Research on Stock Price Prediction Based on Random Forest with Feature Engineering and Rolling Window Evaluation

Zhanhang Gao*
School of Big Data and Statistics, Anhui University, Auhui, China
Corresponding Author: Zhanhang Gao (we23201020@stu.ahu.edu.cn)

Abstract: With the growth of the stock market, equity investment has drawn considerable attention. However, high returns come with high risks, making stock selection and market timing particularly challenging. This paper develops a stock price prediction model based on the Random Forest algorithm with feature engineering. In the feature selection stage, key technical indicators are identified by integrating correlation coefficient filtering, importance evaluation using Random Forest and XGBoost, and K-means clustering. The model is evaluated using a rolling window approach, with a training window of 500 trading days and a test window of 63 trading days, along with 95% prediction intervals for robustness testing. The results demonstrate that the Random Forest model performs well under different market conditions, achieving an overall R^2 of 92.05% and a prediction interval coverage rate of 92.1%, which closely approximates the theoretical 95% confidence level.

Keywords: RF model; Stock price prediction; Feature engineering; Rolling window approach

1. Introduction

Given the large number of data features and the substantial volume of data involved in stock prices of various companies, Random Forest (RF) is capable of handling input samples with high-dimensional features and evaluating the importance of each feature. In 1896, Charles H. Dow first proposed the stock price average index, which is considered the first stock data feature. Zhang et al. proposed a stock prediction model based on RF, balanced learning, and feature selection [1]. Meher Bharat Kumar et al. applied for the first time a high-frequency RF model to predict stock prices of Indian fintech companies [2].

The complexity of the stock market, a nonlinear system, renders predicting its price a particularly arduous task effective feature selection has become a critical factor determining the predictive accuracy of models. With the integration of modern forecasting theories with information technology, statistics, and optimization algorithms,

various prediction techniques have flourished, leading to the development of diverse forecasting models such as the GARCH model [3], LSTM model [4] and grey systems [5], stochastic process models, and Random Forest. Relevant studies have enhanced the accuracy of stock price prediction by selecting dataset features from multiple perspectives, including correlation analysis based on Pearson and Spearman coefficients, as well as calculating importance coefficients of stock features using various methods and averaging the results.

Stock market prediction remains a challenging task due to the non-stationary nature of financial time series and occasional structural breaks [6]. Traditional approaches that train single models on multi-year data may yield overly optimistic results as they fail to account for these regime changes. Furthermore, financial decision-making requires not only accurate point forecasts but also reliable uncertainty quantification. Ignoring prediction confidence may lead to excessive trading during market volatilities. To address these limitations, we enhance our methodology by incorporating rolling window evaluation and prediction intervals. The rolling window approach tests model robustness across different market states, while confidence estimation provides crucial risk management information for practical deployment.

2. Relevant Theoretical Basis

2.1 Random Forest Model

RF is a forest of multiple unrelated decision trees [7], an integrated algorithm for weak learners as decision trees, developed by Breiman proposed machine learning algorithm [8]. RF solves the overfitting problem that can occur from a single decision tree, and can be better generalized to new data sets. In addition, the accuracy of RF is greatly improved compared to a single decision tree model, because RF misjudgment only occurs in more than half of the decision tree judgment errors. The randomness of RF is mainly reflected in two aspects. Firstly, in terms of sampling, bootstrap resampling is used to draw n samples with replacement from the original training set. This process is repeated T times, which reduces the correlation between weak learners. Secondly, some features of the feature set are selected on the splitting node, and then the optimal features are selected from some features for tree splitting, which further enhances the generalization ability of the model.

3. Data Preparation

3.1 Data Sources and Preprocessing

Stock index technical indicators are calculated by the highest price, lowest price, opening price, closing price, and trading volume of the stock index, and are effective expressions of the trading volume of the historical price of the stock index, so they are usually introduced as characteristics in stock forecasting.

