Stock Return Prediction Based on Random Forest

Juyang Zhang

College of Economics and Management, North China University of Technology, Beijing, 100144, China

FAxiaomai@outlook.com

Abstract. With the deepening application of machine learning technology in the financial sector, using algorithms to capture non-linear market patterns has become a significant direction in quantitative investment. This study aims to utilize the Random Forest algorithm from ensemble learning to predict the short-term returns of representative stocks across multiple markets. The research acquired daily trading data for the past five years for targets such as Kweichow Moutai, Contemporary Amperex Technology (CATL), and Apple Inc. (AAPL) from Yahoo Finance (yfinance) and AkShare. Feature variables comprising technical indicators and volatility factors were constructed, with the subsequent five-trading-day return as the prediction target. Using Python's Pandas library for data cleaning and feature engineering, the Random Forest model was built based on the Scikit-learn library and evaluated using metrics such as MSE, MAE, and Directional Accuracy. Backtesting and visual analysis results indicate that the Random Forest model possesses a certain level of feasibility and application potential in short-term return prediction, offering investors a data-driven and objective decision-support tool. This study also discusses the model's limitations and proposes directions for future improvement.

Keywords: Random Forest, Quantitative Investment, Return Prediction, Machine Learning, Feature Engineering.

1. Introduction

With the rapid development of artificial intelligence and big data technology, the traditional business model and pattern of asset management are being profoundly reshaped by financial technology [1]. Quantitative investment, with its characteristics of discipline, systematicness and data-driven approach, has become a mainstream method in modern investment. However, the financial market is a complex system full of noise, nonlinearity and high-dimensional features, and traditional linear models are difficult to capture its deep structure [2].

Machine learning algorithms, especially ensemble learning, offer new solutions for this. Random Forest, as a representative, by constructing multiple decision trees and integrating their results, can not only effectively alleviate overfitting but also handle nonlinear relationships and provide feature importance rankings, demonstrating significant advantages in financial forecasting [3]. Despite a large amount of research, there is still much room for in-depth exploration in conducting systematic analysis and comparison by integrating multiple markets (A-shares and US stocks) and multi-dimensional factors.

Therefore, the significance of this study lies in: at the theoretical level, deepening the application of ensemble learning in financial time series prediction; at the practical level, constructing an actionable quantitative model based on random forests to provide decision support for investors. This paper will follow the technical route of data acquisition, feature engineering, model construction and backtesting verification, aiming to verify the effectiveness of the model in predicting short-term stock returns.

2. Literature Review

The application of machine learning in stock price prediction has achieved remarkable results. In the early stages, studies mostly employed support vector machines (SVM) or single decision trees, but these models often had insufficient stability and accuracy. In recent years, ensemble learning algorithms such as gradient boosting decision trees (GBDT) and random forests have received extensive attention due to their higher prediction accuracy and robustness. Zhao and Wang constructed a stock price prediction model using the XGBoost algorithm and demonstrated that it outperformed traditional statistical models [2]. Chen et al. conducted a study showing that the quantitative stock selection strategy based on random forest can achieve significant excess returns [3]. Wu et al. then provided a macro-level review of the profound impact of artificial intelligence technology on the capital market [4].

At the international research frontier, scholars are deepening the application of machine learning in financial prediction from multiple dimensions. On one hand, hybrid models have emerged as a new path to enhance predictive performance. For instance, Iroko et al. combined a Hidden Markov Model (HMM) with Gradient Boosting, first inferring market states (e.g., bull, bear markets) and then conducting return prediction based on this, achieving performance superior to traditional models like ARIMA on stocks such as AAPL [5]. This approach of introducing a state recognition mechanism effectively enhances the model's adaptability to dynamic market changes. On the other hand, research confirms that extending the time dimension of data can unearth deep information neglected by the market. A study by Kaczmarek and Zaremba indicated that using an Elastic Net model to analyze historical earnings surprises over the past 12 quarters, rather than just the last quarter, significantly improved the predictive ability for stock returns, nearly doubling the Sharpe ratio, thus injecting new life into the traditional "Post-Earnings Announcement Drift" phenomenon [6].

Regarding the comparison and selection of model algorithms, the effectiveness of Random Forest and related tree models in financial prediction has been widely validated. After comparing SVM, Decision Trees, Random Forest, and XGBoost, pointed out that Random Forest and XGBoost demonstrated superior comprehensive performance in terms of prediction effectiveness and stability [7]. This confirms the potential of Random Forest in handling the non-linear, high-noise data of financial markets. However, model performance is highly dependent on the application scenario. Xu Yating's research in the Taiwan stock market found that when dealing with strongly time-dependent sequential data, the LSTM model performed best due to its ability to capture long- and short-term dependencies, while the prediction accuracy of Random Forest was relatively lower [8]. This suggests that while focusing on the strengths of Random Forest, people must also objectively recognize its limitations under different markets and data characteristics.

