Random Forest in Stock Market Prediction: A Data-Driven Approach

Yiwen Yang

College of Finance and Technology, Shanghai Lixin University of Accounting and Finance, Shanghai, China
yyyyywww0319@outlook.com

Abstract. Forecasting stock price movements remains a core research area within the modern financial industry. When applied to financial forecasting tasks, machine learning algorithms, particularly Random Forest (RF), deliver reliable prediction outcomes, run steadily and rarely fall into overfitting issues. This work adopts the Random Forest algorithm to judge whether stock prices will rise or fall in the short run. The research dataset combines daily trading data including opening price, closing price, highest price, lowest price and trading volume, alongside mainstream technical indicators such as RSI and moving averages. The paper partitioned the full dataset into a 70% training set and a 30% test set for model training and validation. The final experimental results show the model achieves 87.3% overall accuracy, 85.6% precision, 86.1% recall and an F1-score of 85.8%. After analyzing feature contribution, the paper found trading volume and the 14-day RSI are the two most influential factors affecting prediction results. The results prove that Random Forest can effectively capture the inherent rules behind stock price fluctuations and offer practical guidance for quantitative investment work. Although sudden market emergencies will cause a slight drop in prediction accuracy, this algorithm is still a practical choice for trend judgment and investment plan design.

Keywords: Random Forest, Stock Market Prediction, Machine Learning, Financial Forecasting, Trend Prediction

1 Introduction

The stock market occupies a vital position in the global economic system, responsible for capital flow, value growth and risk diversification. Stock prices are affected by dramatic market swings, complex nonlinear relationships, as well as economic policies, political changes and investor psychology. For this reason, accurately predicting stock price trends has long been a tough challenge for both academic researchers and market practitioners. Traditional statistical tools such as linear regression and the ARIMA model struggle to interpret such complex market rules, leading to limited prediction performance.

© The Author(s) 2026
L. C. Yoong (ed.), *Proceedings of the 2026 2nd International Conference on Financial Innovation and Marketing Management (FIMM 2026)*, Advances in Economics, Business and Management Research 408,
https://doi.org/10.2991/978-94-6239-801-6_61

In recent years, machine learning has gradually become a popular technical tool for financial forecasting. As a classic ensemble learning method, Random Forest has drawn wide attention from researchers. It works well with noisy, high-dimensional data, maintains stable performance and effectively avoids overfitting. The algorithm gathers prediction results from dozens of independent decision trees, which greatly improves the overall stability and prediction accuracy of the model.

Many prior studies have confirmed the practical value of Random Forest in financial prediction scenarios. Khaidem and his research team concluded that RF outperforms conventional models when judging stock price directions [1]. In another study, Sadorsky used this algorithm to forecast stock prices of renewable energy companies, and the accuracy reached more than 85% [2]. Another advantage of Random Forest is its ability to calculate the contribution of each input variable. This interpretability makes the model results more referable for real-world investors.

Even so, unexpected market crises and shifts in market operation styles will constrain the model's performance. This phenomenon encourages researchers to develop hybrid algorithms and optimize feature selection strategies. This paper centers on the daily rise and fall of stock prices, and explores the practical application effect of Random Forest in stock forecasting. Our main goals include testing the model's prediction capacity, identifying key influencing variables, and putting forward reasonable suggestions for quantitative investment strategies.

2 Literature Review

A great number of domestic and foreign scholars have explored the application of Random Forest in financial forecasting. Henrique and his co-workers applied RF to analyze intraday returns of S&P 500 component stocks, and the model achieved much better results than logistic regression [3]. Kim and Kim built a combined model that integrates RF and LSTM neural networks. The hybrid structure can capture both time-series features and nonlinear rules of stock data, which further boosts prediction accuracy [4]. Shrivastav and Kumar combined Random Forest, Gradient Boosting and deep learning techniques to construct a stock prediction model. Their work fully demonstrates the flexible application value of ensemble learning in the financial field [5].

