Proceedings of the 2nd International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/86/20240938

Research on Stock Price Prediction Based on LSTM Model and Random Forest

Yuxuan Qi[1,a,*]

[1]*College of Information Science and Technology, Beijing University of Chemical Technology,*
Beijing, 102299, China
a. 2023040233@buct.edu.cn
**corresponding author*

Abstract: In this study, cutting-edge methods of applying deep learning techniques to stock market predictions were explored, specifically focusing on the stock data of Tesla Inc. Long Short-Term Memory networks (LSTMs), an advanced form of Recurrent Neural Networks (RNNs) capable of effectively addressing the issues of vanishing and exploding gradients that traditional RNNs face, were employed. This enhances the model's learning capability and predictive accuracy for time series data. The innovation of this research lies in the integration of the LSTM model with the Random Forest algorithm, forming a hybrid model aimed at leveraging the complementary strengths of both models to improve the accuracy of stock price predictions. Through empirical analysis of Tesla's stock data, it was found that the hybrid model outperformed the individual LSTM model. This result not only proved the effectiveness of LSTMs in handling complex time series prediction problems but also demonstrated the potential of enhancing predictive performance by integrating different types of models. The findings offer a new perspective for financial market analysis and prediction, especially in the use of deep learning technologies for stock price forecasting. They provide valuable references for future research and practice in this field. Further investigations could explore the applicability of this hybrid approach to other financial instruments and markets.

Keywords: Tesla stock, long short-term memory (LSTM), random forest, Hybrid model, stock price prediction

1. Introduction

Forecasting the direction of stock market prices presents a significant challenge, given the myriad uncertainties and factors that affect daily market valuations. These include the state of the economy, investor attitudes towards particular firms, and geopolitical developments. As a result, stock markets are vulnerable to rapid shifts, leading to unpredictable fluctuations in stock prices [1]. Kadhem and Thajel used the hidden Markov process to analyze and model prices and found that the price movements show significant swings during various moments of market turbulence [2]. Forecasting involves anticipating future events by analyzing historical data and is applicable in various sectors such as business, industry, economics, environmental science, and finance. Forecasting data can be categorized into two types: univariate and multivariate. Univariate data pertains to data about a single entity, like an individual stock, while multivariate data includes data on stock prices from different companies over different periods. Analyzing time series data aids in recognizing patterns, trends, and the presence of periods or cycles within the dataset [3].

The methodologies used in forecasting have undergone significant transformation alongside advances in computing technology, marking a transition from traditional techniques to more sophisticated analytical methods, revolutionizing the realm of financial forecasting. The surge in computational capabilities spurred demand for approaches capable of leveraging this enhanced power, leading to the integration of machine learning and deep learning into the arsenal of financial analysts.

Machine learning, building upon the foundations of regression analysis, introduces a versatile framework for predicting stock market trends by enabling models to sift through historical data and identify patterns that might elude conventional statistical techniques. Neural networks, a cornerstone of deep learning, have progressed by processing information across layers of interconnected nodes, mimicking the human brain's ability to discern complex patterns and relationships within large data sets. This model is particularly adept at grasping the nuances of market dynamics and offering forecasts that account for a multitude of factors influencing stock prices. These models demonstrate enhanced efficacy in handling multi-collinearity compared to linear regression algorithms [4]. Progress in machine learning and deep learning has significantly enhanced the analytical tools accessible to financial analysts. These technologies provide enhanced precision and the ability to analyze intricate market trends, marking a significant advancement in the pursuit of accurate forecasts for stock performance. This progress showcases the breakthroughs in computing technology and marks the beginning of a new age in financial analysis. Currently, the level of understanding and the ability to predict outcomes are constantly developing.

