Application of machine learning model in stock prediction

Lingyun Gao

Faculty of Information Science and Engineering, Ocean University of China, Qingdao, 266100, China

Abstract. Economic stability and investor decision-making are significantly impacted by the stock market, which plays a crucial role in financial markets worldwide. This study focuses on applying machine learning models, particularly Random Forest, Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM), to predict Tesla stock prices. Various models' performance in stock prediction is evaluated using R-squared (R2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). The results show that Random Forest and LSTM achieve the highest forecast accuracy, outperforming other models. These findings suggest that machine learning models are effective tools for stock prediction, providing valuable insights for investors. The superior performance of Random Forest and LSTM may be attributed to their ability to capture complex patterns in the data, especially in volatile markets. To further enhance the precision and reliability of stock predictions, future research should explore multi-source data integration and advanced model optimization techniques. Additionally, incorporating market sentiment and macroeconomic indicators could improve prediction robustness and offer deeper insights into market trends.

Keywords: learning model, random forest, LSTM prediction

1. Introduction

Stock market forecasting has always been a focus of research in the financial field. Investors can make better investment choices through in-depth analysis of stock price trends. The company's investment returns, risk management, and the long-term stability and growth of the whole economic market are all significantly impacted by this. The stock market is not only a complex financial system but also an environment full of uncertainty and change. Therefore, encouraging the sound growth of the financial market greatly depends on increasing the precision and dependability of stock forecasting.

As a leading brand in the global electric vehicle industry, Tesla's stock price has experienced significant fluctuations in recent years. This volatility is closely related to a variety of factors, including market sentiment, company performance, industry dynamics, and macroeconomic factors. However, stock price-related data usually exhibits characteristics such as high noise, nonlinearity, and time dependence. At the same time, financial markets are constantly evolving, so improving the accuracy of stock predictions has become a top priority for current research. Traditional statistical methods have great limitations in dealing with complex changes, which makes machine learning and deep learning technologies gradually become effective stock prediction methods.

Many machine learning models, including random forests and support vector machines, are currently employed in stock prediction. Research has demonstrated the effectiveness of algorithms like random forests and support vector machines in handling intricate financial data. At the same time, deep learning models, especially LSTM networks and convolutional neural networks, have demonstrated strong capabilities in processing time series data. They can capture long-term dependencies and automatically extract features, thus achieving remarkable results in stock prediction. In addition, research combined with sentiment analysis is also increasing. For example, researchers such as Bollen have demonstrated the effectiveness of sentiment data in short-term market forecasting by analyzing social media sentiment, providing a new perspective for traditional financial forecasting methods (Bollen et al., 2011).

Although existing research has made some progress in machine learning and sentiment analysis, research on Tesla stocks is still relatively limited. First, many studies are limited to a single data source and lack the discussion of multi-source data integration, which limits the applicability of the

model (Gupta & Chen, 2020). Second, deep learning models are prone to overfitting in small sample data, which poses new challenges to highly volatile stocks such as Tesla. Therefore, how to achieve an effective fusion of multi-source data and build a comprehensive prediction model suitable for a specific stock market has become a current research hot spot.

This work aims to compare the effects of different machine learning models in Tesla stock prediction applications, address the issue of inadequate data fusion in existing stock prediction research, and verify the enhancement effect of sentiment analysis on model performance. The purpose of this study is to provide investors and financial professionals with a useful reference for the selection and optimization of Tesla stock market prediction models.

2. Model Overview

2.1. XGBoost

The XGBoost model, formally proposed by Chen et al. (2016), is an improved version of the gradient-boosted decision tree (GBDT) and belongs to the category of ensemble learning models. It uses Classification and Regression Tree (CART) as the base learner and integrates it through the Boosting strategy to optimize predictions by gradually building decision trees. Since its release, the XGBoost model has received widespread attention and praise from machine learning researchers for its remarkable features such as fast computing speed, good fitting effect and flexible configuration, and has achieved remarkable results in many application fields.

