Comparison of Different Machine Learning Prediction about Stock based on Multi-Factor Input-Taking the Nasdaq Index as an Example

Yingshan Guo*

College of Software Engineering, Sichuan University, Chengdu, China

*Corresponding author: 2020141461182@scu.edu.cn

Abstract. Stock investment has always been widely concerned, and the prediction of future stock trends is what many investors look forward to. There are numerous techniques that can be used to predict stocks as machine learning advances. The commonly used method is support vector machine, random forest, linear regression, etc. Recurrent neural networks, multi-layer perceptron, single-layer LSTM networks, naive Bayes networks, convolutional neural networks, back propagation networks, etc. are examples of deep learning techniques. In historical studies, researchers have tended to predict directly from stock prices or used time series as an independent variable to build a forecasting model for stock prices. In this study, we propose input training parameters based on stock indicators and build a sliding window of it to predict the future price. Based on LSTM, ANN, support vector machine regression, Linear Regression to make predictions on price, and analyze the differences between them. Our research is based on the Nasdaq and the evaluation values show that the neural network approach is effective for stock return forecasting.

Keywords: Stock price prediction; LSTM; ANN; SVR; linear regression.

1. Introduction

Stock price forecasting has always been a difficult subject. The development of machine learning has injected new vitality into stock forecasting and provided researchers with new stock research ideas. With the development of the machine learning, there are many methods can apply to the stock prediction, like the linear regression, random forest, support vector machine, etc. In addition, the Neural Network has as well as been popular. Such as Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), and Convolutional Neural Networks (CNN). These years, some studies have shown that using machine learning methods to predict stocks may be effective. The algorithmic trading framework for time-series prediction can be improved using the suggested planar feature representation approaches and deep convolutional neural networks (CNN). [1]. Machine learning can successfully identify the signal that mispriced stocks belong [2]. A deep recurrent neural network model built on the LSTM network with several inputs and outputs. The program is able to predict an opening price, lowest price, and peak price all at once. [3]. Using Paragraph Vector can get a continuous distributed representation of each news report. LSTM can correctly anticipate the stock price based on this vector and the open price. [4]. The LSTM technology was used with the naive Bayesian approach to extract market emotion components in order to increase forecast accuracy. [5]. Using the traditional back propagation (BP) learning process, Yu Song and Mingyue Qiu discovered that some selected input variables may accurately predict stock market returns. [6]. On the basis of data from several firms and each company's annual reports, Hu et al, assessed an SVM model. The objective of the study was to show how SVM is better than prior traditional predictive regression models and how it can predict outcomes using data that had not previously seen. [7]. The back-propagation method was used by Chauhan et al. to analyze ANN on numerous organizations. The goal of this challenge is to use data mining techniques to execute neural networks with back engineering calculations for financial exchange [8]. In order to anticipate future stock prices, Adebiyi AA et al. used a forward multi-layer neural network model with backpropagation algorithm and fundamental analytical variables of stock market indicators. [9]. Tsung-Jung Hsieh et al, proposed a method promising for stock forecasting. It uses the wavelet transform to eliminate noise from stock time - series data and the artificial bee colony algorithm (ABC) to improve RNN in parameter space. [10]. Sudeepa Das et al. provide a prediction system for predicting stochastic nonlinear stock market values using intrinsic time-scale decomposition (ITD), cluster-based modified crow search algorithm (CMCSA), and enhanced artificial neural networks. [11]. An innovative (financial distress prediction) FDP approach based on SVM ensemble was suggested by JieSun [12]. Amin Hedayati Moghaddam et al. used short-term historical stock prices and dates as input to study the performance of ANN in Nasdaq index prediction [13]. For forecasting a financial time series, S.C. Nayak suggests utilizing the chemical reaction optimization based virtual data position (CROVDP) exploratory approach. It could boost the multilayer perceptron's accuracy (MLP). WeiShen et al. apply radial basis function neural networks (RBFNN) to predict market indices. The radial basis function is modified at the same time using the artificial fish swarm method (AFSA), which increases effectiveness and precision.

