NVIDIA Stock Price Prediction by Machine Learning

Jiaqi Tan *

Beijing New Channel International School, Beijing, China

* Corresponding Author Email: 1811000315@mail.sit.edu.cn

Abstract. This study explores how to forecast NVIDIA stock values using machine learning models. The chosen prediction models are Long Short-Term Memory (LSTM), Random Forest, and Linear Regression. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score are only a few of the metrics used to assess the effectiveness of a model. The empirical findings showcase the remarkable prowess of the LSTM model in prognosticating NVIDIA stock prices, exhibiting a heightened level of accuracy and predictive acumen in contrast to Random Forest and Linear Regression counterparts. This investigation not only contributes a potent methodology to the realm of stock market prognostication but also holds promising prospects for prospective real-world implementations. Future implications include enhancing decision-making for investors and promoting financial stability through accurate market trend assessment.

Keywords: NVIDIA; Stock Price Prediction; Machine Learning.

1. Introduction

Accurate prediction of stock prices stands as a pivotal endeavor in financial activities, driving informed investment decisions and risk management strategies. The dynamic and volatile nature of stock markets underscores the significance of robust prediction models. Extensive research has been dedicated to unraveling the intricacies of stock price forecasting, utilizing diverse methodologies to discern market trends and price movements. Notably, the deployment of machine learning [1] algorithms has emerged as a prominent avenue for enhancing predictive accuracy.

Amid the plethora of research, it is evident that the landscape of stock price prediction is rich with exploration. Numerous studies have ventured into this realm, employing an array of models and techniques to unravel the complexities of market behavior. For instance, Jiang Bolin, Jie Yue & Shengjie Lan. (2021) utilized a LSTM (Long Short-Term Memory) [2] model to explore stock price prediction. In this case study, they demonstrated how to construct and train an LSTM model using the Keras [3] library in Python, applied to time series data of stock prices. The results indicated the potential effectiveness of the LSTM model in predicting time series data. These efforts have contributed to the evolution of predictive methodologies, yet the challenge of accurate stock price forecasting remains intricate.

Against this backdrop, the present study contributes to this evolving landscape by employing machine learning models, including Long Short-Term Memory (LSTM), Linear Regression and Random Forest, those statistical learning methods to predict NVIDIA stock prices [4-6]. The comprehensive evaluation involving metrics such as MSE, RMSE, MAE, and R^2 score adds to the precision of model assessment [7-10]. By presenting the results of each model's performance, clarifying the efficacy of using the three models to predict stock prices is the aim of this study, with LSTM demonstrating promising potential in achieving accurate forecasts. The ensuing sections present a detailed account of the project's methodology, data collection and preprocessing, model selection, experimentation, and findings, ultimately culminating in a nuanced understanding of the potential and limitations of machine learning methods in the realm of predicting stock values.

2. Data

The historical dataset of NVIDIA stock prices from 1999-01-22 to 2023-07-24 are used. It contains 6165 rows of trading data with several columns: highest price, closing price and opening price. In this study, the closing price is used, and the trend of the closing price is shown in Fig. 1.



Fig. 1 'Close' price trend.

The data used for LSTM, Linear Regression and Random Forest models is from the year 2016-01-04 to 2023-07-24. To train the model, there is a training dataset; to test the model after training, a test dataset is used. Those two datasets are split from the main dataset of NVIDIA stock price dataset. The 80 percent training dataset and 20 percent test dataset for these three models: LSTM, Linear Regression and Random Forest. Those statistical learning methods are used to predict the stock price.

3. Models

The machine learning models: LSTM, Linear Regression and Random Forest are used in this project.

3.1. Long Short-Term Memory (LSTM)

The LSTM model is an apt choice for prediction due to its inherent ability to ability to recognize long-term trends and dependencies within sequential data. Its strengths lie in effectively handing vanishing gradient problems, thereby preserving information over extended sequences. Furthermore, LSTM's ability to learn from varying time intervals and encode past information ensures robustness in stock price prediction.

