Proceedings of the 3rd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/30/20241093

Predicting Twitter stock price using linear regression, random forest, and MLP regression

Ziyu Huang

Rosedale Global High School, 650032, Kunming, China

196102107@mail.sit.edu.cn

Abstract. Since the chaos and complexity in the stock market, predicting stock prices with machinery approaches helps assert a new perspective for people to reach satisfactory results in stock analysis. The stock price of Twitter reflects its market value and performance, which are influenced by various factors such as user engagement, revenue, news, and sentiment. Predicting the stock price of Twitter is a challenging task that requires sophisticated methods and data analysis. In this essay, the author compares three different machine learning models to predict Twitter's stock price: linear regression, random forest, and MLP regression. The paper uses historical data on Twitter's stock price and various features such as volume, peak value, and trough value. The researcher evaluates the performance of each model using metrics such as Mean-squared Error and R-squared and finds that MLP regression outperforms the accuracy and generalization of the other two models. The author also discusses the limitations and implications of our findings for investors and researchers.

Keywords: Stock Price Prediction, Twitter, Linear Regression, Random Forest, MLP Regression.

1. Introduction

The stock market has always been a complex domain where investors seek accurate predictions to make informed decisions. With the advent of social media platforms, such as Twitter, their impact on stock prices has gained significant attention. The difficult challenge of predicting stock values necessitates the application of highly sophisticated statistical models and machine learning algorithms. There has been an increase in interest in applying different strategies in recent years, such as LSTM, Decision Trees, and Support Vector Machines (SVM), to predict stock prices [1]. Through a thorough analysis of the data and a review of the relevant literature, this report aims to explore the applicability of machine learning algorithms, specifically Linear Regression, Random Forest, and MLP Regression, to predict the stock price of Twitter. By employing these algorithms, we can attempt to capture patterns and correlations from historical data to provide robust predictions.

Predicting stock prices accurately has always been a challenge, as it involves understanding various factors that impact market movements. The majority of stock price predictions primarily utilize historical stock price indicators and neglect the incorporation of sentiment analysis in forecasting [2]. Moreover, many deep learning models struggle with capturing dependencies over long distances, particularly when dealing with extensive datasets. Consequently, there has been insufficient attention given to long-term market sentiment and its ability to trigger unpredictable fluctuations in stock prices, which may fail to fully capture all the available market information.

DOI: 10.54254/2753-8818/30/20241093

Leveraging social media data, particularly Twitter, has emerged as a novel way to gain insights into investor sentiment and market trends. By combining traditional financial analysis with machine learning techniques, the author aims to build models that can enable investors to make well-informed decisions [3].

For this study, historical Twitter data and corresponding stock prices of Twitter Inc. will be collected and utilized. Twitter's stock price data from 2013 Q4 to 2022 Q4, along with relevant Twitter metrics such as tweet sentiment, tweet volume, and user engagement, will be collected as features. This data will enable us to train and evaluate the efficiency of the prediction models. The indicators in this dataset are the Opening Price, Closing Price, Highest Price, Lowest Price, and Volume.

2. Related works

A similar approach was used in Kumar, V., & Rani, M. [4]. This research paper contrasts the outputs of the aforementioned algorithms in predicting stock prices and provides valuable insights into their strengths and limitations. "Stock price prediction with LSTM and random forest" presents a step-by-step guide to build an LSTM and random forest model for predicting stock price percentage change. This conference paper explores the integration of Long Short-Term Memory (LSTM) and Random Forest algorithms for stock price prediction, showing promising results [5]. In "A federated learning-enabled predictive analysis to forecast stock…" [6], The authors suggested machine learning models, such as MLP, to assist investors in making buy-or-sale decisions by fusing big data approaches with fundamental analysis. Cakra practiced the Random Forest and Naive Bayes algorithms to categorize tweets for sentiment analysis. Linear regression was then utilized to forecast the company's stock price [7]. These works collectively highlight the potential of machine learning algorithms in predicting stock prices. However, it's crucial to remember that while these models can capture patterns and correlations from historical data, predicting stock prices accurately remains an uphill task due to the multitude of factors impacting market movements.

3. Application of machine learning method in Twitter stock prediction

3.1. Application of LSTM model in Twitter stock prediction

3.1.1. Introduction of Linear Regression Model. Assuming a linear relationship between both independent and dependent variables, linear regression is a straightforward but effective procedure [8]. By fitting a line to the historical data, this model attempts to predict future stock prices according to historical trends and correlations.

