INTI JOURNAL | Vol.2023:35
eISSN:2600-7320

Predicting Stock Prices Using Data Mining Technique

Thuy Nguyen Thi Thu[1*], Thi-Lich Nghiem[1*]

[1]Thuongmai University, Hanoi, Vietnam

[*]**Email**: thuynguyenthithu@tmu.edu.vn, lichnt72@tmu.edu.vn

Abstract

The stock market, for a long time, has been known as a complicated yet captivating system. It is a mainstream investment platform for both beginners and financially savvy people to grow and hold their assets. While it remains a good way to earn profit, the stock market is often considered as one of the risky approaches, mostly due to the nature of the field, and an enormous number of various factors that not often welcome the naïve investors. Therefore, the demand for using a tool that can support us on an overall view of the market trends, facilitating the financial analysis and strategies to identify the optimal time to purchase stocks and the actual stocks to purchase has risen for many years recently. In this study, we focus on using data mining techniques that can support investors in predicting the stock price with existing data from previous phrases. Given data is taken from Yahoo Finance within the 7-year period from 2015-2022. This data will be used to train the algorithms, then we can decide which one is the most suitable for the data mining tools to give the best suggestions for investors.

Keywords
Data Mining, Predicting techniques, Machine Learning

Introduction

The stock market has been known as a complicated predicting market. This is because there is a mainstream investment platform coming from investors who want to grow and hold their assets. The stock market also is considered as challenging investing approaches as its nature of the field as well as many alternative factors affecting to the investor's decisions. Therefore, the topic of predicting stock prices in stock market always is interested by many researchers.

Machine learning can be seen as a combination discipline between the computer algorithms and alternative statistical models (Zhang et al., 2019; Yeturu, 2020). Its applications contributed to many fields in society, particular to the financial sector. Machine learning in finance is also known as the tools to predict future capital market price trends. For example, to predict the stock price with the stock price sub-correlation in New York Stock Exchange market, they applied the ARIMA model and the LSTM model to experiment and evaluate the accuracy (Xiao & Jinxia, 2022). The results show that the LSTM model performs better in prediction than the ARIMA

Submission: 4 August 2023; **Acceptance:** 8 August 2023

model. In addition, the authors also propose an integrated model of ARMIA-LSTM and show that it is significantly superior to other methods (Xiao & Jinxia, 2022). To predict the future trend of NIFTY 50 share price of National Stock Exchange of India, they used deep learning networks including RNN, LSTM and CNN for analysis. The results show that LSTM performs better than other models (Fathali eta al., 2022). In the NSE stock market prediction is researched by deep learning models (MLP, CNN, LSTM, ARIMA). Other research shows that the obtained results are compared with the ARIMA model, and it found that CNN performs better than the existing model (ARIMA) (Hiransha et al., 2018).

In this paper, the alternative machine learning algorithms such as Extreme Machine Learning (EML) (Cen et al., 2017), Long Short-Term Memory (LSTM) (Greff et al., 2016), and Random Forest (RF) (Breiman, 2001) are used to predict stock prices. The used data is taken from Yahoo Finance within the 7-year period from 2015-2022 with Apple Inc. (code of AAPL). By dividing into training, validating and testing datasets, the machine learning algorithms will be used in order to choose the most suitable one for investors.

Machine learning Algorithms

To decide which is most suitable one for investors in predicting stock prices, alternative machine learning algorithms are used as follows.

Extreme Machine Learning (EML)

The extreme learning machine (EML) has a learning process in where the learning types including batch, sequential, and incremental learning. The advantages of using them because of their benefits for speed, fast convergence, the ability and implementation easily. The EMLs also can be seen as a kind of feed-forward neural networks. Their structure contains a single layer or multiple layers of hidden nodes. Their outputs are resulted from learning process where the hidden node parameters stayed the same during the process. The starting hidden nodes can be setup randomly or inherited from their predecessors. Their weights are learned in the fast learning scheme. Therefore, the EML has a good generalization performance as its learning time quicker than other networks such as backpropagation ones. The EML architect can be seen in Figure 1.

