Proceedings of the 2nd International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/86/20240936

Comparison of Random Forest and LSTM in Stock Prediction

Haoyuan Wu[1,a,*]

[1]The College of Finance and Statistics, Hunan University, 410082, China
a. mark08@hnu.edu.cn
**corresponding author*

Abstract: As an integral component of the financial market, stock prices have attracted the attention of many investors. Due to the frequent fluctuations and sensitivity to market dynamics, predicting stock prices is challenging. The volatility of stock prices and potential significant differences across different periods add to the difficulty of forecasting and reduce its accuracy. The Random Forest model and the LSTM model, as representative models in decision trees and deep learning algorithms respectively, demonstrate high accuracy and adaptability in predicting stock prices. The paper will separately utilize the Random Forest model and the LSTM model to fit the S&P 500 price data from 2013 to 2018 (represented by Apple's stock prices) as training and testing sets, and then compare the fitting results of the two models. The conclusion is as follows: In the absence of white noise in the data, the Random Forest model demonstrates smaller biases in predicting data compared to the LSTM model, and it can also respond more swiftly to price fluctuations.

Keywords: Random forest, LSTM, accuracy, comparison

1. Introduction

The stock market is characterized by its dynamism, unpredictability, and nonlinearity. Predicting stock prices presents a significant challenge due to various factors such as political conditions, the global economy, company financial reports, and performance [1]. Due to the multitude of uncertainties and variables influencing daily market values, including economic conditions, investor sentiments towards specific companies, and political events, predicting stock market prices is a formidable challenge. As a result, stock markets are vulnerable to rapid shifts, leading to unpredictable fluctuations in stock prices [2]. Forecasting involves predicting future events by analyzing historical data and is applicable across various domains such as business, industry, economics, environmental science, and finance. Historical data can be classified as either univariate or multivariate. Univariate data contains information about a single stock, while multivariate data encompasses stock prices from multiple companies at different points in time. Analyzing time series data helps in recognizing patterns, trends, and cycles within the dataset [3].

Since stock market data is time series data, predicting and analyzing historical stock market data to identify patterns is an important method for addressing these issues. And Predicting trends in stock market prices is challenging due to the presence of noise and uncertainties [4]. People invent many algorithms designed for price prediction, such as advanced neural networks, gradient-boosted regression trees, support vector machines, and Random Forest, aim to unveil intricate patterns marked by non-linearity and uncover relationships challenging to discern using linear methods. These models demonstrate enhanced efficacy in handling multi-collinearity compared to linear regression algorithms [5]. This article is going to compare two typical ones of them - the Random Forest and the Long Short-Term Memory (LSTM) together to see which model can work better. Random forests perform prediction or classification tasks by constructing a "forest" consisting of multiple decision trees. Each decision tree serves as a classifier, making decisions based on the analysis of input data features. In a random forest, each tree is generated based on random sampling of the training data, meaning each tree utilizes a different subset of data for training. As the number of trees in the forest increases, the generalization error of the forest approaches a limit almost surely. The generalization error of a forest of tree classifiers depends on both the individual strength of each tree and the level of correlation among them. This approach avoids the problem of modeling the underlying distribution and instead focuses on making accurate predictions through other variables [6]. This approach sidesteps the challenge of explicitly modeling the underlying distribution, instead prioritizing accurate predictions for specific variables based on others [7].

The core idea behind the Long Short-Term Memory (LSTM) architecture is centered on a memory cell that can maintain its state over time. This cell is enhanced by non-linear gating units which regulate the flow of information into and out of the cell [8]. It is a form of recurrent neural network that has demonstrated significant success across various tasks due to its ability to differentiate between recent and earlier examples by assigning distinct weights to each, while disregarding irrelevant memory deemed unnecessary for predicting the next output [9]. Thus, LSTM is a special type of RNN that possesses internal memory and multiplicative gates. It addresses the issues of vanishing or exploding gradients caused by gradient-based weight updates in RNNs [10].

2. Methods

2.1. Data Source

The dataset used in this paper is fetched from the Kaggle website (S&P 500 stock data). It was from 2013 to 2018, collected by Cam Nugent. This dataset contains 619k groups of data, and because the S&P 500 data encompasses price data from many stocks, this paper utilizes Apple Inc.'s stock prices to represent the S&P 500 price data. The original dataset remained in .csv format.

