The Investigation of Stock Price Prediction Based on Machine Learning

Boming Huang *

Department of Physics, Xian Jiaotong University, Xian, China

* Corresponding author: 2101280325@stu.xjtu.edu.cn

Abstract. As the global economy experiences unprecedented growth, the stock market is undergoing a period of expansion and prosperity. Driven by the allure of high risks and high returns, a significant number of individuals have committed themselves to stock trading. In the age of the Internet of Everything, the trajectory of the stock market is intertwined not only with the fortunes of individuals and businesses but also deeply influences the fate of nations and societies. However, stock market forecasting is quite a hard task owing to its volatile nature, where even seemingly unrelated factors can lead to dramatic changes. Therefore, the importance of constructing reliable stock prediction methods is self-evident. In recent years, scholars have mainly focused on machine learning and deep learning methods to complete this task. After literature review, this article summarizes and discussed several commonly used and effective methods, such as Linear Regression, Random Forest, Long Short-Term Memory (LSTM), etc. The advantages, disadvantages and applications of different methods are discussed, like the shortcoming of Linear Regression to handle non-linear models well and the relatively excellent ability of LSTM to predict long sequences. This article would be helpful for emerging scholars to have a grasp of the characteristics, advantages and scope of application of these methods, and facilitate the development of future research.

Keywords: Machine Learning, Stock Price Prediction, Linear Regression, Recurrent Neural Network.

1. Introduction

The economy is a cosmopolitan issue that influences the political situations in various regions and the well-being of people around the world. In different financial systems, the stock markets are always regarded as a determinant component, reflecting its boom and bust. Due to the existence of compound interest and the company's natural economic advantages, owning stocks has the potential to bring huge profits, attracting a large number of investors to participate. Stock price prediction is an exceptionally challenging task due to the influence of a myriad of intricate factors. These factors encompass the operational conditions of companies, the repercussions of macroeconomic policies, industry health, as well as the intricate interplay of psychological factors and information disparities. As a result, the predictability of stock prices becomes highly uncertain, giving rise to unforeseeable risks. Therefore, establishing a relatively accurate and effective stock prediction model to provide investors with a scientific reference has profound theoretical significance and important utilization value.

Over the years, investors have mainly adopted two traditional analysis methods to predict the value and trend of stock prices: Fundamental Analysis and Technical Analysis [1]. While they offer ease of use and possess a degree of reference significance, their significant limitations are unmistakable, largely stemming from the inherent hysteresis that cannot be avoided. Therefore, researchers were motivated to find better solutions for such a volatile non-linear system, and Time Series Forecasting caught their attention. Among related models, Autoregressive Integrated Moving Average (ARIMA) stood out with higher prediction accuracy [2].

At the end of the 20th century, Artificial Intelligence technology developed rapidly, and scholars began to use neural networks for stock prediction, pioneering this application [3]. Feedforward Neural Network (FNN) and Backpropagation (BP) were introduced and widely applied [4]. Investors found that these tools allowed them to swiftly analyze intricate and diverse data, leading to more satisfactory outcomes. Later on, problems of over-fitting, large number of parameters and difficulty in adjustment were observed, which inspired the proposal of Support Vector Machine (SVM) based on Empirical Risk Minimization (ERM) and Structural Risk Minimization (SRM) [5]. With the help of Kernel Function, this model achieved good results in generalization, avoiding local optimal and handling high dimension data.

Furthermore, benefiting from the improvement of data technology and computational power, scholars have started to focus on deep learning technology. It mainly includes the well-known Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) [6, 7]. In 1997, LSTM, a special kind of RNN, was published, solving the gradient vanishing and exploding problems [8]. In 2014, Gated Recurrent Units (GRUs), an LSTM-like mechanism with fewer parameters, was introduced. It performs better on less training data, but Bengio et al. came to no concrete conclusion on which of the two gating units was better [9]. Generally speaking, CNN is kind of more powerful than RNN for the ability to be stacked into a very deep model, although the sacrifice of generalization ability and vulnerability to adversarial attacks are also disputed.