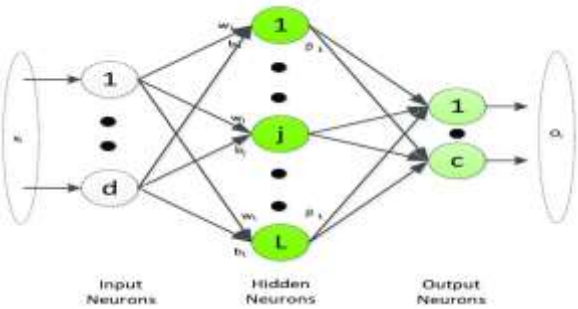

Figure 1: The architecture of the EML

Long Short-Term Memory (LSTM)

Long short-term memory (LSTM) network (Greff et al., 2017) can be seen as a high advantage of the recurrent neural network. The advantage of this network is seen in its structure with the "gate" and the "cell" units. The gates can be used in the time steps for capturing the long-term memory and short-term memory. This can avoid gradient exploding or vanishing problems which can be seen in standard recurrent neural networks. The gates in the LSTM are also known to three gates as forget gate, input, and output one. The LSTM architect can be seen in Figure 2.

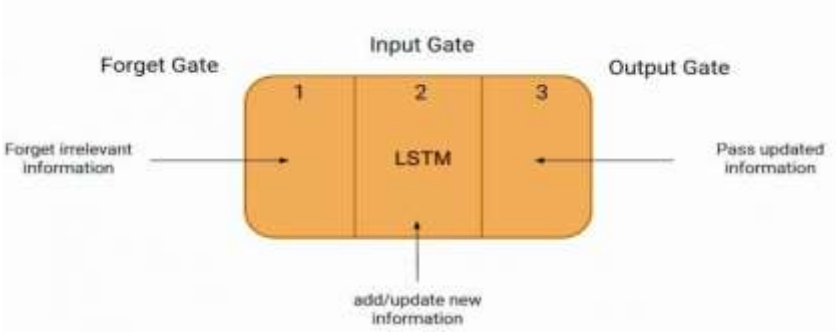

Figure 2: The architecture of the LSTM

Random Forest

A supervised learning algorithm is random forest - an ensemble of decision trees, often trained using the "bagging" approach, to make up the "forest" that it constructs. The bagging method's main premise is that combining learning models improves the end outcome. The process can be seen in Figure 3.

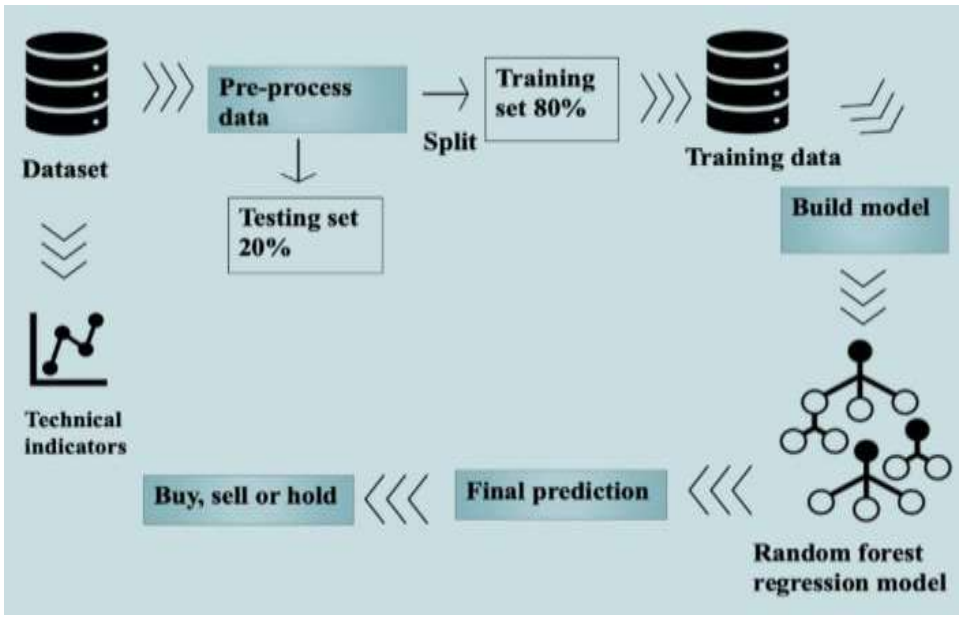

Figure 3: RF process

The random forest algorithm's steps are as follows:

Step 1: From a data collection with k records, n random records are selected at random and used in the Random Forest algorithm.