3.2. Linear Regression

A fundamental statistical learning technique called linear regression is used to represent the connection between a dependent variable and one or more independent variables. Linear Regression [4] presupposes a linear relationship between features, like "Date," and the target variable when used to forecast stock prices. It estimates the best-fit linear line that minimizes the sum of squared residuals.

3.3. Random Forest

Random Forest is an ensemble learning method based on decision trees. It creates a collection of decision trees by randomly selecting features and data points. Each tree contributes to the final prediction, and the ensemble averages prediction combines the individual predictions [5]. Random Forest can record intricate data interactions and mitigate overfitting, making it appropriate for predicting stock values.

4. Model Evaluation

To comprehensively evaluate the prediction performance of those individual models, this paper employed a range of evaluation metrices:

4.1. Mean Squared Error-(MSE)

MSE calculates the mean squared differences between predicted and real data. A lower MSE shows a closer fit of the model's predictions to the real data. However, it is sensitive to outliers, giving greater weight to larger errors.

4.2. Root Mean Squared Error-(RMSE)

RMSE is MSE under a root sign, it scales back mistakes to the original data's scale. It offers a measure of the typical prediction error that is easier to understand. A lower RMSE indicates better predicting accuracy, similar to MSE.

4.3. Mean Absolute Error-(MAE)

The average of absolute differences between projected and real data are calculated using MAE. It offers a clear explanation of the average prediction error and is less sensitive to outliers than MSE.

4.4. R-squared score-(R2)

R-squared, the coefficient of determination, expresses how much of the variation of the dependent variable is accounted for by the models. A higher R2 value indicates that the model and the data are more closely matched. However, R2 may not account for overfitting and doesn't capture the magnitude of prediction errors.

Using a combination of these evaluation metrics provides a holistic view of each model's performance and helps identify their strengths and weaknesses in capturing the underlying patterns in stock price data.

5. Results

5.1. LSTM Model

In the LSTM model, this paper plotted the predicted and actual stock prices over time. The Fig. 2 below showcases the comparison:

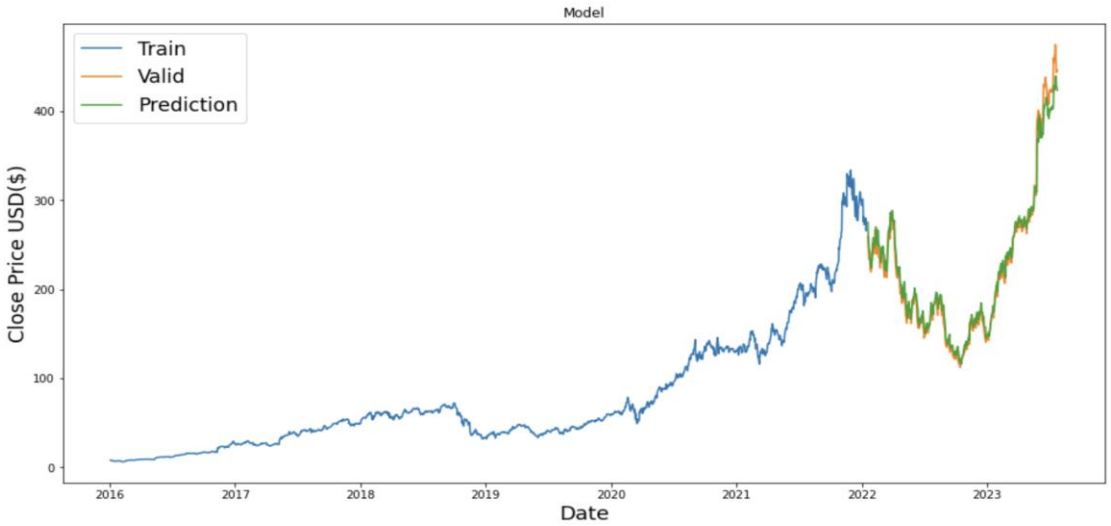


Fig. 2 LSTM model prediction

The Fig. 2 illustrates how the LSTM model captures the trends and fluctuations in the stock prices, it is obvious to see that the predict prices match the actual prices very well.

