Forecast of Stock Price

Zizhan Jiang(✉)

The University of Bath, Bath, UK
jiang030401@gmail.com

Abstract. The stock market is an important part of the financial market. The stock price prediction based on the model has very important practical significance for individuals and enterprises. So this paper uses regression models to fit past stock prices and forecast their future volume. This paper uses the polynomial regression method to regression the stock price from 2012 to 2017, and then uses LSTM to predict the inventory. The data used in this paper is from 2012 to 2017. Training on the data of the past few years, predicting the output in 2017, and then comparing it with the actual output. After training, the result shows that the trend of the predicted volume is similar to the actual volume in 2017. Therefore, LSTM truly forecasts the stock volume.

Keywords: stock price · stock volume · regression model · prediction model · Long- Short Term Memory (LSTM)

1 Introduction

With the continued development of the world, the data collected in daily life becomes more and more continuous. The stock market is a high-risk, high-yield market, and it is also a common market that appears in people's daily lives. Therefore, citizens pay more attention to the stock price. When it will rise or drop, when it should be sold. While no one knows the future, a presumption can be forecast with models and data. More over, stock prices are one of the most important parts of the national financial market, which shows the status of the economy [1]. The stock market is an important part of China's financial market. The running of the stock market affects national economics a lot [2]. Financial data is a typical data series. Many investigators have used time series models in the financial market, like the ARIMA [3][4] model and the GARCH [5] model. LSTM predicts much better than other methods such as RNN, therefore the author chooses LSTM models.

This paper introduces how LSTM works, and how to use it to predict some information by using past data. There are three parts to this paper after a brief introduction. The first part is polynomial regression. It changes data into a continuous series. The second part is the introduction of LSTM, which is a kind of deep learning. The third part is training and results. In this part, the author uses models and data to prove the ideas. The research presented in this paper can help people understand the price volatility and predictability of the stock market from a data level to a certain extent, so as to combine theoretical knowledge with practical problems and increase the practical application of the model.

F. Balli et al. (Eds.): ESFCT 2022, AEBMR 226, pp. 1529–1539, 2022.
https://doi.org/10.2991/978-94-6463-052-7_169

2 Data

All the data used is from Kaggle [6]. Figure 1 shows the open and close price, high and low price, and volume of a stock, from 2012 to 2017. It's a 1232 rows by * 7 column graph after reading it by python.

3 Polynomial Regression

Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x:

$$h(x) \ = w_0 + w_1x_1 + w_2x_2 \tag{1}$$

As shown in Fig. 2, comparing the red line with the blue points. The straight line cannot truly show the trend of a stock price, as two lines do not fit. Adding a quadratic term transforms it from plane to paraboloid.

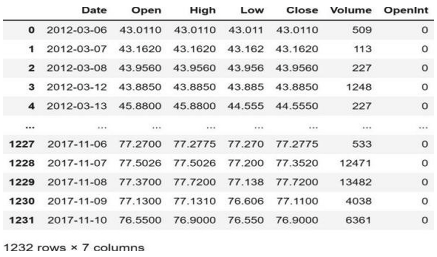
$$h(x) \ = w_0 + w_1x_1 + w_2x_2 + w_3x_1x_2 + w_4x^2 + w_5x^2 \tag{2}$$

changing x by z, it can be shown as following:

$$z = \ [z_1, \ z_2, \ z_3, \ z_4, \ z_5] = [x_1, \ x_2, \ x_1x_2, \ x_1^2, \ x_2^2]$$