During the model training process, XGBoost uses strategies such as greedy algorithms and approximate algorithms. When constructing each subtree, it selects the features and value points that can minimize the objective function as the split points of the leaf nodes. This process continues until the loss function no longer decreases significantly, or the preset tree depth, number of leaf nodes, and other stopping conditions are reached. At this time, the current subtree training is completed and the next subtree training is transferred. The entire XGBoost model training procedure ends when the number of subtrees reaches the predetermined total number of subtrees. In the field of stock prediction, the XGBoost model also shows great potential. By learning multi-dimensional features such as historical stock prices, trading volume, financial data, and macroeconomic indicators, XGBoost can capture the complex patterns and potential laws of the stock market, thereby achieving accurate predictions of future stock prices. This not only helps investors formulate more scientific investment strategies and reduce investment risks but also improves the efficiency and accuracy of investment decisions.

2.2. Random forest

The random forest model is an ensemble learning technique based on multiple decision trees. It improves the accuracy of the overall model by building multiple decision trees and combining their prediction results. The random forest uses the bootstrap method to randomly extract subsets from the data set to build a decision tree, and randomly selects features for decision-making during the splitting process of each node, which effectively reduces the risk of overfitting in the model and improves the model's generalization ability.

In practical applications, especially in the realm of stock forecasting, the random forest model has many advantages. First, by learning multi-dimensional features such as historical stock prices, trading volume, financial data, and macroeconomic indicators, the random forest can capture the complex changes and potential laws of the stock market, thereby achieving accurate predictions of future stock prices. Second, the random forest can automatically filter out the most influential features for stock price prediction from a large number of market features, helping investors focus more on key information and ignore noise and interference, thereby further improving the prediction effect. This not only aids investors in gaining deeper insights into market dynamics and devising more informed financial plans, but also improves the efficiency and accuracy of investment decisions.

2.3. Deep learning model

LSTM is a specialized variant of recurrent neural network (RNN) tailored for handling time series data. LSTM successfully overcomes the limitations of traditional RNN in processing long-term dependencies by creatively introducing memory units and a series of sophisticated gating mechanisms, including input gates, forget gates, and output gates (Oliveira, 2017).

In the field of stock prediction, the LSTM model, with its powerful time series analysis capabilities, can use rich historical price data to accurately predict the future trend of stocks. This feature makes LSTM not only suitable for short-term price prediction, helping investors quickly capture market dynamics and formulate flexible trading strategies; at the same time, it is also good at long-term price prediction, providing investors with a broader market outlook and assisting in formulating long-term asset allocation plans.

2.4. Performance evaluation metrics

1. MAE is the average of the absolute errors between the predicted and actual value. It directly reflects the average error of the predicted value and is suitable for absolute error analysis in stock price prediction.

2. RMSE is the square root of the mean of the squares of the errors between the predicted value and the true value. Compared with MAE, it is more sensitive to large errors and can amplify the impact of significant deviations on model performance. It is suitable for evaluating whether there are extreme errors in the prediction results, which is particularly important for the prediction of highly volatile assets in the stock market.

3. R2 indicates the goodness of fit between the model prediction value and the true value, reflecting the model's ability to explain the variation of the target variable. The R2 value ranges from 0 to 1, with values closer to 1 indicating a better model fit. It is suitable for analyzing the overall prediction ability of the model and is of great significance for evaluating the global performance of the model in stock prediction.

3. Case Analysis

3.1. Data sources and selection criteria

This article uses Tesla stock data as an example for research and discussion. The data source is the Kaggle data set "Tesla stock data", which spans from 2010 to June 2022. When deciding to choose the Tesla stock data set, it was considered that the data set should contain complete historical stock price data, including opening price, closing price, highest price, lowest price, trading volume, etc., to conduct a comprehensive analysis. The data should span a long period and contain records from multiple years to facilitate time series analysis and model training.