2.2. Variable Selection

S&P 500 stock data are shaped by shifts in American economic progress, mirroring the broader trajectory of global stock development and subject to the impact of significant global occurrences. Given the irregular and unpredictable nature of major events, fluctuations in prices can occur frequently and prove challenging to ascertain in terms of their magnitude, as depicted in Figure 1.

From Figure 1, it can be concluded that before the middle of 2015, S&P500 close prices had entered an upward phase, and then experienced a one-year period significant decrease. After the first few months of 2016, the price started to increase Its highest value is almost 3 times the lowest value.

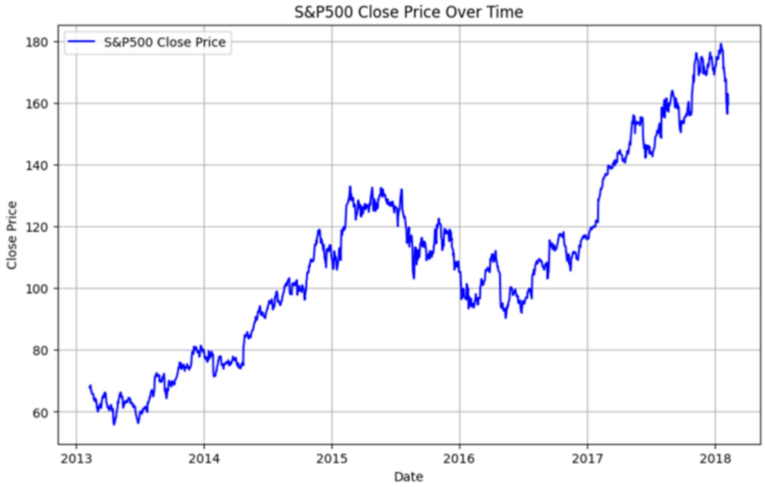

Figure 1: S&P500 price plot.

2.3. Model Selection

This article selects the Long Short-Term Memory (LSTM) model and the Random Forest model separately to predict the stock price and then combine the results together. The method used to combine these two models are the Feature Extraction and Balancing Model Weights.

The Feature Extraction is to integrate advanced sequential features from the LSTM model with structured features from the Random Forest model. The Balancing Model Weights is to assign weights to LSTM and Random Forest separately to balance their contributions in the fusion model and the weights can be adjusted based on performance on a validation set.

3. Results and Discussion

3.1. Data Processing

To assess the consistency of the first-order difference results with a smooth series, this paper employs the Augmented Dickey-Fuller test (ADF test) to examine data smoothness. The primary objective is to ascertain whether the first-order difference of the data exhibits smoothness by evaluating the presence of a unit root in the time series data (Figure 2). The presence of a unit root indicates a lack of smoothness in the time series data, whereas its absence suggests the opposite. The result is shown in table 1.

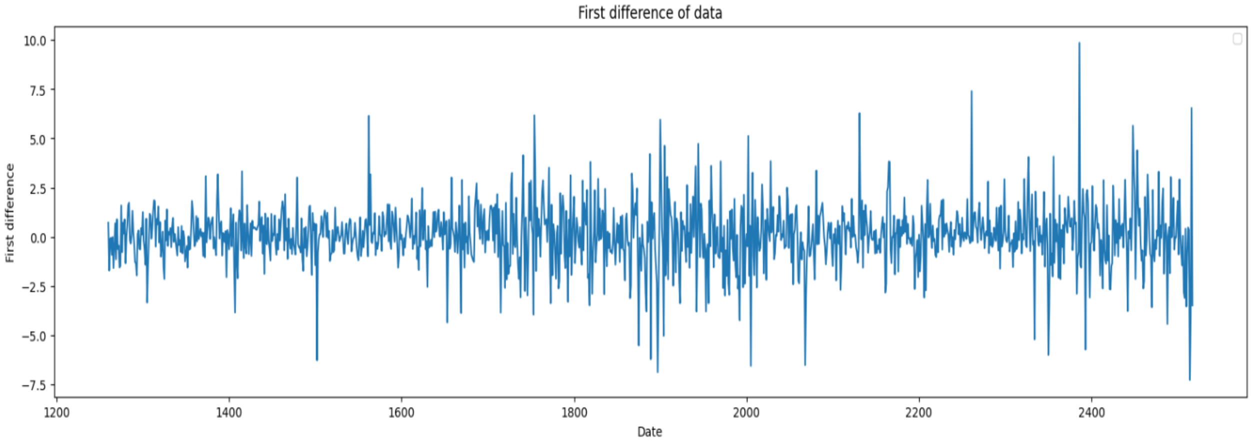

Figure 2: First difference of data.

