Proceedings of ICFTBA 2025 Symposium: Data-Driven Decision Making in Business and Economics
DOI: 10.54254/2754-1169/2025.BL29181

Can the Multi-dimensional Random Forest Lacking Macroscopic Technical Features Improve the Accuracy of Stock Prediction in Index Model?

Yuxin Fang

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu,China
20231057@aufe.edu.cn

Abstract. Traditional Index Models have long been constrained by oversimplified assumptions limiting their accuracy in portfolio construction. This paper proposes a machine learning-enhanced Index Model framework, leveraging Random Forest algorithms whether this way can improve the accuracy of stock prediction or not. Using daily data of the OEX index (and its constituent stocks) from 2004 to 2024 as the empirical sample, this paper conducts the empirical research. The results show that the machine learning-enhanced Index Model can achieve a lower test-set MSE and higher R^2 compared to the traditional Index Model, indicating superior return prediction accuracy. The ability of the random forest to optimize the index model portfolio depends on the compatibility of stock features with the model. Therefore, the original random forest has the characteristics of low interpretability and high bias failure, which leads to the model being prone to failure. Moreover, this prediction constructed based on the model's predicted returns outperform the benchmark OEX index in risk-adjusted returns and risk control. The random forest model captures the complex situations overlooked by the traditional index model, and provide more reliable input for stock prediction.

Keywords: Random forest, Index model, Multi-dimensional feature, Macro factor.

1. Introduction

The accuracy of predicting stock future returns becomes one of the most significance issues in financial market. Owing to the constant increases of complexity in financial investment, traditional signal investment models can hardly decrease multiple risk factors in accuracy. How to realize scientific Investment Portfolio Construction becomes global pressing scheme.

Previous literature explored several models to study investment portfolio. Capital Asset Pricing Model (CAPM) focuses on quantifying the relationship between risk and return to guide investors in asset allocation [1]. As for the application of machine learning in asset allocation and portfolios, focuses on the specific application of machine learning in portfolio allocation, studying how to optimize asset weight allocation through machine learning algorithms to improve portfolio performance [2,3]. Using machine learning, identifies macro factors driving tail risks through the random forest algorithm, analyzes their impact on the risk-adjusted returns of stock/bond portfolios [4]. The random forest method can predict stock market trends [5,6]. Some scholars combine Markowitz's traditional portfolio theory (mean-variance model) with machine learning algorithms such as random forest and XGBoost to study the optimal diversification of South African stock portfolios, exploring the integration of traditional theories and modern algorithms in asset allocation [8]. Some studies propose a hybrid model of LSTM and Graph Neural Network (GNN), integrating time-series analysis and graph structure analysis for stock price prediction, capturing market dynamics through multi-dimensional features to improve prediction accuracy. Some literature combines large language models to analyze qualitative risk factors disclosed by companies, and constructs a financial risk assessment framework in combination with traditional quantitative techniques, studying its impact on market volatility and exploring the application of new risk factors in financial forecasting [9]. Machine learning models also has some limitations [10].

This paper mainly focuses on improving the accuracy of stock prediction using random forest. It collects daily returns of various stocks from OEX Bloomberg over recent years, builds a random forest model, and verifies the accuracy of predicting future returns from three aspects: model performance evaluation, economic logic validation, and business scenario application based on the output results. By applying machine learning techniques such as random forest to mine effective features, an enhanced Index Model is constructed to optimize the risk-return performance of the stock prediction. Through the financial data cleaning and feature engineering in Python, the construction of machine learning models (random forest), the screening of effective verification dimensions and model validation, the application of stock prediction and the analysis of effectiveness. This helps to combine theories such as momentum effect and volatility to analyze the actual operation logic of the market, and support the optimization of stock prediction for quantitative investment decisions.

The innovation of this research lies in adopting multi-dimensional feature engineering, including lagging return features, volatility features, and technical indicator features (constructed based on daily returns and lacking macroeconomic factors). It has achieved the transformation from a single raw data to multi-dimensional features, without relying on the influence of external data such as macro and company characteristics. Solely by deeply exploring the feature engineering of daily returns, a feature set that can reflect multi-dimensional information such as "return inertia", "risk level", and "trend signal" of assets has been constructed. This provides rich input for the random forest model, thereby optimizing the stock prediction.