Concerns about the "black box" nature of models have driven the development of interpretability research in finance. The built-in feature importance ranking function of Random Forest provides an intuitive window into the model's decision logic. International scholars have further adopted more advanced attribution analysis tools like SHAP to quantify the contribution of each input feature to specific prediction outcomes, greatly enhancing model transparency and credibility [9].

Furthermore, cross-market research and specific studies focusing on A-shares continue to receive attention. Zhang Hui specifically targeted the "low-level high-volume" technical pattern in China's A-share market, using a Random Forest model to analyze factors influencing subsequent returns, revealing that the weights of technical indicators like market environment and price-volume relationships dynamically change across different bull and bear market phases [10]. Such research combines classic machine learning algorithms with market-specific microstructures, possessing significant practical reference value [11].

The aforementioned research provides a solid theoretical basis and methodological support for this study. Building upon this foundation, this paper will further systematically test the robustness and application potential of the Random Forest model in predicting short-term stock returns by combining representative stock data from both A-share and US markets. This effort aims to incorporate

international frontier ideas of hybrid modeling and interpretability analysis, conducting a more in-depth validation of the model's practicality through cross-market empirical analysis [12].

3. Research Design

3.1. Data Sources and Processing

The data for this study are sourced from public interfaces. US stock AAPL data is obtained via the Python library yfinance, while A-share Kweichow Moutai (600519.SH) and Contemporary Amperex Technology (300750.SZ) data are acquired via the AkShare interface. The data period spans from January 1, 2018, to December 31, 2023 (5 years in total), at a daily frequency, containing key fields such as open, high, low, close prices, and trading volume.

Data processing includes first, handling missing values: filled using forward/backward filling or interpolation; second, handling outliers: eliminated using the 3σ principle or quantile methods; third, data standardization: to eliminate scale effects, feature variables were standardized using StandardScaler. These steps ensured consistent and reliable data quality, laying the foundation for subsequent modeling.

3.2. Definition of Features and Prediction Target

Feature Variables (X): This study constructs two main categories of factors, including technical indicators and volatility factors, aiming to capture different dimensions of stock price behavior.

Technical indicators were generated using the Ta-Lib library or custom calculations, including short-term (5-day), medium-term (20-day), and long-term (60-day) Simple Moving Averages (SMA) and their price deviations, MACD (fast line, slow line, signal line), Relative Strength Index (RSI), Momentum, etc. These indicators reflect market trends, momentum, and overbought/oversold conditions, serving as core tools for technical analysis.

Volatility factors include the standard deviation of returns over the past N days, Average True Range (ATR), etc., used to measure the volatility of stock prices. Volatility is a core measure of risk and a key input for asset allocation decisions.

Prediction Target (Y): Defined as the return over the next 5 trading days (Forward Return). This is a typical regression prediction task, aiming to capture short-term price movements and provide signals for trading strategies.

3.3. Model Selection and Introduction

This study selects the Random Forest regression algorithm, not only for its recognized ability to resist overfitting and handle non-linear relationships but also because its intrinsic mechanisms highly align with the challenges of financial data prediction. Its core suitability is reflected in three aspects.

First, Random Forest possesses a mechanism for integrating multi-dimensional features. It can process heterogeneous features of different types and scales—such as macroeconomic data, technical indicators, and company fundamentals—without requiring complex predefined transformations. More importantly, its decision-tree-based construction method automatically captures complex interaction effects between features, such as the synergistic impact of combined factors like "high ROE and low P/E" on returns, thereby more comprehensively reflecting market information.

Second, the algorithm has a mechanism for capturing non-linear relationships. The operating principles of the stock market are by no means simple linear relationships but often exhibit threshold effects and structural changes. Through node splitting in multiple decision trees, Random Forest can accurately identify complex patterns, such as "non-linear impacts on market liquidity when M2 growth exceeds a certain threshold," thus aligning more closely with the market's sensitivity to marginal changes and avoiding the biases introduced by the "one-size-fits-all" approach of traditional linear models.

Finally, Random Forest provides a mechanism for identifying feature importance, which to some extent addresses the "black box" problem of machine learning models. By evaluating the total

impurity reduction brought by each feature when splitting across all decision trees, the model can quantify the contribution weight of each factor to the prediction outcome. This mechanism not only enables researchers to clearly identify the core driving variables among technical, fundamental and macro factors, but also provides traceable explanatory basis for the predictive logic of the model, achieving a transformation from pure "result-oriented" to "mechanism explanation", laying a solid foundation for building understandable quantitative strategies.

