Stock Price Prediction Using Machine Learning Techniques

Zeyu Li *

Faculty of Science, McGill University, Montreal, Canada

* Corresponding author: Zeyu.li2@mail.mcgill.ca

Abstract. Large variations in stock price and unstable fluctuations will not only bring huge losses but also might influence the decisions made by investors. Machine learning in financial areas has been widely used in the past few years, and they are especially superior in forecasting tasks. This article will talk about how machine learning techniques could be used to predict stock prices and some possible ways of improvements. The main method used in this article is the LSTM structure, and the building blocks are constructed into a full Recurrent Neural Network to predict the price for several famous technology companies from NASDAQ. According to the results in this article, a forecasting algorithm based on the LSTM model can minor the error down to a few dollars. It is overall not accurate enough and cannot be used to draw a financial conclusion about investment, but still possible to explain some personal confusion and give solutions to individual difficulties. In addition, this study can be further upgraded in terms of structure and logic. This article hopes to provide a broader research perspective for the application of machine learning in the financial field.

Keywords: Machine Learning, Stock Price, Long Short-Term Memory.

1. Introduction

The stock market plays a significant part in the whole financial system, so variations in prices could lead to personal monetary loss and even national economic corruption. Multiple factors, such as merchandise, business deals, hostile takeovers and short selling, trades between retail investors, or even speeches of internet celebrities can cause stock price shock. Many investors or fund managers experience huge losses in the NASDAQ Index every year due to the fluctuation in various market behaviors. To assist investors in mitigating their losses and offering suitable guidance to inexperienced shareholders, it is crucial to predict stock movements promptly and accurately.

As time goes on, several mature machine learning models have been applied to research related to finance and the stock market, like decision trees, deep learning, and support vector machines [1]. In some previous studies, deep learning is the most commonly used way for forecasting and prediction [2, 3]. Datasets for the stock market are not only large and nonlinear but also can be vague and memorable. These features correspond to the advantages of a Recurrent Neural Network, while a traditional RNN has a single hidden state that is passed through time. However, this can also make it difficult for the network to learn long-term dependencies, while the LSTM network is capable of learning long-term dependencies in sequential data or time-series forecasting. So, LSTM is considered to have a better development prospect in finance and accounting. The model has also been used before in previous finance and accounting research such as detecting financial statement fraud [4]. LSTM is a type of Recurrent Neural Network (RNN) with multiple enhancements.

This research explores various methodologies to predict market trends utilizing computer science and machine learning techniques. This research is mainly about a stock price prediction based on the LSTM method, which could enable researchers to obtain an approximate market situation and price estimation for specific stocks. The study also aims to identify patterns in stock price fluctuations over time and assess whether there are reliable methods for individuals to predict these trends, potentially facilitating profit generation through machine learning models.

2. Methods

2.1. Data

The datasets in this project are from the Yahoo Finance website (https://aroussi.com/post/python-yahoo-finance). The 'finance' library is built for Python to store the market data from NASDAQ, containing the previous records of prices and volume of sales annually. It is easy to download and analyze the marketing data in this way. Using the information from the data such as closing price, the volume of sales, moving average, and daily return, it would be easy to calculate the average return, correlation, value of risk, etc, indeed, these calculated values are helpful for the prediction, as it is well known that technology stocks are the ones that have the most increase in price and sales in the past ten years. Tech companies seem to be leading the human development progress as the research is carried out and the products that are being made. Any small action or decision made by them, or rules and restrictions related to them can have a huge impact on the market and investor's confidence. They also brought up the market index to a high level, which created a prosperous scene for investors.

2.2. Model

LSTM was originally designed to deal with the vanishing gradient problem that could appear in traditional RNNs. The LSTM cell can process data sequentially and keep its hidden state through time. Its relative insensitivity to gap length is an advantage over other RNNs, hidden Markov models, and other sequence learning methods. The most widely used LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over a specific time interval and the gates control the flow of information when going into and going out of the cell. Forget gate decides what information to ignore from previous states by assigning a previous state, compared to a current input, a value between 0 and 1. Input gates decide what kind of new values will be stored in current states, similar a forget gates, whereas output gates control information in the current state to output by assigning a value from 0 to 1 to the information, depending on the previous and current states.

The sequential model built in this research is composed of four kinds of blocks, using 'keras', as shown in Table 1, generated by the summary function in keras. It is a two-layer LSTM network followed by two Dense layers, which is a common architecture for time series prediction tasks. A 128-unit LSTM layer is used on the top will return the full sequence output, after it accepts the training data. Then a 64-unit LSTM layer will return the last output in the output sequence. The 25-unit dense layer, which is also recognized as the fully connected layer is used. It serves as a transition between the high-dimensional LSTM outputs and the final output. Each of its 25 neurons will learn to represent different features in the input data. The 1-unit Dense layer is the final layer of the model. The purpose of this layer is to reduce all the learned high-dimensional representations into a single output. In a regression problem, this could be a single continuous value. In this case, it would be the closing price of the stock.

