LSTM-Based Time Series Prediction Model: A Case Study with YFinance Stock Data

Yimiao Yu

Department of Information Engineering, Zhejiang Ocean University, 316022, China

Abstract. The goal of this study is to anticipate the time series of stock data that YFinance provides using a Long Short-Term Memory (LSTM) model, with a particular emphasis on the closing prices and daily returns of Apple Inc. (AAPL). The historical closing price data from January 1, 2010, to September 20, 2021, was used as the training set, while the data from September 21, 2021, to August 22, 2024, was employed as the validation set to test the model's predictive capability. The experimental results demonstrate that the LSTM architecture performs excellently in handling data with long-term dependencies and trends, attaining a root mean square error (RMSE) of 5.2129 and a coefficient of determination (R^2) of 0.94, thus accurately forecasting the stock price movements of Apple Inc. However, the model exhibits poor performance in predicting high-frequency fluctuations, with an R^2 of only -0.1, indicating a weak ability to capture high-frequency volatility.

1 Introduction

For investors and institutions, accurately anticipating stock data is essential due to the unpredictability and volatility of financial markets [1]. The advancement of deep learning and machine learning algorithms has provided new avenues to address these challenges.

Although the AutoRegressive Integrated Moving Average (ARIMA) is a traditional time series model, models based on Long Short-Term Memory (LSTM) have recently proven to be effective tools for financial time series prediction due to their remarkable capacity to handle sequential data [2]. For example, Duan applied an LSTM model on financial data obtained from Eastmoney.com, achieving a coefficient of determination (R^2) of 0.9235, which largely explained the fluctuations in stock data [3]. Furthermore, He proposed a stock index prediction model based on Singular Spectrum Analysis (SSA) and LSTM, which improved the R^2 by 0.2 compared to the standard LSTM model [4].

This study focuses on the stock data of Apple Inc. (AAPL), analyzing its closing prices and daily returns. These metrics reflect the company's operational performance and are influenced by factors like the macroeconomic environment and market sentiment. The study uses data from January 1, 2010, to September 20, 2021, for training the model, and data from September 21, 2021, to August 22, 2024, for testing, to comprehensively evaluate the performance of the LSTM model.

Corresponding author: Yuyimiao123@zjou.edu.cn

2 Methods

2.1 Dataset description

YFinance (Yahoo Finance API) is a popular Python library that provides real-time and historical data on various financial assets, including stocks, funds, bonds, and commodities, sourced from Yahoo Finance.

The YFinance dataset includes key market data fields such as the Date, which represents the trading session's date; Open: the opening price of the stock for trade on a certain day; High: the highest price the stock was traded at that day; Low: the closing price of the day at which the stock was exchanged; close, the price at which the stock closes at the end of the trading day; Adj Volume is the total number of shares traded in a particular day; close is the modified closing price that accounts for corporate events such as stock splits and dividends.

2.2 Long short-term memory (LSTM) architecture

One kind of deep learning model used for processing and forecasting time series data is the LSTM network. LSTM effectively captures and retains long-term correlations within time series through its internal gating mechanisms [5]. It is widely applied in fields such as signal processing, speech recognition, and financial time series analysis. Because of its ability to effectively capture and preserve critical information within long-term dependencies, LSTM has been chosen as the primary prediction model for this study.

2.3 Model training and design

With an emphasis on the closing prices and daily returns of the stock, the dataset utilized in this study includes important stock data for Apple Inc. (AAPL) from January 1, 2010, to August 22, 2024. In terms of dataset partitioning, 80% of the data (from January 1, 2010, to September 20, 2021) was chosen as the training data, and the remaining 20% (from September 21, 2021, to August 22, 2024) served as the test data for assessing the LSTM model's prediction ability.

Regarding the model design, this study specifically chose the Huber loss function to balance the issues associated with Mean Squared Error (MSE) and Mean Absolute Error (MAE), effectively smoothing data points, particularly in the fluctuating and uncertain environment of monetary markets. The Huber loss function enhances the model's robustness by better handling fluctuations and outliers, thereby improving the overall predictive accuracy [6].

2.4 Evaluation metrics design

To systematically and scientifically assess the predictive quality of the LSTM model on closing prices and daily returns, this study has designed the following evaluation metrics: R^2 and RMSE (root mean square error), and Directional Accuracy (DA). The roles of these metrics in this study are explained below.

