Improving Machine Learning's Performance in Predicting Stock Price in Unexpected Situations

Dingjun Wang

School of Business and Management
Shanghai International Studies University
Shanghai, China
0201147073@shisu.edu.cn

Abstract
Machine learning, known as deep learning, enables computers to learn from data sets, and more importantly, to think and make decisions like humans. In latest years, this idea has become rather significant in the field of finance for being capable of handling complex tasks, such as predicting stock price. Although various algorithms have been designed to make machines learn better and decide better, problems still exist. Particularly, in the case of predicting stock price, there are still factors that have been ignored by researchers. For instance, stockholders may have miss cognition over the market when great events occur. However, that is not an impossible problem to solve. By implementing the algorithm and the logic of processing the data, factors that used to be considered unmeasurable can be cognized by machines. We will further introduce this in the paper. In addition, during the research, we majorly used the Scikit-Learn library of Python and particularly the SVM regression method.

Keywords: *Machine Learning; Python, Stock Price; Scikit-Learn; Data Dividing; Regression; Support Vector Machine*

1. INTRODUCTION

Machine Learning and Deep Learning are concepts that have been popular for a few decades. With the fast development of Computer Science and the implementation of computing devices, these concepts, which aim at offering the human possibility to solve complicated problems without being explicitly instructed, now have great vision. In most fields that require computing and mathematical analysis, machine learning could be applied to handle various tasks. The application of machine learning has already been an important part of our lives. For instance, most internet video platforms have adapted machine learning to analyze their users to send users videos that they might prefer to watch.

Except for application in user analysis, machine learning has also been considered useful in research involving finance. One typical example, which is also a major issue that we would like to research, is predicting stock price by machine learning using tools in the python library.

Due to the overwhelming data scale, predicting stock price has been long considered impossible. What's more, Based on Efficient Market Hypothesis[1], stock prices can't be predicted. Last century, some mathematicians proposed various math functions to roughly predict stock price. Unfortunately, the predictions can't be called a success. Even though they considered more factors, they still can not further improve the accuracy. But now, the idea of machine learning has made the huge gap smaller and smaller. Machine Learning, composed of various algorithms, enables computers to find intuitive information by carrying on repeated learning tasks. This means the huge data that is considered impossible to handle by a human is pretty much an easy task from the perspective of artificial intelligence.

This feature makes predicting stock prices no longer an illusion. The exact reason people regard predicting stock prices as impossible is the massive data that may be involved in learning the fluctuations in the stock market. And that is also what machine learning is good at.

To make the prediction accurate, it is rather essential to choose the appropriate algorithm. This involves understanding both computer science and finance. From researching papers on that topic, we find that many of the teams put their most focus on computer science: data,

Y. Jiang et al. (Eds.): ICEDBC 2022, AEBMR 225, pp. 1509–1514, 2022.
https://doi.org/10.2991/978-94-6463-036-7_224

function, relationships between factors[2]. They manage to balance all the parameters in their functions and try to achieve the best accuracy in one single ideal model.

However, the stock price is not determined only by the factors in the stock market but also by factors in numerous industries. At present. there are already some researches that focus on the general framework of how machine learning could be used to predict stock. But to when outer forces influence the stock market, like pandemics, political issues, environmental events, war, etc., machine learning would always collapse into failure because none of these influences are measurable in a function. To make machine learning a truly useful tool in real life, where outer factors on markets can never be avoided, such influences should not be neglected. Instead, even though we can't take these abstract factors into our algorithms, we still can figure out ways to refer to them. Dividing data sets that are influenced by these factors from those not involved is one of the best solutions. By further implementing the specific algorithm on the divided data set, we could easily manipulate the logic of how our computers run the learning process[3]. These outer influences would no longer be a barrier for learning machines. This would be further explained in the paper.

In this paper, we will first make a brief introduction to the logic of Machine Learning and its general framework for predicting stock prices. Afterward, we would propose possible ways to make machine learning focus not only on calibrating the parameters in functions but also on the outer factors that usually could not be seen in any function.

2. RELATED WORKS

Machine Learning and Deep Learning are concepts that have been popular for a few decades. With the fast development of Computer Science and the implementation of computing devices, these concepts, which aim at offering the human possibility to solve complicated problems, now have witnessed brilliant success in both academic areas and real-life applications.

Predicting stock price was one of those impossible problems. Researchers have proposed various kernels under various methods to solve problems like this. One of the most commonly used methods in the python library is SVM or SVR for simply the regression problem Under the SVR method, common kernels like Radial Basis Function, Polynomial, Linear, Sigmoid are all useful in different cases[4]. These functions focus on revealing the logic behind the facts through balancing parameters in functions. On most occasions, they can achieve accuracy higher than 90%.