Ensemble Learning is another promising research direction. Rather than relying on a single algorithm, it leverages multiple learning algorithms to achieve superior predictive performance compared to what any individual learning algorithm could achieve in isolation [10], and it also reduces the spread or dispersion of the predictions. Different Ensemble Learning Algorithms include RF (Random Forest), XG-Boost + LSTM, Blending Ensemble, E-SVR-RF, etc.

The rest of this article is organized as follows: Section 2 introduces a series of commonly used algorithms and corresponding basic principles in the field of stock price prediction. Section 3 discusses advantages and disadvantages of these algorithms. Finally, some possible research ideas are provided and the whole paper is concluded in Section 4.

2. Methodology

2.1. Linear Regression

Linear Regression aims to find a straight line that best represents the general trend of a sequence by estimating the unknown parameters of linear predictor function from the data. Its formulation, as is shown below, is rather simple.

$$y = X\beta + \varepsilon. \tag{1}$$

After reasonably preparing the data, approaches like Ordinary Least Squares are used to evaluate the result. If not satisfying, Gradient Descend with a proper learning rate can be used to optimize the model by updating the parameters in the direction towards minimizing the error.

In the work of Chauhan and Sharma [11], the researchers used Linear Regression Model to predict American stock market. The data authors have used is from National Stock Exchange of India (NSE) and yahoo finance. They applied three different algorithms, and concluded that Linear Regression has the highest accuracy. Another work presented by Shakhla et al came to similar conclusions [12].

2.2. Random Forest

Random Forest, as the name suggests, is an algorithm that integrates multiple decision trees using the bagging idea of ensemble learning, originating from and generally outperforming Decision Trees. It is a derivative of the bagging method and was improved by adding random selection of features, creating an uncorrelated forest of decision trees. The basic ideas can be summarized as Fig 1.

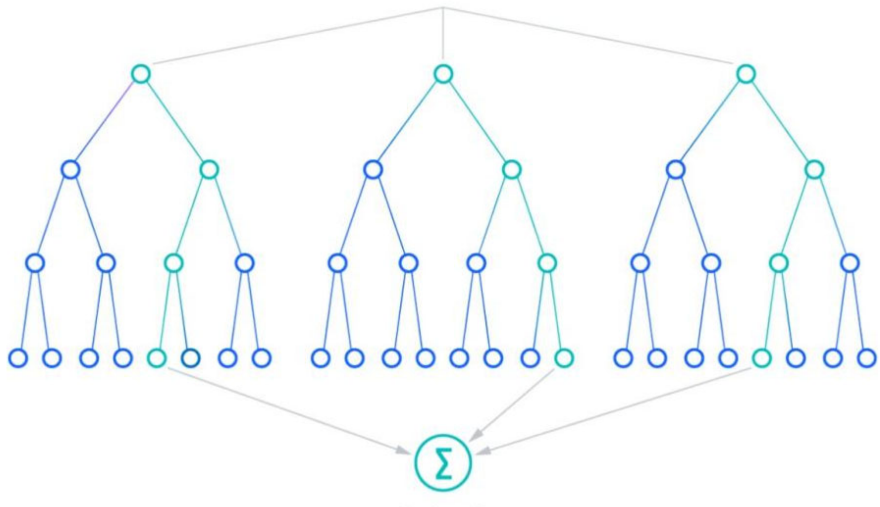

Figure 1. The structure of random forest [13].

In training, one third of the training sample is usually set aside as a test set. Values of cell states and hidden states can be updated with the help of forgetting and memorizing. The results will be passed through the activation function to get the output. Finally, the test set preserved is used for cross-validation, finalizing the prediction.