3.1.2. Application analysis of Linear Regression Model in Twitter stock prediction. Using Data Pre-processing describes the data pre-processing steps undertaken in the analysis [9]. It explains how the 'close data' was copied, the 'Date' column was removed, and the data was scaled using the MinMaxScaler. The train-test split process is explained, where there are training and testing sets generated from the scaled data. The proportion of data allocated to each set is mentioned. The Model Training explains the model training process using the Linear Regression algorithm. It outlines the use of the 'X_train' and 'Y_train' data for training the model.

The predictions made by the trained model on both the training and testing sets are discussed. The values for 'y_train_pred' and 'y_test_pred' are displayed, indicating the model's ability to predict stock prices.

The performance metrics that were employed to evaluate the model's performance are explained. The R-squared values ('train_r2' and 'test_r2') are interpreted, indicating the proportion of variance explained by the model. The mean squared error ('test_mse') is discussed, which measures the mean squared difference between the values that were anticipated and those that were observed.

DOI: 10.54254/2753-8818/30/20241093

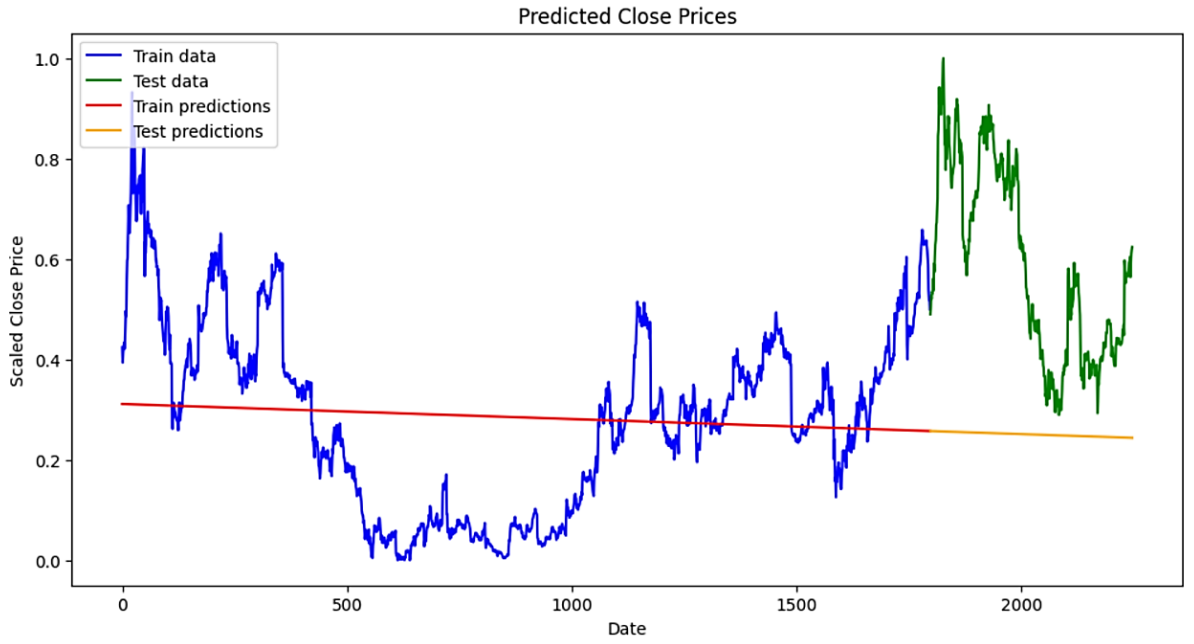

Figure 1. Performance of the Predicted Close Prices Using Linear Regression Model

Therefore, the visualization techniques employed in the analysis are explained as shown in Figure 1.

The graph generated using the matplotlib library is discussed, showing the original close prices, training predictions, and testing predictions. The significance of the visualization in understanding the model's performance is emphasized [10].

3.2. Application of Random Forest Model in Twitter stock prediction

3.2.1. Introduction of Random Forest Model. The ensemble approach called Random Forest uses several decision trees combined to provide predictions [11]. By leveraging the power of aggregation, Random Forest can capture complex relationships and interactions between variables, enhancing the accuracy of stock price predictions.