The formula can be written into:

$$h(x) \ = w_0 + w_1z_1 + w_2z_2 + w_3z_3 + w_4z_4 + w_5z_5 \tag{3}$$



	Date	Open	High	Low	Close	Volume	OpenInt
0	2012-03-06	43.0110	43.0110	43.011	43.0110	509	0
1	2012-03-07	43.1620	43.1620	43.162	43.1620	113	0
2	2012-03-08	43.9560	43.9560	43.956	43.9560	227	0
3	2012-03-12	43.8850	43.8850	43.885	43.8850	1248	0
4	2012-03-13	45.8800	45.8800	44.555	44.5550	227	0
...	...	...	...	...	...	...	...
1227	2017-11-06	77.2700	77.2775	77.270	77.2775	533	0
1228	2017-11-07	77.5026	77.5026	77.200	77.3520	12471	0
1229	2017-11-08	77.3700	77.7200	77.138	77.7200	13482	0
1230	2017-11-09	77.1300	77.1310	76.606	77.1100	4038	0
1231	2017-11-10	76.5500	76.9000	76.550	76.9000	6361	0

1232 rows × 7 columns

Fig. 1. Stock price dataset

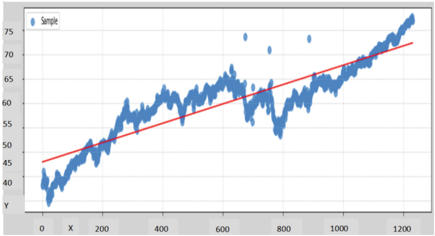


Fig. 2. The regression of the stock price

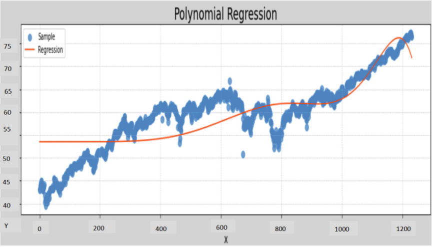


Fig. 3. The Polynomial regression of the stock price

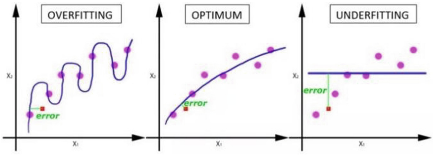


Fig. 4. Three types of regression

Then it has changed into linear regression model.

Figure 3 shows the graph regressed by polynomial regression. The regression line (red line) in Fig. 3 is much similar to the line in Fig. 2. Especially in the last part, it almost 100% fitted with the data. The problem must be avoided. When doing regression, three results might be obtained, over-fitting, optimum and under-fitting. Optimum is the best status for regression.

Figure 4 shows the three types. Over-fitting means the regression line passes all the points, while under-fitting means the trend is not fit. Figure 2 shows an example of under-fitting.

4 LSTM

LSTM is called long short-term memory, which is a kind of artificial RNN(recurrent neural network) that is used in DL(deep learning). It can process not only single data points (such as images), but also entire sequences of data (such as speech or video). A common LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell. Comparing RNN with LSTM.

Recurrent neural network(RNN) is a kind of neural network that can deal with data series. It was proposed by Hopfield [7] in 1982. RNN which he proposed, is a ring structure that can connect input information together, but it is difficult to achieve. In 1986, Jordan [8] proposed a new RNN. In 1990, Elman [9] improved Jordan's RNN and proposed the Elman network, which is nowadays the most common RNN and has been widely used.

Hochreiter and Schmidhuber [10] proposed the LSTM model in 1997, as RNN could not solve the long order dependence problem of data. LSTM model has gates inside, and it can solve exploding and vanishing gradient problems by calculating. In 2000,

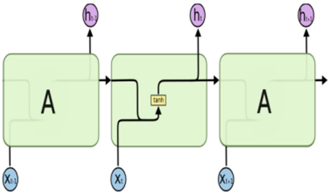
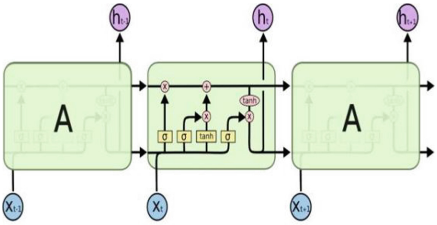


Fig. 5. RNN



The repeating module in an LSTM contains four interacting layers.