Table 1: Results of ADF test.

The ADF test results	Value
ADF Statistic	-34.8201
P-value	0.0000
99% confidence interval	-3.4356
95% confidence interval	-2.8638
90% confidence interval	-2.5680

Based on the results of the ADF test, the author observes that the original time series data falls below the ADF test value corresponding to the 95% confidence interval after first differences. Therefore, the data post first differences can be considered as smoothed data.

Since the white noise in the data may affect the results of LSTM fitting, it is necessary to conduct a white noise test on the data beforehand. The white noise test is conducted to ascertain whether the data in the time series remain correlated. If all the information of the series has been extracted after model calculation, the series will be transformed into a white noise series that cannot be predicted anymore. Passing the white noise test means no further information is available for extraction. In this paper, the Ljung-Box test (LB test) is employed to compute the autocorrelation of the outcomes in the time series by analyzing the information (Table 2).

Table 2: Results of Ljung-Box statistic.

	lb_stat	lb_pvalue
1	1254.095196	1.069E-274
2	2500.869266	0.000
3	3741.091513	0.000
4	4974.393758	0.000
5	6199.835630	0.000
6	7417.288662	0.000
7	8626.900813	0.000
8	9828.267435	0.000
9	11021.306438	0.000
10	12205.759484	0.000

Based on table 2, all p-values are significantly smaller than 0.05, even smaller than machine precision. This indicates that under all lag orders, this paper can reject the null hypothesis and conclude that the sequence does not exhibit white noise characteristics.

3.2. Random Forest Results

Based on the Random Forest model, The training set and the test set are split in an 8:2 ratio, and the results are shown in Figure 3.

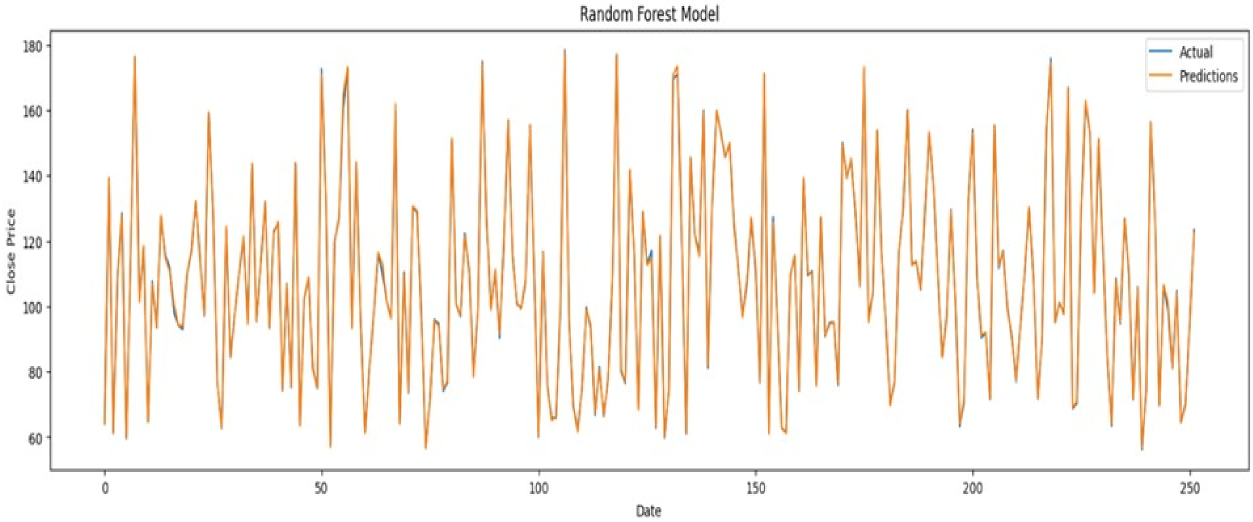

Figure 3: Results of random forest.

The blue line shows the actual prices and the orange line shows the predicted data. From the findings above, it is evident that the line of predicted stock price is almost the same as the line of the actual stock price line shows, means that the Random Forest algorithm is relatively effective in predicting stock prices.

3.3. LSTM Results

Based on the LSTM model, The dataset was initially normalized, then split into training and test sets in an 8:2 ratio. The obtained results were subsequently subjected to reverse normalization to obtain the final prediction outcomes (Figure 4).