3.2.2. Application analysis of Random Forest Model in Twitter stock prediction. The author's methodology revolves around the implementation of the Random Forest model. This model exhibits exceptional proficiency in handling complex data and capturing non-linear relationships. By leveraging an ensemble of decision trees, the Random Forest model offers robust prediction capabilities.

To train and test our model, the author curated a diverse dataset encompassing various features. These features include historical stock data, market sentiment, trading volume, and financial indicators. To ensure unbiased evaluation, the dataset was carefully split into training and testing sets. The designated training data was used to train the Random Forest model. By executing the appropriate code snippet, the model was able to learn from the provided dataset and establish patterns and correlations. After assessing the model's performance, the author employed several evaluation metrics. The Mean Squared Error (MSE) and R-squared values for the training and testing sets were computed. The corresponding code snippet was utilized to generate these metrics.

The analysis revealed promising results in terms of the model's predictive capabilities. The MSE and R-squared values indicated a favorable performance in predicting Twitter stock prices [12]. A comprehensive comparison between how well the model performs on the training and testing sets was conducted to gauge its generalization abilities.

DOI: 10.54254/2753-8818/30/20241093

This paper therefore showcased the predicted close prices of Twitter stock using the trained Random Forest model.

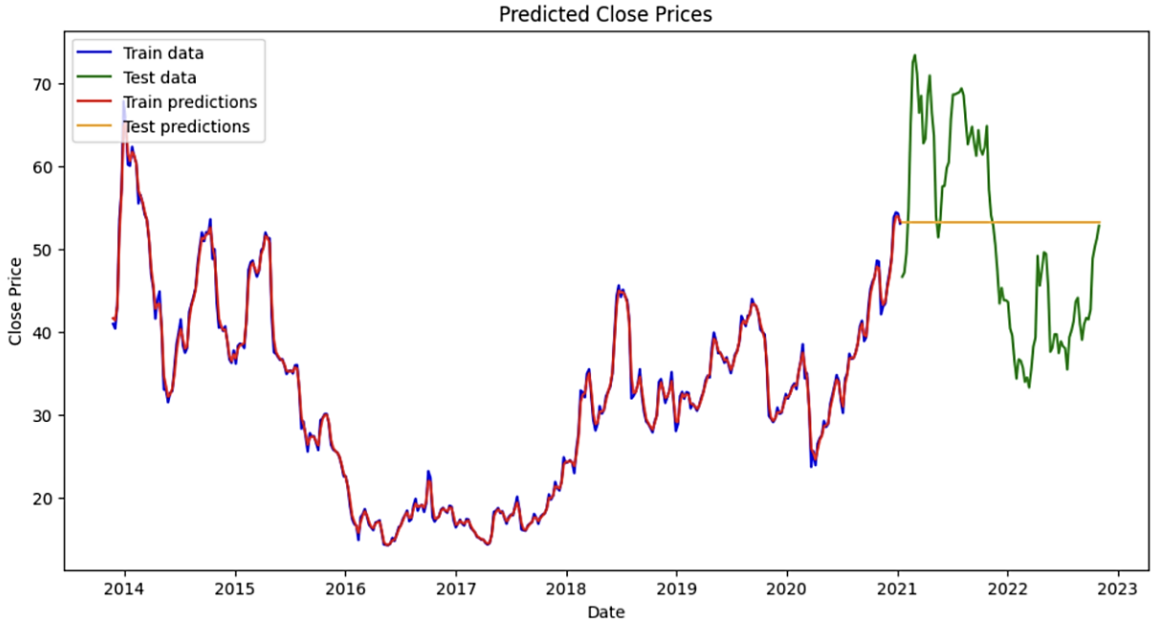

Figure 2. Performance of the Predicted Close Prices Using Random Forest Model

By including the relevant code snippet, the visual representation allowed for better comprehension of the model's predictions. By incorporating machine learning techniques, investors can leverage these predictions to make informed decisions. However, it is imperative to recognize the constraints of the model and investigate possible avenues for enhancement.

3.3. Application of MLP Regression Model in Twitter stock prediction

3.3.1. Introduction of MLP Regression Model. An artificial neural network type called Multilayer Perceptron (MLP) Regression is able to identify non-linear correlations between variables. [13]. By using multiple layers of interconnected nodes, MLP Regression can model complex patterns in the data, potentially improving the accuracy of stock price predictions [14].