In historical studies, researchers have tended to predict directly from stock prices or used time series as an independent variable to build a forecasting model for stock prices. In addition, there are few direct studies on the rate of return at present, and most of the classification problems that convert the rate of return into the direction of change. Due to the fat-tailed distribution and high vitality of the stock market, the trained model is prone to overfitting or poor performance when evaluating returns. In this study, to address the shortcoming of small fluctuations and lack of information based on price predictions, this research proposes input training parameters based on stock technical indicators (MACD signals, RSI, Bollinger Bands, Stochastic Oscillator, Commodity Channel Index, etc.). Based on LSTM, linear regression, support vector machine regression, and ANN to make predictions on stock price, and analyze the differences between them. We plan to use the evaluation parameters (RMSE, MAE, R2, etc.), and perform a back-test of the return rate to find a better method which is useful on stock prediction.

2. Methodology

2.1 Data Description

The research data are acquired from https://hk.finance.yahoo.com/quote/%5EIXIC/history. We have collected the. IXIC historical data and it included the fundamental information (Open, High, Low, Close, Volume). The data range is from 31.12.2009 to 15.11.2017 and contains 1985 rows. The Fig. 1 shows the fluctuations of the close price. The test set (30%) is represented by the orange line, while the train set is represented by the blue line (70%).

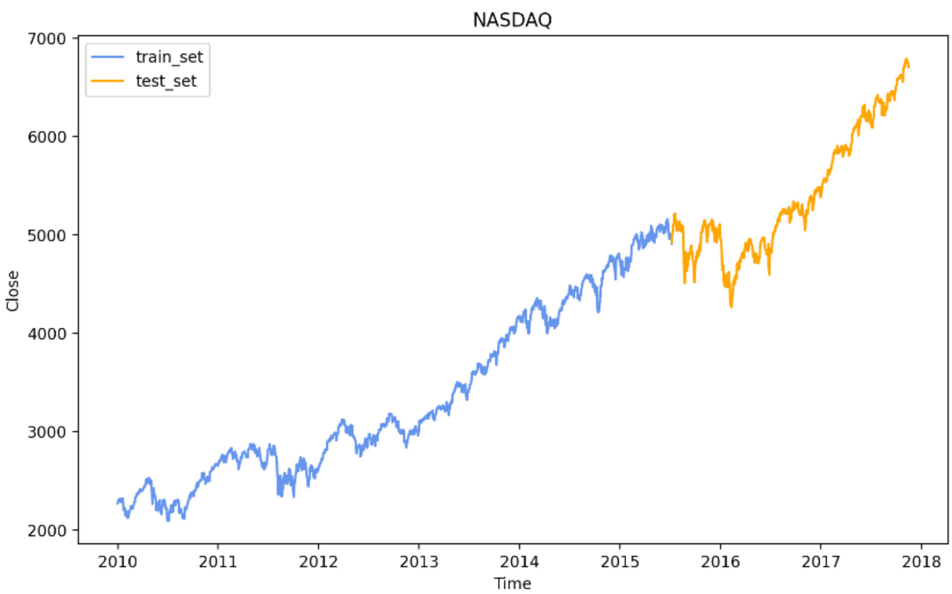

Figure 1. Fluctuation of NASDAQ

To make prediction on stock price, the following stock indicators are calculated. Momentum measures how quickly a stock price rises or falls, and it can be calculated by eq. 1.

$$\mathrm{mom} = \mathrm{V_p} - \mathrm{V_{p-10}} \tag{1}$$

Where V_p denotes the price at the present and V_{p-10} denotes the price from 10 days ago.

The Price Rate of Change (ROC) calculates the percentage difference between the price currently and the price a predetermined number of periods ago. Overbought and oversold scenarios can be recognized using it. The five-day indicator can be calculated by eq. 2.

$$ROC_5 = \left(\frac{V_p - V_{p-5}}{V_{p-5}}\right) \times 100 \tag{2}$$

Where V_p denotes the price at the present and V_{p-5} denotes the price from 5 days ago.