Step 2: For each record set, a unique decision tree is built.

Step 3: An output will be produced by each decision tree.

Step 4: For the outcomes of classification and regression, the final result is evaluated basing on a majority vote or an average.

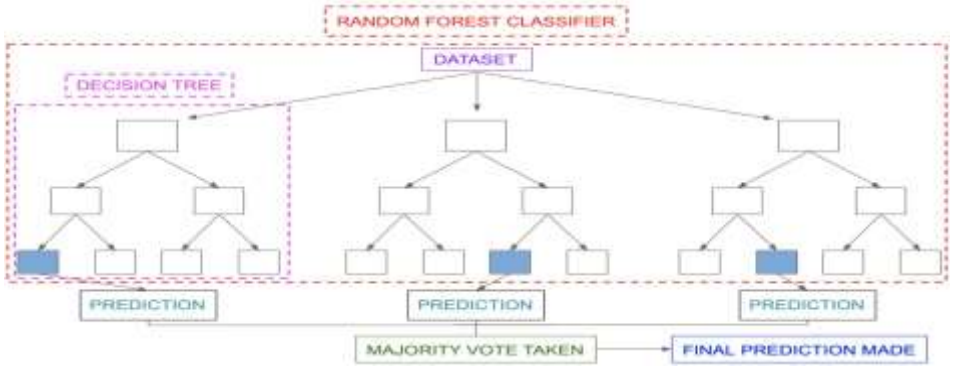

Figure 4: The architecture of the RF

Experiments

Data: The Data source is extracted from Yahoo website (finance.yahoo.com) with YFinance tool for Apple Inc. (AAPL). The duration data is 7 years (2015-2022).

Algorithms: Stock prediction is considered a regression problem since it requires numerical prediction. The algorithms are used here as extreme learning, long short term memory and random forest.

The inputs: The input attributes are taken from historical data in Yahoo websites. They are: Date, Open price, High, Low, Close Price, Adj Close Price, and Volume

The output: Predicting Stock Close Price

Results

To visualize the predicted prices, both predicted stock price and the real stock price are shown in graph by using the Python Matplotlib. From the resulting chart above, it is clear to show that the trend of the real stock prices closely. This shows the effectiveness of using LSTM to work with time series or sequential data like stock prices. The historical data is used for predicting Apple's stock price in 2023 with LSTM algorithm in Figure 6. This shows that the price has been slightly and fluctuant decreased compared to the one in 2022.

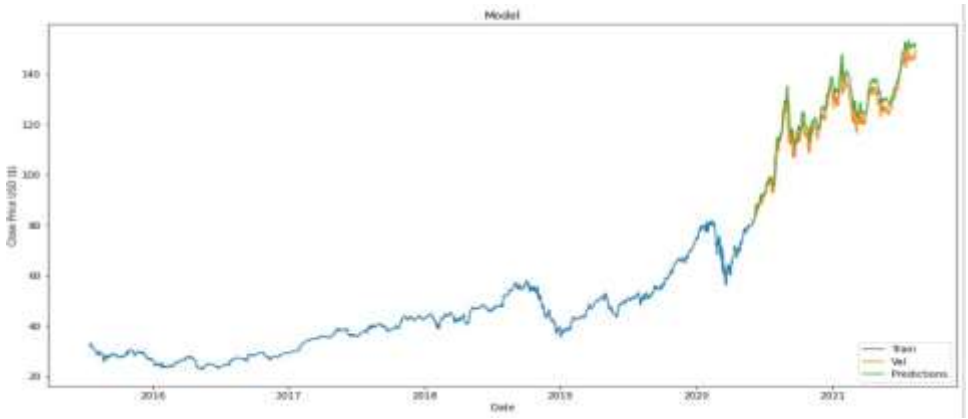

Figure 5: Visualized predicted Stock Price by using LSTM.