Fig. 6. LSTM

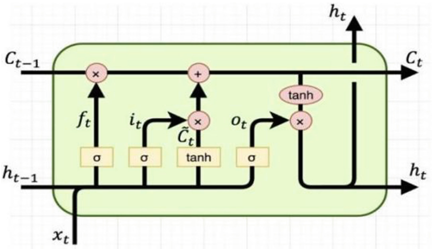


Fig. 7. The main part of LSTM

Felix [11] added a forget gate in LSTM to solve the network crash problem, as the LSTM memory storage increased by the length of the series. LSTM has 4 common structures [12], including: many-to-one, one-to-many, many-to-many(same size) and many-to-many(different size).

σ is the gate.

The formula is as following:

$$f_t = \sigma(W_f \cdot [h_{t-1},\ x_t] + b_f) \tag{4}$$

$$i_t = \sigma(W_i \cdot [h_{t-1},\ x_t] + b_i) \tag{5}$$

$$C_t = tanh(W_c \cdot [h_{t-1},\ x_t] + b_C) \tag{6}$$

Sigmoid is a value between 0 and 1, which means how much information can pass through this gate. 0 means no value can pass through the gate, while 1 means any value

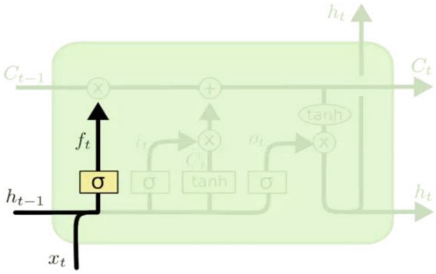


Fig. 8. The first path of LSTM

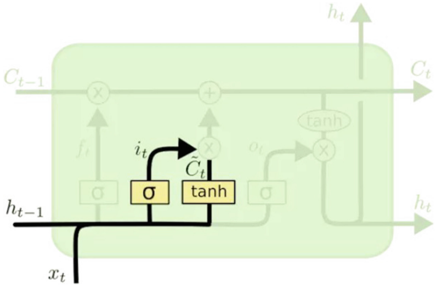


Fig. 9. The Second path of LSTM

can pass. According to Fig. 7, LSTM has three gates to protect and control the cell. Analyze Fig. 7 by changing it into four independent parts. This part of LSTM decides whether the information should be abandoned or not.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{7}$$

f_t is the forget gate. It is used to screen old memories. For example, if a student has already passed the math exam, he will face physics tests. LSTM shows the way he reviews. Assume that a student has finite cranial capacity and can easily forget what he does not need. The knowledge he needs to forget is formulas and theorems, but he must also obtain calculating ability. Therefore, when all his memories pass through the first gate, he will only remember how to calculate. The second part of LSTM ensures that the information is up to date.

As shown in Fig. 9, i_t is the update gate.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{8}$$

$$C \sim t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_C) \tag{9}$$

x_t is the update gate. x_t means physical knowledge. While among all the knowledge, some is important. By passing this gate, it filters the updated information. The third part is updating cell status.

$$C_t = f_t * C_{t-1} + i_t * C_t \tag{10}$$

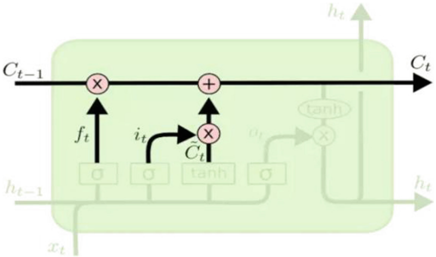


Fig. 10. Combination of two paths

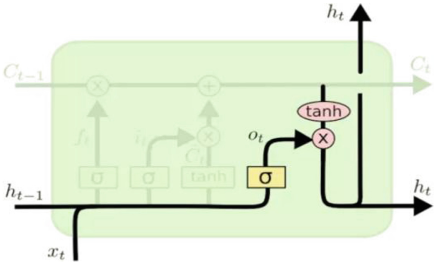


Fig. 11. The last path and output of LSTM

By adding the obtained parts ($f_t * C_{t-1}$) and updating parts ($i_t * C_t$), it will cause a new status (C_t). Keep on that example, student held onto his calculating ability, and learned some new physical formulas. Then he had already prepared for the exam. The last part is used to output information.

$$o_t = \sigma(W_o)[h_{t-1},\ x_t] + b_o \tag{11}$$

$$h_t = o_t * \tanh(C_t) \tag{12}$$

o_t is the output gate.

