Compare of Linear Regression Model and LSTM Neural Network in Machine Learning

Yixuan Pan
Faculty of Arts and Science, Concordia University, Montreal, Quebec, H3G1M8, Canada

Keywords: Machine Learning, Tesla Stock, Linear Regression, LSTM.

Abstract: Stock forecasting involves analysts leveraging their profound knowledge of the stock market to predict the future trajectory of the stock market and the extent of price fluctuations based on the evolution of stock prices. This predictive activity relies solely on presumed factors and set conditions. Numerous investors employ mathematical models and algorithms to sift through vast datasets, producing stock price predictions. The adoption of machine learning and artificial intelligence technology in this domain is increasingly prevalent. Comparing linear regression models and Long Short-Term Memory Networks (LSTM) in stock market analysis involves evaluating their effectiveness in predicting stock trends. Linear regression, known for its simplicity and ease of interpretation, is suitable for datasets with linear relationships. However, it might not effectively capture complex patterns in financial markets. On the other hand, LSTM, a type of recurrent neural network, excels in handling time-series data and can model complex relationships by learning from long-term dependencies in the data. This makes LSTM more adept at understanding and predicting the often non-linear and volatile nature of stock prices, albeit at the cost of increased computational complexity and a need for more data.

1 INTRODUCTION

The stock market has always been a place where investors are keen and gather. After investing in stocks, investors expect that the funds they invest in can bring them high returns, so they also hope to obtain the same profits when buying and selling stocks. Most investors will rely on experience and their own known knowledge to make stock selections, so improving accuracy is a very important goal in predicting stocks (Kumbure et al 2022). In recent years, with the advancement of machine learning and algorithms, more and more people and even investment banks are using machine learning methods to predict stock prices as an important judgment tool (Kofi et al 2020 & Naqa and Murphy 2015). Therefore, to better fit the requirements, improving the accuracy of the model becomes a top priority. Machine learning is used because predictions use much data and time (Jordan and Mitchell 2015). As the name suggests, machine learning is to enable machines to learn like humans by collecting, storing, analyzing data, and making decisions on their own and more, due to the huge amount of data, manual processing is not feasible. It is therefore particularly important to use machine learning algorithms to analyze and predict such values.

Linear regression is a machine learning method employed to learn or establish patterns (functions) from a labeled training set and make predictions for new instances based on these patterns (Maulud and Abdulazez 2020 & Monner 2012). The training set comprises a series of training examples, each containing an input object (typically a vector) and an expected output. The output of the function can be a continuous value (referred to as regression analysis), or it can predict a categorical label (known as classification). Long Short-Term Memory (LSTM) belongs to the realm of deep learning (LeCun et al 2015), standing for Long Short-Term Memory. As its name suggests, LSTM is a neural network capable of memorizing both long-term and short-term information. It serves as a crucial algorithm for time series analysis and represents a specialized type of Recurrent Neural Network (RNN) adept at learning extended dependencies (Sherstinsky 2020). Its primary purpose is to address issues such as gradient disappearance and gradient explosion encountered during the training of long sequences. In simpler terms, LSTM tends to exhibit superior performance

Pan, Y.
Compare of Linear Regression Model and LSTM Neural Network in Machine Learning.
DOI: 10.5220/0012871700004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 218-223
ISBN: 978-989-758-690-3

in handling longer sequences compared to conventional RNNs. This paper will conduct a comparative analysis of the strengths and weaknesses of the linear regression algorithm in machine learning versus the LSTM algorithm in deep learning, with a focus on utilizing Tesla stock data as the dataset.

2 LSTM NEURAL NETWORK PART

2.1 Analyze Data

This data set is a set of historical stock price data for specific stocks (Table 1).

The dataset spans from June 1, 2021, to July 16, 2021, encompassing information across multiple trading days. Each row corresponds to a trading day, and each column provides data regarding the stock's performance on that particular day. The dataset includes the following columns:

Date: The transaction date.

Open: The opening stock price on the trading day.

