Predicting Stock Fluctuations: Comparative Analysis of Advanced Machine Learning Models Using Tesla Stock

Mingjie Zhu[1,a], Yuan Cheng[2,b], Lin Huang[2,c], Ningjia Duan[1,d]

[1]Department of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China

[2]Department of Methematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China

[a]Mingjie.Zhu22@student.xjtlu.edu.cn, [b]Yuan.Cheng22@student.xjtlu.edu.cn,
[c]Lin.Huang22@student.xjtlu.edu.cn, [d]Ningjia.Duan22@student.xjtlu.edu.cn

Keywords: Deep Learning, LSTM, XG Boost, Random Forest, Sentimental Analysis

Abstract: This essay explores the use of advanced machine learning models to predict stock fluctuations, focusing on Tesla's stock performance. It emphasizes the advantages of these models in handling complex datasets and non-linear relationships. The experiment evaluates three models: Random Forest Regressor, LSTM, and XGBoost, using historical stock prices and sentiment analysis of Twitter data. The results show that XGBoost performs the best, followed by LSTM, while Random Forest Regressor exhibits lower accuracy. XGBoost and LSTM align more frequently with real stock trends, while Random Forest matches only a small percentage. Overall, this essay provides valuable insights into the application of advanced machine learning models for predicting Tesla's stock trends, benefiting traders, analysts, and investors in the stock market.

1. Introduction

Recent years, there has been increasing interest in using advanced machine learning models to predict stock growth or decline. Advanced machine learning refers to the application of the latest machine learning techniques and algorithms to solve complex problems and challenges. These advanced models offer several advantages. Firstly, they can handle large-scale and complex datasets, allowing us to extract valuable information and patterns from massive amounts of data. Secondly, advanced machine learning methods such as deep learning and reinforcement learning have strong learning and adaptability capabilities, enabling automatic learning and adaptation to evolving environments. Additionally, advanced machine learning methods can handle non-linear relationships and high-dimensional features, providing more accurate predictions and decision-making capabilities. Importantly, advanced machine learning techniques can leverage existing knowledge and models through transfer learning and ensemble learning, accelerating the learning process and improving model performance. These advantages make machine learning a valuable tool for predicting stock fluctuations.

Tesla, a leading electric vehicle (EV) manufacturer, has established itself as a prominent player in the stock market. With its innovative technology, strong brand recognition, and visionary leadership of CEO Elon Musk, Tesla has captivated investors and garnered significant attention worldwide. The company's stock, traded under the ticker symbol TSLA, has witnessed remarkable growth and volatility, making it a focal point for traders, analysts, and enthusiasts alike. Tesla's market capitalization has soared, and its stock price movements often influence the broader EV industry and even the overall market sentiment. Overall, the significance of Tesla's stock stems from its position as an innovative leader, advanced technology, and strong brand influence. Its standing in the electric vehicle industry and its vision for the future of sustainable energy make Tesla a hot topic of interest for investors and market participants.

This article primarily introduces three different models for predicting the trend of Tesla stock in the market and compares them based on the analysis of their application methods and principles to determine which model has a more accurate predictive effect on Tesla stock.

1.1. LSTM

It is generally considered that the task of predicting the stock market is challenging due to its volatility and noise characteristics. In modern social economy and organization, the question of how to accurately predict stock movement remains an open one. Due to investor concerns and the attraction of high returns, many related studies have emerged in economic science. The main component of conventional stock market prediction methods is time-series analysis.

Li et al. conducted an extensive analysis of investor sentiment predictability and developed an LSTM model based on extracted investor sentiment to predict the opening and closing prices of the Chinese stock market. They used out-of-sample forecasts of the CSI 300 index and found that the model could predict the opening prices of the next trading day more accurately compared to the closing price [1]. In another study by Maqbool et al. [2], MLP-Regression was used to accurately predict stock prices of Tata Motors and Tata Steel. The model achieved a 0.90 accuracy for both the current trend and future trends over a 10-day period. Zhang et al. [3] developed a sentiment-guided adversarial learning model for predicting stock prices, employing a conditional GAN (CGAN), which is a popular variation of GAN. Their approach incorporated sentiment analysis to enhance the predictive models. Qi et al. [4] utilized LSTM to accurately predict the price of single stocks by leveraging their historical price index data as experience. Another hybrid model that combines deep learning and sentiment analysis was proposed in a paper [5]. This model incorporated long short-term memory (LSTM) neural networks and technical indicators from the stock market. The results showed that the proposed model outperformed a single model and a model without sentiment analysis in terms of predicting stock prices and classifying investor sentiments. Thormann et al. [6] introduced the concepts of financial feature engineering and the architecture of LSTM for highly nonlinear stock price forecasting. They combined sentiment analysis with technical financial indicators and used past tweets for sentiment analysis to forecast Apple stock prices 30 minutes and 60 minutes ahead. In a recent study by Wilksch et al. [7], an improved LASSO-LASSO model was applied to forecast stock direction by combining technical analysis and sentiment analysis. The performance of this model was found to be superior to previous models such as VADER, NTUSD-Fin, FinBERT, and Twitter RoBERTa when evaluated on Twitter posts. Additionally, the training and inference costs of the model were significantly lower compared to BERT-based large language models. Furthermore, an important contribution to the field of stock price prediction is mentioned in [8], which describes an improved LASSO-LASSO approach that combines technical analysis and sentiment analysis. These studies, along with other influential works [9], have advanced the understanding and application of sentiment analysis and deep learning techniques in predicting stock prices.