RMSE: To assess the prediction error of regression models, RMSE is a frequently used statistic. It measures the discrepancy between the actual and anticipated values, indicating how accurate the model was in predicting. A lower RMSE number denotes an elevated degree level of model prediction precision. [7, 8]. Using RMSE, the model's prediction performance on Apple Inc. (AAPL) closing prices and daily returns is evaluated in this study.

R^2: R^2, a statistical measure, is the percentage of the dependent variable's fluctuation that can be anticipated from the independent variables. Greater values signify an improved correspondence between the model and the data. It has a 0 to 1 range [9]. R^2 is computed in this study to represent the model's capacity to explain closing prices and daily returns for Apple Inc.

DA: The DA measure is employed to assess the accuracy with which the model can depict the time series' trend and direction. Higher numbers correspond to more precise directional predictions; the range is 0 to 1 [10]. In this study, the functionality of the network in forecasting the direction of daily returns is evaluated in this study using DA.

3 Experimental results and analysis

This study conducted predictions on the closing prices and daily returns of Apple Inc. (AAPL) and evaluated the performance of the LSTM model. The results indicate significant differences in the performance of LSTM across these two prediction tasks. The details are as follows:

In the prediction of closing prices, the LSTM model achieved relatively favorable results. Figure 1 displays the model's predictions for Apple Inc.'s closing prices visually, with the x-axis representing time and the y-axis displaying the closing price (in USD). The blue line indicates the data from the training set, the real closing prices are displayed in the red line, and the green line depicts the predicted closing prices by the LSTM model. As illustrated in Figure 1, the projected and actual results are almost the same, with an overall upward trend. This demonstrates that the LSTM model accurately captured the fluctuations and trends in Apple Inc.'s stock price data.

Fig. 1 LSTM Model's predictions for Apple Inc. (AAPL) stock closing prices (Photo/Picture credit: Original).

In particular, the model's generated RMSE of 5.2129 showed a comparatively tiny average variation between our expected and actual outcomes, proving the model's good accuracy in capturing closing prices. Furthermore, 94% of the variance in the closing prices can be explained by the model, according to the model's R2 of 0.94. This demonstrates how well the model matches the data despite significant fluctuations and long-term patterns.

The LSTM model's performance in predicting daily returns was comparatively inadequate. The model's forecasts for Apple Inc.'s daily returns are visualized in Figure 2, in which the y-axis shows daily returns, and time is represented by the x-axis. The line in blue indicates the training data, the crimson line shows the daily returns of Apple Inc. in reality, and the expected daily returns of the LSTM model are shown by the green line. As shown in Figure 2, the LSTM model's forecasts are overly smooth compared to the real information, leading to notable differences between the expected and measured readings.

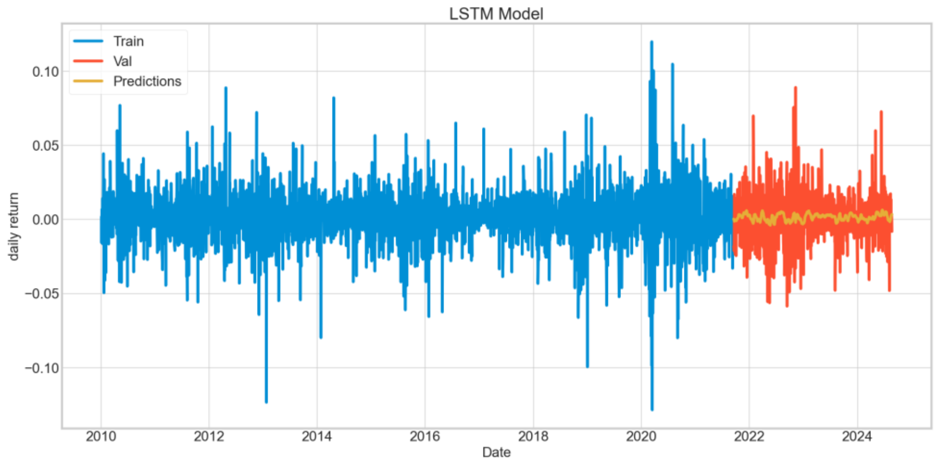

Fig. 2. LSTM model's predictions for Apple Inc. (AAPL) daily returns (Photo/Picture credit: Original).

