The effectiveness of stock prediction models: evidence from time series analysis and machine learning scenarios

Qinglin Chen[1, †], Shuya Ma[2, *, †], Ruochen Yang[3, †]

[1]School of Business, University of New South Wales, Sydney New South Wales, Australia
[2]Faculty of Science, University of Ottawa, Ottawa, Canada

[3]School of Science, The Ohio State University, Columbus, Ohio, Unite State

*Corresponding author: Z5261204@ad.unsw.edu.au

[†]These authors contributed equally

Keywords: Stock price predict, Deep learning, ANN, DNN, Random Forest, XGBoost, LightGBM, Time Series, ARIMA.

Abstract: In the era of big data, stock prediction models usher in a new era. In this paper, we will discuss several state-of-art stock prediction models, including decision tree-based model, neural network model and time series model. Based on the analysis we figure out the applicability and limitations of each model, as well as demonstrate the future prospect. Overall, these results shed light on how the application of machine learning and big data can improve the accuracy and reliability of stock price predictions.

1. Introduction

A stock is a financial instrument with both high risk and reward, and also a wide range of trading options. Contemporarily, big data technology has been more and more widely applied to stock price prediction. Reasonable and accurate forecasts are feasible to obtain remarkable return with a controllable risk [1]. So far, numerous methods have been proposed to predict stock prices. Based on different modeling theories, stock price forecasting models can be divided into two categories: the traditional models based on statistical theory (including time series model) and the machine learning models based on neural networks and gray theory. The traditional statistical model includes ordinary least squares, ridge regression, lasso regression, and Stochastic Gradient Descent. These traditional models are linear models, which can be explained easily and simply. However, it is poor for nonlinear data sets and is not ideal for recognizing complex patterns.

The machine learning models are more advanced, including some model based on decision tree and time-series models based on big data and some models based on deep neural networks (DNNs). Applying machine learning algorithm with the aid of numerous tools and libraries has made technical analysis of the stock market easier [2]. The purpose of a decision tree is to find rules buried in a mountain of data, whereas it is also unstable with the variations in the training sets [3]. As for deep neural networks (DNNs), it can effectively solve the problem of complex, incomplete, and ambiguous financial data information, and can include more changing correlation factors to better predict the fluctuation of financial data. Nevertheless, the explanation and interpretation for the results are difficult, i.e., it is usually a black-box [4].

With regard to time series models, they have been widely used in previous studies more as forecasts of volatility than as forecasts of stock prices, which forecasts value with previous data. Autoregressive Integrated Moving Average (ARIMA) models are a general type of time series data forecasting model. In this case, they employ differencing to obtain a stationary time series obeyed the ADF tests, then use historical data to forecast future values. To anticipate future values, these models employ “auto” correlations and moving averages over residual errors in the data [5]. Due to its simplicity, the model is easy to establish and provide useful insight in forecasting stock prices. On the other hand, the model is poor at long-term forecasting and additional methodologies can be implemented to assist prediction.

This paper will analyze the above several prediction models currently used in the financial market, as well as reasonable suggestions. We hope it can help forecast stock price better, so as to realize the better application of the models in the market. Due to the obvious shortcomings of the traditional models, the relevant models will not be discussed in this paper. The rest part of the paper is organized as follows. The Sec. 2 will talk about the machine learning models including DT and NN models. The Sec. 3 will discuss Time series model. Eventually, a brief summary will be given in Sec. 4.

2. Machine learning approaches

2.1 Decision Tree based model

It has long been thought that stock prices are unpredictable. However, based on the efficient-market hypothesis (EMH), the market is efficient for the current set of information, and that without making economic gains in this market, we tend to think of stock movements as completely unpredictable [6].

As it known to all, stock market price sequence is usually dynamic, non-parametric, chaotic and noisy, which makes investment has inherent risk. In addition, given the model specifications to be followed shortly, we note that the price movements can be regarded as Brownie motions, with volatility more pronounced in the short term. Apparently, a good understanding of recent share price movements can be helpful for minimizing the risk. Accurate prediction of stock market price trends to maximize capital gains and minimize losses [7]. On this basis, we want to focus on Decision tree-based Machine learning Methods, including, random forests (RF), forests of gradient boosted decision trees (GBDT) and extreme Gradient Boosting (XGboost).