Still using that example, how well you review does not equal how many scores you may get on the exam. It is also the same in this system. $\mathrm{C_t}$ is the information you may remember during review. $\mathrm{H_t}$ is the final score you get after the exam. Although there is no direct relationship between review and score, the better your review, the higher your score.

5 Training and Result

Figure 12 shows the stock volume from 2012 to 2018. From the graph, the volume is maintained at a section, but sometimes its volume causes a deviation. This graph cannot clearly show any useful information or results just by looking at it.

Figure 13 shows the process of epoch works, and how to train the model.

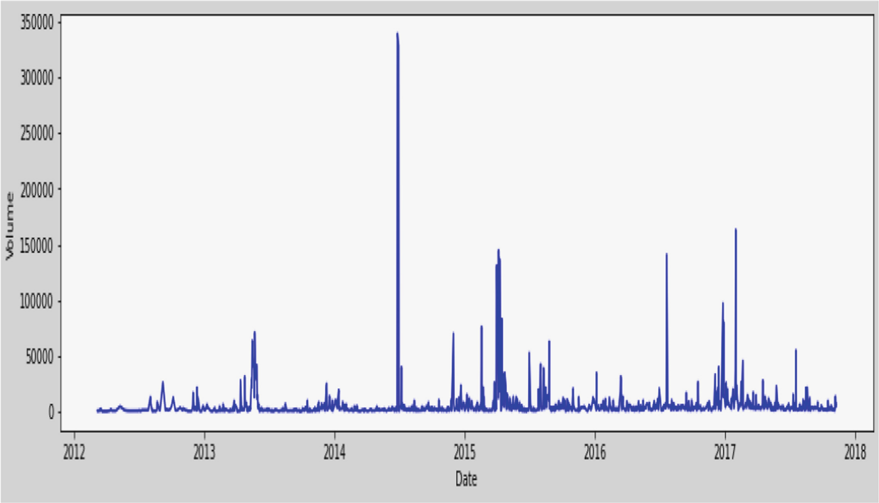
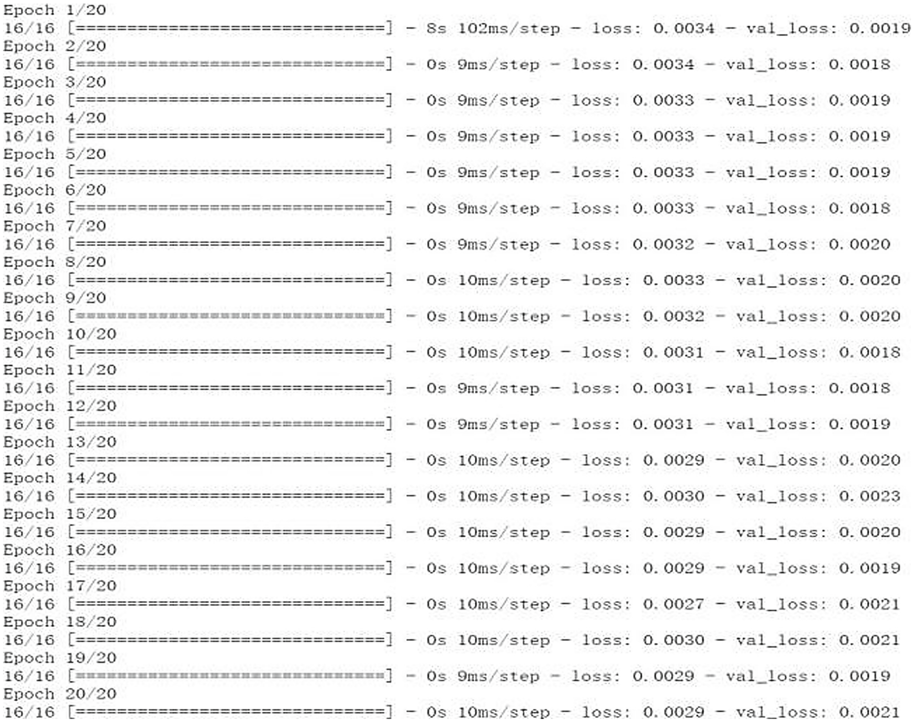