Table 1. Model Structure

Layer (type)	Output Shape	Parameter #
Lstem_14 (LSTM)	(None, 60, 128)	66560
Lstem_15 (LSTM)	(None, 64)	49408
Dense_14 (Dense)	(None, 25)	1625
Dense_15 (Dense)	(None, 1)	26

3. Result

This experiment collected and analyzed data from the year 2012 to now, which is more than ten years. The 'Adam' optimizer is used, which stands for Adaptive Moment Estimation. As the dataset contains approximately 3000 numbers for each stock, it is large enough and needs to use different learning rates for different weights and specific updates separately. It can also accelerate the

convergence to a global minimum of the loss function. The loss function used here is MSE, which is widely known as the mean square error.

The first task of the model was to predict the trending price of AAPL, also known as the famous company Apple. The training loss is 0.001. The result is shown in Fig. 1. Fig. 1 summarizes the result of the test when the training is done with one epoch and a batch size of one. The blue line shows how much training the experiment used to help the model do the prediction, labeled as 'Train'. The blue line goes from the start of the year 2012 and continues to March 29, 2023. The price starts from lower than 25 dollars and goes high up to 180 dollars per stock at the start of 2022. And it can be seen from the graph that after the year of 2018, AAPL experienced a rapid increase. According to time series theory [5], this feature counts a much larger proportion during training than the previous years, which indeed represents a better overall trend for the stock. The red line is the true price of the stock, starting from March 29th, 2023, labeled as 'Val'. The price goes up and down in a time interval of approximately seven months, for several times. The yellow line is generated from the model, recognized as the predicted stock price for AAPL, labeled as 'Prediction'.

Fig. 1 shows that the overall trend of prediction for the fluctuating price is correct, as the line of predicted price is almost the same as the real-world price line. Table 2 shows the actual values of these predictions. From the actual values in the chart, it is easy to see that the predicted values are always approximately around the Close price. There is a fly in the ointment that the differences between the Close price and prediction value seem to grow larger as time goes on. More data for this stock will be needed to prove that the model is precise enough to afford prediction jobs in a longer period.



Figure 1. The Prediction Price of AAPL with batch size one

Table 2. The actual values of prediction for AAPL

Date(YYYY-MM-DD)	Close Price(US$)	Predicted Price(US$)
2023-03-29	160.770004	164.310059
2023-03-30	162.360001	164.758301
2023-03-31	164.899994	165.672272
...	...	...
2023-10-25	171.100006	180.667984
2023-10-26	166.889999	179.220764
2023-10-27	168.220001	177.525894

As the experiment mentioned in the previous paragraph set an epoch to one on our first try, to save the training time, it can also have a risk of causing a low performance. With an increased number of epochs, it is expected that the model to perform better. However, this is not true according to our experiment.

It can be seen in Fig. 2, that the training took 20 epochs but got a worse result. It seems like an overfitting has happened in this case, while the calculated loss continues to go down the result shows that the performance has a negative improvement. To compare the results generated by different tests, it is possible to carry out root mean square error for evaluation, which would be a common technique to measure the distance between predicted values and actual values. The root means square error for the one epoch try is 4.302, while it is more than 20 with the 20-epoch training. Even if we tried techniques such as dropout, still can't solve the problem.

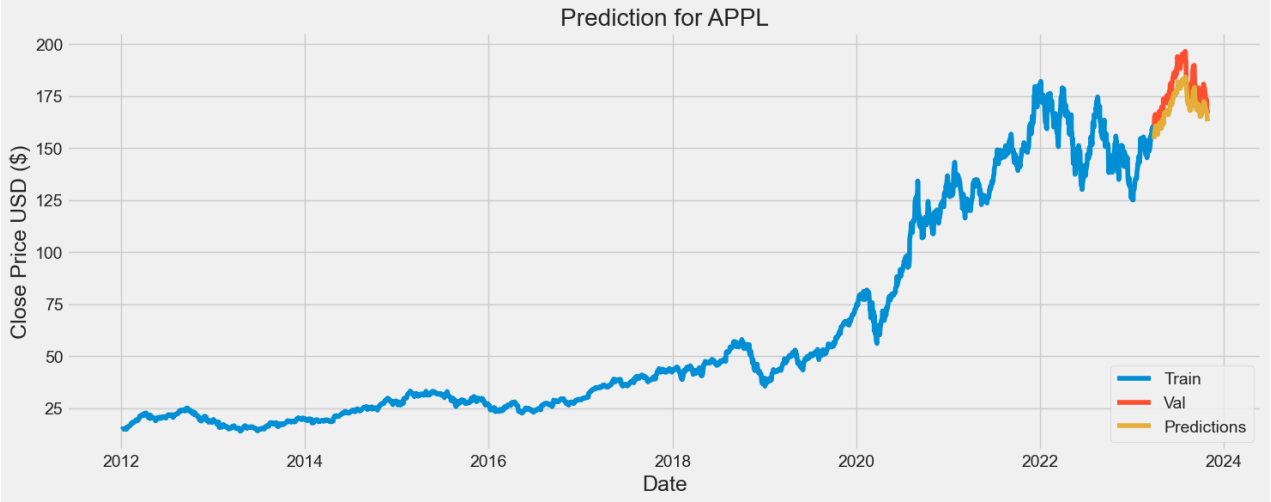

Figure 2. The Prediction Price of AAPL with a batch size of twenty.