Loke K.S. applied Random Forest on data collected from Hong Kong Stock Exchange from 2011-2014 [14]. He found that the results were inconsistent, with some being satisfying and some very poor. Polamuri et al combined Random Forest and Extra Tree Regression, achieving the highest accuracy compared with other algorithms [15].

2.3. ARIMA

Borrowed from statistics, the ARIMA model can be applied to most stock series forecast based on historical prices solely. The task is usually fulfilled with analysis of relations of various values in stock markets. Its feature of pure statistical analysis gives it an edge to obtain satisfying accuracy while preserving model simplicity.

An ARIMA model can be divided into three procedures. Firstly, Autoregressive model captures the main trends in a long-term series and ignores temporary sudden changes. Next, the difference is taken d times to eliminate the impact of the overall trend. After that, a moving average model handles abnormal data changed that are noisy. Combining these mutually complementary models above gives the resulting ARIMA(p,d,q) model.

$$y_t = c + \varphi_1 y'_{t-1} + \varphi_2 y'_{t-2} + \ldots + \varphi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (2)$$

A. Adebiyi et al compared the results of ARIMA (p, d, q) Model on Zenith Bank Index and other valid data-set, demonstrating ARIMA's great potential to prediction on short-term basis [16]. Hybrid model of ARIMA and LightGBM was studied by Zheng et al, obtaining better results than solely using each model [17].

2.4. RNN

Previous neural network models can only process each input independently, and the two inputs before and after are not related. This makes them perform less than satisfactory on many tasks, including those related to stock forecasting. Therefore, the RNN model is developed to cope with sequence information better.

A simple RNN has an input layer, hidden layer and an output layer, as shown in Fig. 2. In each circulation, the output of the hidden layer is stored in its memory for subsequent prediction.

Theoretically, RNN can utilize information in arbitrarily long sequences, although in practical applications it usually considers the output of the first few cycles.

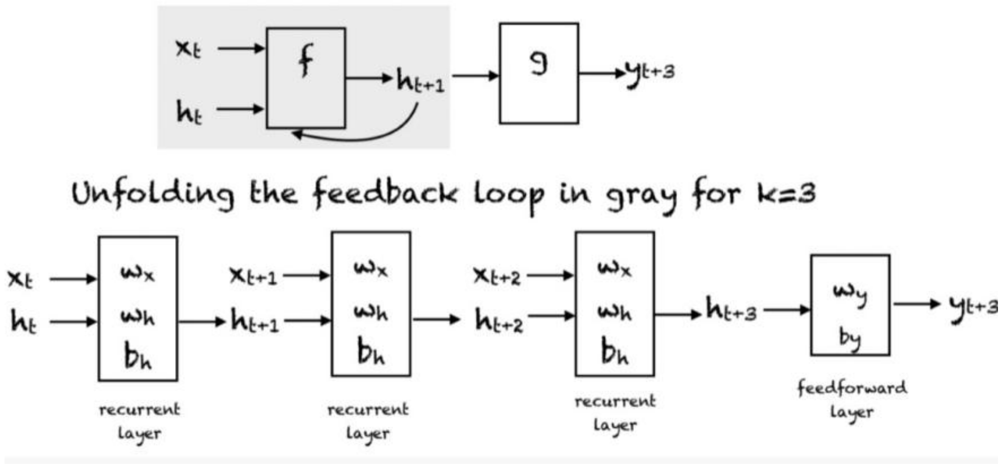

Figure 2. The structure of RNN [18].

Purohit et al trained a RNN model on Google data-set and found it capable enough in predicting the near future [19]. Dey et al carried out a comparative analysis of RNN in stock prediction, concluding that it performed best when the pattern was relatively gentle without obvious abnormal fluctuations [20].

2.5. LSTM

Long Short-Term Memory network is a recurrent neural network aiming to deal with the vanishing gradient problem presented in traditional RNNs. The relative sensitivity to time gives it an edge on learning patterns in time series over other machine learning models.