Fig. 12. Stock volume from 2012 to 2017



```
Epoch 1/20
16/16 [==============================] - 8s 102ms/step - loss: 0. 0034 - val_loss: 0. 0019
Epoch 2/20
16/16 [==============================] - 0s 9ms/step - loss: 0. 0034 - val_loss: 0. 0018
Epoch 3/20
16/16 [==============================] - 0s 9ms/step - loss: 0. 0033 - val_loss: 0. 0019
Epoch 4/20
16/16 [==============================] - 0s 9ms/step - loss: 0. 0033 - val_loss: 0. 0019
Epoch 5/20
16/16 [==============================] - 0s 9ms/step - loss: 0. 0033 - val_loss: 0. 0019
Epoch 6/20
16/16 [==============================] - 0s 9ms/step - loss: 0. 0033 - val_loss: 0. 0018
Epoch 7/20
16/16 [==============================] - 0s 9ms/step - loss: 0. 0032 - val_loss: 0. 0020
Epoch 8/20
16/16 [==============================] - 0s 10ms/step - loss: 0. 0033 - val_loss: 0. 0020
Epoch 9/20
16/16 [==============================] - 0s 10ms/step - loss: 0. 0032 - val_loss: 0. 0020
Epoch 10/20
16/16 [==============================] - 0s 10ms/step - loss: 0. 0031 - val_loss: 0. 0018
Epoch 11/20
16/16 [==============================] - 0s 9ms/step - loss: 0. 0031 - val_loss: 0. 0018
Epoch 12/20
16/16 [==============================] - 0s 9ms/step - loss: 0. 0031 - val_loss: 0. 0019
Epoch 13/20
16/16 [==============================] - 0s 10ms/step - loss: 0. 0029 - val_loss: 0. 0020
Epoch 14/20
16/16 [==============================] - 0s 10ms/step - loss: 0. 0030 - val_loss: 0. 0023
Epoch 15/20
16/16 [==============================] - 0s 10ms/step - loss: 0. 0029 - val_loss: 0. 0020
Epoch 16/20
16/16 [==============================] - 0s 10ms/step - loss: 0. 0029 - val_loss: 0. 0019
Epoch 17/20
16/16 [==============================] - 0s 10ms/step - loss: 0. 0027 - val_loss: 0. 0021
Epoch 18/20
16/16 [==============================] - 0s 10ms/step - loss: 0. 0030 - val_loss: 0. 0021
Epoch 19/20
16/16 [==============================] - 0s 9ms/step - loss: 0. 0029 - val_loss: 0. 0019
Epoch 20/20
16/16 [==============================] - 0s 10ms/step - loss: 0. 0029 - val_loss: 0. 0021
```