4. Discussion

This experiment shows a rough estimation of the stock price using the machine learning technique LSTM, and the result shows that under this specific model structure, training using one epoch with a batch size of one appears to have a better performance than further training ones. Considering that an error within a few dollars can still make a big change to the investors' decisions, there is still much that needs to be improved. There might be more problems other than overfitting that have already been recognized, and besides overfitting could be caused by multiple reasons such as insufficient feature representation which is caused by the model structure itself. Here are a few points that might help to enhance the accuracy and at the same time minor the mistakes.

For a general individual investor, this model might work well as they consider only the superficial phenomenon that could be reached on their own. Similarly, this model can only interpret the fundamental mechanism in the market due to the concise and clear structure. More machine learning techniques could be used to enrich the ability and understanding of the model, for example, the Convolution Attention Block Module, also known as CBAM [6]. CBAM can be recognized as an improvement of SENet[7], while SENet focuses on recalibrating channel-wise feature responses, CBAM tries to consider spatial attention, which means it recalibrates the spatial distribution of each channel. This feature allows CBAM to adaptively adjust the importance of each feature map based on both the inter-channel and spatial relationships, which could lead to a more accurate and robust feature representation that we needed in this study.

Another thing to note is that real investors will usually do lots of investment analysis, including the financial statements of companies, profit margin, trading volume, or P/E ratio to decide whether a certain company is going to gain profit in the following years, or if the stock is worth to be kept other than sale it out. The stock market is such a complex composition of all kinds of numbers and factors involved. It is a game between people and strategies, but not a game for simple numbers. Analyzing the closing price in this article is not sufficient enough for a suggestion, and indeed also not enough to do forecasting. Further improvements could be focusing on the elaboration of input data, not restricted to the closing price, but can also be related to numbers like daily return, the

correlation between stocks in specific clusters, or volume of sales. With more data to be used during the analysis, a closer prediction could be made using a similar machine-learning technique.

5. Conclusion

This paper builds a model based on LSTM to predict the stock closing price of specific stocks, the ones that have a significant impact on a broad market index. This article uses data from the actual market, the trending for five large companies in NASDAQ, to predict the future price for these five stocks. The predictions can show a rough trend about how the price will change, with an error of only a few dollars. It could not be integrated into an investment suggestion as any mistakes during a decision can cause a huge loss, much more than what numbers could show on the screen. The main reason for a low performance, as shown in the discussion part, would be a simple model that saves time for training eliminates too much feature representation inside the model itself, and lacks different types of input variables that might also help or manipulate the trade. However, this result sends a good signal to researchers and investors that these trends are not random or capricious, but with traces that could be followed. Starting from the basic ideas in this article, the task of forecasting the price is still worth talking about, able to improve, and possible to deal with problems in the real world.

References

[1] Jinan Zou, Qingying Zhao, Yang Jiao, Haiyao Cao, Yanxi Liu, Qingsen Yan, Ehsan Abbasnejad, Lingqiao Liu, Javen Qinfeng Shi. Stock Market Prediction via Deep Learning Techniques: A Survey.2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), 24 Dec 2022.

[2] Sohrab Mokhtari, Kang K. Yen, Jin Liu. Effectiveness of Artificial Intelligence in Stock Market Prediction based on Machine Learning. nternational Journal of Computer Applications, 30 Jun 202. 1

[3] Nusrat Rouf, Majid Bashir Malik, Tasleem Arif, Sparsh Sharma, Saurabh Singh, Satyabrata Aich, Hee-Cheol Kim. Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. Electronics 2021, 10 (21): 2717.

[4] Jan, C.-L. Detection of Financial Statement Fraud Using Deep Learning for Sustainable Development of Capital Markets under Information Asymmetry. Sustainability 2021, 13: 9879.

[5] Mahdy Shirdel, Reza Asadi, Duc Do, Micheal Hintlian. Deep Learning with Kernel Flow Regularization for Time Series Forecasting. arXiv: 2109. 11649 [cs. LG], 23 Sep 2021.

[6] Sanghyun Woo, Jongchan Park, Joon-Young Lee & Kweon Inso. CBAM: Convolutional Block Attention Module. In Computer Vision – ECCV.European Conference on Computer Vision, 2018, 11211: 3 - 19.

[7] Hu Jie, Shen Li, Albanie Samuel. Squeeze-and-Excitation Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 7132 - 7141.