A moving average called an exponential moving average (EMA) prioritizes and provides more weight to the most recent data points. It can be used as an important indicator of the stock trend. The 10-days EMA indicator can be calculated by eq. 3.

$$EMA_{10} = (Smoothing \times (C - P)) + P \tag{3}$$

Where C denotes the price at the present, P denotes the previous periods EMA, and $Smoothing$ is the constant used to smooth exponential data.

Relative Strength Index (RSI) evaluates the frequency and size of recent price swings in an investment to determine whether or not that security is overpriced or undervalued. It provides the short-term signal of buying or selling. The RSI can be calculated by eq. 4.

$$RSI = 100 - \left(\frac{100}{1+RS}\right) \tag{4}$$

Where $RS = Average\ Gain\ /\ Average\ Loss$. In the default situation, the average gain and loss are based on the 14 periods.

Bollinger Band was created by famous technical trader John Bollinger. It includes a set of trendlines (upper band, middle band, and lower band) and can indicate when the market is overbought or oversold. The Bollinger Band can be calculated by eq. 5 and eq. 6.

$$BBANDS_{upper} = MA(TP, n) + m * \sigma[TP, n] \tag{5}$$

$$BBANDS_{lower} = MA(TP, n) - m * \sigma[TP, n] \tag{6}$$

Where MA represents moving average, $TP = (High + Low + Close)/3, n$ refers to Numbers of days in smoothing period, m represents Numbers of standard deviations, $\sigma[TP, n] = Standard\ Deviation\ over\ last\ n\ periods\ of\ TP.$

2.2 Machine Learning Methods

2.2.1 Long short-term memory (LSTM)

A specific kind of recurrent neural network (RNN) is called long short-term memory network (LSTM), often known as a group of neural networks capable of processing sequential input. LSTM has three "gate" structures (shown in Fig. 2): an input gate, a forgetting gate, and an output gate. When information enters the LSTM network, it may be selected by rules. The forgetting gate will remove any data that does not follow the algorithm, leaving just the data that does.

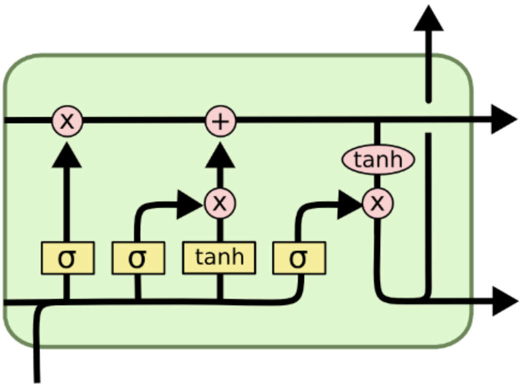

Figure 2. Structure of LSTM

The gate allows for the selective passage of information. The LSTM may add and remove information for neurons via the gating unit. It consists of a layer of a Sigmoid neural network and a pair multiplication method to decide whether information moves. The sigmoid layer generates each element as a real number in the range [0, 1], indicating the weight that is applied to the relevant piece of input. The tanh activation function is placed in a layer of the LSTM neural network.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{7}$$

$$tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{8}$$

The LSTM neural network's forgetting gate, which reads h_{t-1} and $x_t$and assigns the neuron stateC_{t-1}, decides what information needs to be discarded. Equation 3 demonstrates how to calculate probability without ignoring it.

$$ft = \sigma(w_f \cdot [h_{t-1}, x_t] + b_f \tag{9}$$

Where x_t is the current neuron's input and h_{t-1} is the preceding neuron's output. How much fresh information is added to the neuron state depends on the input gate. As demonstrated in eq. 12, After a $tanh(x)$ layer generates candidate vectors C_t, which information needed to be updated is determined by the input layer using the sigmoid activation function, and the state of the neuron is modified as a result.