Three models which includes Random Forest, XG Boost and LSTM Long short-term memory are used to predict the stock price of Tesla. The models can build a relationship between the stock price and sentiment coefficient. To describe the accuracy of fitting results we use the mean squared error (MSE) and mean absolute percentage error (MAPE). The result shows that Random Forest (RF) has a lower variance. Moreover, in this case, we also compare whether the models can accurately predict the trend of stock price changes. Therefore, we compared the increasing and decreasing nature of the actual stock prices over time with those predicted by each model and found that the XG Boost predict more accurately compared with other models while RF performed the worst. Experimental results indicate that XGBOOST scored the highest, possibly because our research is beneficial for improving existing models and developing new ones. It may provide financial professionals with excellent tools.

2. Methodology

2.1. LSTM

Currently, deep learning advances have improved the accuracy of recurrent neural networks (RNNs) when learning long sequences. It aims to improve deep learning networks' effectiveness by increasing the convergence of models during gradient descent. However, gradients in RNNs are prone to disappearing or exploding. In order to solve this problem, we propose a LSTM neural network. LSTMs are neural networks that use a particular gating mechanism to assess learning sequences, save sequence attributes, and modify the present situation based on the sequence properties. LSTM utilize three types of gates: input gates, output gates, and forget gates. There is a set of these three gate in each cell of the sequential model.

1) Forget Gate

Forget gate updates the long-term memory by multiplying it with an index (f_t) generated by the layers input and the short-term memory to determine what percentage of the long-term memory will be remained. The output value ranges between 0 and 1. When f_t is 1, the long-term information will be totally reserved. Equation (1) is the specific equation of the forget gate.

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_0) \tag{1}$$

The activation function σ here is a $tanh$ function, $W_f$and b_f are the matrics of weightings and biases. x_t is the current input of this layer.

2) Input Gate

Input gate updates the long-term memory by adding it up by i_t, which is the product of function σ and function γ. Which determines the potential long-term memory and the memorizing rate of it. The eventual long-term memory value that will be inherited by next layer is generated by this gate. Eq.2 shows how the adder i_t comes.

$$i_t = \sigma(W_t * [h_{t-1}, x_t] + b_t) * \gamma(W_t * [h_{t-1}, x_t] + b_t) \tag{2}$$

Where the σ function is a *sinh* function, and the γ is a *sigmoid* function. While x_t is the input of this layer.

3) Output Gate

The second output of the entire LSTM model layer is the short-term value (l_t), which are generated eventually by the output gate. Eq. 3 and 4 shows how to calculate the intermediate value s_t and the equation that generates l_t.

$$s_t = \sigma(W_o * [h_{t-1}, x_t] + b_t) \tag{3}$$

$$l_t = \gamma(i_t) * s_t \tag{4}$$

As mention previously, the σ function is a *sinh* function and γ is a *sigmoid* function.

2.2. XG Boost

A first-time suggestion for XGBoost was made in 2014 by Chen et al.21. The system implements a gradient boosting algorithm and a decision tree based on gradient lift. The XGBoost model extends the objective function by second-order Taylor, and solving the quadratic function is the optimization issue of the objective function. Furthermore, to improve generalization performance, tree complexity is simultaneously introduced to the goal function. Models have the following primary functions (eq.5).

$$O = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{i=1}^{k} \Omega(f_k) \tag{5}$$

Where *n* is the sample size, y_i is the *i*-th sample's true value, $\widehat{y_i}$is its predicted value, $y_i$is the *i*-th sample's true value, $L(y_i, \hat{y}_i)$represents the difference between these two values. $\Omega(f_k)$ is the tree complexity and *k* is the number of features.

Taylor's second-order objective function expansion is as following.

$$O^t = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_k) + C \tag{6}$$

The loss function's first and second derivatives are defined as g_i and h_i, respectively; $f(x_i)$is the structure value x_i of tree in the *t*-th iteration.