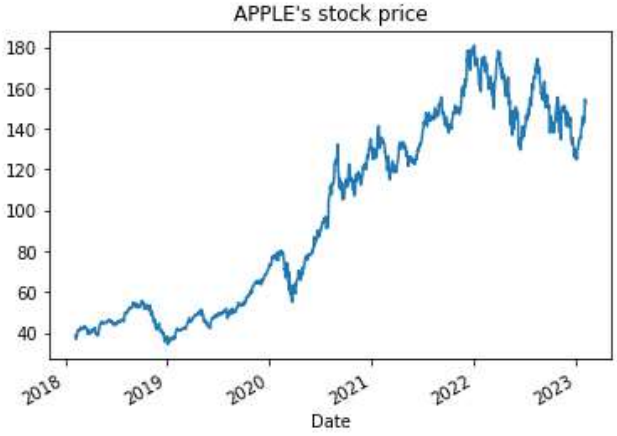

Figure 6: Predict Apple's stock price using LSTM algorithm

To compare the advantage of using alternative machine learning algorithms the alternative measured factors are used such as MAE, MSE, RMSE, R^2 score and accuracy rate. Detail values of comparisons can be seen in Table 1.

Table 1: Measure the accuracy of the model

	MAE	**MSE**	**RMSE**	**R^2_Score**	**Accuracy**
LSTM	**0.0545**	**0.0378**	**0.6148**	**0.8532**	**83.1%**
EML	**0.0396**	**0.0245**	**0.0493**	**0.876**	**91.56%**
RF	**0.0771**	**0.0273**	**0.1652**	**1.0**	**99.87%**

From table 1, RF algorithm has the highest accuracy rate with 99.87%. The random forest model is created following the conventional approach. The RF settings are used as default of RF algorithm as it is believed that it can have enough capacity for the data domain including financial dataset.

The constructing process of RF shows that for each decision tree, the randomly generated subset is used as a sample of training data. Then, a decision tree can be built based on the sub-random forest features. By repeating many times of using random forest algorithm for original dataset, a voted output is produced.

Conclusion

The experimental results clearly show the effectiveness of the proposal to use machine learning in forecasting time series data. The method of analyzing and processing data before entering the model has improved the normalization of data, making it more suitable for the model and minimizing the influence of noise, and enhancing uniformity. The paper also used the open gate forecasting model of stocks and showed that the time series data model is highly effective, especially when the data is not regular. Experimental results show that the RF model with flexible architecture has the potential in predicting trends. Therefore, it has the highest accuracy. From that, it can be concluded that using machine learning model in forecasting time series data is an effective and potential method. The stock price data need to be consider with other factors such as economical issues, or trend of sale, etc. This means that other effecting factors should be collected with historical stock price data, and more parameters should be added into the models during training process. All these should be left for further research.

References

Breiman, L. (2001). Random Forests. Mach. Learn. 45 (1), 5–32. doi:10.1023/A:1010933404324

Cen C., Li K., Duan M., & Li K. (2017). Chapter 6 - Extreme Learning Machine and Its Applications in Big Data Processing. Intelligent Data-Centric Systems. Pages:117-150

Fathali, Z., Zahra K., & Lamjed B. S. (2022). Stock Market Prediction of NIFTY 50 Index Applying Machine Learning Techniques. Applied Artificial Intelligence. 36 (1).

Greff K., Srivastava R. K., Koutník J., Bas R. Steunebrink, & Schmidhuber J. (2017). LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks and Learning Systems. Vol 28 (10). Pages: 2222 - 2232

Hiransha M, Gopalakrishnan E.A., Vijay Krishna Menon, & Soman K.P. (2018). NSE stock market prediction using deep-learning models. Procedia computer science 132. Pages: 1351-1362.

Xiao, D., & Jinxia S. (2022). Research on Stock Price Time Series Prediction Based on Deep Learning and Autoregressive Integrated Moving Average. Scientific Programming 2022.

Yeturu K. (2020). Chapter 3 - Machine learning algorithms, applications, and practices in data science. Handbook of Statistics, Vol. 43. Pages: 81-206.

Zhang, F., Liu, J., Wang, B., Qi, Z. & Shi, Y. (2019). A Fast Algorithm for Multi-Class Learning from Label Proportions. Electronics. Vol. 8 (6). https://doi.org/10.3390/electronics8060609