This study utilizes daily frequency stock price data obtained from the RESSET database. The sample is from the Chinese A-share market, covering the period from January 4, 2016, to March 29, 2024. The final analyzed dataset comprises 2088 trading days after aligning the data across all selected stocks and indicators. The original data for each stock includes the following daily metrics: Opening price (Open), Closing price (Close), Highest price (High), Lowest price (Low), Price-to-Earnings ratio (PE), Price-to-Book ratio (PB), Trading volume (Volume), and Turnover amount (Trdsum). The Close price is the target variable for prediction. In addition to these 8 raw indicators, 20 technical indicators are constructed, resulting in a total pool of 28 potential features. Their detailed explanations are provided in Table 1.

Table 1: Detailed Explanation of Technical Indicators.

Indicators	Meaning
SMA_5	5-day Simple Moving Average
SMA_10	10-day Simple Moving Average
EMA_5	5-day exponential moving

Table 1: Detailed Explanation of Technical Indicators.

Indicators	Meaning
EMA_14	14-day exponential moving average
EMA_26	26-day exponential moving average
EMA_50	50-day exponential moving average
ADX_14	14-day Average Directional Index
MOM	Momentum
RSI	Relative Strength Index
ATR	Average True Range
OBV	On-Balance Volume
CCI	Commodity channel index
Trix	Triple Exponential Average
DIF	Deviation value
DEA	Signal Line
MACD_Hist	MACD Histogram
ROC	Rate of Change
ADOSC	Chaikin A/D Oscillator
WillR	William Indicator
BIAS	Bias Ratio

Data preprocessing involved two key steps:

1. Handling Missing Values: Given the minimal number of missing values, a forward-filling method was employed. This approach preserves the data size while maintaining time series continuity.

2. Feature Scaling: To eliminate dimensional imbalances, all feature values were normalized to the [0, 1] range using the Min-Max scaling method. The scaling parameters were fit exclusively on the training data within each rolling window to avoid look-ahead bias.

3.2 Feature Engineering

3.2.1 Filtered Feature Selection

Calculate the correlation coefficient of each technical indicator and the closing price, and select the technical indicator with a high correlation coefficient for prediction. The Pearson correlation coefficient can analyze the linear relationship between the dependent variable and the response variable, while the Spearman correlation coefficient can analyze the nonlinear relationship between the two. Figure 1 is based on the example data to obtain the Pearson correlation coefficient graph and the Spearman correlation coefficient graph, which clearly show that the correlation coefficient between the closing price (Close) and SMA_5, SMA_10, EMA_5, EMA_14, EMA_26, EMA_50, and Open reaches more than 0.8. At the same time, the correlation coefficient between the closing price and High and Low is 1, which shows a high correlation, so the above indicators are preliminarily

selected to be introduced into the model for prediction. If highly correlated technical indicators are introduced at the same time, it may increase the complexity of the model, and even lead to a decrease in the prediction performance of the model, and the computational cost will also increase, which needs to be further screened. In this paper, High, Low, and Open are added to the model as feature indicators, and it is found that they do not significantly improve the performance of the model, so the High, Low, and Open indicators are removed.

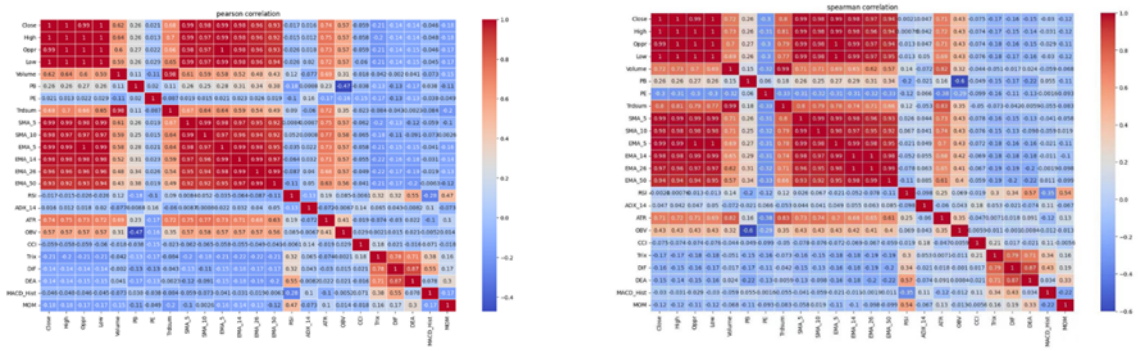

Figure 1: Pearson and Spearman Correlation Coefficient Diagram.