3.2. Results

Table 1. Performance comparison of each model

	Model	MAE	RMSE	R^2
0	Random Forest	1.811886	5.168800	0.999545
1	XGBoost	2.201952	6.272740	0.999330
2	LSTM	0.007560	0.013029	0.995717

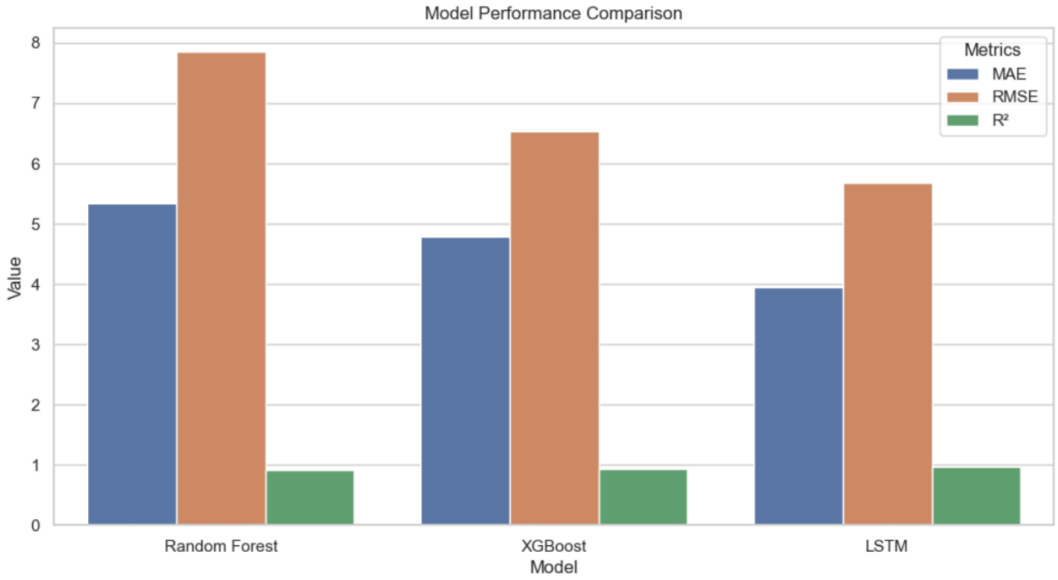

Figure 1. Visualization of performance comparison results of each model (Photo/Picture credit: Original).

Alt Text for Graphical Figure:

A bar chart or line chart visualizes the performance comparison of different models. Each model is represented by a distinct color or label, with the performance metrics (such as MAE, RMSE, R^2) plotted on the y-axis and the models on the x-axis.

Table 1 and Figure 1 present the performance comparison of each model, showing that Random Forest outperforms XGBoost and LSTM in terms of MAE, RMSE, and R2.

4. Discussion

4.1. Main findings

1. Analysis of advantages and disadvantages of the model

In this study, the LSTM model outperformed other models significantly in terms of RMSE and MAE, indicating that it has strong adaptability in capturing price trends. This aligns with the findings of Li et al. (2022), indicating that LSTM has advantages in time series forecasting. However, the LSTM model also has certain limitations, such as slightly insufficient stability when dealing with extreme market fluctuations, which may be related to its dependence on data (Zhang, & Liu, 2021).

In contrast, the XGBoost model performs better in terms of R^2 value but is inferior to LSTM in terms of MAE and RMSE indicators. This shows that although it has a strong ability to explain data variation, in actual prediction, the complexity of the model structure may result in a decrease in prediction accuracy.

The random forest model performed poorly in all selected performance indicators, especially when dealing with highly volatile data, and the stability of the prediction results was poor, which shows that it is not capable of adapting to complex market conditions.

2. Suggestions for improvement

To further improve the performance of LSTM, the introduction of the attention mechanism can be considered. According to Zhang (2021), the attention mechanism has been shown in similar studies to enhance the model's ability to capture important features, which may effectively address the limitations of LSTM in processing complex time series. The XGBoost model can improve its robustness and accuracy by adjusting hyperparameters and adopting ensemble learning methods.

3. Research limitations:

The shortcomings of this study are mainly reflected in the limitations of model selection and data set. For example, the data set used only covers a specific period, which may lead to some market

changes and trends not fully reflecting the actual situation. Yu (2020) pointed out that further research should be supplemented in the diversity of data sets and model selection to more fully understand the complexity of market forecasting.