However, the stock price is not determined only by the factors existing in the stock market, in other words, the data itself, but also by factors in numerous industries. At present, researchers that focus on the general framework of how Machine Learning could be used to predict stock find it hard to achieve higher accuracy. In our opinion, to achieve better accuracy and make this concept truly appliable in the stock market, we also focus on factors beyond the data.

3. METHOD

3.1 Data Set

Since it is Force Majeure that we are concerning, first we need to search for data on the opening stock price of a certain company in a certain period that is distinctly influenced by outer factors to test our model. Besides, we also need to specify what exact kind of outer factor to be our consideration.

Given that Covid-19 has applied a great impact on our world for a rather long time, taking this tremendous pandemic as a research background would make our conclusions typical and realistic. Meanwhile, it is favorable to forecast stock price to analyze data from a universal company as the stock price is considered most likely to be influenced by the pandemic.

Eventually, we chose the stock price of Apple Company during the time from 21st Jan. 2020 to 17th Dec. 2020 as the training set and also the predictive target. To explain the choice of the time, the period was exactly when the first great pandemic wave hit the global and American markets. Afterward, the world witnessed a growing and overwhelming influence of covid-19 to the stock market.

3.2 Comparison

When predicting stock price, researchers usually use supervised machine learning to carry on the learning task. As unsupervised machine learning tends to be clustering and random, which doesn't fit our needs. And for supervised machine learning, one of the most efficient and commonly used methods to solve regression problems is Supper Vector Regression, abbreviated as SVR. Specifically, in python SVR, there are 5 types of algorithms or kernels. The default and most used function of the SVR method is Radial Basis Function[5]. We'll choose RBF to take the comparison forecasting.

To verify whether our proposed model truly works better when global incidents occur, we would make two predictions with data from the same company. One prediction would use the traditional way to train, which makes the learning machine analyze the full data set. While another prediction would instead make computer learning from divided data. A different division of the data is divided based on how severe the pandemic was in America at that time. This would make the machine better understand that the logic behind the market greatly changes when the pandemic happens[6]. At last, we

would compare the general performance of the two predictions.

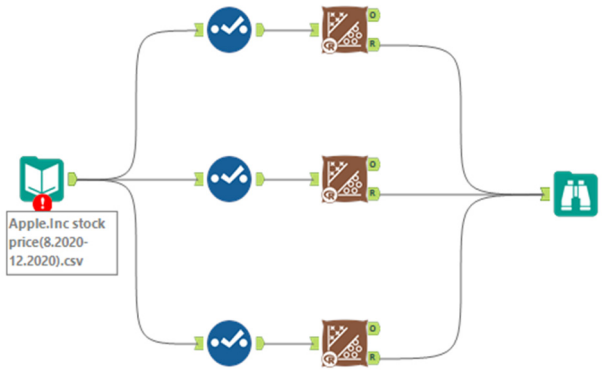

Figure 1. Dividing Data

3.3 Processing the Data.

As described, we first manage to train the learning machine with the full data set and made a prediction. For the divided data set, corresponding to the situation of covid-19 all over the year, we chose April and August for knots and divided the whole year into three periods.

According to statistics from New York Times, American Market first witnessed an increment in pandemic from January to April. Correspondingly, the stock price of Apple company dropped by approximately 25% and reached its lowest point around the 4th.April.After April, the daily increase of covid remained a stable value and made the market regain confidence. But in August, the pandemic again flooded American domestic land, and so the market was impacted again. From August to December, the number of new cases fluctuated and so did the stock price[7].

We can see from statistics that the market behaved differently in that three periods[8]. Mixing the data all together would rather confuse the computer and surely lower the accuracy of prediction. This is the very reason for us to divide data up.