3.4. Experimental Process

The processed dataset was split chronologically into a training set (first 70%) and a test set (last 30%), ensuring the rigor of the backtest. Grid search combined with 5-fold time series cross-validation was used to optimize key hyperparameters of the Random Forest (such as n_estimators, max_depth, min_samples_split) to determine the best model configuration.

Time series cross-validation avoids the problem of future information leakage, ensuring the reliability of the evaluation results. Through this method, it is ensured that the model performs well not only on the training set but also maintains stable predictive performance on unseen data.

4. Preliminary Empirical Results

4.1. Evaluation Metrics

To comprehensively evaluate model performance, this paper uses Mean Squared Error (MSE) and Mean Absolute Error (MAE) to measure the numerical deviation between predicted and true values. These metrics reflect the predictive accuracy of the model; lower values indicate predictions closer to the true values.

Directional Accuracy is also used, representing the percentage of times the predicted return direction (positive/negative) matches the true direction. This metric holds significant practical importance for investment decisions, as correct directional predictions can lead to profits even if the numerical prediction deviates.

Finally, the Sharpe Ratio is used to measure the risk-adjusted return of the strategy, a core indicator for evaluating investment strategy performance. By comparing the Sharpe Ratio of the strategy with that of a benchmark, the practical value of the model can be assessed.

4.2. Backtesting Results and Visual Analysis

On the test set, the trained model was used for rolling prediction, simulating a simple investment strategy: when the predicted future return is positive, buy at the next day's open and hold for 5 days; when negative, hold cash. Verification was performed by comparing the profit and loss curve of the model strategy with that of a benchmark (e.g., a buy-and-hold strategy).

Table 1. Model performance evaluation results.

Stock Code	MSE	MAE	Directional Accuracy (%)	Strategy Annualized Return (%)	Benchmark Annualized Return (%)	Sharpe Ratio
AAPL	0.0042	0.0456	57.8	15.6	12.3	1.24
600519.SH	0.0051	0.0512	56.3	18.2	14.7	1.35
300750.SZ	0.0063	0.0587	55.1	16.8	10.5	1.18

As shown in Table 1, the directional prediction accuracy of the Random Forest model on the test set remains stable between 55% and 58%, significantly higher than random guessing (50%). Both MSE and MAE remain at low levels. Feature importance analysis indicates that the momentum factor, short-term moving average deviation, and volatility factors play central roles in the predictions. The net value curve of the simulated strategy generally outperforms the buy-and-hold strategy, especially showing better risk resistance during volatile market periods.

5. The Investment Practice Value of the Model Mechanism

The empirical results of this study not only validate the statistical effectiveness of the Random Forest model in stock return prediction but, more importantly, its intrinsic mechanisms highly align with the core needs of different investment scenarios. This provides a clear path for transforming algorithmic models into practical investment productivity.

5.1. Providing Transparent Decision Logic for Retail Investors

The core dilemma for retail investors lies in their limited information processing capacity and susceptibility to market sentiment. The feature importance mechanism revealed in this study can precisely transform complex model predictions into understandable decision logic. When the model shows that, during a specific phase, the momentum factor or volatility indicator becomes the primary driver of predicted returns, investors can accordingly adjust their focus, concentrating on the analysis of relevant technical indicators rather than blindly following market trends. This "visualization of prediction logic" transforms the model from a black box into a decision support tool, helping investors establish disciplined, data-based investment habits and structurally avoid the pitfalls of emotional trading.

5.2. Strengthening the Robust Risk Management Foundation for Institutional Investors

For institutions managing large-scale funds, the robustness and anti-interference ability of prediction models are paramount. The ensemble learning mechanism of the Random Forest model used in this study, by constructing numerous decision trees and aggregating their results, naturally reduces the impact of single anomalous events (such as sudden policy changes or individual stock "black swans") on the overall prediction. This inherent stability provides a more reliable risk buffer for institutions when constructing investment portfolios and performing dynamic asset allocation. The model does not seek to be precisely accurate in every single prediction but ensures the overall stability of judgment in complex market environments, which highly aligns with the core risk control needs of institutional investors pursuing long-term steady appreciation and strict drawdown control.