Fig. 13. The result after epoch

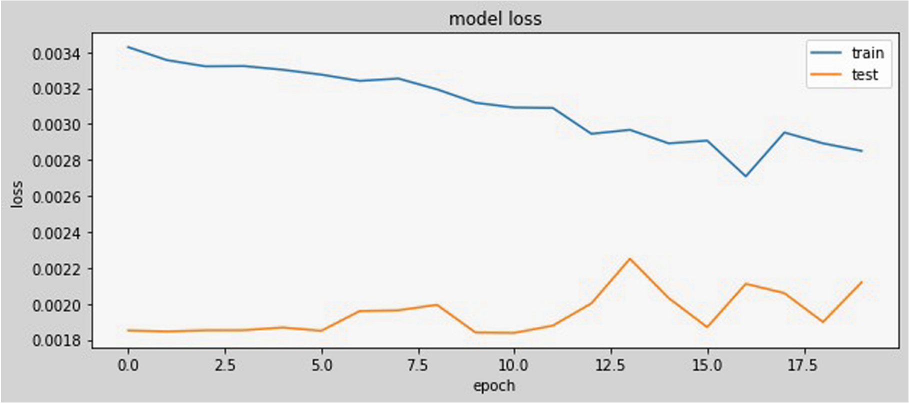


Fig. 14. Model loss

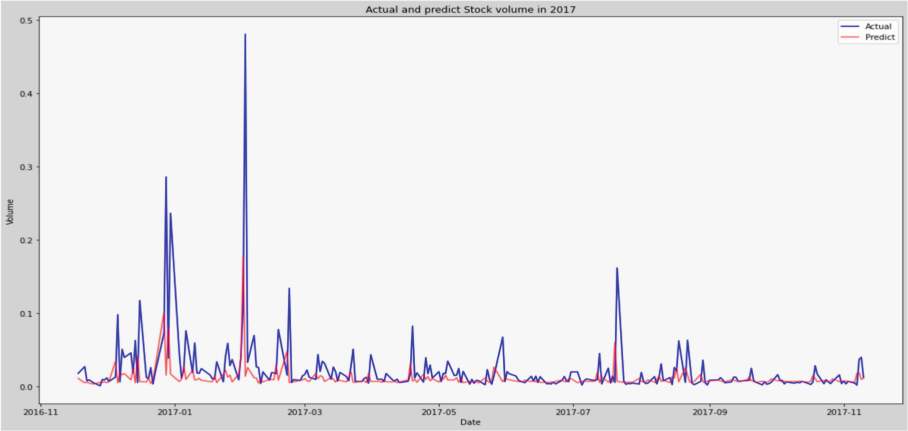


Fig. 15. The predict and actual stock volume

As shown in Fig. 14, the blue line represents loss, which is also called train loss. The orange line is the value loss, which is also called a test loss. The trend of these two lines shows the model status. There are five different statuses.

Train loss decreases and test loss decrease, indicating that the model keeps learning. Train loss decreases as testing becomes more unstable, implying over-fitting.

Train loss is stable, but test loss has decreased, indicating that the dataset 1000% has issues that need to be investigated.

Train loss and test loss stable means model facing difficulties, which needs to decrease the study rate.

Train loss increases and test loss increases mean the structure is untrue, which is the worst circumstance.

In Fig. 14, the train loss decreases, and the test loss fluctuates. It is the first and second situation, which means the model is still learning (a good status), while being cautious it might be over-fitting. Figure 15 shows the predicted and actual stock volume in 2017. The trends of the two lines are the same.

6 Discussion

An epoch is a term used in machine learning and indicates the number of passes through the entire training dataset that the machine learning algorithm has completed. The more we train the model, the more accurate, it will be.

Figure 17 shows the model of predicted volume, but the epoch is larger than it was in Fig. 16. When comparing the two figures, the predicted line in Fig. 17 is more similar

