The progress and state-of-art applications of machine learning for stock price prediction

Yixin Gong *

School of International Studies, Zhejiang University, Hangzhou, China

*Corresponding author: 3190105926@zju.edu.cn

Abstract. Stock price is an unstable time series affected by plenty of factors. Since various noises have significant impacts on its trend, the way to realize more accurate forecasts in terms of multidimensional data features has become a concern for scholars worldwide. Among all the methods, machine learning approaches are one of the popular ideas in recent years. This paper introduces the meaning of stock price prediction and the development of machine learning in this field for the past few years. Theoretical background of Random Forest, XGBoost and LSTM are provided and the state-of-art researches based on the above methods are also summarized. It concludes with a discussion of these models and the limitations of this paper, as well as an outlook for future work. The study aims to synthesize the scattered sources of information for the reference of later scholars. As a result, human beings can find better ways to maximize investment benefits and warn of stock market crises in years to come. Overall, these results shed light on guiding further exploration of stick price forecasting.

Keywords: machine learning, stock price prediction, Random Forest, XGBoost, LSTM.

1. Introduction

Stock price is an unstable time series influenced by various factors. One would like to analyze the most important ones in order to maximize investment benefits and warn of stock market crises. Since stock price inherently contains noises during volatility, how to make more accurate forecasts with multidimensional data features has become a concern worldwide. Investors usually use traditional fundamental analysis, technical analysis and evolutionary analysis. These methods are too theoretical and cannot fully reflect the correlation between data [1]. With the turbulent international situation and recurring epidemics in recent years, stock price forecasting has become more difficult than ever. Contemporarily scenarios based on machine learning techniques are widely used in stock price prediction with its powerful algorithmic functions and learning ability.

Researchers have proposed different ensemble learning and deep learning models to observe stock price changes. Booth et al. proposed the trading system based on a performance-weighted random forest ensemble to improve the profitability [2]. Various regression techniques were analyzed and the merits for weighting were explored. Weng et al. incorporated features extracted from the Web to predict short-term stock prices [3], which showed that the use of these features did not replace traditional financial indicators. Nabipour et al. selected four stock market groups for experimental evaluation and compared nine machine learning models [4]. Jiang et al. take the technical factors into consideration for the sake of enhancing the accuracy of stock trend forecasting (especially the direction) [5]. Logistic regression was used as meta-classifiers in the second layer. Bao et al. constructed a complex composite model with wavelet transforms (WT), stacked autoencoders (SAEs) and LSTM [6]. The SAEs with hierarchical extraction of deep features were first introduced into stock price forecasting. Kim and Won proposed a new hybrid model based on the combination of LSTM model to various GARCH-type approaches [7]. To be specific, it combined excellent sequential pattern learning and significantly improves the forecasting performance.

A considerable amount of academic research has accumulated in this field. This study is an attempt to synthesize the key research insights and unveil major research trends for the reference of later scholars. The remaining part of the paper is as follows. The Section 2 provides the theoretical background of the methodologies mentioned. The Section 3 summarizes state-of-art Applications,

while the Section 4 discusses the limitations of this paper and forecasts the future research direction of stock price prediction. Finally, the Section 5 draws a conclusion.

2. Basic model description

2.1 Random Forest

In general, Random Forest is widely known as an ensemble approach for classification, regression, etc., which is a collection of many strong trees. Random Forest uses bootstrapping to take samples from the training set randomly with backtracking, and random features are selected for each tree based on bagging [8, 9]. Typically, the procedure can be described as follows

(1) The training set is separated to multi-subsets

(2) Some of the factors are stochastically selected.

(3) The subset is fitted and the optimal cut of the chosen factors are recorded, obtaining k trees result.

(4) Combining all the results by averaging (for a regression) or voting (for a classification).

2.2 XGBoost

Contemporarily, boosting methods is widely adopted due to the high efficiency and effectiveness [10]. Among various boosting approaches, XGBoost is a machine learning tool that are used mostly nowadays. The core of XGBoost is the optimization of the objective function [11]. The procedure for this model can be given as follows. To begin with, the objective function is

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

Based on Taylor expansion, one derives

$$obj^{(t)} \approx \sum_{i=1}^{n} l\left[\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) + C \tag{2}$$

$$obj^{(t)} = \sum_{j=1}^{t} \left[\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T \tag{3}$$

$$G_j = \sum_{i \in I_j} g_i \, , H_j = \sum_{i \in I_j} h_i \tag{4}$$

and introduced the Eq. (5) to Eq. (3), one obtains

$$obj^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{5}$$

where $l(y_i, \hat{y})$ is the training error of sample x_i, $\Omega(f_k)$ represents the regular term of the first tree, K represents the total number of trees, f_k represents k the first tree and C is a constant.