2.2 Random forest

The algorithm proposed by Breiman constructs a DT forest [8]. If a new instance must be classified, the characteristics of that instance are presented to each DT in the forest. Each DT returns a classification value, which serves as the vote for that class. Finally, the classification values given by RF are those associated with the states with the most votes for class variables in all DT's in the forest [9]. If M is the number of features in a data set, specify a number M << M whose value remains constant during forest construction and is used to randomly select the features of each node. For each RT, if N is the number of instances in a data set, then RF selects a random sample from the original data that replaces N instances. This sample is going to be the training set for building DT. As a matter of fact, RF is fundamentally different from other ensemble learners (e.g., Boosting [10]), as the sizes of individual base learners are fluctuated. On this basis, individual trees are usually large when implement the RFs [11].

2.3 XGBoost and LightGBM

The gradient ascending decision tree (GBDT) is regarded as a wonderful optimization method for tree algorithms [12]. Although it must be one of the most popular models at present, but it also has many defects, e.g., single variable owing to the leaf feature. Because current GBDT methods are designed for single outputs, they are capable of handling multiple outputs individually [13].

The most famous GBDT improvement is XGBoost [14], which makes a lot of improvements to GBDT [15]. However, it still faces the problem overfitting and time costing due to the large quantity of changeable variables. For the sake of overcoming the shortages of XGBoost, a Light Gradient Boosting Machine (LightGBM) is proposed in 2017 [16]. Compared to XGBoost, it possesses many advantages:

(1) LightGBM also provides higher accuracy and shorter training time than XGBoost.

(2) Support parallel tree enhancement operations, even on large data sets (compared with XGBoost) can provide faster training speed.

(3) Histogram -esquealgorithm is used to transform continuous features into discrete features, thus achieving extremely fast training speed and low memory utilization rate.

(4) High accuracy can be obtained by using leaf-wise split instead of level-wise split, which leads to very fast aggregation phenomenon and can capture the underlying pattern of training data in a very

complex tree structure. Overfitting can be controlled by using the hyperparameters 'num_leaves' and 'max_depth'.

Algorithm 1: Gradient Boosting

Input: training set $\{(x_i, y_i)\}_{i-1}^{n}$, the differentiate loss function L(y, F(x)), as well as number of iterations M.

1. Initializing model with a constant value

$$F_0(x) = argmin \sum_{i=1}^{n} L(y_i, \gamma)$$

2. For m = 1 to M do

(1) Computing pseudo-residuals:

$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for i= 1,…,n.

(2) Fitting a base learner $h_m(x)$ to $\{(x_i, r_{im})\}_{i-1}^{n}$.

(3) Computing multiplier γ_m

$$\gamma_m = argmin \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

(4) Updating with

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

2.4 Neural Network Model

Stock prices fluctuations are often affected by a combination of internal and external market factors, so linear models have significant limitations in stock price forecasting. There is a desire to predict stock prices more accurately by using a neural network-based model that includes as many factors as possible that affect stock prices. A neural network identifies potential relationships between data through algorithms that simulate the processes of the brain. It can quickly adapt to changing inputs to produce optimal results. On this basis, neural networks are becoming massively popular in trading systems. This paper will introduce ANNs (Artificial Neural Networks) models based on neural networks.

ANNs are self-adaptive, data-driven algorithms with little assumptions [17], which is effective in addressing nonlinear problem. With the data provided from stock market. The methodology of using ANNs to predict the price for a certain stock usually adopted into four steps as given in Ref. [18]. We will give the PETR4 stock as an example as following.

First step is to issue domain comprehension. An analysis of domain was carried out in order to acquire information about the financial market and to identify the factors that influence stock prices [18]. By acquiring both tacit and explicit knowledge and constructing conceptual maps based on this knowledge, this information can be materialized in the price fluctuations of the stock.