2. Data and method

2.1. Data

2.1.1. Data source

The data for this study comes from the OEX Bloomberg Data financial database in recent years provided by professors at the University of Colombia in the advanced investment science research project. The database covers representative data from the U.S. financial market, including daily market data of the OEX index, constituent stock information, etc., providing authoritative and comprehensive basic data support for subsequent data processing, feature engineering and stock effectiveness analysis. The investigation employed the method of using Python to clean the database and build a machine learning model with effective feature engineering. The investigation focused on the opening dates and daily returns of different stock codes from the OEX Bloomberg Data table (mainly the "returns" section). Using the recent accurate Bloomberg data is representative, and through multi-dimensional feature engineering to optimize the input and enhance the generalization ability of the random model, the accuracy of the US OEX index model was improved.

2.1.2. Data processing

Data preparation and feature engineering are carried out in four steps: data cleaning, core processing (seasonal adjustment/ lagging), multi-dimensional feature construction, and feature selection:

Data preparation. Unify the time dimension and handle missing values, and remove outliers.

Seasonal adjustment (eliminating periodic fluctuations). Use the STL (seasonal decomposition) tool in the stats models' library of Python to separate "trend term + seasonal term + residual term", and take "trend term + residual term" as the adjusted data. Example: For monthly OEX index returns, use a 12-period moving average to eliminate the "January effect" and "year-end effect", and obtain the seasonal-adjusted returns.

Delayed processing (capturing time series dependencies). For Daily Returns, construct the return characteristics for 1-day lag (Lag1), 5-day lag (Lag5, which refers to the past 1 week), 20-day lag (Lag20), 30-day lag (Lag30), and 60-day lag (Lag60). For volatility (such as the daily volatility in Daily Statistics), construct the volatility characteristics for 1-week lag (Lag7) and 1-month lag (Lag30), reflecting the historical risk level.

Multi-dimensional feature construction (based on daily returns). Lagged return characteristics: Lag1, Lag5, Lag20, etc.; Volatility characteristics: Rolling volatility over the past 5 days and 20 days (annualized) and lagged volatility. Technical indicator characteristics: Based on daily returns;

Divide the training set and test set, use mean square error and coefficient of determination to evaluate the prediction accuracy of the model for "future returns"; then verify whether the above three characteristics are reasonable.

Lagged returns: Historical returns have momentum or reversal effects on the future, it is necessary to observe whether the sign and importance of the lagged characteristics conform to market laws.

Volatility: High volatility usually accompanies high risk, it is necessary to verify whether the negative correlation between "volatility_5 days" and future returns is significant.

Technical indicators: For example, when RSI is overbought (>70), returns decline, and when RSI is oversold (<30), returns increase, it is necessary to verify through group statistics whether such patterns exist. If the mean of future returns in the "oversold" group is > that in the "overshot" group, it indicates that the logic of RSI's overbought and oversold in the data is valid.

2.2. Random forest model

Random forest is a supervised learning algorithm based on ensemble learning. Its core is to enhance prediction accuracy and stability through "voting of multiple decision trees", proposed by Leo Breiman in 2001. Its core principle is "random sampling + multi-tree integration", and the process can be simplified into 3 steps.

2.2.1. Sample random sampling (bootstrap sampling)

From the original training set, hundreds of different sub-training sets are generated through "with-replacement sampling". Each subset has a size comparable to the original data set, and approximately 63.2% of the original samples are selected. The remaining 36.8% are "out-of-bag samples OOB", which can be used for model validation without additional data.

2.2.2. Feature random selection

When building each decision tree, for each decision node, a portion of features from all features are randomly selected. The optimal splitting rule is found based on this part of the features to avoid over-reliance of a single tree on a certain strong feature.

2.2.3. Ensemble voting of multiple trees

For classification tasks: The output of all decision trees is the category, and the final result is the "category with the highest number of votes"For regression tasks: The output of all decision trees is the prediction value, and the final result is the "average of all prediction values".