2.3 LSTM

The output values of RNN depend on the historic and current input sequences, which makes it suitable for predictive mining tasks [12]. LSTM is a special type of RNN model, which can store relevant information in both short and long term, avoiding the long-term dependency problem of traditional RNN [13] A typical sketch of such networks is demonstrated in Fig. 1. With the inputs of x_t and h^{t-1}, the forgetting gate calculates the value of f_t as given in Eq. (6):

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{6}$$

Once the forgetting rate f_t is determined, the input gate determines the information to be updated. The forgetting rate i_t is multiplied by the candidate value $\hat{C}t$ to obtain the updated data Ct according to the relationships presented in Eqs. (7)-(9):

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{7}$$

$$\widetilde{C_t} = \tanh(W_C[h_{t-1}, x_t] + b_C) \quad (8)$$

$$C_t = f_t C_{t-1} + i_t \widetilde{C_t} \quad (9)$$

The state of the unit is processed by tanh function to obtain a value between [−1, 1], resulting in the final output.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (10)$$

$$h_t = o_t \tanh(C_t) \quad (11)$$

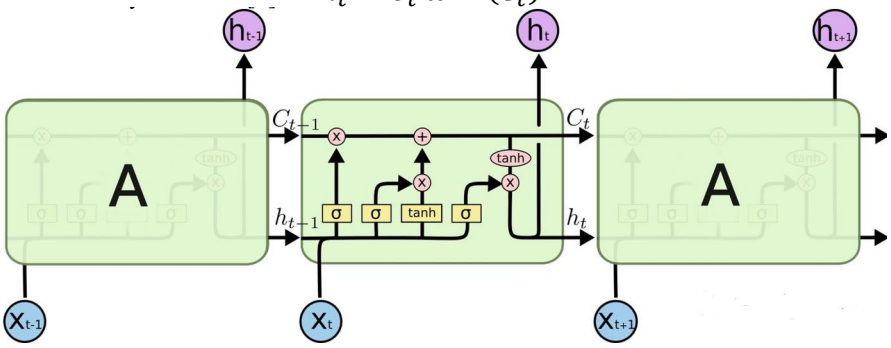


Figure 1. A sketch of LSTM [13].

3. The state-of-art applications

3.1 Random Forest

Zhang et al. proposed a novel stock price trend prediction system, Xuanwu [14]. The models were trained using imbalance learning, feature selection and random forests implemented in WEKA with parameters set to default. Simple accuracy was used as a performance metric, and the experiments were repeated ten times [15]. A real-world evaluation of seven years of trading data from the Shenzhen Growth Enterprise Market showed an average accuracy of 67.5% when the features for each trading day in each training instance included only four values. Roy et al. addressed the problem of predicting whether stock prices would increase by 25% in the same quarter of the following year. To make the cut points of tree node splits and the attributes selected for the tree more random, they used extremely randomized trees.

Nguyen et al. evaluated the impact of various parameters of Random Forest on stock price prediction performance through the TAIEX dataset [16]. As the results show, the best performance was achieved in 11 training years, and the validated RMSE and MAE values obtained were 86.2171 and 63.8813, respectively. Moreover, for the Taiwan stock price prediction problem, the number of max-leaf nodes was more significant than the number of trees.

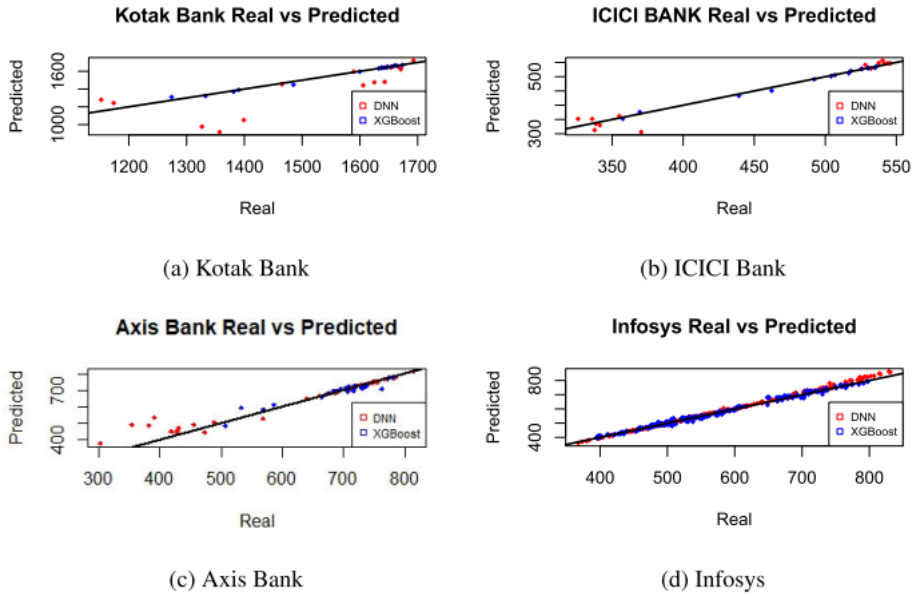
3.2 Xgboost

Naik and Mohan designed a price forecasting system in terms of Hybrid Feature Selection (HFS) technique based on XGBoost and DNN regression scenarios [17]. The experiments considered 42 various stock financial parameter retrieved from the Bombay Stock Exchange (BSE), India. The results of both models are significant. From Fig 2, it can be observed apparently that HFS-based XGBoost has a better performaces.