A tree is defined as follows:

$$f_t(x) = w_{q(x)}, w \in R^T, \quad q: R^d \to \{1,2,3, \dots, T\} \tag{7}$$

In this application, q is the tree's structure: Map input samples $x_i \in R^d$ to leaf nodes. T is the number of leaf nodes. And w is a one-dimensional vector representing the weight of the leaf nodes. Using the node weight vector and the L_2 norm of its leaf nodes, the complexity of the tree can be calculated as follows:

$$\Omega(ft) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{8}$$

The objective function may be written as follows:

$$O^t = \sum_{j=1}^{t} \left[G_j w_i + \frac{1}{2}(H_j + \lambda) w_j^2 \right] + \gamma T = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{9}$$

Where I_j is the sample set of the j-th leaf node, λ and γ is the weight factor.

- Metrics of importance

A feature's relevance is calculated through three methods in the XGBoost model: weight (the number of times a feature has been used to split data across all trees), gain (the average gain across all splits), and cover (the average coverage across all splits). This study uses weight to measure feature relevance.

2.3. Random Forest

The Random Forest model (RF) was created as a solution to the drawbacks of the traditional Decision Tree model. RF reduces model bias and variation by simultaneously training several decision tree learners. Based on N bootstrap samples, an unpruned regression tree is trained on each sample of the original dataset to create a random forest model. In order to avoid multiple potential splits, K randomly selected predictors are used instead of all possible predictors. By averaging the projections of the C trees, fresh data is calculated by repeating the procedure until C trees are constructed. With RF, multiple training datasets are utilized in building trees in order to boost variety and decrease variance. It is possible to represent a regression RF model mathematically as following.

$$\hat{f}_{RF}^{C}(x) = \frac{1}{C} \sum_{i=1}^{C} T_i(x) \tag{10}$$

An output is generated by aggregated predictions of C decision trees using a vectored input variable x, where T_i is a single regression tree that is created using a selection of input variables and bootstrapped samples. When trees are created using RF, out-of-bag error estimates are potentially conducted by recycling training instances that weren't used for individual trees. A random sample subset with no external validation dataset is used to assess generalization error. Through the determination of the relevance of input characteristics, RF might help improve model performance on high-dimensional datasets. In this method, a mean drop in prediction accuracy is calculated when one input variable changes while the others remain the same. It determines which characteristics are most significant for the final model by assigning a relative relevance score to each variable.

3. Experiment Evaluation

3.1. Dataset Description

Stocks closing prices history during 2016 and 2019 and the sentimental index (polarity) from twitter are analyzed comprehensively by three different algorithms, and three models for prediction are established based on which. The history stock prices of Tesla (TSLA) and Apple (AAPL) are acquired from yahoo finance. 70% of the data is used to train the Random Forest Regressor, Long Short-term memory and Extreme Gradient Boosting model respectively, the rest is involved in the test set for fitting goodness test. Twitter sentimental polarity generated by Natural Language Process during the same time period could be exploited to indicated the stock price to some extend. Our experiment exploited these datasets for the comparison of deep learning models.

3.2. Intuitive Comparison

The graph below compared the real closing price with predicted stock price generated by different models in terms of date. It's obvious that there are some flat parts in the line graph representing Random Forest Regressor. While the XGBoost method shows the most accurate fitting result during the whole period. (See Figure 1)

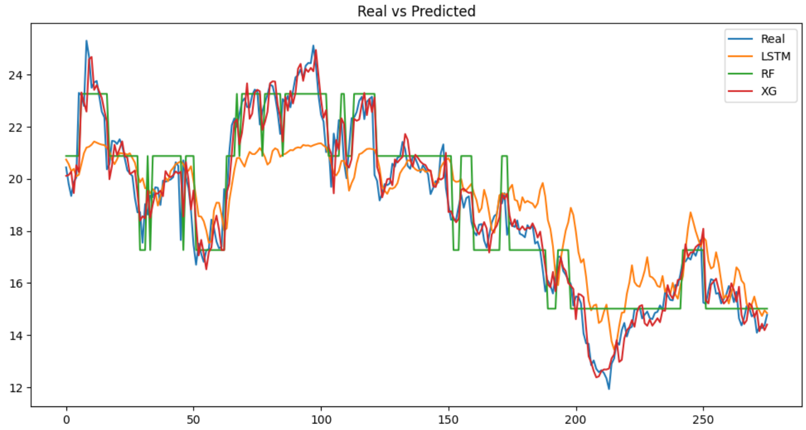

Figure 1 Comprehensive Comparison

3.3. Detailed Comparison

The mean squared error (MSE) and mean absolute percentage error (MAPE) of the result output by the three different models are shown in Table 1.