3.3.2. Application analysis of MLP Regression in Twitter stock prediction. Using the train_size and test_size variables, the dataset was divided into training and testing sets. The training data (train_data) consisted of 80% of the close_data, while the testing data (test_data) comprised the remaining 20%. For both the training and testing sets, the X_train and X_test lists were populated with the close_data values up to the second-to-last element. The corresponding y_train and y_test lists were populated with the subsequent close_data values. The MLP Regressor model was utilized for training the Random Forest model. The model was configured with a single hidden layer of 10 neurons, a 'tanh' activation function, and the 'adam' solver. The X_train and y_train data were used to fit the model.

The trained model was used to predict the stock prices for the X_test data. The predicted values (y_pred) were then transformed back to their original scale using the scaler.inverse_transform() function. The R-squared and Mean Squared Error (MSE) metrics were calculated using the y_test and y_pred values.

DOI: 10.54254/2753-8818/30/20241093

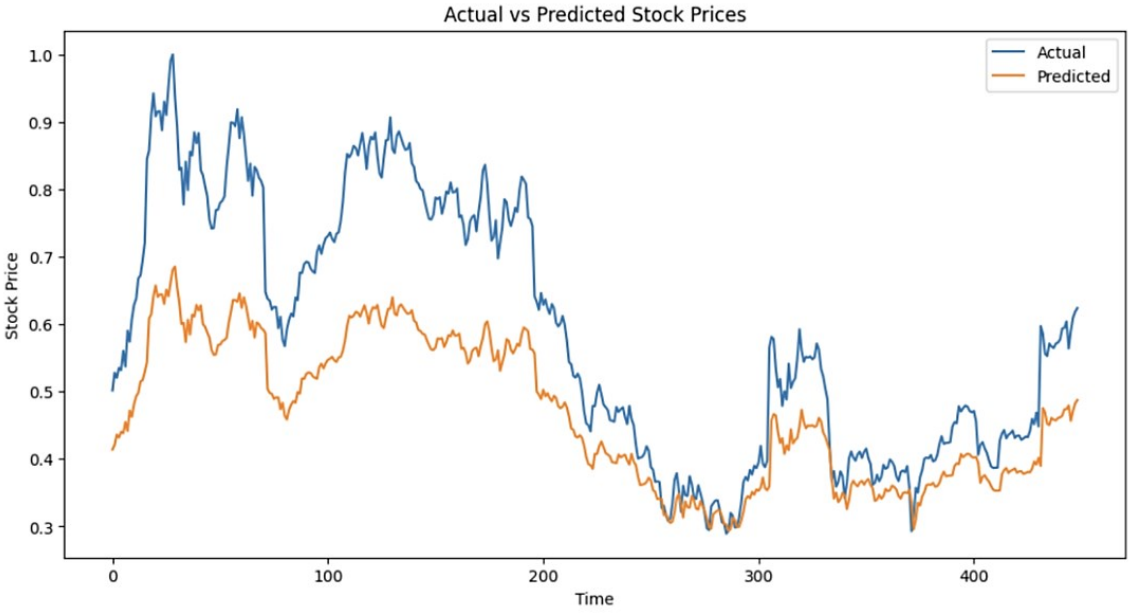

Figure 3. Comparison between the Predicted and Actual Close Prices using MLP Regression

The actual stock prices (y_test) and the predicted stock prices (y_pred) were plotted using the matplotlib library. The resulting line plot displayed the actual and predicted stock prices over time, providing a visual representation of the model's performance.

4. Discussion

The MLP Regression model, as presented in Table 1, exhibited the smallest average squared difference between the projected and actual stock values, with an MSE value of 0.0222. On the other hand, both Linear Regression and Random Forest models had higher MSE values, with Random Forest having the highest MSE of 0.3337. Therefore, the MLP Regression model performed the best in terms of minimizing the prediction errors.

Table 1. MSE and R-Squared values comparison in three models

	Linear Regression	Random Forest	MLP Regression
MSE	0.149203235199881	0.3336964317509105	0.0222112712308604
R-Squared	0.007121053374803	0.00712105337480	0.3829143840600147

R-squared Comparison: The percentage of the variance in the dependent variable (stock prices) that can be accounted for by the independent variables (features) is expressed as the R-squared value [15]. The MLP Regression model in this instance had the greatest R^2 value of 0.3829, meaning it could account for about 38.29% of the variation in the stock prices. On the other hand, both Linear Regression and Random Forest models had lower R^2 values of 0.0071, suggesting that they explained a very small portion of the variance. Therefore, the MLP Regression model performed the best in terms of capturing the variability in the stock prices.