Kim conducted XGBoost to predict stock returns for the next rebalance date [18]. Historical data of Korean Composite Stock Price Index constituents from 2010 to 2016 were considered. XGBoost was implemented using Scikit-learn [19]. 1260 daily returns were used for training and 252 returns were used for validation. Testing was performed using 154 daily returns. On each day, 10 previous values were used as the feature. The prediction horizon was set to 14 days. Table 5 and Table 6 show the MAE and RMSE errors for predicting stock returns on these rebalancing days, respectively. Both errors are small and alike for all stocks.

Raubitzek and Neubauer used three different machine learning algorithms to test the predictability of selected underlying indicators [20]. 100 runs were performed in each data using different algorithm

memory to predict one step into the future. Overall, XGBoost performed slightly worse than the Lasso, but better than linear SGD regression. The results also show a clear decline in predictability and an increase in model complexity as time goes by.



(a) Kotak Bank (b) ICICI Bank

(c) Axis Bank (d) Infosys

Figure 2. DNN and XGBoost prediction [17].

3.3 LSTM

Chen and Ge [21] explored the attention mechanism in LSTM-based stock price movement prediction. A two-layer LSTM network with dropout in each layer was used, and the attention layer was before the LSTM layer. The LSTM layer took the output of the attention layer as input and applied dropout in the LSTM layer. At the end of the two LSTM layers, a dense layer was added to obtain prediction classes. Finally, a BatchNormalization layer was applied. The output was passed into a softmax classification layer. The attention LSTM (AttLSTM) model was compared with the LSTM model and the prediction accuracy improved for 56 of the 72 stocks.

Budiharto developed a model for price forecasting in Indonesian exchange based on LSTM [22]. They compared results by adjusting epoch and historical data, and obtained the best prediction result of 94.59% at 100 epochs using 1 year data. The conclusion was drawn that for LSTM, short-term historical data should be used to obtain the best accuracy. Bathla et al. proposed a neural network model to investigate whether deep learning could predict in high volatility situation as the stock market in 2020 [23]. Various stock index data were extracted through the Yahoo Finance API.

4. Limitations and Future prospects

With its powerful algorithmic functions and learning ability, machine learning has performed well in stock price forecasting tasks. Nevertheless, each model has its own shortcomings. For Random Forest, it does not require feature selection and always performs well on datasets. Nevertheless, such good performance is sometimes reflected in over-fitting, especially in cases with large noise. XGBoost also suffers from a similar problem. In addition, it needs to traverse the dataset in the process of node splitting, which is time-consuming. The spatial complexity of the pre-sort process also causes XGBoost to consume twice as much memory. Whereas ensemble learning methods mostly predict general trends and are not capable enough to predict very high variations, LSTM is more suitable for temporal sequence issues, e.g., stock price prediction. It accounts for dependencies across observations and do well on volatile temporal sequence with stationary components. Nevertheless, as each cell has several multilayer perceptron (MLP) inside, if the time span is large and the network is deep, the computation can be very large.

There is a large room for performance improvement in the future. First, the performances of the models can be improved by increasing the size of the dataset and applying different fundamental stocks and technical parameters (i.e., add more hidden layers and experiment with different architectures). In addition, more learning algorithms will be applied to the training sets to study the prediction performance and more complicated feature selection methods will be examined to select better combinations of features. It is also possible to combine different ensemble learning methods with neural networks to form new hybrid models to predict stock price fluctuations more comprehensively. There is still a large scope to improve and fine-tune the methods with different optimizers and some external factors should be considered to help predict highly non-linear movement. Last but not least, the way to reduce the noise of temporal sequence will be the focus of future research.

Based on the analysis and summary of the literature, this paper has strong subjectivity. Most of the research results involved are from studies published in mainstream journals in the past five years, which has limited coverage. Therefore, more perfect and comprehensive summary of the previous results is needed in the future. The research also noticed that the research in this area is uneven: there is a lot of discussion based on LSTM, and few scholars talking about XGBoost. It is expected that future researchers can make appropriate adjustments in the choice of direction so as to achieve the balanced development of this field.