At the same time, more and more scholars have found that a single model is not ideal for predicting stock prices in most cases, and they are more willing to use more complex combination models to predict stock prices. Pan utilized an enhanced Fruit Fly Optimization Algorithm (FOA) model to forecast prices and found that the prediction model combining the Modified Fruit Fly Optimization Algorithm (MFOA) and Diagonal Recurrent Neural Networks (DRNN) possessed the greatest capacity to predict closing oil and gold prices [5]. Wu et al. used a combination of Ensemble Empirical Mode Decomposition (EEMD) and LSTM to predict and found that the model can still predict prices accurately despite varying decomposition outcomes [6]. Huang et al. creatively used Empirical Mode Decomposition (EMD), which breaks down the initial information into a few Intrinsic Mode Functions (IMF) and rest components at different frequencies [7]. The Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) model used by María et al. can avoid modal confusion while excluding residual noise in the IMF [8].

This study seeks to investigate the integration of long short-term memory neural networks (LSTM) and random forest models for feature extraction from market data to forecast price trends. The LSTM neural network is a robust deep learning model designed for time series data, particularly adept at capturing extended relationships in sequential data. The LSTM network enables the extraction of valuable aspects from historical stock data, such as price patterns, trading volume, and other relevant information. Efficiently extracting these attributes will form the basis for future prediction tasks. Some experts have used LSTM to predict the price of crude oil and have obtained promising results. Pan and Li used the Long Short-Term Memory (LSTM) method to apply it to short-term price prediction and the finding is that using daily data from the first 24 months to predict daily data from the next 3 months was the best [9].

Random Forest is an ensemble learning method, during the prediction phase, each tree within the forest autonomously generates a prediction, and the final output is determined through a majority vote (for classification tasks) or an average calculation (for regression tasks) of the individual tree predictions [10]. Random Forest exhibits strong resilience and precision.

This study aims to delve into the efficacy of this amalgamation of deep learning and ensemble techniques, offering fresh perspectives and technologies for the realm of stock market forecasting. It seeks to contrast the proposed model against conventional singular models for stock prediction, validating its efficacy with the anticipation of pioneering novel methodologies in predicting stock trends within the financial domain.

By combining the strengths of LSTM in capturing long-term dependencies and the robustness of Random Forest in handling complex data, the hypothesis is that the proposed hybrid model will yield superior performance compared to traditional single models. The LSTM component will be responsible for extracting meaningful features from the historical stock data, while the Random Forest will leverage these features to generate accurate predictions.

To evaluate the effectiveness of this approach, extensive experiments using real-world stock market data will be conducted. The performance of the hybrid LSTM-Random Forest model will be compared against baseline models such as individual LSTM and Random Forest models. Various evaluation metrics, including mean squared error, mean absolute error, and directional accuracy, will be employed to assess the predictive capabilities of each model. Furthermore, the impact of different hyperparameter settings and data preprocessing techniques on the performance of the hybrid model will be investigated. This will provide insights into the optimal configuration for achieving the best prediction results. The findings of this study are expected to contribute to the advancement of stock market forecasting techniques by showcasing the potential of combining deep learning and ensemble methods. The proposed hybrid LSTM-Random Forest model has the potential to assist investors, financial analysts, and decision-makers in making more informed and accurate predictions regarding stock price movements.

In conclusion, this study explores the integration of LSTM and Random Forest models for stock market forecasting. By leveraging the strengths of both techniques, it aims to develop a robust and accurate prediction model that can outperform traditional approaches. The results of this research will have significant implications for the financial industry and contribute to the ongoing advancements in the field of stock market analysis and prediction.

2. Methods

2.1. Data Source

The dataset used in this paper was fetched from the Kaggle website (TESLA Stock Data) [1]. This dataset provides historical data on TESLA INC. stock (TSLA). The data is available at a daily level, and the currency is USD. The dataset contains 2,956 groups of data, all of which were selected as samples for this research. The original dataset was in .csv format.

2.2. Variable Selection

Tesla's stock displays high volatility, impacted by internal updates, market sentiments, and industry shifts. Despite this, it demonstrates a persistent upward trajectory due to innovation in electric vehicles and renewable energy. Being technology-focused, Tesla's stock trends are shaped by technological advancements and new products. Market sentiment, influenced by future outlook and Elon Musk's impact, also plays a role in stock performance. Investors must differentiate short-term fluctuations from long-term progress, as illustrated in Figure 1.