High: The peak price attained by the stock during that day's trading.

Low: The lowest price reached by the stock during that day's trading.

Close: The closing stock price on the trading day.

Adj Close: The adjusted closing price, accounting for factors like dividends and stock splits.

Volume: The trading volume, representing the total number of shares traded on that day.

This dataset is valuable for conducting stock price analysis, trend analysis, technical analysis, and other research related to the stock market.

2.2 Model Selection

This article will utilize both the linear regression model and the LSTM neural network method. Linear regression, a widely employed statistical method, predicts the relationship between a variable (dependent variable) and one or more independent variables. In this context, linear regression aims to predict the closing price (Close) of the stock based on other data columns (such as open price, high price, low price, volume, etc.). For the LSTM neural network, the initial step involves data preprocessing, including data standardization, creating time series data, and ultimately splitting the dataset into training and test sets. Subsequently, using libraries like Keras or other deep learning tools, an LSTM model is constructed, typically comprising an LSTM layer and one or more dense layers. The model is trained using the training set, with options for different optimizers, loss functions, and multiple training iterations to enhance performance.

Table 1. The structure of the dataset.

	Date	Open	High	Low	Close	Adj Close	Volume
3014	2022-12-22	136.00	136.63	122.26	125.35	125.35	210090300
3015	2022-12-23	126.37	128.62	121.02	123.15	123.15	166989700
3016	2022-12-27	117.50	119.67	108.76	109.10	109.10	208643400
3017	2022-12-28	110.35	116.27	108.24	112.71	112.71	221070500
3018	2022-12-29	120.39	123.57	117.50	121.82	121.82	221923300
3019	2022-12-30	119.95	124.48	119.75	123.18	123.18	157304500
3020	2023-01-03	118.47	118.80	104.64	108.10	108.10	231402800
3021	2023-01-04	109.11	114.59	107.52	113.64	113.64	180389000
3022	2023-01-05	110.51	111.75	107.16	110.34	110.34	157986300
3023	2023-01-06	103.00	114.39	101.81	113.06	113.06	220575900

Table 2. The structure of the descriptive statistics.

	Open	High	Low	Close	Adj Close	Volume
count	3024.00	3024.00	3024.00	3024.00	3024.00	3.024000e+03
mean	61.39	62.77	59.87	61.34	61.34	9.673891e+07
std	96.89	99.12	94.37	96.76	96.76	8.174686e+07
min	1.45	1.48	1.41	1.46	1.46	3.594000e+06
25%	11.57	11.86	11.21	11.56	11.56	4.700618e+07
50%	16.61	16.85	16.38	16.60	16.60	7.826100e+07
75%	29.05	30.11	28.39	28.70	28.70	1.198309e+08

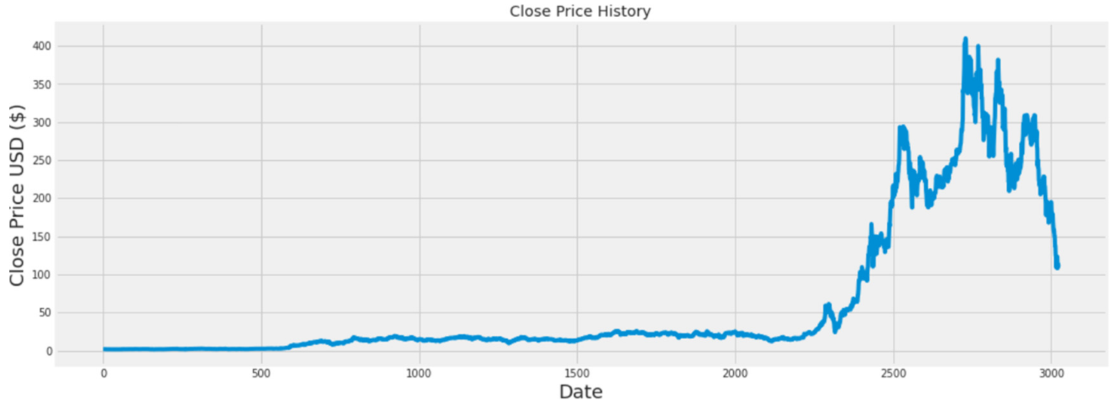

Figure 1. The diagram of information about Tesla stock (Photo/Picture credit: Original).