5.2. Linear Regression

The Fig. 3 showcases the predictions made by the Linear Regression model compared to the actual stock prices:

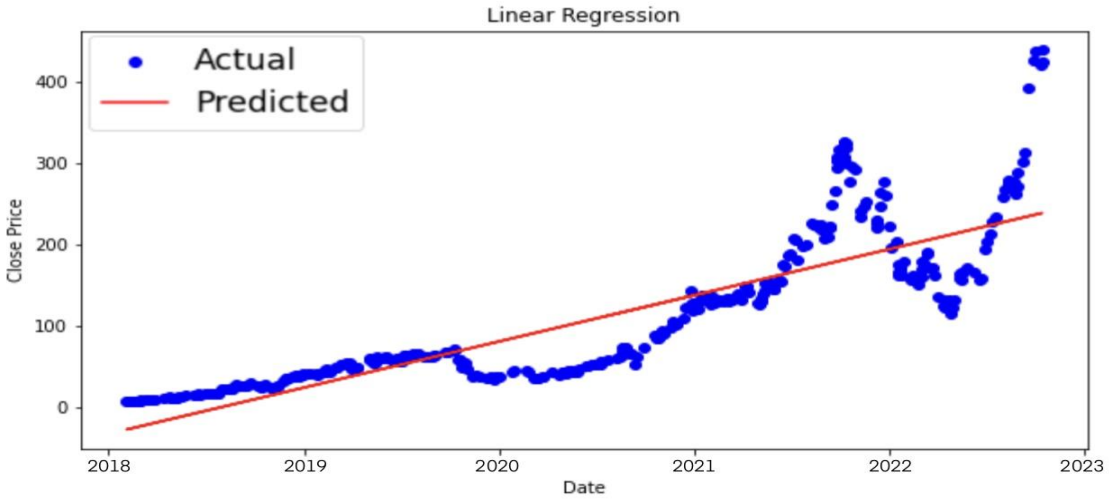


Fig.3 Linear regression model prediction.

Fig. 3 highlights the linear nature of the predictions, which might not fully capture the complex dynamics of stock price movements.

5.3. Random Forest

Fig. 4 shows the prediction errors of the Random Forest model relative to the actual stock prices:

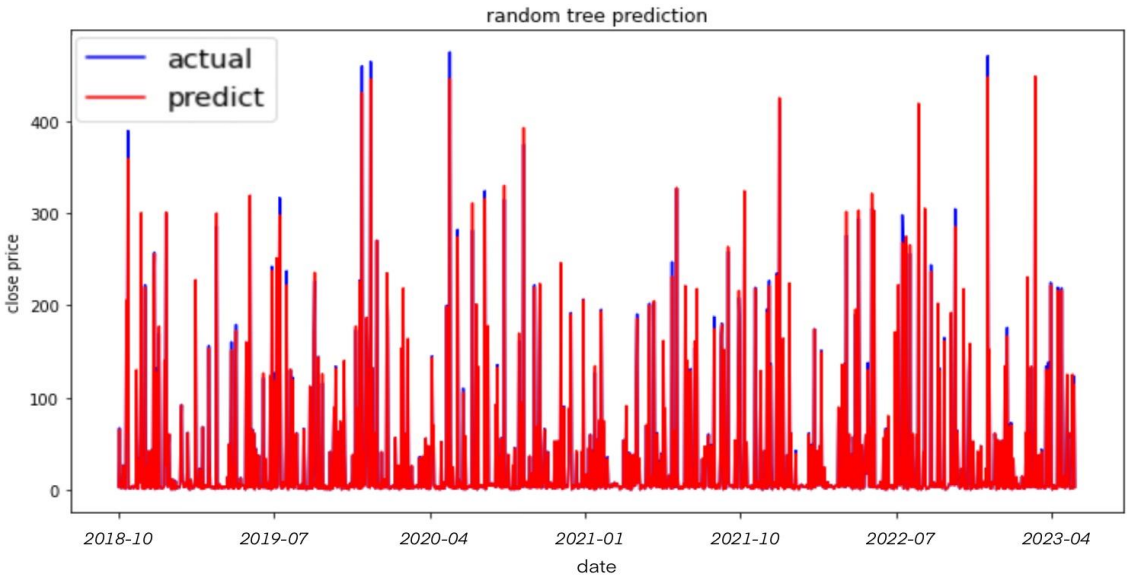


Fig.4 Random Forest model prediction.