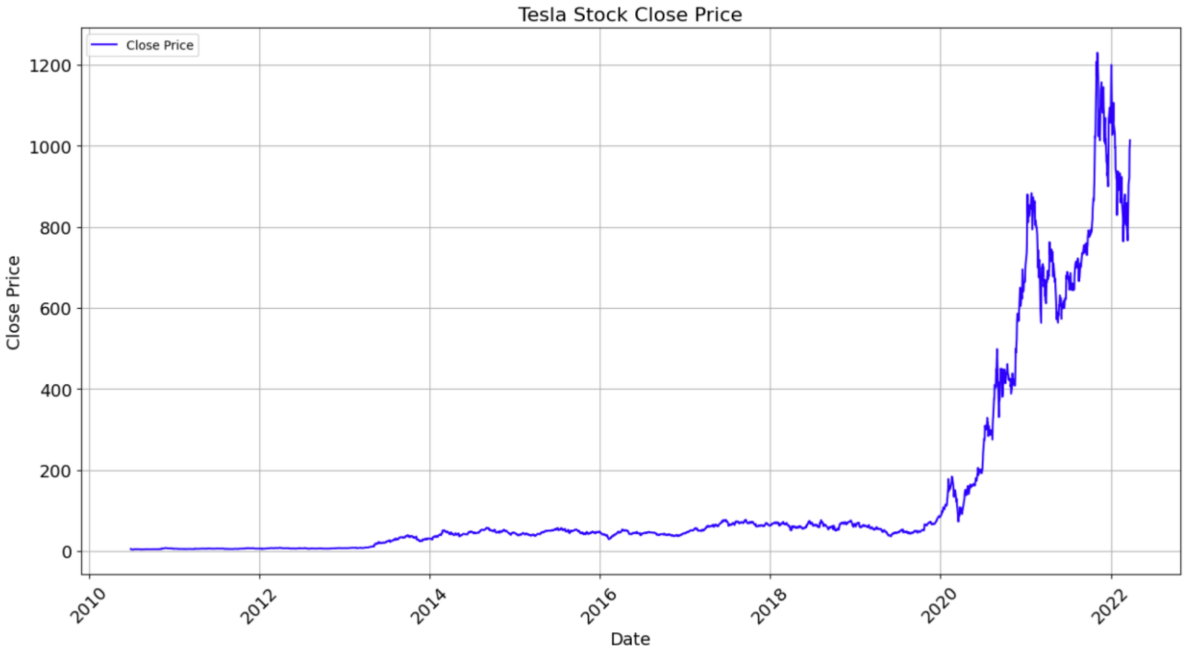

Figure 1: Tesla stock price.

2.3. Model Selection

This study used three machine learning models to predict Tesla stock prices: Random Forest, LSTM, and Linear Regression (as a meta-model). Initially, LSTM was used for its ability to handle time-series data and solve long-term dependency issues. Then, the Random Forest model was applied. It enhanced accuracy and stability by combining multiple decision trees. Finally, linear regression served as a meta-model. It integrated predictions from the other models to improve overall accuracy. Together, these models provided a robust prediction framework for Tesla stock prices.

3. Results and Discussion

3.1. Data Processing

The data is from the Kaggle website platform. The details of the data set are shown in the following table 1. The data set consists of 2,596 time points, covering Tesla's stock information from June 29, 2010, to March 24, 2022. Each time point is measured in days.

Table 1: Tesla dataset information.

Variable	Value
Total number of time points	2596
Sample start time	On June 29th, 2010
Sample end time	On March 24, 2022
Time interval unit	day
Sample variables that were collected	Date, Open, High, Low, Close, Adj Close, Volume

3.2. Univariate LSTM model

The LSTM model in this paper is constructed on the Keras platform, which is a high-level neural network API written in Python, with backends such as Tensorflow and Theano. It can seamlessly switch between CPU and GPU. In the modeling process, only the column of closing price is used. The hyperparameters of the LSTM model are set as follows (Table 2):

Table 2: Parameter setting of the univariate LSTM model.