A common LSTM unit consists of one cell and three gates, responsible for input, output and forgetting respectively. The cell will store information at any time interval, and the three gates will help the cell process the information and control the corresponding input and output. Information is assessed for retention, storage, or output by assigning a value between 0 and 1 according to both past and current states. With these mechanisms, the LSTM network is able to preserve primary, enduring dependencies for making predictions in both present and future time steps.

C. Sullivan et al applied LSTM on data collected from Kaggle and conclude that simpler model performed better on this data [21]. Zou and Qu established the Attention-LSTM forecast framework based on LSTM, which has the ability to choose features automatically by weights assignment and outperforms other models [22].

3. Discussion

3.1. Linear Regression

As an algorithm borrowed from statistics, the basic idea of Linear Regression is to predict future sequences with characteristics of existing sequences. However, if some disordered attributes in the stock market are made serial, order relationship will be inappropriately introduced, causing interference to subsequent processing such as distance calculation. Moreover, Linear Regression Model has many basic assumptions like Linearity and weak exogeneity, which mean that predictor variables are supposedly error-free, and mean of response variable can be expressed linearly with parameters and predictor variables. Obviously, that’s not always the case when it comes to stock price, thus Linear Regression is not excellent in terms of performances.

3.2. Random Forest

As a classifier composed of multiple decision trees, its output is determined by voting. It overcomes the shortcoming of over-fitting of decision trees and reaches higher accuracy, while sacrificing the intrinsic interpretability. This sacrifice is not worth it in some cases because developers cannot track the training of the model, and credibility of this model for users is also reduces. Meanwhile, Random Forest may not serve as a better solution for problems in which multiple classification tasks are involved [23]. In addition, RF model has no memory, making it unable to take past stock prices and fluctuations into account with current inputs and outputs.

3.3. ARIMA

Generalized from ARMA, the ARIMA model stands out among many models because of its particularly simple structure, and constantly outperforms complex structural models in short-term prediction. On the other hand, similar to Random Forest, it lacks the ability to memorize. Since stock prices are usually greatly affected by past stock prices, this feature limits its good results to short-term investment and prediction.

3.4. RNN and LSTM

The RNN model is computationally powerful and can be used in many temporal processing models and applications, but as mentioned in Section 2, it has problems of vanishing or exploding gradient. More specifically, the gradient on a layer of RNN comes from the product of previous gradients, and too many layers will lead to inherent instability. LSTM, however, tackles this problem based on the delicately designed hidden layer nodes [24]. In a nutshell, LSTM is a non-linear model with the ability to learn patterns and memorize past prices during training and application, making it a more suitable and widely adopted model at present.

4. Conclusion

In this review several classical machine learning methods that have been widely employed in stock price prediction are investigated. This review provides a brief introduction to each method, talks about their respective advantages and disadvantages, and also describes the corresponding application scenarios. Overall, this article summarized the main literature in this field and provides a more comprehensive reference for subsequent scholars' research work and algorithm selection. At the same time, it can easily be seen from the discussion above that most of the current existing algorithms only focus on past prices during training and application, ignoring the fact that stock market itself is extremely susceptible to external conditions, especially those hard to quantify, like macroeconomic policies and international situations. These factors play a more decisive role in stock trend and contribute significantly to making stock prices more unstable and unpredictable. Therefore, more algorithms with the ability to take into account external factors are expected to emerge and achieve better prediction results.

References

[1] Haoran Xu, et al. A Review of the Application of Machine Learning in Stock Prediction. Compute Engineering and Application, 2020, 56 (12): 19 - 24.

[2] Emmanual Dava, Albert Leonardo, et al. Forecasting Indonesia Exports using a Hybrid Model ARIMA-LSTM. Procedia Computer Science, 2021, 179 (1): 480 - 487.