3.2.2 Embedded Feature Selection

Embedded feature selection is a very efficient method that can closely integrate the feature selection process with the training process of the learner [9]. To ensure the accuracy and effectiveness of feature selection, we utilize two different ensemble learning models, RF and XGBoost [10], to calculate and rank the importance of features separately, followed by cluster analysis to assist feature selection.

After constructing the RF model, feature importance was calculated and ranked, as shown in Figure 2. The EMA_5 index exhibited the highest importance at 0.9743, exceeding the 0.25 threshold, indicating a substantial impact on predictions. Features with importance below 0.25 were considered negligible and eliminated. The MSE for this ranking was 0.0013, confirming the reliability of the RF results. Notably, EMA_5 was also retained in the initial filter feature selection, justifying its inclusion in the final prediction model.

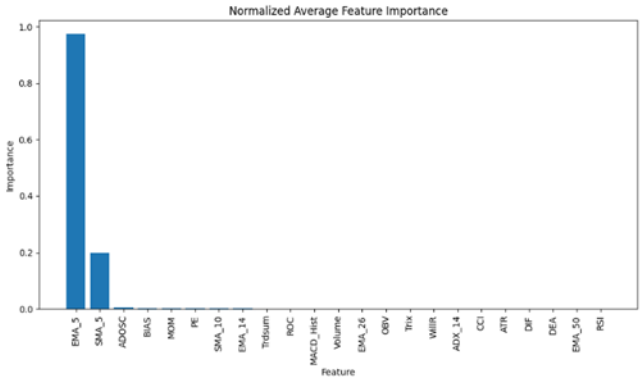

Figure 2: Ranking Chart of the Importance of Each Feature of the RF Model.

As shown in Figure 3, EMA_5 and SMA_5 achieved feature importance scores exceeding the 0.25

threshold, indicating their substantial impact on predictions, while other indicators scored below 0.05 and were eliminated. Both indicators passed the initial filter feature selection. Although SMA_5 showed inconsistent selection between RF and XGBoost, further validation confirmed its effectiveness in improving model performance. Therefore, both EMA_5 and SMA_5 were introduced into the final model.

Figure 3: Importance Ranking Diagram of Each Feature of the XGBoost Model.

3.2.3 Cluster Analysis

As shown in Figure 4, clustering divided the indicators into five clusters, with a mean squared error of 0.004, indicating reliable classification. The highest-scoring feature from each cluster—SMA_10, MOM, EMA_14, RSI, and ADX_14—was selected. Among these, SMA_10 and EMA_14 had already passed the filter feature selection. After sequentially adding the remaining three to the model, only MOM significantly improved performance; RSI and ADX_14 showed negligible gains and increased complexity, so they were excluded. Thus, MOM was retained alongside SMA_10 and EMA_14.

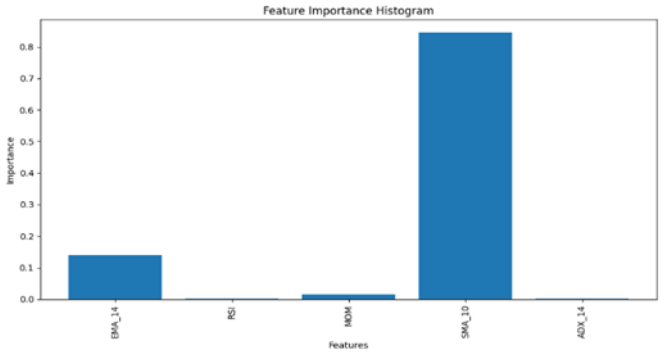

Figure 4: Cluster Feature Importance Score Table.