The tacit knowledge is the "tips" and "gut-feelings" that the financial analysts and investors earn from their experiences. Using questionnaire to gather professional opinions on models, strategies, and indicators used in stock pricing that may aid in identifying short, medium, and long-term changes in stock prices [18]. The explicit knowledge is the reading review that describe what factors may help develop the PETR4 stock. In this case, a conceptual map was developed to bring together a combination of fundamental and technical analytical factors [18].

Then it can move to the step for pre-selection and collection. Based on the concept map in the first step, variables with sufficient samples and available historical series are selected based on the availability and accessibility of variable data in the concept map. To discover the link and statistical significance, the variables from the set of samples were investigated based on the cross-correlation function between the series and other variables [18]. By computing the autocovariance of each variable to determine the window size for each variable. These steps are important as the window size could influence the learning and prediction ability of neural networks.

Gathering the data from step 2, do a normalization for those data to adjust those data into a same range. The last step is the modeling and close price prediction of the stock. By establishing the training and tests set and network architecture, the model chose the log-sigmoidal function as the activation function to predict the closing price for PETR4. We can do the stock price prediction from then on using this function. Using four methods: MAPE, RMSE, THEIL Coefficient e POCID to estimate the model accuracy.

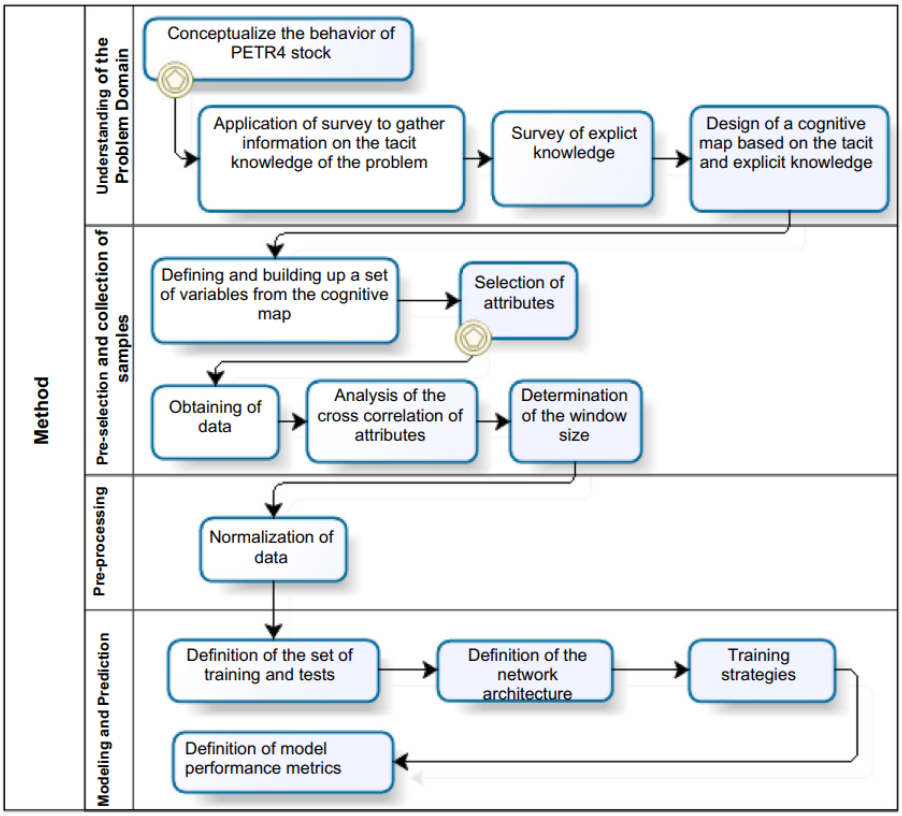

Figure 1. Prediction Process Step [18].

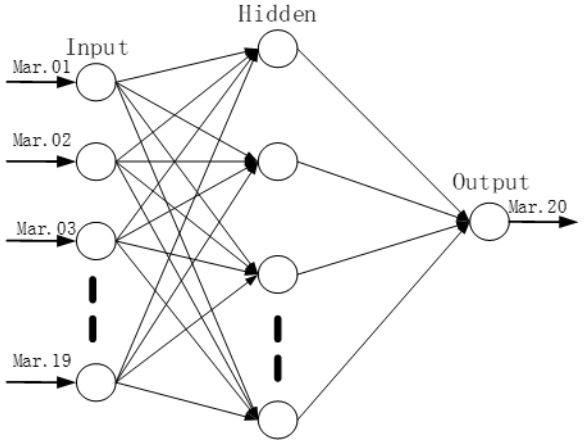

Figure 2. The process of neural networks.