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \tag{10}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{11}$$

$$\hat{C}_t = tanh(Wc \cdot [h_{t-1}, x_t] + b_c) \tag{12}$$

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{13}$$

$$h_t = O_t * tanh(C_t) \tag{14}$$

2.2.2 Linear regression

Linear regression is an important method for modeling the multiple variables and outputs a dependent variable. Most often, the least squares method is used to fit the independent and dependent variables. By reducing the sum of the squares of the vertical deviations, this approach finds the line that best fits the data. Regression algorithms make predictions on continuous value types. The weight vector is solved using the linear regression technique by minimizing the loss function, which is defined as the sum of the sample set's minimal square errors. The eq. 15. Shows the basic form of linear regression. In this experiment, multiple linear regression was used. With linear regression, outliers can significantly affect regression and may result in inaccurate predictions which should be aware of. Inputting MOM, RSI, ROC, Bollinger Bands, EMA as dependent variables to output the stock price.

$$y_i = x_i^T \beta + \varepsilon_i \tag{15}$$

2.2.3 Support vector machine regression (SVR)

SVM, which has origins in Vapnik-Chervonenkis (VC) theory and statistical learning, generalizes well to new data. The SVM principle may be generalized to solve regression problems. Support vector regression (SVR), like classification, is characterized by the ability to modify the margin and quantity of support vectors using kernels, sparse solutions, and VCs. Using a symmetric loss function that equally penalizes high and low miscalculations, SVR is taught as a supervised learning approach. Support vector regression can be either linear or non-linear by utilizing the appropriate kernel functions. Vapnik's insensitive method enables absolute values of errors below a set threshold to be ignored both above and below the estimated value by constructing a flexible pipe with a minimum radius symmetrically around the estimated function. While points inside the tube are not penalized in this method, neither are points above or below the tube. The fact that SVR's computational complexity is unaffected by the input space's size is one of its main advantages. Additionally, it has a strong potential for generalization and great prediction accuracy.

2.2.4 Artificial Neural Network (ANN)

One of a method for processing continuous or intermittent information input is ANN, which uses data to find a fundamental trend and extrapolates from it. Based on the research results of the brain, ANN abstracts and simulates the information processing function of the brain with hundreds of millions of neurons. Compared to the majority of conventional approaches, ANN is able to simulate and interpret complicated patterns in unstructured data. An input layer of neurons (or nodes, units), one or more hidden layers of neurons, and an output layer of neurons make up an artificial neural network. In this experiment, the ANN model was constructed by three layers: input layer, hidden layer, and output layer. Weights from each input load (MOM, ROC, EMA, RSI, BBANDS_upper, BBANDS_middle, BBANDS_lower) are multiplied, added, and then sent to the hidden neurons. The third layer, also known as the output layer, receives the estimated total weight after that. In the output layer, there is only one neuron, and it will provide the anticipated stock price. The ANN architecture is seen in detail in Fig. 3.