3.2.4 Feature Selection Results

In order to prevent the computational cost and increase the complexity of the model that may be brought about by introducing too many indicators, we introduce the other 4 indicators into the model in turn, and test the effectiveness of the indicators by judging whether to improve the performance of the model, and the test finds that the other 4 indicators have improved the performance of the model. Therefore, for the example data, the SMA_5, SMA_10, EMA_5, EMA_14, EMA_26, EMA_50, and

MOM models were finally selected for subsequent prediction.

3.2.5 Interpretation of Key Features from a Financial Perspective

Beyond statistical selection, the top-ranked features can be interpreted through established financial theories. Based on the importance rankings from RF and XGBoost, as well as their prevalence in the filtering and clustering stages, the five most influential technical indicators are identified as: EMA_5, SMA_5, EMA_14, SMA_10, and MOM.

1. EMA_5 and SMA_5 (Short-term Moving Averages): These indicators represent the short-term trend and are central to trend-following strategies. Their high importance suggests that recent price movements exert a strong influence on future prices, consistent with the phenomenon of momentum effect in behavioral finance, where assets performing well (poorly) in the recent past tend to continue performing well (poorly) in the short term.

2. EMA_14 and SMA_10 (Medium-term Moving Averages): These reflect the medium-term trend. Their selection, alongside short-term averages, allows the model to capture multi-scale market trends. The relationship between short-term and medium-term averages often signals trend strength or potential reversals.

3. MOM (Momentum Indicator): This indicator directly measures the rate of price change over a specific period. Its significance strongly aligns with the **momentum effect** theory, indicating that the model effectively captures this well-documented market anomaly where securities with strong past returns continue to outperform.

4. Case Analysis

4.1 Evaluation Indicators

The predictive performance of the stock prediction model is evaluated using RMSE, MAE, and R^2 [10].

RMSE root mean square error index

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

MAE average absolute error indicator

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$$

R^2 fit goodness metrics

$$R^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2}$$

4.2 Random Forest with Rolling Window Validation

Given the non-stationarity of stock data and potential structural breaks (between bull and bear markets), training a single model on multi-year data may yield overly optimistic results. To more realistically assess model performance akin to real-world deployment, we employ a rolling window forecasting approach instead of a static train-test split for robustness evaluation.

The RF model was evaluated using the rolling window approach with the following enhanced configuration:

Training Window: 500 trading days (≈2 years)

Testing Window: 63 trading days (≈3 months)

Uncertainty Quantification: 95% prediction intervals

Market State Identification: Automatic bull/bear market classification

Table 2 results demonstrate that the model maintains strong performance across different market regimes, with prediction intervals providing reliable coverage.

Table2: Results under the Rolling Window Method Considering Market Conditions.

Market Condition	R^2	MAE	RMSE
Overall	92.05%	0.041	0.029
Bull Market	90.71%	0.049	0.035
Bear Market	93.17%	0.035	0.023

4.3 Empirical Analysis Based on the Random Forest Model

In this study, we use the RandomForestClassifier class from the sklearn library to build an RF model. Before building the RF model, key parameters are set as follows. The number of trees is set to 75 to ensure model stability, balancing improved accuracy with computational cost. The maximum tree depth is limited to 5, a value determined by cross-validation to prevent overfitting while learning complex patterns. Additionally, the minimum samples for splitting an internal node is set to 2, also chosen via cross-validation to control tree growth. Finally, parallel processing is enabled by setting n_jobs to-1, utilizing all CPU cores to expedite training. Based on the selected important features and combined with the rolling window method, the results with 95% confidence intervals are presented in Figure 5.