The results of this process show a high accuracy of the prediction. The test set's POCID index of accurate direction predictions were 93.62 %, whereas the validation set's POCID index was 87.50 % [18]. ANNs have therefore shown to be a viable alternative to traditional methodologies for predicting stock behavior and trends, with excellent predictive accuracy. Whereas, there are some critical point ANNs should be carefully analyzed [19]. The input variable of this model can directly influence the accuracy of the results. And the hidden layer, as an adjustable part of the model, cannot be fixed for a particular problem.

In addition to ANNs, there are several other neural network-based models used to predict stock prices, including CNNs and LSTMs. CNNs are being developed based on solving the weakness of ANNs. Researchers have lately began employing CNNs with stock price graph pictures solely because to the difficulties of obtaining a sufficient diversity of data [20]. The information contained in stock price charts can help extract information that is highly relevant to stock price forecasting and forms the basis of stock forecasting. According to research, stock price forecasts may be performed with a high accuracy using simply picture data, although no numerical information is available [20]. Nevertheless, the choosing of the input graph is still a issue need to be solved in the further study.

LSTMs is an advanced version of ANNs. To assess the irrational component of stock price, the LSTM neural network includes investor mood and market data [21]. However, different from ANNs, a memory cell in an LSTM is a new structure that consists of four major elements: an input gate, an output gate, a forget gate, and a neuron unit [22]. This particular structure determines the maximum amount of information that can be added to a nerve cell and how much can be output, and also how much of this book-eating information can be retained for the next computation. LSTM NNs can collect both short- and long-term data, and they don't have an issue with time scale gradient disappearing [15].

2.5 Pros and cons

Stock price prediction using neural network method could be more reliable since any data that can be expressed numeric ab be used in such model. Neural network methods are good at dealing with the nonlinear data such as images. It’s a more precise method to forecasting the stock price in both short and long term. However, neural networks is also a black box for us. Due to the complexity within the neural network algorithm, we cannot precisely define the degree of interaction between variables in the prediction model. Besides, neural network methods are highly dependent on training data, which lead to the over-fitting problem.

3. Time Series model

3.1 Basic descriptions

Time series analysis is mostly used for short-term forecasting. The theoretical foundation of time series analysis is simple: if every element of a time series (or stochastic process) has a connection with its predecessor etc., we may forecast the time series' future value based on its past observations. The autoregressive model is a direct manifestation of this concept. A stochastic process of the following type is known as an autoregressive (AR) process of order p, abbreviated as AR(p), where p signifies the number of autoregressive terms that need to be found; is the white noise, i.e., the random error term that fulfils the standard econometric model; and is the number of lags.

It is noted that AR(p), MA(q) and ARMA (p, q) are all stationary stochastic processes. However, in the practice of econometrics, the time series obtained often exhibit trends such as systematically rising or falling. Some time series also exhibit cyclical fluctuations from week to week. Such time series must arise from non-stationary stochastic processes, and thus cannot be directly simulated by applying a stationary stochastic process such as AR(p), MA(q) or ARMA (p, q). For non-stationary time series, they should first be stabilized. Among them, the difference transform is the most used method of smoothing. The smoothed stochastic process is then simulated using AR(p), MA(q), or ARMA (p, q). This is known as the differential Autoregressive Integrated Moving Average Model (ARIMA), abbreviated as ARIMA (p, d, q). where d is the number of differential transformations implemented.