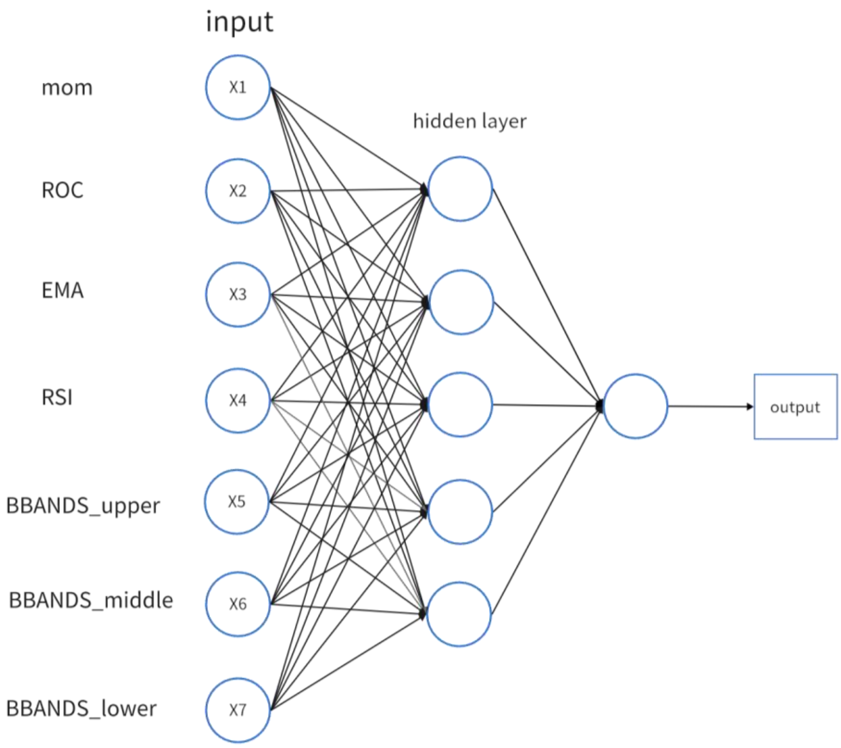

Figure 3. Structure of ANN

3. Results and Discussion

By comparing NASDAQ projections made using LSTM, ANN, linear regression, and SVR models, the models' efficacy was assessed. The predicted data are assessed using the mean-square error (MSE), the root mean square error (RMSE), the coefficient of determination (R2), and the mean absolute error (MAE).

MSE is computed using eq.16. RMSE has been used to evaluate the model and computed using eq. 17. MAE has also been computed using eq. 18.

$$MSE = \frac{\sum_{i=1}^{m}(O_i - P_i)^2}{m} \tag{16}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}(O_i - P_i)^2}{m}} \tag{17}$$

$$MAE = \frac{\sum_{i=1}^{m}|O_i - P_i|}{m} \tag{18}$$

Where O_i denotes the original price, P_i denotes to the predicted price. R^2 has also been computed using eq. 19.

$$R^2 = \frac{SS_{regression}}{SS_{total}} \tag{19}$$

Where $SS_{regression}$ refers to the sum of squares due to regression, SS_{total} refers to the total sum of squares. Figure 4 shows the results of predicting the test set using LSTM, ANN, Linear regression, and Support Vector Regression.

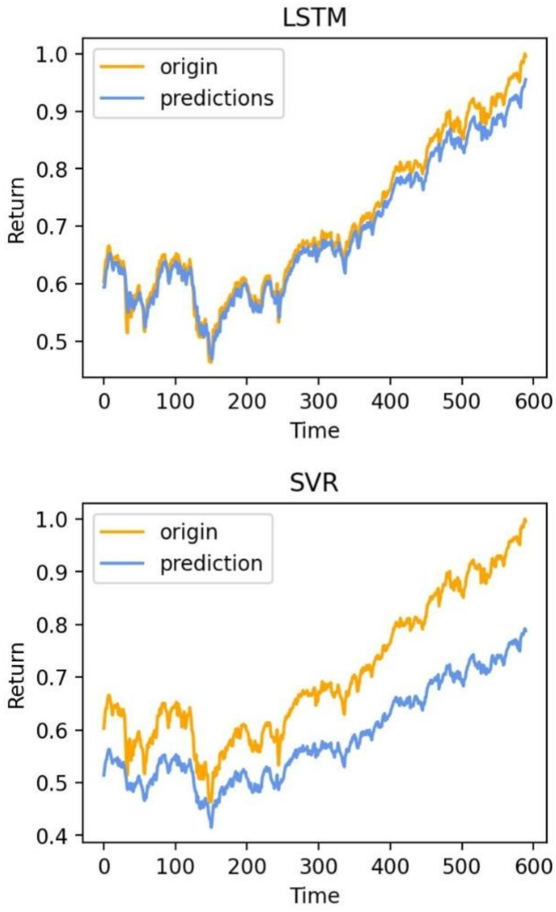

Figure 4. difference between predicted values and original values

Table 1. Evaluations of different methods

	MSE	RMSE	MAE	R^2
LSTM	0.000	0.006	0.004	0.998
ANN	0.000	0.007	0.005	0.997
Linear Regression	0.000	0.006	0.004	0.998
SVR	0.016	0.125	0.118	0.116

From the visualization of the above forecast value, LSTM, ANN, and Linear Regression are better than the SVR method for inputting the seven dependent variables to predict future price. The performance of LSTM on the test set is that the data before 300 points are well fitted, and then the predicted value starts to be slightly lower than the original data, however, there isn't a wide difference between the original number and the anticipated value. The performance of the ANN fits well in the first 400 data points of the test set, and then the anticipated value is lower than the original value. Compared with LSTM, it is more volatile after deviating from the original value. In the forecast of SVR, the anticipated value is always lower than the original data, but the overall fluctuation trend is roughly the same as the original data. Linear regression has the best prediction results for the Nasdaq index. In the test set, from the initial data to the end data, the anticipated value curve and the original value curve fit well.