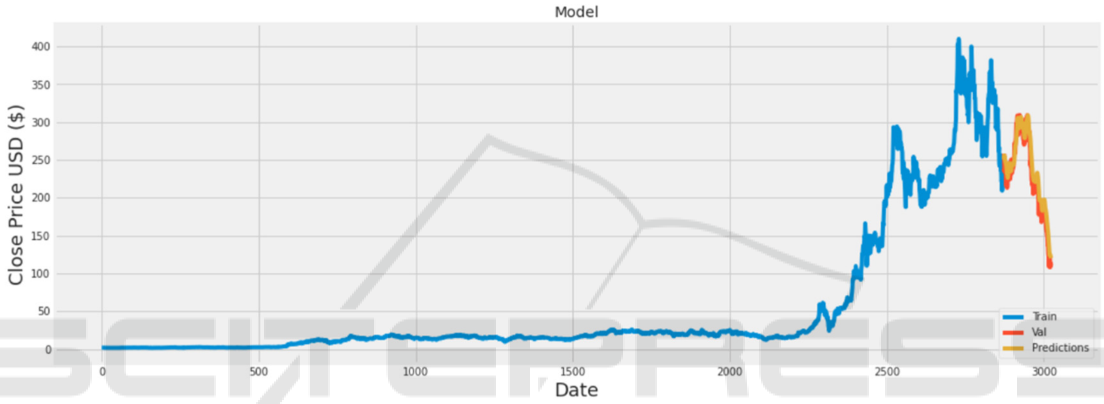

Figure 2. The diagram of the Forecasted data and actual data (Photo/Picture credit: Original).

Evaluation of the model is performed on the test set, employing metrics like mean square error (MSE) and mean absolute error (MAE) (Willmott and Matsuura 2005 & Marmolin 1986). The first step includes importing the necessary packages and downloading the data for analysis. Upon successful download, the first ten data entries are displayed (Table 1). Following a check for obvious errors, descriptive statistics are employed to obtain more detailed data (Table 2) (count, mean, std, min, 25%, 50%, 75%, max), providing a quick overview of the data's basic characteristics.

Then use the corresponding function (which is used to obtain the column name /column label information in the data frame) to get the properties of the Pandas data frame (DataFrame). The second step is to visualize some information about Tesla stock, as shown in (Figure 1), which respectively visualizes Tesla's stock price, trading volume chart, structural technical indicators, and yield rate.

2.3 Data Standardization

To standardize, normalize, and preprocess data using sklearn during model training, the sklearn.preprocessing module is utilized. The objective is to achieve faster model convergence. Normalization is considered a form of standardization, mapping data to the interval [0,1], while standardization scales data to a specific interval. Standardized data has a mean of 0 and a standard deviation of 1, allowing it to be positive or negative. Detailed operational steps are provided in the code link at the end of the article.

The training set and test set are divided using specific steps outlined in the link code.

The establishment and training of the LSTM model follow the steps presented in the link code.

Prediction and obtaining prediction results involve specific steps detailed in the link code. The final result of the Root Mean Square Error (RMSE) is 13.968, and the visualization of results is depicted in Figure 2.