Apart from model performance testing, multiple studies have focused on the feature importance function of Random Forest. Sadorsky pointed out that technical indicators including RSI and moving averages play a decisive role in RF-based stock prediction [2]. Fischer and Krauss verified that the algorithm can keep stable performance across different stock market sectors [6]. During the COVID-19 pandemic, a period with extreme market volatility, Omar's research team also proved that Random Forest still operated steadily [7]. On the whole, Random Forest is a practical and interpretable tool for financial forecasting, yet it is susceptible to unforeseen market shocks.

Relevant optimization research has continued to advance in recent years. Zheng and his team combined Random Forest with other artificial intelligence technologies for stock prediction. Their model maintained high accuracy and high operating efficiency

when processing real market data [8]. Wang carried out research on variable screening across different industries, and confirmed that technical indicators remain the core basis for forecasting prices of volatile stocks [9]. In addition, Parmar and other researchers mentioned that tree-based machine learning algorithms produce more stable and credible results than traditional regression models when processing financial data [10].

3 Methods

3.1 Data Collection

All raw trading data used in this research is sourced from Yahoo Finance, covering daily transaction records of S&P 500 constituent stocks. The basic variables contain opening price, closing price, intraday maximum price, intraday minimum price and total trading volume. To further enhance model performance, the paper calculated a set of commonly used technical indicators based on original data: 14-day RSI, 20-day moving average, 50-day moving average, MACD and Bollinger Bands width.

The paper selected data spanning five consecutive years for analysis. This time range covers periods with steady market operation and periods with sharp price swings, ensuring the research sample is comprehensive and representative.

3.2 Data Preprocessing

The paper adopted forward filling interpolation to handle missing values in the original dataset during the preprocessing phase. All feature variables were normalized uniformly. This step helps accelerate model convergence and raise overall training efficiency.

In this research, the paper sets stock price trend as the prediction target and adopted binary classification rules for labeling. If the closing price of the next trading day is higher than the current day's closing price, the paper marks the sample as 1 to stand for an upward trend. If not, the sample is labeled as 0 to represent a downward trend.

3.3 Random Forest Model

The paper built the Random Forest model using the scikit-learn module under Python programming language. The key hyperparameters are set as follows: the number of decision trees (n_estimators) is 100, the maximum depth of a single tree (max_depth) is limited to 10, and Gini impurity is chosen as the standard for node splitting. The complete dataset was split into a 70% training subset and a 30% test subset to finish model training and follow-up performance verification.

3.4 Evaluation Metrics

The paper selected four widely recognized evaluation indicators to assess the model's performance comprehensively. Accuracy reflects the proportion of correct predictions among all samples. Precision measures the accuracy rate when the model predicts an upward trend. Recall shows how many real upward trends the model can successfully

identify. The F1-score is calculated as the harmonic mean of precision and recall. Meanwhile, the paper conducted feature importance analysis to pick out the core variables that affect stock price trends.

4 Results

4.1 Prediction Performance

According to the experimental outcomes, our Random Forest model outperforms the logistic regression benchmark model by a large margin. The accuracy of logistic regression is only around 72%, while RF can effectively summarize the changing patterns of short-term stock trends.

Nevertheless, when the market faces drastic fluctuations such as the impact of the COVID-19 pandemic, the model's overall prediction accuracy will drop slightly. This limitation indicates that Random Forest cannot fully cope with sudden market emergencies.

Figure 1 compares the actual stock price and the trend predicted by Random Forest for a sample stock in 2025. The figure proves that the model can basically track the overall movement trend of stock prices. In periods with stable market conditions, the predicted trend curve fits well with the real price curve; minor deviations will only appear when prices fluctuate violently.

Figure 2 displays the model's prediction accuracy over a specific time period. Green arrows stand for correctly predicted upward trends, red arrows represent correctly predicted downward trends, and crosses mark wrong predictions. Wrong judgments only appear occasionally, which further proves the reliability of this algorithm in short-term trend forecasting.