5. Conclusions

In summary, this paper investigates stock price prediction based on three state-of-art machine leaning methods (Random Forest, XGBoost and LSTM). Specifically, the theoretical background and state-of-art applications of each algorithm are demonstrated. According to the analysis, all models have relatively good results. Random Forest is better suited for predicting broad trends. XGBoost has a mediocre performance among machine learning methods, but it performs well when combined with deep learning methods in a hybrid model. LSTM can predict stock prices more accurately in the daily period. In addition, various studies put different factors into the model to determine their correlation with stock price changes. In the future, the performances of the models can be improved by applying different technical parameters or combining different ensemble learning methods with neural networks. It is also expected that future researchers can adjust in their choice of direction. Overall, these results offer a guideline for scholars to build more practical stock price prediction models.

References

[1] Wang Y, Guo Y K. Application of Improved XGBoost Model in Stock Forecasting[J]. Computer Engineering and Applications, 2019, 55(020): 202-207.

[2] Booth, Gerding, McGroarty. Automated trading with performance weighted random forests and seasonality[J]. EXPERT SYST APPL, 2014, 2014, 41(8): 3651-3661.

[3] Weng B , Lin L, Xing W, et al. Predicting Short-Term Stock Prices using Ensemble Methods and Online Data Sources[J]. Expert Systems with Applications, 2018, 112(DEC): 258-273.

[4] Nabipour M, Nayyeri P, Jabani H, et al. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis on the Tehran stock exchange[J]. IEEE Access, 2020, pp(99):1-1.

[5] Jiang M, Liu J, Zhang L, et al. An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms[J]. Physica A: Statistical Mechanics and its Applications, 2020, 541.

[6] Wei B, Jun Y, Yulei R, et al. A deep learning framework for financial time series using stacked autoencoders and long-short term memory[J]. PLoS ONE, 2017, 12(7): e0180944.

[7] Kim H Y, Won C H. Forecasting the Volatility of Stock Price Index: A Hybrid Model Integrating LSTM with Multiple GARCH-Type Models[J]. Expert Systems with Applications, 2018, 103(aug.):25-37.

[8] Yan Z, Qin C H, Song G. Random Forest Model Stock Price Prediction Based on Pearson Feature Selection[A]. Computer engineering and Applications, 2021, 57(15).

[9] Roy S S, Chopra R, Lee K C, et al. Random forest, gradient boosted machines and deep neural network for stock price forecasting: a comparative analysis on South Korean companies[J]. International Journal of Ad Hoc and Ubiquitous Computing, 2020, 33(1):62.

[10] Nan, Zhou, Wen, et al. Evolution of high-frequency systematic trading: a performance-driven gradient boosting model[J]. Quantitative Finance, 2015, 15(8).

[11] Tian L, Feng L, Sun Y, et al. Forecast of LSTM-XGBoost in Stock Price Based on Bayesian Optimization[J]. Intelligent Automation and Soft Computing, 2021, 29(3):855-868.

[12] Li M, Zhu Y, Shen Y, et al. Clustering-enhanced stock price prediction using deep learning[J]. World Wide Web, 2022.

[13] Tian L, Feng L, Yang L, et al. Stock price prediction based on LSTM and LightGBM hybrid model[J]. The Journal of Supercomputing, 2022:1-26.

[14] Jing Z A, Sc A, Yan X A, et al. A novel data-driven stock price trend prediction system[J]. Expert Systems with Applications, 2018, 97:60-69.

[15] Pfahringer B, Reutemann P, Witten I H, et al. The WEKA data mining software: an update[J]. Acm Sigkdd Explorations Newsletter, 2009, 11(1):10-18.

[16] Nguyen H T, Tran T B, Bui P H D. An effective way for Taiwanese stock price prediction: Boosting the performance with machine learning techniques[J]. Concurrency and Computation-Practice & Experience, 2021.

[17] Naik N, Mohan B R. Novel Stock Crisis Prediction Technique-A Study on Indian Stock Market[J]. IEEE Access, 2021, PP(99):1-1.

[18] Kim H. Mean-Variance Portfolio Optimization with Stock Return Prediction Using XGBoost[J]. Economic Computation and Economic Cybernetics Studies and Research, 55(4):5-20.

[19] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python[J]. 2012.

[20] Raubitzek S, Neubauer, T. An Exploratory Study on the Complexity and Machine Learning Predictability of Stock Market Data[J]. Entropy, 2022, 24(3).

[21] Chen S, Ge L. Exploring the attention mechanism in LSTM-based Hong Kong stock price movement prediction[J]. Quantitative Finance, 2019, 19.

[22] Budiharto W. Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM)[J]. Journal of Big Data, 2021, 8(1).

[23] Bathla G, Rani R, Aggarwal H. Stocks of year 2020: prediction of high variations in stock prices using LSTM[J]. Multimedia Tools and Applications, 2022.