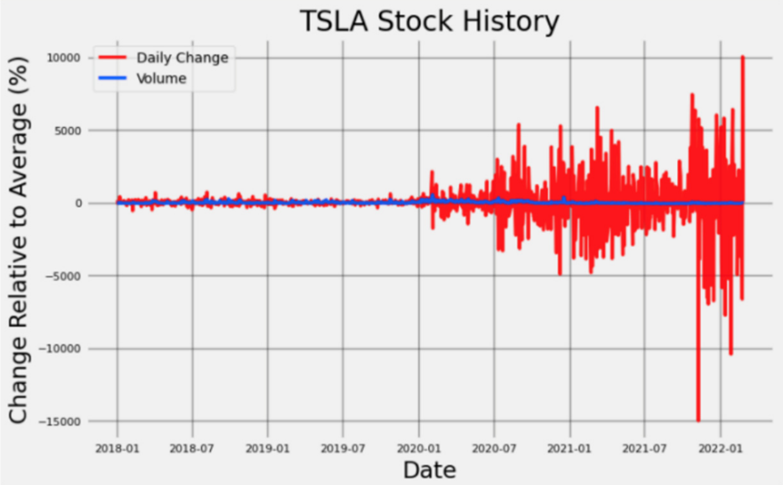

Figure 3. The diagram of the range of stock price (Photo/Picture credit: Original).

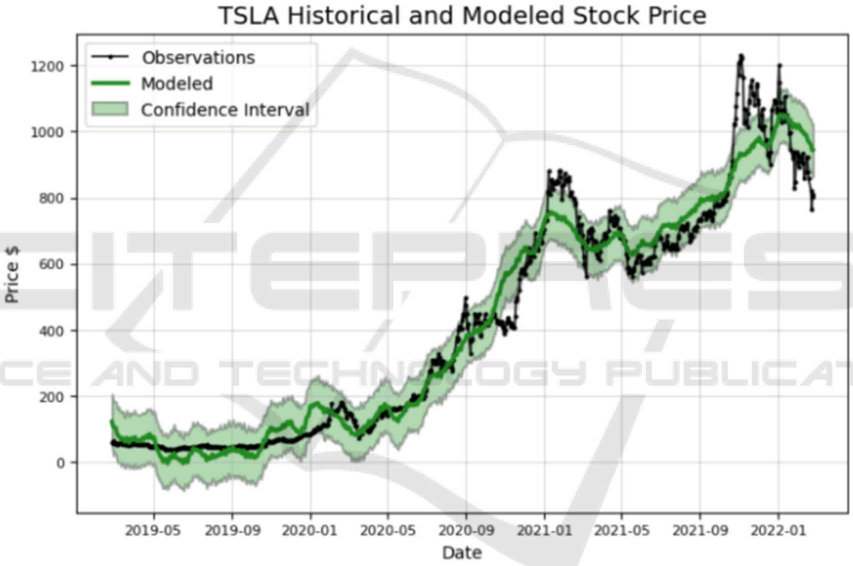

Figure 4. The diagram of the confidence intervals (Photo/Picture credit: Original).

3 LINEAR REGRESSION MODEL PART

3.1 Profit Based on Model Prediction

The idea is as follows, design a model that will lead to better results than "the profits of long-term investment and the profits of setting up the time machine to the maximum extent". This is the desire to identify the best buying and selling points for a stock. To control transaction risks as much as possible. The specific operation is shown in the link code at the end of the article. In Figure 3, it shows the most violent range of stock price fluctuations.

For those who hold it for a long time, it does not matter if the stock keeps rising, but for those who invest in the short term, it means that this stock There is more room for maneuver, but of course, the risks are also greater. By calling the class written by the model, it can see the relationship between the actual changes in stocks, model predictions, and confidence intervals (Figure 4). This can better adjust the model to make predictions.

3.2 Stock Evaluation and Prediction

It can be seen from (Figure 5) that the average error of the test data is $228.87 and the average error of the data predicted by the trained model is $52.24. This means that the error generated by the trained model has been greatly optimized. In addition, it can be seen that the accuracy rate of predicting stock price increases is 47.95% and the accuracy rate of predicting stock price decreases is 38.27%.

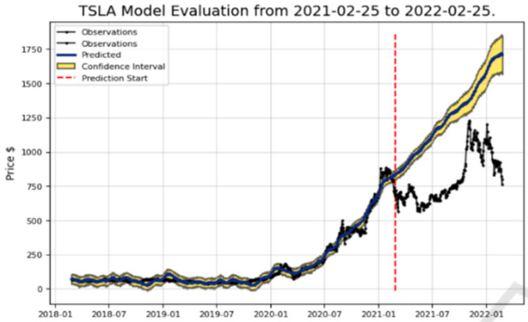

Figure 5. The diagram of the prediction result (Photo/Picture credit: Original).