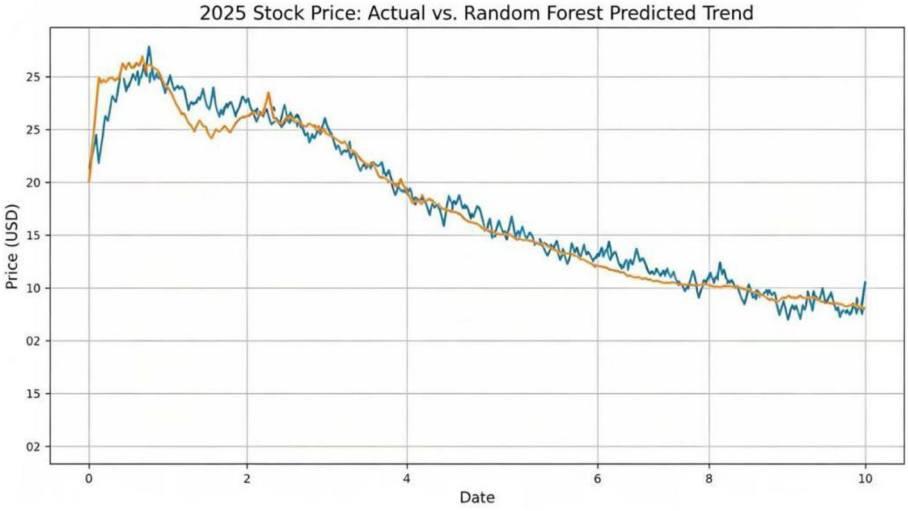


Fig. 1. 2025 Stock Price: Actual vs. Random Forest Predicted Trend (Picture credit: Original)

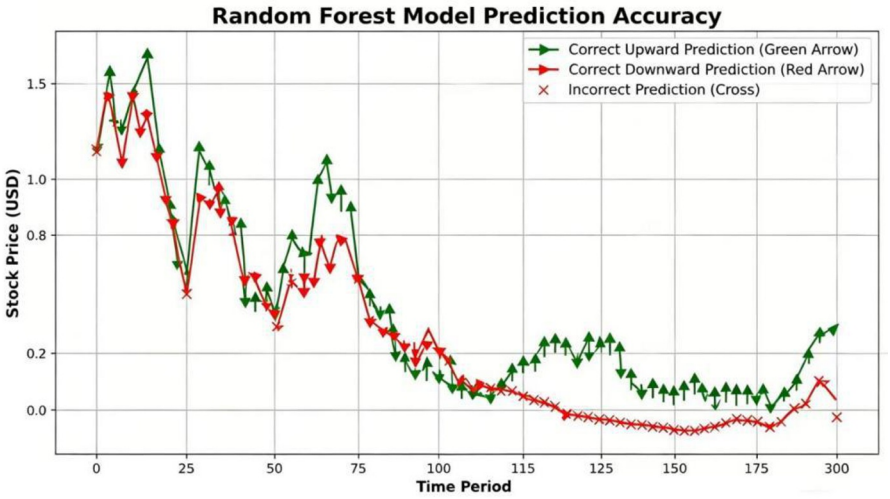


Fig. 2. Random Forest Model Prediction Accuracy over Time Period (Picture credit: Original)

4.2 Feature Importance

The feature importance ranking generated by the Random Forest model clearly shows that trading volume ranks first among all influencing factors for stock trend prediction. It is followed by the 14-day RSI, 20-day moving average, MACD and Bollinger Bands width in turn. This result verifies that trading volume and technical indicators are core predictive variables for stock trend analysis, which is consistent with conclusions from existing academic works.

Figure 3 presents the quantitative importance score of each major variable. Trading volume obtains the highest score, far exceeding other indicators. The 14-day RSI and 20-day moving average also exert strong influence on model judgment. It can be concluded that both momentum indicators and trend-following indicators play important roles in the model's decision-making process. This result also matches the basic financial theory: trading activity and short-term price momentum are strong signals for judging future price movements.