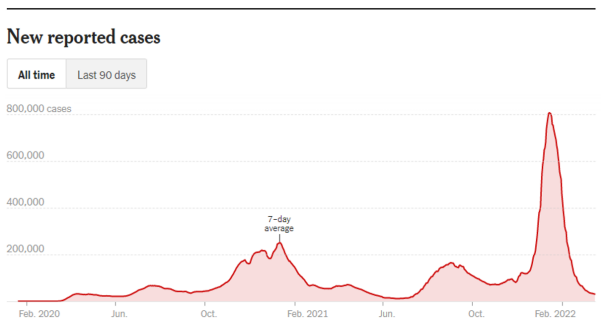

Figure 2. Pandemic statistics in America

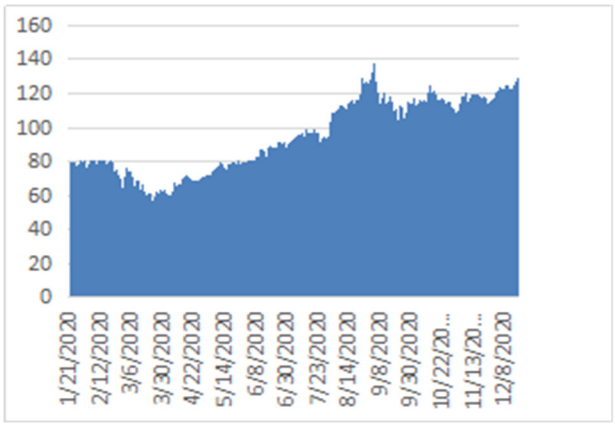

Figure 3. Stock Price of Apple Inc. in 2020

3.4 Prediction Performance

We'd name the prediction that analyzed the general data as Prediction A, and the three sub-predictions based on divided data as prediction B.1, B.2, and B.3. To compare performance, we would take Root Mean Square Error, R-Squared, Mean Absolute Error, and Median Absolute Error of the two models as standard[9]. Also, we would consider the residual plot of the difference between the predicted value and the real value.

Table 1. General Performance of Model A and B

Performance	Measurement			
	Root Mean Square Error	R Squared	Mean Absolute Error	Median Absolute Error
A	20.56022	0.01599	18.23874	20.07188
B.1	6.56235	0.02789	5.49011	6.68611
B.2	8.36076	0.05185	6.86798	6.02000
B.3	7.13547	0.01485	4.57807	2.69324