While the model achieved an RMSE of 0.01825, showing that the anticipated outcomes and the actual data are comparatively close, the R^2 was -0.1, suggesting that the model has almost no explanatory power for the volatility in daily return data. Additionally, the DA was 52.51%, slightly above random chance, indicating that the LSTM model has some capability in predicting the trend of daily returns, but this capability is quite limited.

4 Discussion

4.1 Model advantages

The LSTM model demonstrates significant advantages in capturing long-term trends. A high R2(R2 = 0.94) and a poor RMSE (RMSE = 5.2129) in the closing price prediction indicate that the LSTM model well captures the long-term upward trend in stock closing prices.

As shown in Figure 1, despite the fluctuations in Apple Inc.'s (AAPL) closing price data, the LSTM model consistently maintains a strong predictive performance for the overall trend. As pointed out by Febriansyah, our use of the LSTM model for predicting Apple's closing prices and daily returns further confirms the model's superior performance in handling financial data characterized by long-term trends and dependencies [11].

4.2 Model limitations

In the prediction of daily returns, it is evident that the LSTM model performs poorly when handling financial data characterized by high volatility, low serial correlation, and randomness. As shown in Figure 2, the LSTM model fails to respond accurately and promptly to the fluctuations in Apple Inc.'s daily returns. The primary reason for this is the high-frequency data with significant noise, rendering it unmanageable for the LSTM model to differentiate between noise and meaningful data. This results in suboptimal performance when dealing with short-term or intraday data. This finding is consistent with the observations made by Bao, Yue, and Rao, who noted that LSTM models have limitations in predicting data with high noise levels, high frequency, and high volatility [12].

5 Conclusion

Overall, this study further confirms LSTM's capacity to identify long-term patterns within the monetary markets, particularly within the domain of stock price prediction, where it demonstrates outstanding accuracy and stability. However, when it comes to predicting high-frequency, low-amplitude financial data such as daily returns, LSTM shows relative limitations and struggles to respond promptly to short-term, sharp fluctuations.

To address the limitations of the LSTM model in handling high-frequency data with significant noise, future research could focus on data processing techniques and hybrid model optimization to enhance LSTM's performance in noisy environments. In terms of data processing, methods such as wavelet transform or empirical mode decomposition could be employed for data filtering and noise reduction, thereby improving the decomposability of the signal. On the modeling side, combining LSTM with other models that are adept at handling high-frequency noise, such as ARIMA or Generalized Autoregressive Conditional Heteroskedasticity (GARCH), could mitigate the limitations of LSTM.

References

1. S. Raza, B. Sun, P. K. Khine, M. A. Kemal. Uncovering the effect of news signals on daily stock market performance: An econometric analysis. Int. J. Financ. Stud. 11(3), 99 (2023).
2. G. E. P. Box, G. M. Jenkins. Time series analysis: Forecasting and control. Holden-Day (1976).
3. C. Duan, W. Ke. Advanced Stock Price Prediction Using LSTM and Informer Models. Deleted J. (2024).
4. X. T. He. A Stock Index Prediction Model Based on LSTM: Taking New Energy Vehicles as an Example. Front. Bus. Econ. Manag. 2(1), 45-60 (2024).
5. S. Hochreiter, J. Schmidhuber. Long short-term memory. Neural Comput. 9(8), 1735-1780 (1997).
6. P. J. Huber. Robust estimation of a location parameter. Ann. Math. Stat. 35(1), 73-101 (1964).
7. K. Pearson. On the mathematical theory of errors of judgment, with special reference to the personal equation. Philos. Trans. R. Soc. Lond. A. 198, 235-299 (1902).
8. C. F. Gauss. Theory of the motion of the heavenly bodies moving about the sun in conic sections. Hamburgh: Friedrich Perthes und I. H. Besser (1809).
9. K. Pearson. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. Philos. Trans. R. Soc. Lond. A. 187, 253-318 (1896).
10. M. H. Pesaran, A. Timmermann. A simple nonparametric test of predictive performance. J. Bus. Econ. Stat. 10(4), 461-465 (1992).
11. F. Febriansyah, A. Sujjada, F. Sembiring. Prediksi Harga Bitcoin Menggunakan Algoritma Long Short-Term Memory (LSTM). J. Inovtek Polbeng Seri Inform. 9(1), 1-10 (2024).
12. W. Bao, J. Yue, Y. Rao. A deep learning framework for financial time series using stacked autoencoders and long short-term memory. PLoS ONE 12(7), e0180944 (2017).