4. Future research directions

Given the above shortcomings, future research can consider expanding the data set, especially introducing more macroeconomic indicators and market sentiment data to improve the generalization ability and practical application value of the model. This is consistent with the suggestions of Gupta and Chen (2020), which emphasize the significance of sentiment analysis in financial forecasting. Or combine more advanced machine learning and deep learning methods to further develop a more accurate model for stock forecasting.

4.2. Influencing factors and research limitations

The choice of test data will affect the prediction results. Common features of stock prices include opening price, highest price, lowest price, closing price, and trading volume. For a specific market product such as Tesla, its specific features may also include company financial reports, industry competition trends, etc. The scale differences of different features may affect the training of the model (Zhong, Enke, 2019).

Market conditions will also have an impact on the forecast results. Tesla's stock is highly volatile and will be affected by market sentiment, news events, and other market sentiments, as well as policy changes, industry developments, and other factors. When external challenges become greater, the accuracy of the model's forecast may decrease (Nelson, 2017).

In view of the limitations of this study, in the future, we can further develop a more accurate model for stock prediction by further exploring external factors such as market sentiment and news analysis, combined with more advanced machine learning and deep learning methods. This will not only provide investors with more effective decision-making support but also promote the sustainable development of the economy and technology.

5. Conclusion

This study systematically analyzes the effectiveness of machine learning models in stock prediction, especially for the performance of Tesla stock. By comparing three models, random forest, XGBoost, and LSTM, the study found that the LSTM model performed best in capturing the dynamic characteristics in time series data, and its prediction accuracy was significantly higher than other models. This result not only verifies the application potential of deep learning in the financial field but also emphasizes the importance of Tesla stock as a research object. Tesla's innovation and market leadership in the electric vehicle industry have made it the focus of investors' attention. Therefore, accurate prediction of its stock price has important practical value.

The application prospects of machine learning technology in the financial field are broad and cover many aspects. For instance, machine learning can be employed for portfolio optimization, helping investors optimize asset allocation by analyzing historical data and market trends. In addition, in terms of risk management, machine learning models can identify potential market risks and credit risks and provide more accurate risk assessment tools. These applications not only enhance the efficiency of investment decisions but also aid financial institutions in better coping with market fluctuations and uncertainties.

In the future, the development trend of machine learning technology in stock market analysis will focus on several aspects. First, multi-source data fusion combined with market sentiment and news analysis will become an important research direction, and the accuracy of predictions will be improved through natural language processing and sentiment analysis. Secondly, with the continuous advancement of deep learning technology, the complexity and computing power of the model will be further enhanced, making real-time trading and dynamic decision-making possible. In addition, the introduction of cross-market analysis and a global perspective will open up new space for the

application of machine learning in the financial field. Looking ahead, researchers can explore more innovative methods to improve the interpretability and applicability of the model and promote the sustainable development of financial technology.

References

[1] Bollen, J., Mao, H., & Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8.

[2] Chen, T., & Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[3] Gupta, R., & Chen, M., 2020. Sentiment analysis for stock price prediction. International Journal of Computer Applications, 178(19), 36–41.

[4] Li, X., Zhang, Y., & Wang, H., 2022. Research and application of stock market prediction model based on LSTM. Financial and Economic Research, 45(3), 123–135.

[5] Nelson, D. M., Pereira, A. C. M., & de Oliveira, R. A., 2017. Stock market's price movement prediction with LSTM neural networks. In 2017 International Joint Conference on Neural Networks (IJCNN), 1419–1426. IEEE.

[6] Oliveira, N., Cortez, P., & Areal, N., 2017. Stock market sentiment analysis using news articles. Journal of Business Research, 68(5), 1017–1028.

[7] Yu, P., 2020. Dataset selection and influencing factors in stock market prediction. Modern Economic Analysis, 38(5), 98–110.

[8] Zhang, H., 2021. Improved LSTM model based on attention mechanism and its application in stock market. Artificial Intelligence and Finance, 13(1), 89–101.

[9] Zhang, J., & Liu, M., 2021. Application and challenges of LSTM model in stock market volatility prediction. Data Science and Engineering, 12(4), 215–227.

[10] Zhong, X., & Enke, D., 2019. Predicting the daily return direction of the stock market using hybrid machine learning algorithms. Financial Innovation, 5(1), 1–20.