The Fig. 4 provides insights into the of the Random Forest model by showcasing the magnitude of prediction errors.

5.4. Comparison

The detailed indicators are summarized in the following Table 1.

Table 1. Results for four models.

	MSE	RMSE	MAE	R2 Score
Random Forest	10.6	3.3	1.0	0.997
LSTM	110.2	10.5	8.3	0.984
Linear Regression	2353.9	48.5	35.6	0.731

Interestingly, the Random Forest model exhibited the highest scores across all metrics. LSTM and Linear Regression also demonstrated competitive performance, albeit with relatively higher errors.

6. Discussion

6.1. Random Forest's Strong Performance

The higher performance of Random Forest can be attributed to its capacity to recognize intricate correlations in the data and reduce overfitting. Random Forest's ensemble nature combines various decision trees to generate reliable predictions that are in good agreement with the variety of patterns in stock price data.

6.2. LSTM's Sequence Learning

LSTM, as a deep learning model, demonstrated its capacity to capture sequential dependencies in stock price sequences. Although it didn't outperform Random Forest, its performance highlights its potential to capture more intricate temporal patterns if provided with larger datasets and appropriate hyperparameter tuning.

6.3. Linear Regression's Simplicity and Limitations

Linear Regression's comparatively higher errors suggest its limitations in capturing the complex relationships present in stock price data. While simple and interpretable, this model may struggle to model non-linear trends that influence stock price movements.

7. Conclusion and Future Work

This paper employed a range of statistical learning methods, including Random Forest, LSTM and Linear Regression, to predict stock prices. The evaluation results indicated that Random Forest emerged as the top-performing model, showcasing its effectiveness in capturing underlying trends and patterns. These findings underline the importance of employing diverse modeling approaches and highlights avenues for potential future improvements.

Future work could explore tuning those models that are used in this project and employ some boosting methods. Other machine learning models will also be employed to predict the stock price.

References

[1] Géron, A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. "O'Reilly Media, Inc.", 2022.

[2] Hsieh, T. J., Hsiao, H. F., & Yeh, W. C. 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.

[3] Ketkar, N., & Ketkar, N. Introduction to keras. Deep learning with python: a hands-on introduction, 2017: 97-111.

[4] Shaoming Bai. Research on Railway Passenger and Freight Volume Based on Linear Regression Model. Western Communications Technology, (2021), (10): 141-144.

[5] Zhengxu Yan, Chao Qin & Gang Song. Stock Price Prediction Based on Pearson Feature Selection with Random Forest Model. Computer Engineering and Application, 2021, (15): 286-296.

[6] Hang Li. Statistical Learning Methods. Tsinghua University Press, 2019.

[7] Marmolin, H. Subjective MSE measures. IEEE transactions on systems, man, and cybernetics, 1986, 16(3): 486-489.

[8] Yanhong Guo. Evaluation of Beijing Land Resources Carrying Capacity Based on Mean Square Error Analysis. Resources and Industry, 2011, 13(6): 62.

[9] Bin Li, Dingxiang Du, Xingguo Wang, Shuyang Wang, Qi Cheng, & Yarong Guo. The AC submarine cable longitudinal protection of the offshore wind power transmission system based on the mean absolute error. Electrical Measurement and Instrumentation, 2022, 59(6): 122 -129.

[10] Ozer, D. J. Correlation and the coefficient of determination. Psychological bulletin, 1985, 97(2): 307.