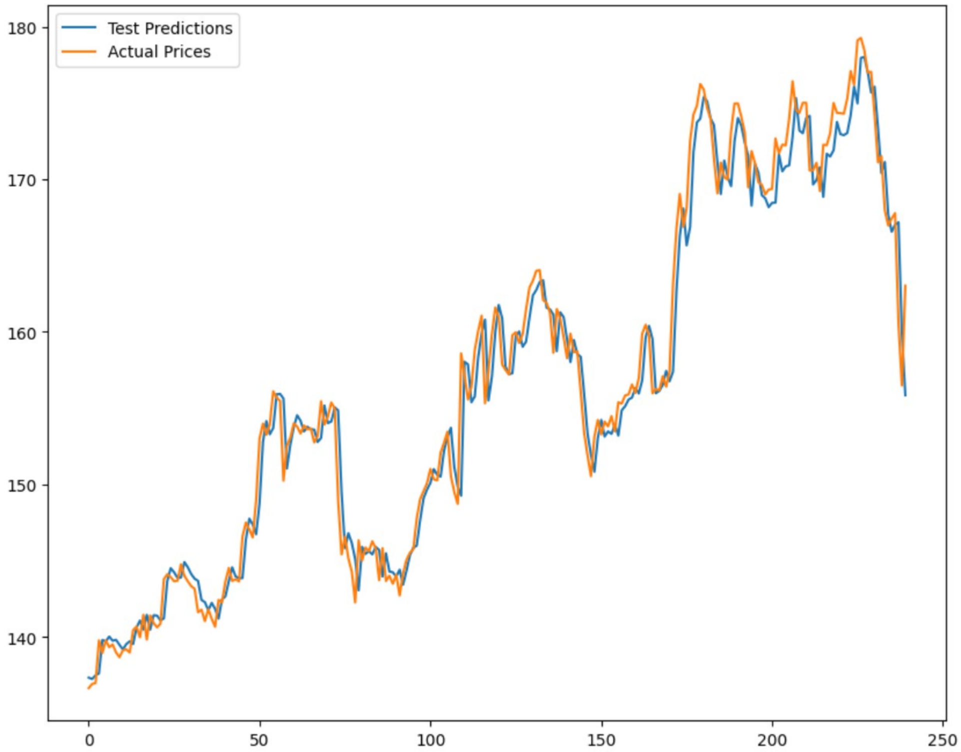

Figure 4: Results of LSTM.

The blue line stands for the actual prices and the orange line represent the predicted data. From the graph, it can be observed that the predicted price variation doesn't align closely with the actual price line, rather, it appears to be a result of a translation of the real price. This indicates that the variation in predicted values consistently lags behind the changes in actual values, this phenomenon illustrates the lagging nature of LSTM models. Therefore, when an LSTM model encounters sudden price shifts, it might struggle to accurately predict prices.

3.4. Comparison Results

In order to compare the accuracy of these two models, this article first use Root Mean Square Error (RMSE) to evaluate it. The result is shown in table 3. Based on the result of RMSE, the RMSE of Random Forest is lower than that of the LSTM. Means that the Random Forest model can predict the stock price more accurately

Following that, the article proceeds to plot scatter graphs using the test set and predicted data as the x and y coordinates respectively, and create a line where the x-coordinate equals the y-coordinate as the ideal line. The result is shown in figure 5.

Table 3: Results of Ljung-Box statistic.