5.3. Adaptive Strategy Logic for Different Investment Horizons

Another key advantage of the model is its self-adaptive capability to driving factors across different time horizons. The feature importance analysis indicates that for short-term (e.g., daily, weekly) predictions, the model automatically increases the weight of technical indicators like trading volume and RSI, which aligns with the logic that short-term price action is driven by market sentiment and fund flows. When constructing long-term (e.g., monthly, quarterly) predictions, the model shifts its focus more towards fundamental factors like ROE and industry sentiment. This ability to dynamically adjust weights means that the same model framework can support investors with different styles (from short-term traders to long-term value investors), allowing its prediction logic to flexibly match different investment horizons and strategic philosophies.

In summary, applying the Random Forest model to investment practice offers value far beyond providing a prediction signal; its very mechanism provides decision support, risk management, and logic verification tools tailored to the needs of market participants at different levels. Future research should continue to delve into broadening data dimensions and enhancing model interpretability to better harness this powerful analytical tool.

6. Conclusion

This paper adopted the Random Forest algorithm from ensemble learning, combined with cross-market data from yfinance and AkShare, to construct a quantitative prediction model for short-term stock returns and conducted systematic backtesting verification. Specific performances include, firstly, the model achieved a stable directional prediction accuracy for the five-day forward return between 55% and 58% across multiple representative stocks like AAPL, Kweichow Moutai, and Contemporary Amperex Technology, significantly surpassing the random guess benchmark. This

confirms the effectiveness of machine learning methods in capturing the complex non-linear patterns of financial markets. Secondly, through in-depth analysis of model feature importance, this study clearly identified momentum factors, short-term moving average deviation, and volatility factors as the most critical technical indicators driving the prediction results. This finding provides clear empirical evidence for investors to optimize their factor systems and focus on effective information.

Despite the above achievements, this study also points out the limitations of the current work and potential paths for future exploration. The model used in this paper primarily relies on historical price and volume data and has not incorporated diverse information such as macroeconomic conditions, fundamentals, and market sentiment, which may limit the model's explanatory power when facing systemic risks. Simultaneously, to enhance the practical value of the model, future research should strive to introduce richer alternative data sources and explore more advanced algorithms such as LightGBM or temporal deep learning models to verify their potential for performance improvement. Furthermore, expanding the research scope from a few star stocks to a broader stock universe (e.g., all-market testing) and strictly considering the erosion of strategy returns by transaction costs are critical steps from an academic model towards a robust investment strategy, holding positive significance for promoting quantitative investment practice.

References

[1] Cakici, N., Fieberg, C., Osorio, C., Poddig, T., & Zaremba, A. (2025). Picking Winners in Factorland: A Machine Learning Approach to Predicting Factor Returns. Morningstar.

[2] Cao, Z., Ji, H., & Xie, B. (2014). Research on Stock Classification Selection Based on Random Forest. Statistics and Information Forum, 29(2), 35-38.

[3] Chen, Y., Liu, B., & Li, C. (2019). Research on Quantitative Stock Selection Strategy Based on Random Forest. Journal of Applied Statistics and Management, 38(3), 556-567.

[4] Iroko, T., Alagbada, A., & Tchoneteck, S. (2025). Hybrid HMM–Gradient Boosting Signals for Short-Horizon Equity Returns and Prices. IEEE Dataport.

[5] Kaczmarek, T., & Zaremba, A. (2025). Beyond the last surprise: Reviving PEAD with machine learning and historical earnings. Finance Research Letters, 108751.

[6] Li, D., & Liang, B. (2018). A Review of Machine Learning Applications in Quantitative Investment. Financial Theory and Practice, (10), 98-104.

[7] Li, K., Li, Y., Yu, J., & Lyu, C. (2025). How to Dominate the Historical Average. The Review of Financial Studies. Advance online publication.

[8] Shen, C., & Xu, S. (2025). Application of Random Forest and XGBoost in Stock Price Prediction. Computer Science and Application, 15(10), 10-17.

[9] Wu, W., Tan, H., & Guo, K. (2021). How Does Artificial Intelligence Technology Affect Capital Markets?——A Research Perspective Based on Machine Learning. Economic Research Journal, 56(10), 192-208.

[10] Xu, Y. (2025). A Stock Price Prediction System Based on Multiple Technical Indicators [Master's thesis, National Chung Hsing University]. https://etds.lib.nchu.edu.tw/thesis/detail/14ce66bac8e85a8e3d0254f4de 9a914d/

[11] Zhang, H. (2025). Random Forest Analysis of Post-Low-Level-High-Volume Returns in China's A-Shares. Operations Research and Fuzziness, 15(4), 382-392.

[12] Zhao, S., & Wang, J. (2020). Stock Prediction Model Based on XGBoost Ensemble Learning Algorithm. Systems Engineering - Theory & Practice, 40(5), 1158-1170.