3.2 Model development

In most of the model development procedures, the ARIMA model estimation includes the three steps which is detailed discussed in Ref. [23]. Ayodele illustrated ARIMA model development by fitting Nokia stock index and Zenith Bank index [24]. Both indices were nonstationary and first-difference was required to perform. Since stock prices are also mostly nonstationary due to inflation and economic growth, first-difference would likely need to perform for most stocks in the market. To

estimate parameters, different Bayesian and standard error criterion (Tables 1 and 2) were considered and the smallest values were chosen to be the best fit of model - ARIMA (2,1,0) for Nokia stock index and ARIMA (1,0,1) for Zenith stock index. After observing ACFs and PACFs, there was only white noise residuals left which means best model can be selected without further estimation [24]. According to the results, the level of accuracy are mostly identical in the short run and actual value starts shifted away with the increase in the time. It concluded that the model development performance was satisfactory and there were closely related predicted and actual prices.

Table 1. Nokia Stock results for different ARIMA parameters

ARIMA	BIC	Adjusted R^2	S.E of Regression
(1,0,0)	5.3936	0.9907	3.5824
(1,0,1)	5.3950	0.9907	3.5817
(2,0,0)	6.1061	0.9811	5.1157
(0,0,1)	8.8324	0.7126	19.9942
(1,1,0)	5.3956	0.0002	3.5860
(1,1,2)	5.3941	0.0035	3.5800
(2,1,0)	**5.3927**	**0.0033**	**3.5808**
(2,1,2)	5.3947	0.0031	3.5812

Table 2. Zenith Stock results for different ARIMA parameters

ARIMA	BIC	Adjusted R^2	S.E of Regression
(1,0,0)	2.4385	0.9970	0.8151
(1,0,1)	**2.3736**	**0.9972**	**0.7872**
(2,0,0)	3.3682	0.9925	1.2974
(0,0,1)	6.9285	0.7372	7.6951
(1,1,0)	2.3659	0.0708	0.7860
(1,1,2)	2.3714	0.0701	0.7873
(2,1,0)	2.4370	0.0031	0.8144
(2,1,2)	2.4412	0.0036	0.8142

In Jeffrey's research, the Chinese stock market price index was modelled using ARIMA-Intervention analytic methods, which created a fit allowing one to evaluate and derive conclusions about how the index behaves over time [14]. The database for the research was from PACAP-CCER Chinese database which is mainly for study and research purpose. The paper developed ARIMA models with and without intervention with the three steps mentioned above. Comparative study was produced by using AIC, Schwarz Criterion [25] and Hanna Criterion [26]. ARIMA without intervention obtained larger values than existing criterion for the intervention ARIMA model. It indicates that intervention model was better fit for the time series data. It was drawn to the conclusion that exogenous forces that impact market on great scale can be interpreted with intervention.

3.3 Analysis of experimental results Advantages and Disadvantages

Statistical forecasting in time series using ARIMA models have been successful and reliable decision-making method. It is simple and efficient in short term forecasting. Autoregression is a linear model, which is difficult for the model to capture non-linear stock fluctuation. Some researchers proposed to mix the ARIMA and GARCH model to overcome the disadvantages of linear models while keeping ARIMA's high performance in short-term forecasting.

4. Conclusions

In summary, we discuss the performances of different stock price prediction models from time series models to machine learning scenarios including RF and ANN. In addition, the development methodologies for each model as well as the shortcomings of each and state-of-art advanced versions are demonstrated. For example, from traditional DT to GBDT to XGoost and lately LightGBM

proposed by Microsoft; from ANN to CNN to Long short-term Memory (LSTM); from ARIMA to ARIMA-intervention etc. In the future, research projects will need to be able to address the complexity and dynamic nature of the environment by integrating multiple models and exploiting diverse types of data to produce better results. Time series forecasting takes into account a variety of natural variable (e.g., trends, seasons). Therefore, parameter estimation becomes major challenge for this model. It can be believed that new algorithms capable of handling both seasonality and trends are expected to emerge in the future years. Machine learning and deep learning-based technologies may be used to improve and speed up feature extraction procedures. Overall, these results offer a guideline for the effectiveness of different stock price prediction models. In terms of their advantages and limitations, analysts and investors should be aware when making investment decisions with these prediction models.

References

[1] Ji X, Wang J, Yan Z. A stock price prediction method based on deep learning technology[J]. International Journal of Crowd Science, 2021, 5(1): 55-72.