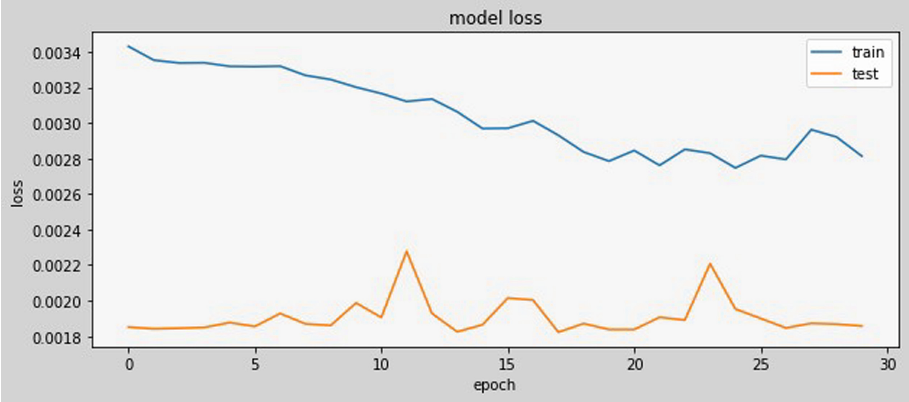


Fig. 16. The model loss with larger epoch

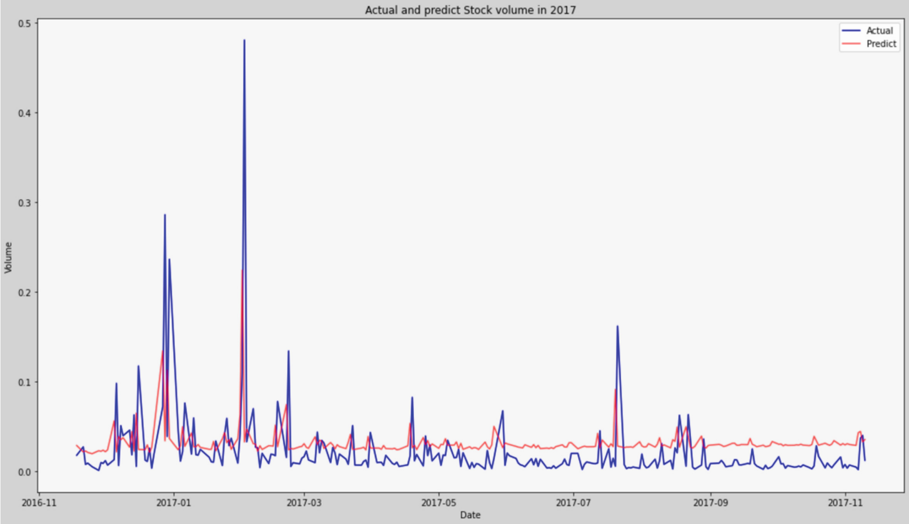


Fig. 17. The predict and actual stock volume

to the actual volume than in Fig. 16. The only way to make the predicted line fit with the actual line is to use larger data. The more data we have, the more datasets used to train the model, the more accurate model it will provide.

7 Conclusion

This paper uses polynomial regression to regress the stock price from 2012 to 2017. Then it uses LSTM to predict stock volume. The model has been trained by the past years' data and predicts the volume in 2017 and then compares it with the actual volume. Two lines are similar. In fact, the stock price is not only affected by time series. It might also be affected for other reasons, like politics, the environment, and so on. Therefore, the predicted volume might not match the actual. Future research will try to make a model with larger dimensions, which means more factors will be considered, and the predicted result in this model will be more accurate than the actual result.

References

1. Wang Ping, Zhang Hongwu. An empirical study on the relationship between China's stock market and national economy. Special Zone Economy. (10) (2006) pp.70-72.
2. Zhang Shanshan. Research on the "barometer" effect of China's stock market on the national economy. Jilin University, 2011.
3. G. Peter Zhang. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 50, 2003, pp. 159–175.
4. Gong Congcong. An empirical study on the relationship between China's stock market and national economy. Shandong University, 2012.
5. Helmut Herwartz. Stock return prediction under GARCH - An empirical assessment. International Journal of Forecasting, 33(3), 2017.
6. Kaggle. Data source. www.kaggle.com.
7. J.J. Hopfield. Neural Networks and Physical Systems with emergent Collective Computational Abilities. National Academy of Sciences of the United States of America, 79(8), 1982.
8. MichaelI. Jordan. Chapter 25 Serial order: A parallel distributed processing approach. Advances in Psychology, 121. 1997, pp. 471–495.
9. Elman JL. Finding Structure in Time. Cognitive Science,14(2) (1990) pp.179-211.
10. Horchreiter S, Schmidhuber J. Long Short-Term Memory. Neural computation. 9(8) (1997) pp.1735-1780.
11. Felix A.Gers. Ju.rgen Schmidhurber, Fred Cummins. Learning to Forget: Continual Prediction with LSTM. Neural Computation, 12(10), 2000.
12. Yang Li, Wu Yuxi, Wang Junli, Liu Yili. Review on the research of recurrent neural networks. Journal of Computer Applications, 38(S2):1–6+26, 2018.