Based on the provided values, the MLP Regression model outperformed both Linear Regression and Random Forest models in terms of both MSE and R^2 metrics. It achieved the lowest MSE, indicating better prediction accuracy, and the highest R^2, indicating a better fit to the data. However, it's crucial to remember that these results are purely based on the values that were supplied; additional research and analysis may be needed to reach more firm conclusions.

DOI: 10.54254/2753-8818/30/20241093

5. Conclusion

Considering there are so many variables influencing market movements, precisely predicting stock prices is still a difficult undertaking. However, by leveraging the power of machine learning algorithms such as Linear Regression, Random Forest, and MLP Regression, the author can attempt to capture patterns and correlations from historical data to provide robust predictions. Through the analysis of the literature, it is evident that these algorithms have shown promise in the field of stock price prediction. Further experiments and improvements in the models can lead to enhanced prediction accuracy and offer valuable support to investors in making informed decisions.

References

[1] Hamoudi, H. Elseifi, M. A. 2021, Stock Market Prediction using CNN and LSTM, (Stanford University).

[2] Orsel, O. E. Yamada, S. S. 2022, Comparative Study of Machine Learning Models for Stock Price Prediction. (ArXiv.org).

[3] Huang, Y. Capretz, L. F. Ho, D. 2021, Machine learning for stock prediction based on fundamental analysis. (In 2021 IEEE Symposium Series on Computational Intelligence), pp. 01-10.

[4] Kumar, V. Rani, M. 2020, A Comparative Performance Analysis of Linear Regression, Random Forest, and MLP Regression Models for Stock Price Prediction, (International Journal of Scientific Research in Computer Science and Engineering, vol. 8), no. 1, pp. 528-533.

[5] Wang, X. Mao, Y. Zhu, Z. Tao, X. 2017, Stock price prediction with LSTM and random forest, (IEEE International Conference on Big Data), pp. 233-238.

[6] Saeid, P. A. Du, N. Cai, L. Yang, J.-C. Bi, Z. Chen, L. 2023, A federated learning-enabled predictive analysis to forecast stock market trends. (Journal of Ambient Intelligence and Humanized Computing, vol. 14), no. 4, pp. 4529–4535.

[7] Cakra, Y. E. Trisedya, B. D. 2015, Stock price prediction using linear regression based on sentiment analysis. (In 2015 international conference on advanced computer science and information systems), pp. 147-154.

[8] Skiera, B. Reiner, J. Albers, S. 2022, Regression Analysis. In: Homburg, C., Klarmann, M., Vomberg, A. (eds) Handbook of Market Research, (Springer, Cham).

[9] García, S. Ramírez-Gallego, S. Luengo, J. Benítez, J. M. Herrera, F. 2016, Big data preprocessing: methods and prospects. (Big Data Analytics, vol. 1), no. 1.

[10] Shahril Khuzairi, N.M. Che Cob, Z. 2021, A Preliminary Model of Learning Analytics to Explore Data Visualization on Educator's Satisfaction and Academic Performance in Higher Education. (Lecture Notes in Computer Science, vol. 13051.).

[11] Breiman, L. 2001, Random Forests. (Machine Learning, vol. 45), no. 1, pp. 5–32.

[12] Chai, T. Draxler, R. R. 2014, Root mean square error (RMSE) or mean absolute error (MAE). (Geoscientific model development discussions, vol. 7), no. 1, pp. 1525-1534.

[13] Pardo, C. Rodríguez, J.J. García-Osorio, C. Maudes, J. 2010, An Empirical Study of Multilayer Perceptron Ensembles for Regression Tasks. In: García-Pedrajas, N., Herrera, F., Fyfe, C., Benítez, J.M., Ali, M. (eds) Trends in Applied Intelligent Systems. IEA/AIE 2010. (Lecture Notes in Computer Science, vol. 6097).

[14] Janković, R. Amelio, A. 2018, Comparing multilayer perceptron and multiple regression models for predicting energy use in the balkans. (arXiv preprint arXiv:1810.11333).

[15] Kingsmore, K. M. Puglisi, C. E. Grammer, A. C. Lipsky, P. E. 2021, An introduction to machine learning and analysis of its use in rheumatic diseases. (Nature Reviews Rheumatology, vol. 17), no. 12, pp. 710-730.