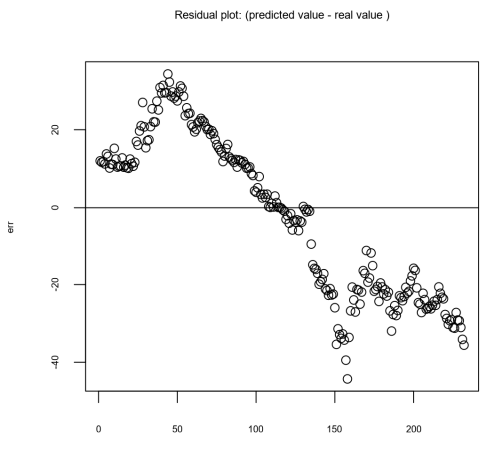

Figure 4. Residual plot of A and B

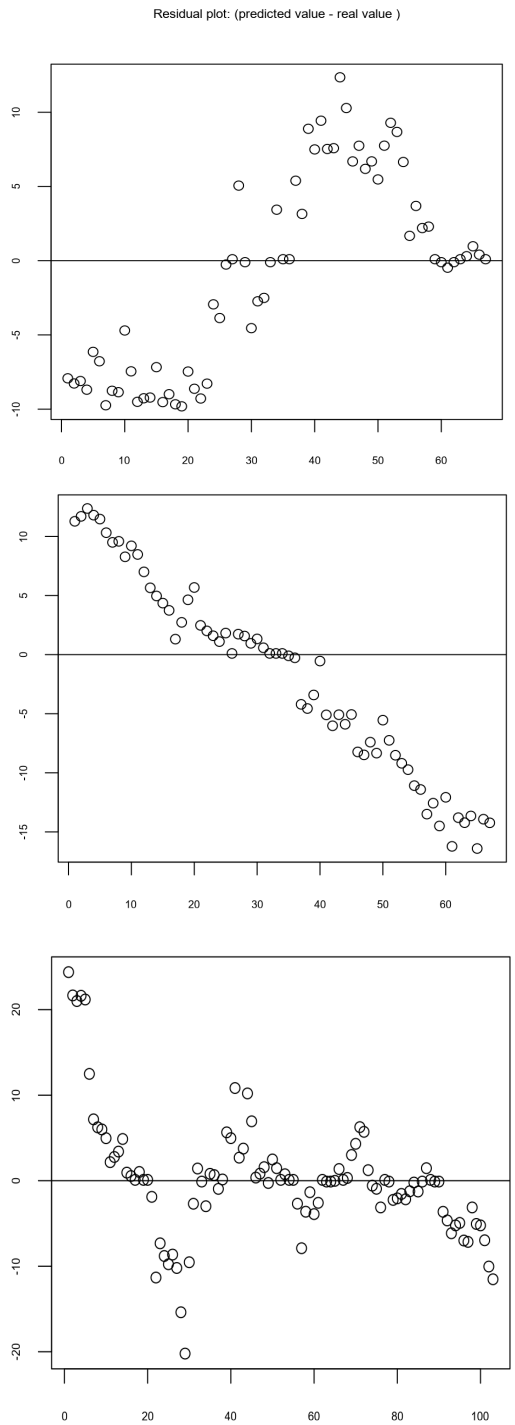

Figure 5. General Residual Plot of Model A

4. Discussion

Recent research has found that in the case of uncontrolled multiple crises, such as war, pandemic, or disaster, machine learning would lose its accuracy when making a prediction, especially when predicting stock price. Now, the stock market facing a greater impact from these influences and enduring bigger fluctuations, improvements have to be made to achieve a better application of machine learning in the future.

Talking of the stock market, economists prefer to follow the assumption that participants in the market would react to fluctuations in the most efficient way. However, we have to cognize that in real life, fear and other emotions would pretty much drive people to take reasonable reactions. Thus, when using computers to predict stock price, we have to help learning machines consider that. Our method is to divide data mined from the market and categorized it from the market behavior. For example, we would mark data that unusually increased as group A. And when we meet a similar trend while making a prediction, we would train the learning machine with data A.

Here is the final performance of the two models in our experiment. In model A, we trained the machine with a whole set of data. In B, we divided the data according to how the market behaved during the pandemic and made three predictions on different stock price figures. Compared with model A, we can see that model B has better performance in almost every aspect and the improvement is rather significant. The logic behind the success was clear: Cutting the giant task into small cubes is way easier for the machine to handle[10].

However, we are also aware that our method of processing the data has its limitation. First, dividing data would probably break the inner connection between data sets. For instance, sample point #a may play important role in influencing the upcoming trend, but such a relationship is cut off if data is divided. Second, at what exact point should we consider the joint between divisions? By dividing data, we have to make sure that the market behaved similarly during that period, but that would require a heavy job to achieve accuracy[11].

5. Conclusion

In this research, we attempted to attain better accuracy in stock price prediction using machine learning under international crises. We found that researchers tend to train machines without processing the data, which makes predictions lack accuracy, especially under global issues. Upon that, we generated the method of further dividing the data from the target period to make the machine better understand the logic behind it.

After the experiment, we confirmed that in certain conditions, our method does greatly improve the final performance. Since the global crisis is applying a bigger impact on the stock market, we believe that our study would achieve considerable value in the future. Furthermore, we would leave how exactly dividing data effect prediction results and whether there is a better method to improve accuracy as future work.

REFERENCES

[1] B G Malkiel. "Efficient market hypothesis." *Finance*.Palgrave Macmillan, London, 1989. 127-134.

[2] P Sodhi, N Awasthi, V Sharma, "Introduction to machine learning and its basic application in python, "Proceedings of 10th International Conference on Digital Strategies for Organizational Success. 2019.

[3] J I Larsen., "Predicting stock prices using technical analysis and machine learning," Institutt for datateknikk og informasjonsvitenskap, 2010.

[4] B Weng, L Lu, X Wang, et al., "Predicting short-term stock prices using ensemble methods and online data sources". Expert Systems with Applications, 2018, 112: 258-273.

[5] S Tiwari, A Bharadwaj, et al., "Stock price prediction using data analytics".2017 International Conference on Advances in Computing, Communication and Control (ICAC3). IEEE, 2017: 1-5.

[6] Y Song, W Jae, and J Lee. "A study on novel filtering and relationship between input-features and target-vectors in a deep learning model for stock price prediction." *Applied Intelligence* 49.3 (2019): 897-911.

[7] S Ramelli, and F Alexander. "Feverish stock price reactions to COVID-19." *The Review of Corporate Finance Studies* 9.3 (2020): 622-655.

[8] D Karmiani, et al. "Comparison of predictive algorithms: backpropagation, SVM, LSTM and Kalman Filter for stock market." *2019 Amity International Conference on Artificial Intelligence (AICAI)*. IEEE, 2019.

[9] W Budiharto. "Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM)." *Journal of big data* 8.1 (2021): 1-9.

[10] S Mehtab, J Sen. "Stock price prediction using convolutional neural networks on a multivariate timeseries."*arXiv preprint arXiv:2001.09769* (2020).

[11] S Tiwari, B Akshay, and Sudha Gupta. "Stock price prediction using data analytics." *2017 International Conference on Advances in Computing. Communication, and Control (ICAC3)*. IEEE, 2017.