[3] Hasan, Syeda Shabnam, Rahman, et al. Improved Stock Price Prediction by Integrating Data Mining Algorithms and Technical Indicators: A Case Study on Dhaka Stock Exchange. Computational Collective Intelligence, ICCCI 2017, PTI, 2017, 10448 (1): 288 - 297.

[4] Arima, Shun'ichi. A Theory of Adaptive Pattern Classifier. IEEE Transactions, 1967, EC (16): 279 - 307.

[5] Vapnik, Lerner, et al. Learning System for Pattern Recognition Using Generalized Portraits. Engineering Cybernetics, 1965, 1 (1): 63 - &.

[6] Zhou, Xiaokang, Yue Li, and Wei Liang. CNN-RNN based intelligent recommendation for online medical pre-diagnosis support. IEEE/ACM Transactions on Computational Biology and Bioinformatics 18.3, 2020, 912 - 921.

[7] Qiu, Yuhang, et al. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control 72, 2022, 103323.

[8] Hochreiter, Schmid Huber. Long Short-Term Memory. Neural Computation, 1997, 9 (8): 1735 - 1780.

[9] Chung Junyoung, Benjio Yoshua, et al. Empirical Evaluation of Gated Recurrent Neural Network on Sequence Modeling. eprint arXiv: 1412. 3555.

[10] Gengguo Cheng, et al. Research on Stock Prediction Model based on Neural Network Ensemble Learning. Compute Engineering and Application, 2020, 55 (8): 238 - 243.

[11] Karan Chauhan, Nitin Sharma. Study of Linear Regression Prediction Model for American Stock Market Prediction. IEEE, 2017.

[12] Shruti Shakhla, Bhavya Shah, et al. Stock Price Trend Prediction Using Multiple Linear Regression. International Journal of Engineering Science Invention, 2018, 7 (10): 29 - 33.

[13] Random Forest https://www.linkedin.com/pulse/random-forest-kirubasagar-v?trk=public_post_feed-article-content 2023.

[14] Loke kar Seng. Impact of Financial Rations and Technical Analysis on Stock Price Prediction Using Random Forest. International Conference on Computer and Drone Application, 2017.

[15] Kudipudi Srinivas, et al. Stock Market Prices Prediction Using Random Forest and Extra Tree Regression. International Journal of Recent Technology and Engineering, 2019, 8 (3): 1224 - 1228.

[16] Ayodele Ariyo Adebiyi, et al. Stock Price Prediction Using the ARIMA Model. UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, 2014.

[17] Xiuyan Zheng, JiaJing Cai, Guangfu Zhang. Stock Trend Prediction Based on ARIMA-LightGBM Hybrid Model. 3rd Information Communication Technology Conference, 2022.

[18] An Introduction to Recurrent Neural Networks and the Math That Powers Them https://machinelearningmastery.com/an-introduction-to-recurrent-neural-networks-and-the-math-that-powers-them/ 2023.

[19] Avanish Kumar, Kaustubh Purohit, Krishan Kumar. Stock Price Prediction Using Neural Network and Long Short-Term Memory. International Conference on Deep Learning, Artificial Intelligence and Robotics, 2019, 153 - 160.

[20] Polash Dey, et al. Comparative Analysis of Recurrent Neural Networks in Stock Price Prediction forr Different Frequency Domains. MDPI, 2021, 14 (8): 251.

[21] Jason C. Sullivan. Stock Price Volatility Prediction with Long Short-Term Memory Neural Network. Stanford University, 2019.

[22] Zou, Zhichao, and Zihao Qu. Using lstm in stock prediction and quantitative trading. CS230: Deep Learning, Winter, 2020, 1 - 6.

[23] Madeh Piryonesi. The Application of Data Analytics to Asset Management: Deterioration and Climate Change Adaption in Ontario Roads. ProQuest LLC, 2019.

[24] Yilin Ma, Ruizhu Han, Xiaoling Fu. Stock Prediction Based on Random Forest and LSTM Neural Network. 19th International Conference on Control, Automation and Systems, 2019.