[2] Shakhla S, Shah B, Shah N, et al. Stock price trend prediction using multiple linear regression[J]. Int. J. Eng. Sci. Inven.(IJESI), 2018, 7(10): 29-33.

[3] Lai R K, Fan C Y, Huang W H, et al. Evolving and clustering fuzzy decision tree for financial time series data forecasting[J]. Expert Systems with Applications, 2009, 36(2): 3761-3773.

[4] Liu H, Long Z. An improved deep learning model for predicting stock market price time series[J]. Digital Signal Processing, 2020, 102: 102741.

[5] Wei L Y. A hybrid ANFIS model based on empirical mode decomposition for stock time series forecasting[J]. Applied Soft Computing, 2016, 42: 368-376.

[6] Fama E F. Efficient capital markets: A review of theory and empirical work[J]. The journal of Finance, 1970, 25(2): 383-417.

[7] Basak S, Kar S, Saha S, et al. Predicting the direction of stock market prices using tree-based classifiers[J]. The North American Journal of Economics and Finance, 2019, 47: 552-567.

[8] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[9] Abellán J, Mantas C J, Castellano J G. A random forest approach using imprecise probabilities[J]. Knowledge-Based Systems, 2017, 134: 72-84.

[10] Buschjager S, Chen K H, Chen J J, et al. Realization of random forest for real-time evaluation through tree framing[C]//2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018: 19-28.

[11] Qian H, Wang B, Yuan M, et al. Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree[J]. Expert Systems with Applications, 2022, 190: 116202.

[12] Zhang Z, Jung C. GBDT-MO: Gradient-boosted decision trees for multiple outputs[J]. IEEE transactions on neural networks and learning systems, 2020, 32(7): 3156-3167.

[13] Poyarkov A, Drutsa A, Khalyavin A, et al. Boosted decision tree regression adjustment for variance reduction in online controlled experiments[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 235-244.

[14] Jarrett J E, Kyper E. ARIMA modeling with intervention to forecast and analyze Chinese stock prices[J]. International Journal of Engineering Business Management, 2011, 3(3): 53-58.

[15] Yu P, Yan X. Stock price prediction based on deep neural networks[J]. Neural Computing and Applications, 2020, 32(6): 1609-1628.

[16] Hannan E J, Quinn B G. The determination of the order of an autoregression[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1979, 41(2): 190-195.

[17] Adebiyi A A, Adewumi A O, Ayo C K. Comparison of ARIMA and artificial neural networks models for stock price prediction[J]. Journal of Applied Mathematics, 2014, 2014.

[18] De Oliveira F A, Nobre C N, Zárate L E. Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index–Case study of PETR4, Petrobras, Brazil[J]. Expert systems with applications, 2013, 40(18): 7596-7606.

[19] Göçken M, Özçalıcı M, Boru A, et al. Integrating metaheuristics and artificial neural networks for improved stock price prediction[J]. Expert Systems with Applications, 2016, 44: 320-331.

[20] Jin G, Kwon O. Impact of chart image characteristics on stock price prediction with a convolutional neural network[J]. PloS one, 2021, 16(6): e0253121.

[21] Gu Y, Shibukawa T, Kondo Y, et al. Prediction of stock performance using deep neural networks[J]. Applied Sciences, 2020, 10(22): 8142.

[22] Gao T, Chai Y. Improving stock closing price prediction using recurrent neural network and technical indicators[J]. Neural computation, 2018, 30(10): 2833-2854.

[23] Chung R C P, Ip W H, Chan S L. An ARIMA-intervention analysis model for the financial crisis in China's manufacturing industry[J]. International Journal of Engineering Business Management, 2009, 1: 5.

[24] Ariyo A A, Adewumi A O, Ayo C K. Stock price prediction using the ARIMA model[C]//2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. IEEE, 2014: 106-112.

[25] Schwarz G. Estimating the dimension of a model Annals of Statistics, 6, 461-464[J]. MathSciNet zbMATH, 1978.

[26] Hannan E J, Quinn B G. The determination of the order of an autoregression[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1979, 41(2): 190-195.