Set up parameters	Description	Set the value
Batch size	Batch size	17
Activation function	activation function	RELU function
Epochs	Number of iterations of the training set	200
Optimizer	optimizer	Adam optimizer
Look_back	Time window	1
Units	Output dimension	8

The expression of the RELU function is as follows:

$$f(x) = \begin{cases} x \ \ if \ x > 0 \\ 0 \ \ if \ x \leq 0 \end{cases} \tag{1}$$

The Adam (Adaptive Moment Estimation) optimizer dynamically adjusts the learning rate of each parameter by using the first-moment estimate and the second-moment estimate of the gradient. It has the following formula:

$$\Delta\theta_t = -\frac{\widetilde{m_t}}{\sqrt{\widetilde{n_t}}+\varsigma} \times \eta \tag{2}$$

Where, $\mathrm{m_t}$, $\mathrm{n_t}$ are the first and second-moment estimates of the gradient, respectively. $\widetilde{\mathrm{m_t}}$, $\widetilde{\mathrm{n_t}}$ is the correction of $\mathrm{m_t}$, $\mathrm{n_t}$, hence the desired unbiased estimate. And $-\frac{\widetilde{\mathrm{m_t}}}{\sqrt{\widetilde{\mathrm{n_t}}}+\varsigma}$ is a dynamic constraint on the learning rate, with a clear range.

3.3. Model Evaluation

The accuracy of the model is reflected in the deviation between the fitted values and the real values. The less the deviation is, the better the fitting degree is, and the more accurate the model is. This article selects three indicators: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics offer a comprehensive evaluation of model errors, which is crucial for assessing the model's performance amidst stock market volatility.

RMSE highlights the impact of significant errors, which is essential for spotting large prediction biases in extreme market conditions. MAE gives an intuitive average of errors, which is useful for gauging the model's average performance in normal market situations. MAPE focuses on relative errors, which is key for assessing the model's consistency and accuracy across various stock price levels. All these metrics combined provide a nuanced evaluation of the model's reliability.

The performance of the model on the sample data is evaluated after fitting the model. The RMSE, MAE, and MAPE of the data set were 13.84,9.74, and 49.56%, respectively. Figure 2 indicates that the test results exhibit a 'translational shift' compared to the original data. This discrepancy is primarily attributed to the model's inability to capture the 'seasonal' characteristics of the original data. Adjusting the look_back parameter is proposed as a solution. The table below presents the errors in the test data with various look_back settings.

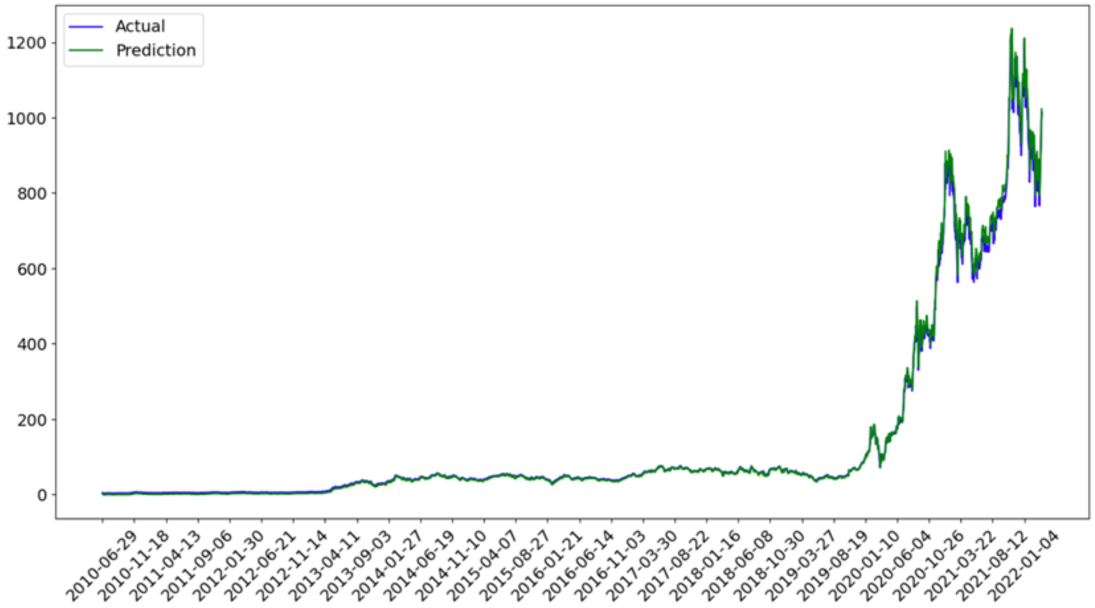

Figure 2: Predicted and true values of the univariate LSTM model.