Table 1 shows the results of evaluating each model using MSE, RMSE, MAE, and R-squared. The results show that the fitting effect of LSTM, ANN, Linear Regression is good, and the correlation reflected by R-squared of SVR is weak. The R-squared of LSTM, ANN, and Linear Regression all reach 0.99, and only the R-square of SVR is 0.116. In addition, the MSE (0.016), RMS (0.125), and MAE (0.118) of SVR are much larger than the other three machine learning methods. The results show that the machine learning method based on the neural network architecture and the traditional linear regression method are more suitable for the prediction of the Nasdaq index based on the multiple indicators of stock input.

4. Conclusion

Stock market forecasting is a difficult undertaking since shifting stock prices depend on many different elements as well as external variables like market mood, corporate value, macroeconomic policies, etc. This experiment innovatively added the characteristics of the stock indicators (MOM, ROC, RSI, EMA, Bollinger band), instead of directly using a few features such as high, low, opening, closing, and stock trading volume in the data set. Based on a comparison of MSE, RMSE, MAPE, and R-squared values, it is obvious that ANN is better at forecasting stock price than SVR. The findings demonstrate that the ANN model's best values are MSE (0.000045), RMSE (0.0067), MAE (0.0045), and R-squared (0.9974). In future research, we can try to use the deep learning model to predict the multi-dimensional stock market. For example, we can consider analyzing news information identification, integrating company financial statement information, etc., to predict market trends, to obtain better investment returns.

References

[1] Chen J F, et al. Financial Time-Series Data Analysis Using Deep Convolutional Neural Networks. International Conference on Cloud Computing and Big Data (CCBD), 2016, 87 – 92.

[2] Avramov D, Cheng S, Metzker L. Machine Learning Versus Economic Restrictions: Evidence from Stock Return Predictability. SSRN Electron. J. 2019.

[3] Ding G, Qin L. Study on the prediction of stock price based on the associated network model of LSTM. Int. J. Mach. Learn. Cybern, 2020, 1307 – 1317.

[4] Rasekhschaffe K C, Jones R C. Machine Learning for Stock Selection. Financ. Anal. J., 2019, 70 – 88.

[5] Qiu M Y, Song Y. Fumio Akagi Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market. Solitons & Fractals, 2016, 85: 1 - 7.

[6] Hu Z, Zhu J, Tse K. Stocks market prediction using Support Vector Machine. Proc 2013 6th Int Conf Inf Manag Innov Manag Ind Eng ICIII, 2013, 2: 115 – 118.

[7] Adebiyi A A, Ayo K C, Adebiyi M O, Otokiti S O. Stock price prediction using neural network with hybridized market indicators. J Emerg Trends Comput Inf Sci 1, 2012, (3): 1 – 9.

[8] Tsung-Jung Hsieh, Hsiao-Fen Hsiao, Wei-Chang Yeh. Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. Applied Soft Computing, 2011, 11 (2): 2510 - 2525.

[9] Sudeepa Das, Tirath Prasad Sahu, Rekh Ram Janghel. Stock market forecasting using intrinsic time-scale decompositin in fusion with cluster based modified CSA optimized ELM. Journal of King Saud University-Computer and Information Sciences, 2022, 34 (10): 8777 - 8793.

[10] Sun J, Li H. Financial distress prediction using support vector machines: Ensemble vs. individual. Applied Soft Computing, 2012, 12 (8): 2254 - 2265.

[11] Amin H M, Moein H M, Morteza E. Stock market index prediction using artificial neural network. Journal of Economics, 2016, 21 (41): 89 - 93.

[12] S. C. Nayak, B.B. Misra, H.S. Behera, Artificial chemical reaction optimization based neural net for virtual data position exploration for efficient financial time series forecasting, Ain Shams Engineering Journal, Volume 9, Issue 4, 2018, Pages 1731 - 1744, ISSN 2090 - 4479.

[13] Wei S, et al. Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm. Knowledge-Based Systems, 2011, 24 (3): 378 - 385.