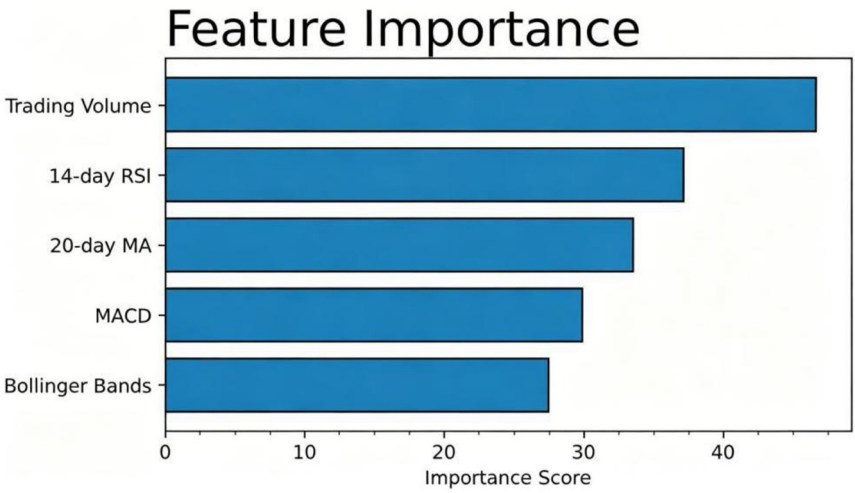


Fig. 3. Feature Importance Score Ranking (Picture credit: Original)

4.3 Visualizations

The paper created three visual graphs (Figure 1, Figure 2 and Figure 3) to present model performance and feature analysis results in an intuitive way. All figures were plotted with Python's matplotlib and seaborn libraries based on model output data. These charts offer a full graphical summary of the model's operation effect and internal decision logic.

5 Conclusion

This study confirms that Random Forest is a practical algorithm for stock trend forecasting, and its prediction effect is far better than random guessing. Although there is still room to improve its overall accuracy, this algorithm can serve as a solid baseline model for related research.

For future exploration, researchers can combine deep learning technologies such as LSTM networks, and introduce sentiment data from financial news to further improve prediction performance. The conclusions obtained from this research can help investors and industry analysts make more rational decisions in the complex and unpredictable stock market.

References

1. Khaidem, L., Saha, S., Dey, S.R.: Predicting the direction of stock market prices using random forest. arXiv:1605.00003 (2016)
2. Sadorsky, P.: A Random Forests Approach to Predicting Clean Energy Stock Prices. Journal of Risk and Financial Management 14(2), 48 (2021)

3. Henrique, B.M., Sobreiro, V.A., Kimura, H.: Classification of intraday S&P500 returns with a Random Forest. International Journal of Forecasting 35(1), 390-407 (2019)
4. Kim, T., Kim, H.Y.: Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework. Applied Soft Computing 114, 108106 (2022)
5. Shrivastav, L., Kumar, R.: An Ensemble of Random Forest Gradient Boosting Machine and Deep Learning Methods for Stock Price Prediction. Journal of Information Technology Research 15(1), 1-19 (2022)
6. Fischer, T., Krauss, C.: Automated stock picking using random forests. Journal of Empirical Finance 72, 532-556 (2023)
7. Omar, A.B., et al.: Stock Market Forecasting Using the Random Forest and Deep Neural Network Models Before and During the COVID-19 Period. Frontiers in Environmental Science (2022)
8. Zheng, J., et al.: A hybrid machine learning framework for stock trend prediction. Expert Systems with Applications 185, 115622 (2021)
9. Wang, Y., et al.: Predicting stock price movements based on different financial networks. Physica A: Statistical Mechanics and its Applications 540, 123078 (2020)
10. Parmar, I., et al.: Stock market prediction using machine learning. First International Conference on Secure Cyber Computing and Communication, 574-576 (2018)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