When conducting time series forecasting using the LSTM univariate model with its time window set to either 3 or 12, the test outcomes indicate a reduced 'translational shift' compared to the original data, alongside smaller errors, leading to more accurate prediction results. This suggests that at these specific time window settings, the LSTM model is better able to capture the temporal dynamics of the data, thereby enhancing the accuracy of its forecasts. However, it's important to note that when the time window is set beyond 12, there is a significant decline in prediction accuracy. This decline could be attributed to the larger time window incorporating more historical information, which may not all be entirely relevant, thus increasing the model's complexity and the uncertainty in predicting future values. Therefore, selecting an appropriate time window size is crucial for optimizing the performance of the LSTM model (Table 3).

Table 3: Evaluation indicators of different time window models

Look_backs	RMSE	MAE	MAPE
1	13.04	9.42	45.47%
2	13.64	9.01	45.71%
3	12.49	9.02	44.36%
4	13.16	9.19	45.33%
5	13.16	9.68	45.45%
6	13.15	9.17	45.43%
7	13.48	9.67	45.68%
8	13.75	9.42	45.66%
9	13.94	9.93	45.93%
10	13.5	9.77	45.32%
11	13.21	9.78	45.36%
12	12.28	8.9	42.61%
13	12.29	9.44	45.46%
14	12.71	9.78	45.28%
15	12.87	9.22	45.48%

3.4. Model Fusion (Random Forest + LSTM)

The analysis of the combined random forest and LSTM model proceeded by first training the random forest model to generate predictions for the dataset, followed by training the LSTM model to also predict outcomes on the same dataset. Subsequently, the predictions from both models were utilized as new features to train a meta-model, such as a linear regression model, effectively integrating these features for enhanced analysis. After fitting the model, finally, look at the performance of the model on the sample (Figure 3):

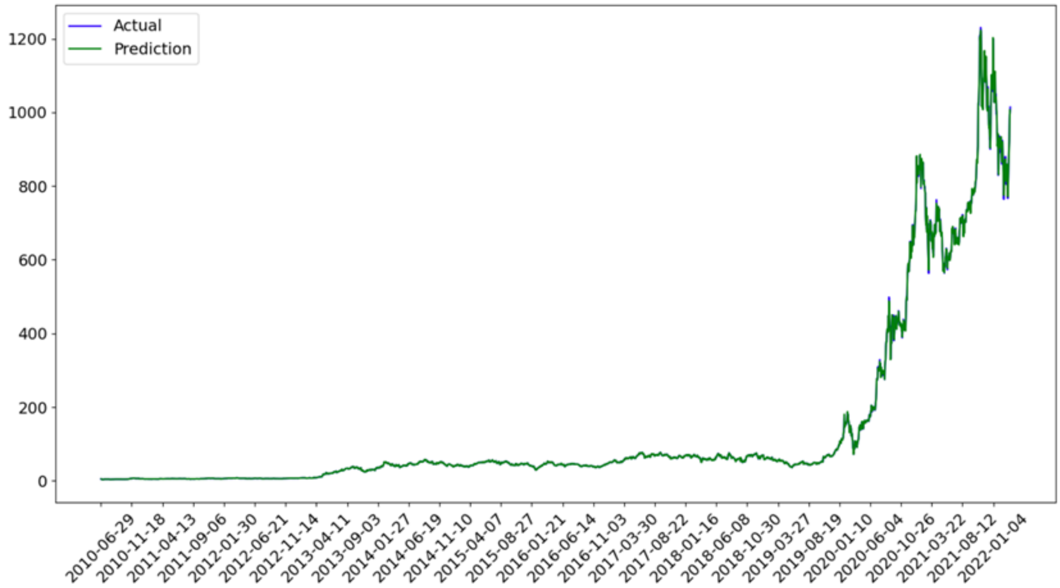

Figure 3: Predicted and true values of the random forest + LSTM model.