Table 1 MSE and MAPE.

	Mean Squared Error	Mean Abs-percentage Error
LSTM	0.5441	0.0275
Random Forest	2.4058	0.0682
XG Boost	1.3295	0.0495

These metrics are prevalent used to indicate the fitting goodness by statistic scientists. It's obvious that the XG Boost model predicts the data with the lowest error in terms of both metrics. However, by it is said that the LSTM presents less accuracy that Random Forest Regressor in terms of both of the two metrics, which is a bit counter intuitive as there are many *arbitrary* flat sessions. These plains cause inaccuracy when predicting the stock price trend.

To describe and measure how well these models represent the trend of stock prices, we extracted the increase-decrease characteristics of the real stock price, and compared it whit the equivalent data of three predicting models. The result is shown as below.

It is obvious that over 10 percent of predict trend by XG Boost model and LSTM cater the real trend. While only 1.5% of the predict trend data generated by Random Forest model fit the real condition. Which indicating that the random forest model makes bad prediction on trend of stocks. Therefore, the accumulative metric is one of the good ways to compare the accuracy of models when there is significant inaccuracy on a certain model, such as the Random Forest Regressor.

4. Conclusion

In conclusion, this essay delves into the application of advanced machine learning models for predicting stock fluctuations, with Tesla's stock serving as a prominent case study. The study demonstrates the advantages of advanced machine learning techniques in handling complex datasets and capturing non-linear relationships, making them valuable tools in predicting stock trends.

Through an experiment comparing three models—Random Forest Regressor, Long Short-Term Memory (LSTM), and Extreme Gradient Boosting (XGBoost)—based on historical stock prices and sentiment analysis of Twitter data, valuable insights are gained. The results showcase the superior predictive accuracy of the XGBoost and LSTM models, which closely align with the actual stock trends. On the other hand, the Random Forest Regressor model exhibits lower accuracy due to its inability to account for certain "arbitrary" flat sessions, resulting in less precise predictions.

These findings underscore the significance of advanced machine learning models in stock prediction, providing traders, analysts, and investors with valuable tools for informed decision-making. The success of models like XGBoost and LSTM in accurately forecasting stock trends holds promise for enhancing market performance and minimizing risks.

As the stock market continues to evolve and become increasingly complex, leveraging advanced machine learning techniques can offer a competitive edge in predicting and understanding stock fluctuations. It is crucial for market participants to stay updated with the latest advancements in machine learning and continue exploring innovative approaches to optimize their predictive models.

Overall, this essay contributes to the growing body of knowledge on the application of advanced machine learning models in stock prediction, particularly focusing on the dynamic landscape of Tesla's stock performance. By embracing these advanced techniques, investors can navigate the complexities of the stock market more effectively, potentially yielding better returns and informed investment strategies.

Chrissanthi, H.D., Charalabos, P.C., Andreas, R.D., George, P.N. and Christos, C.N. (2010) Cointegration of Event-Related Potential (ERP) Signals in Experiments with Different Electromagnetic Field (EMF) Conditions. Health, 2, 400-406.

References

[1] Li, Y., Bu, H., Li, J., & Wu, J. (2020). The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. International Journal of Forecasting, 36(4), 1541-1562.

[2] Maqbool, J., Aggarwal, P., Kaur, R., Mittal, A., & Ganaie, I. A. (2023). Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach. Procedia Computer Science, 218, 1067-1078.

[3] Zhang, Y., Li, J., Wang, H., & Choi, S. C. T. (2021). Sentiment-guided adversarial learning for stock price prediction. Frontiers in Applied Mathematics and Statistics, 7, 601105.

[4] Qi, Y., Yu, W., & Deng, Y. (2021, April). Stock prediction under COVID-19 based on LSTM. In 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC) (pp. 93-98). IEEE.

[5] Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Systems with Applications, 178, 115019.

[6] Thormann, M. L., Farchmin, J., Weisser, C., Kruse, R. M., Säfken, B., & Silbersdorff, A. (2021). Stock price predictions with LSTM neural networks and twitter sentiment. Statistics, Optimization & Information Computing, 9(2), 268-287.

[7] Wilksch, M., & Abramova, O. (2023). PyFin-sentiment: Towards a machine-learning-based model for deriving sentiment from financial tweets. International Journal of Information Management Data Insights, 3(1), 100171.

[8] Yang, J., Wang, Y., & Li, X. (2022). Prediction of stock price direction using the LASSO-LSTM model combines technical indicators and financial sentiment analysis. PeerJ Computer Science, 8, e1148.

[9] Wojarnik, G. (2021). Sentiment analysis as a factor included in the forecasts of price changes in the stock exchange. Procedia Computer Science, 192, 3176-3183.