4 DISCUSSION

It is not difficult to see that the overall accuracy of the LSTM model is higher than the linear regression model. This is evident from the accuracy provided in predicting stock price increases and decreases. The ability of LSTM models to capture nonlinear patterns and dependencies in data likely contributes to their accuracy. The RMSE of the LSTM model is 13.9676, which falls in the range of 10 < RMSE < 20, indicating that its performance is moderate. RMSE measures the average error between predicted and observed values. Lower RMSE means better model performance. While the RMSE of the LSTM model shows that it can provide useful predictions, the modest performance indicates that there is still room for improvement. This may involve fine-tuning model hyperparameters, adjusting the network architecture, or adding more features. The accuracy of the linear regression model in predicting stock price increases and decreases is 47.95% and 38.27% respectively. These accuracy rates are lower relative to the LSTM model. Linear regression models assume a linear relationship between input features and target variables, which may not fully capture the complexity in stock price movements. In summary, although the LSTM neural network model is better than the linear regression model in accuracy, there are opportunities for improvement and perfection in both methods. Choosing a model may depend on the specific requirements of the task, such as the trade-off between interpretability and predictive accuracy, and the computational resources available for training and deployment.

5 CONCLUSION

Based on the comparison of machine learning models for two different types of tasks, linear regression model and LSTM neural network using Tesla stock, this article concludes that each has some advantages and disadvantages. In general, linear regression is suitable for simple linear. For relational problems, the model is simple and easy to understand, while LSTM is suitable for processing complex nonlinear time series data and has stronger modeling capabilities. The choice of model to choose should be determined by the nature of the specific problem and the type of data. Using linear regression and LSTM for predicting stock market trends involves the analysis of historical stock data. Linear regression, although straightforward, may not capture the complex patterns inherent in stock data. LSTM, a type of recurrent neural network, is better suited for time-series data like stock prices, as it can remember and leverage long-term dependencies. However, there are still some research limitations like market volatility (both models may struggle in the face of unpredictable market behavior), data quality (the accuracy of predictions heavily relies on the quality and completeness of the historical data used), overfitting (especially with LSTM, there is a risk of overfitting to historical data, which can reduce its effectiveness in forecasting future trends).

For future research directions, there are hybrid models (investigating the combination of linear regression with LSTM or other machine learning techniques to create more robust prediction models), feature engineering (experimenting with various input features, such as economic indicators or social media sentiments, to enhance the prediction accuracy.), real-time analysis (Incorporating real-time data streams for more current and relevant predictions), algorithmic improvements (continuously refining and adapting algorithms to better suit changing market conditions).

REFERENCES

M. M. Kumbure, C. Lohrmann, P. Luukka, J. Porras. Expert Systems with Applications, 197, 116659, (2022).

N. I. Kofi, A. F. Adekoya, B. A. Weyori. The Artificial Intelligence Review, 53(4), 3007-3057, (2020).

El Naqa, M. J. Murphy. Springer International Publishing, 3-11, (2015).

M. I. Jordan, T. M. Mitchell. Science, 349(6245), 255-260, (2015).

D. Maulud, A. M. Abdulazeez. Journal of Applied Science and Technology Trends, 1(4), 140-147, (2020).

D. Monner, J. A. Reggia. Neural Networks, 25, 70-83, (2012).

Y. LeCun, Y. Bengio, G. Hinton. Nature, 521(7553), 436-444, (2015).

Sherstinsky. Physica D: Nonlinear Phenomena, 404, 132306, (2020).

C. J. Willmott, K. Matsuura. Climate Research, 30(1), 79-82, (2005).

H. Marmolin. IEEE Transactions on Systems, Man, and Cybernetics, 16(3), 486-489, (1986).