Model	RMSE
Random Forest	0.7967
LSTM	2.0830

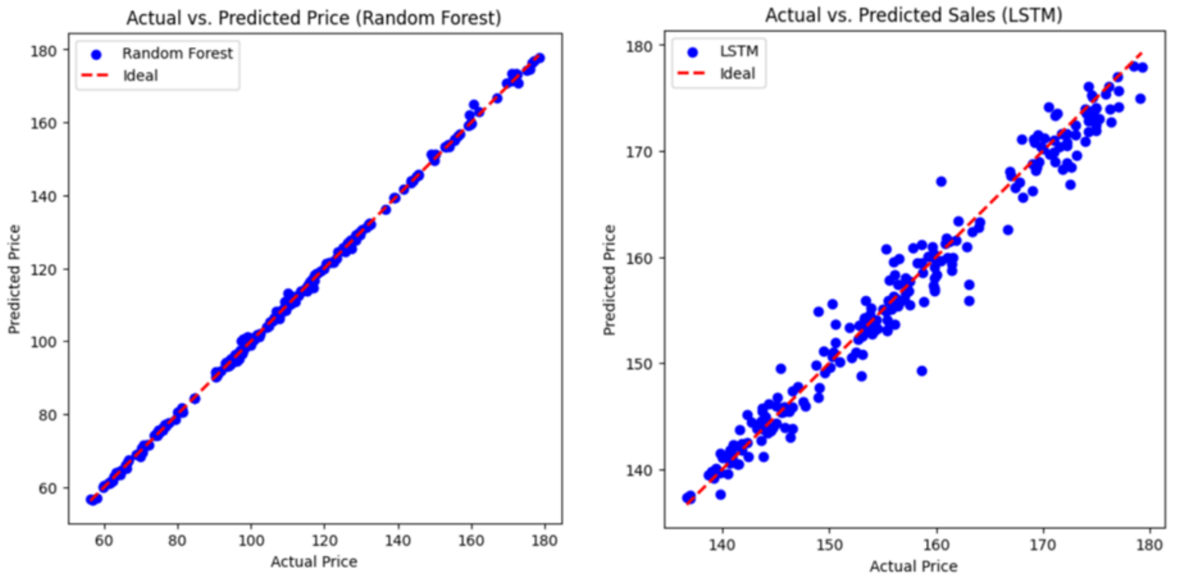

Figure 5: Scatter plot.

The left chart displays the results of Random Forest, while the right chart displays the results of LSTM. It is evident that the scatter points on the Random Forest chart are closer to the ideal line. This suggests that using Random Forest for prediction yields results that are closer to the actual values.

4. Conclusion

The article compares two models, Random Forest and LSTM, for predicting stock prices. It utilizes price data from the S&P 500 index spanning from 2013 to 2018. The variables used for prediction include volume and closing prices. The goal is to evaluate the performance differences between the two models in terms of stock price prediction accuracy.

Continuing, the article proceeds to conduct white noise detection on the dataset used. It is found that the dataset can be considered free from white noise. Following this, the article fits the dataset using both models with a training-to-testing set ratio of 8:2. It then generates comparative graphs of the testing set predictions and actual values for both models. It can be observed that Random Forest performs well in fitting stock prices, while LSTM predictions exhibit a certain degree of lag effect. Specifically, the changes in predicted values tend to occur slightly slower than those in actual values. Finally, the article employs RMSE and scatter plots to compare the accuracy of the predicted data from the two models. The RMSE of the predicted data from the Random Forest model is smaller than that of the LSTM model, indicating that the residuals of the Random Forest predictions are smaller, making them more accurate. At the same time, the scatter plot generated using the predicted data from Random Forest shows points closer to the ideal line, indicating that the predictions from Random Forest are closer to the actual values compared to LSTM.

Overall, as a common and efficient decision tree model, Random Forest performs relatively better in predicting stock prices compared to deep learning models like LSTM, particularly in a noise-free data environment. This provides valuable decision-making guidance for investors.

References

[1] *Vijh, M., Chandola, D., Tikkiwal, V.A., et al. (2020) Stock closing price prediction using machine learning techniques. Procedia computer science, 167, 599-606.*

[2] *Zhang, R.X. and Hao, Y.T. (2023) Research on Stock Price Prediction Based on Deep Learning. Computer Knowledge and Technology, 33, 8-10.*

[3] *Selvin, S., Vinayakumar, R., Gopalakrishnan, E.A., et al. (2017) Stock price prediction using LSTM, RNN and CNN-sliding window model. 2017 international conference on advances in computing, communications and informatics (icacci). IEEE, 1643-1647.*

[4] *Pawar, K., Jalem, R.S. and Tiwari, V. (2019) Stock market price prediction using LSTM RNN. Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018. Springer Singapore, 493-503.*

[5] *Moghar, A. and Hamiche, M. (2020) Stock market prediction using LSTM recurrent neural network. Procedia Computer Science, 170, 1168-1173.*

[6] *Breiman, L. (2001) Random forests. Machine learning, 45, 5-32.*

[7] *Kumar, M. and Thenmozhi, M. (2006) Forecasting stock index movement: A comparison of support vector machines and random forest. Indian institute of capital markets.*

[8] *Sherstinsky, A. (2020) Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.*

[9] *Nelson, D.M.Q., Pereira, A.C.M. and De Oliveira, R.A. (2017) Stock market's price movement prediction with LSTM neural networks. 2017 International joint conference on neural networks (IJCNN), 1419-1426.*

[10] *Smagulova, K. and James, A.P. (2019) A survey on LSTM memristive neural network architectures and applications. The European Physical Journal Special Topics, 228(10), 2313-2324.*