The RMSE, MAE, and MAPE were 1.71,0.659, and 1.37%, respectively. This study is based on Tesla stock data and applies the recently popular deep learning techniques to stock prediction. The LSTM model is adopted to solve the problems of gradient vanishing and exploding, which demonstrates strong learning capability and leads to accurate predictions. The innovation of this study lies in the integration of the random forest model and the LSTM model, which is rarely seen in related research. The results show that the performance of the hybrid model outperforms that of the single models.

3.5. Critical Thinking

Although the preceding conclusions have each undergone scrutiny and the projected outcomes hold a degree of credibility, the statement implies the coherence of the given data. Nonetheless, within this study, certain unresolved issues persist.

Primarily, a significant fluctuation in prices around the year 2020 is evident from the depicted images. Prices remained relatively stable before 2020, exhibiting no pronounced alterations; however, post-2020, prices surged dramatically with frequent oscillations. It is posited in this study that disregarding preceding data could enhance the accuracy of predictions and improve testing outcomes. However, such an approach may compromise the integrity of the time series data, leading to a loss of embedded information. Moreover, Tesla stands as a technology-driven enterprise, where its stock trends are often influenced by technological advancements and product innovations. The trajectory of Tesla's stocks is also frequently swayed by market sentiments. Investors' responses to the company's prospects, industry trends, and the words and actions of its founder, Elon Musk, may impact stock performance. The predictions within this study failed to incorporate a broader spectrum of data from such external perspectives to extract features, relying solely on stock data for forecasting. Thus, the predictions within this text persist with inherent limitations. Future research could consider integrating more external data sources to improve the predictive power of the models.

4. Conclusion

In this study, the LSTM model, a deep learning technique, is effectively applied to predict Tesla stock prices. LSTM helps mitigate the problems of vanishing gradients and exploding gradients during the training process, which makes it a significant predictor of performance, demonstrating the model's powerful learning potential.

The innovation of this paper lies in the rare integration of the random forest method and LSTM in related fields. The resulting hybrid model outperforms the single model in performance, demonstrating the possibility of achieving synergies by leveraging the strengths of both algorithms. This study indicates a feasible direction to effectively integrate various machine learning techniques to improve the accuracy of stock market prediction in future research. The findings could provide valuable insights for investors in making investment decisions and for researchers in exploring more advanced hybrid models for financial market forecasting.

References

[1] Khaidem, L., Saha, S. and Dey, S. R. (2016) Predicting the direction of stock market prices using random forest. Procedia Computer Science.

[2] Kadhem, S. and Thajel, H. (2023) Modelling of crude oil price data using hidden Markov model. Journal of Risk Finance, 24(2), 269-284.

[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. IEEE, 1643-1647.

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

[5] Pan, W. (2014) Mixed modified fruit fly optimization algorithm with general regression neural network to build oil and gold prices forecasting model. Kybernetes, 43(7): 1053-1063.

[6] Wu, Y. X., Wu, Q. B. and Zhu, J. Q. (2019) Improved EEMD-based crude oil price forecasting using LSTM networks. Physica A: Statistical Mechanics and its Applications, 516(15), 114-124.

[7] Huang, N. E., Shen, Z., Long, S. R., et al. (1998) The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis. Proceedings of The Royal Society A, 454, 903-995.

[8] María, E. T., Marcelo, A. C., Gaston, S., et al. (2011) A complete ensemble empirical mode decomposition with adaptive noise. Prague, Czech Republic, 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4144-4147.

[9] Pan, S., Li, H., Wang, Y. and Cai, W. (2021) Crude oil price prediction with LSTM neural networks. Computer Technology and Development, 31(5): 180-185.

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