness Test and White Noise Test

Smoothness in this context implies that the expectation (mean) is constant, the variance remains constant, and the covariance does not exhibit temporal changes. The initial hypothesis of the Augmented Dickey-Fuller (ADF) test assumes non-stationarity. The ADF test yields a p-value less than 0.05, leading to the rejection of the initial hypothesis. Therefore, it can be concluded that the stock volume data in the training set represents a smooth time series.

The initial hypothesis of the Ljung-Box test posits that the sequence follows a white noise pattern. However, the test result yields a p-value less than 0.05, leading to the rejection of the original hypothesis. That is, the stock volume data of training set is a non-white noise sequence. Therefore, the stock volume data of training set is a non-white noise and smooth sequence.

3.4. Prediction using LSTM

3.4.1 Training set normalization

The LSTM network is sensitive to the "scale" of the data, which needs to be scaled from 0 to 1. Therefore, it is also necessary to normalize the training set data. In machine learning, data standardization has many benefits, which can improve the accuracy of the model and also improve the convergence speed. So it is a very important step. The initial hypothesis of the Ljung-Box test posits that the sequence follows a white noise pattern. However, the test result yields a p-value less than 0.05, leading to the rejection of the original hypothesis.

3.4.2 Constructing the sequences and labels

If predicting the next data with the first 30 data means predicting the next day with the previous month, it is equivalent to using a 30th-order lag model and treating the first 30 samples as 30 features. First, the data is divided into X and Y, which means the eigenvalues and target values, or features and labels. Set the look-back parameter to 30, which represents the training window size. Then the paper constructs a function that splits the data into X and Y.

At this point, the paper gets a-train-XY (list type) which contains two arrays: train-X and train-Y. Next, the data needs to be transformed into tensors. The tensor, in math, is a high-dimensional array.

It can be seen that there are 481 training samples. Each input sample contains 1 row and 30 columns, indicating that there is only 1 time step and there are 30 features. Each output sample has only 1 time step and 1 label.

3.4.3 Performing model training

The input shape comes from the shape of the data-X (481,1,30), which becomes (481,1,100) after going through the LSTM layer. After passing through the dense layer, the shape becomes (481,1,20). In addition to the LSTM and dense layers, the paper also sets up the repeat-vector and time-distributed layers. The parameter settings of the repeat-vector layer can change the output step, and the time distributed layer is usually used in conjunction with the dense layer to change the output attribute length.

3.4.4 Training results

The values obtained from the experiments show that the MSE of the LSTM algorithm in training data set is 28333951.75 and the RMSE is 4722325.29.

3.4.5 Applying the model to test set

The logic of testing can be done using a recursive method, where each time a new value is predicted, the predicted value is used as an input for the next prediction. First, the at-test-XY can be generated first using the a-train-XY generation method. Once the prediction results are obtained, the results are then normalized for the training set data.

3.4.6 Plotting the comparison

Red refers to the original data, blue refers to the training data, and green refers to the test data. It can be seen that the approximate trend is very consistent (Figure 3). But there are some places where the peaks are not predicted enough. All in all, the overall effect is quite good.

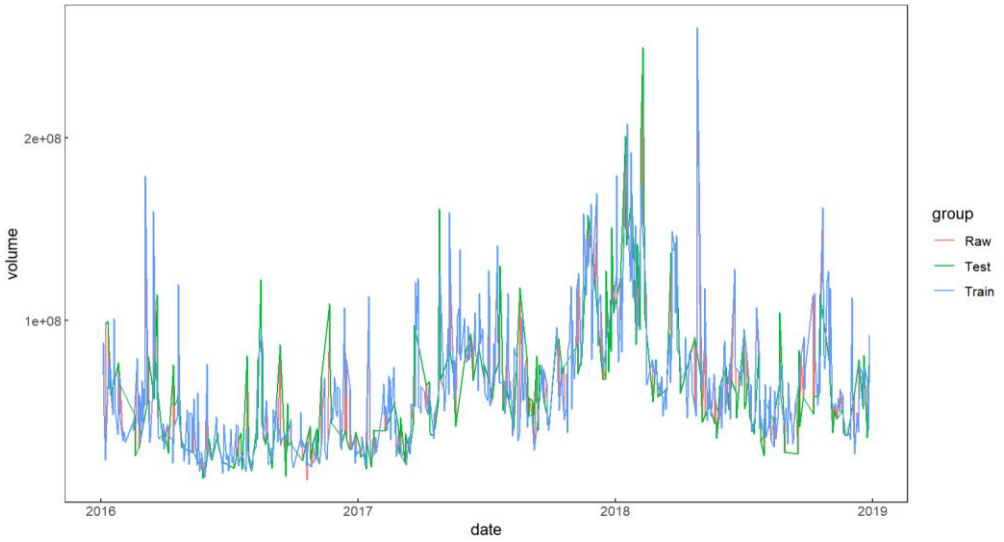

Fig. 3 Time series plot of volume of stock in raw data, test set, and training set

4. Conclusion

This paper analyzes the LSTM algorithm for time series prediction. Specifically, i.e., the value of the next day was predicted based on the information of the previous 30 trading days. The dataset was first divided in the ratio of 7:3, where 70% was used as the training set and 30% as the test set. The values obtained from the experiments show that the training MSE of the LSTM algorithm is 28333951.7500 and the RMSE is 4722325.2917. The final prediction comparison graph proves the potential of the LSTM algorithm to be used in market prediction.

Automated stock trading platforms utilizing LSTM algorithms have the potential to offer substantial value to brokerage firms engaged in day trading. Nevertheless, additional scrutiny and experimentation are essential before these models can be effectively integrated into production. It is evident that the analysis presented in this paper will contribute to improving the prediction of stock price using LSTM algorithms.

References

[1] Fama E F. Efficient capital markets: A review of theory and empirical work. Journal of Finance, 1970.

[2] Kamalov F, Smail L, Gurrib I. Stock price forecast with deep learning. Working paper, 2021.

[3] Borovkova S, Tsiamas I. An ensemble of LSTM neural networks for high-frequency stock market classification. Journal of Forecasting, 2019.

[4] Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 2018, 270(2): 654-669.

[5] Jacobs H. What explains the dynamics of 100 anomalies? Journal of Banking & Finance, 2015.

[6] Green J, Hand J R M, Zhang X F. The supra-view of return predictive signals Review of Accounting Studies, Journal of Finance, 2013.

[7] Gurrib I, Kamalov F. The implementation of an adjusted relative strength index model in foreign currency and energy markets of emerging and developed economies. Macroeconomics and Finance in Emerging Market Economies, 2019, 12(2): 105-123.

[8] Kamalov F, Leung H H. Outlier detection in high dimensional data. Journal of Information & Knowledge Management, 2020.

[9] Kim K. Forecasting significant stock price changes using neural networks. Neural Computing and Applications, 2019.

[10] Lee J, Kang J. Effectively training neural networks for stock index prediction: Predicting the S&P 500 index without using its index data, 2020.