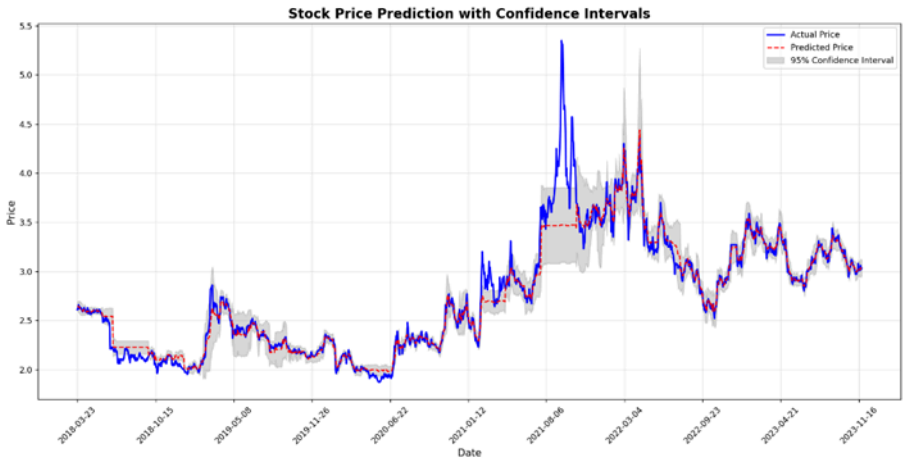

Figure 5: Random Forest Model Prediction Diagram.

We use R^2, RMSE, and MAE evaluation models. R^2 of 0.92 also indicates that the accuracy of the model regression prediction is very high, with RMSE of 0.029 and MAE of 0.041, which is very small, indicating that the model has good performance.

5. Summary

This study establishes a stock price prediction model based on Random Forest through a multi-stage feature selection framework and a rolling window evaluation mechanism. Empirical results show that the model maintains stable performance under both bull and bear market

conditions, with slightly better performance in bear markets (R^2 = 93.17%), validating its adaptability to changing market conditions. By integrating feature importance analysis and uncertainty quantification, the model not only provides point forecasts but also delivers reliable prediction intervals, offering valuable support for risk management decisions. Future research could deepen feature engineering by incorporating additional technical indicators and fundamental data, and by exploring deep learning methods for automatic feature extraction. Further exploration of new techniques in time series forecasting could also enhance the modeling of long-sequence dependencies, thereby improving prediction accuracy and practical applicability.

References

[1] Zhang J, Cui S, Xu Y, et al. A novel data-driven stock price trend prediction system. Expert Systems with Applications, 2018, 97: 60-69. DOI:10.1016/j.eswa.2017.12.026

[2] Meher Bharat Kumar, Singh Manohar, et al. Forecasting stock prices of fintech companies of India using random forest with high-frequency data. Journal of Open Innovation: Technology Market and Complexity, 2024, 10(1): 78-99. DOI:10.1016/j.joitmc.2023.100180

[3] Keren He, Qian Jiang. Research on stock prediction algorithm based on CNN and LSTM. Academic Journal of Computing and Information Science, 2022, 5(12): 23-35. DOI: 10.25236/AJCIS.2022.051215

[4] Fischer, Thomas, Krauss, Christopher. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 2018, 270(2): 654-669. DOI:10.1016/j.ejor.2017.11.054

[5] Zhang Y, Lin X. Empirical analysis of china stock market based on grey system. Academic Journal of Business and Management, 2023, 5 (12): 112-115. DOI:10.25236/AJBM.2023.051219

[6] Yu B. Is the Chinese stock market efficient? Evidence from a combined liquidity trading strategy. China Finance Review International, 2026, 16 (1): 61-96. DOI:10.1108/CFRI-01-2024-0011

[7] Ladjmil N, Benzerra A, Bosseler B. Predicting the structural condition of sewer pipes: a comparative analysis of Random Forest and logistic regression models. Urban Water Journal, 2026, 23 (3): 445-459. DOI:10.1080/1573062X.2025.2571908

[8] Breiman L. Random forests. Manchine Learning, 2001, 45(1): 5-32. DOI:10.1023/A:1010933404324

[9] Semnani M A, Kordrostami S, Sheikhani R A, et al. A hybrid framework for stock price forecasting using metaheuristic feature selection approaches and transformer models enhanced by temporal embedding and attention pruning. Applied AI Letters, 2026, 7(1): 1-17. DOI:10.1002/AIL2.70018

[10] Huisi H, Yiming Y, Jianlong H, et al. Construction of a clinical path discrimination model for stroke patients based on the XGBoost integrated learning algorithm and its application analysis in the MOP under the DIP payment model. Journal of Clinical and Nursing Research, 2025, 9 (4): 291-298. DOI:10.26689/JCNR.V9I4.10484