Advantages of the Random Forest model: High accuracy and strong generalization ability. Through multiple decision trees, the risk of overfitting of a single decision tree can be effectively reduced. The fitting and prediction effect on complex data is better than a single decision tree; It has outstanding anti-overfitting ability. Through sample Bootstrap sampling and feature random selection, multi-trees can avoid "homogenization", and at the same time, out-of-bag samples (OOB) can be directly used for evaluating model performance; It is insensitive to noise and outliers and supports feature importance assessment.

3. Results and discussion

The random forest model obtained through Python code is shown in the chart as follows (This basic random model was established by studying 90 stocks over the past two decades. Taking the random forest chart of NRFR as an example.):

3.1. The MSE and coefficient of determination of random forest

As shown in table 1, the validity of the model is first determined using a dual criterion: the coefficient of determination (R^2) ≥ 0.6 indicates that the model can explain more than 60% of the stock returns/price fluctuations [9]; the mean squared error (MSE) ≤ 0.0005 means that the prediction deviation is within an acceptable range [9,11].

Table 1. The MSE and coefficient of determination of random forest

Stock	MSE of the Model	Coefficient of Determination	Model validity determination
NRFR	0.0000	0.9819	Valid

From the output results, it can be seen that the effectiveness of the original random forest in the investment portfolio optimization of the index model is highly dependent on the stock characteristics and the model's adaptability. Among the ninety stocks, only the NRFR stock simultaneously meets the dual conditions of the determination coefficient and the mean square error. The failure reasons for the remaining eighty-nine stocks are as follows:

Firstly, in terms of the determination coefficient, the main reason is that the model did not input external variables such as policies and macroeconomics to strengthen the technical characteristics, resulting in the determination coefficient being far below the threshold and unable to capture the key fluctuation patterns[6].

Secondly, in terms of the mean square error, the MSE deviation range is too large, exceeding the acceptable range, making it difficult to capture the complex patterns of the financial market.

3.2. Analysis of NRFR

3.2.1. The MSE and coefficient of determination of random forest in validity

As shown in table 1. NRFR's determination coefficient reaches 0.9819, indicating that the model can explain 98.19% of the stock return fluctuations. At the same time, its mean square error is 0, indicating that the predicted values have no deviation from the true values, which can provide a reference for the investment portfolio, but it is necessary to consider whether there is an overfitting situation (Table 2).

Table 2. The MSE and coefficient of determination of random forest in validity

Random Forest Model	Values
MSE of the Model	0.0000
Coefficient of Determination	0.9819

3.2.2. The top 5 feature importance in random forest in validity

The top 5 feature importance in random forest in validity include, SPX Volatility 5-day, Mean Return 20-day, Lag20, Lag60 and Lag 1 (Table 3).

Table 3. The top 5 feature importance in random forest in validity

Top 5 Feature Importance in Random Forest	Values
SPX Volatility 5-day	0.970531
Mean Return 20-day	0.015072
Lag20	0.006972
Lag60	0.005526
Lag 1	0.001898

4. Conclusion

This paper employs machine learning methods to establish a multi-dimensional feature random forest model without external economic factors. At the same time, it investigates whether this model can optimize the stock prediction under the index model. Through research, it was found that although technical features such as lack of policies and company macro factors can enhance the prediction ability of stocks, in most cases, they will lead to model failure and fail to achieve the expected results.

Clarifying model boundaries: Its effectiveness highly depends on the inclusion of external technical features such as macroeconomic factors. This indicates that its performance can only be exerted with the support of specific feature engineering. Promoting model integration research: It points out the direction for subsequent research, that is, it is necessary to deeply explore how to more efficiently integrate external factors such as macroeconomics and policies with the random forest model to break through the current performance bottleneck of the model.

Practical implications: It emphasizes the crucial role of multi-dimensional feature engineering in prediction optimization, especially the technical features of macroeconomic factors. Investors and researchers must pay attention to the mining and integration of external features such as macroeconomic indicators and policy variables when building an investment model based on random forests, and cannot rely solely on internal features.

This model can be effectively applied to active stock selection and monthly portfolio rebalancing scenarios, achieving investment returns that exceed the market benchmark by quantitatively screening for high-value stocks. A thorough study on why random forest is prone to failure in the absence of external conditions such as policies and macroeconomic factors of the company.

References

[1] Sharpe, W. F. (1963). A simplified model for portfolio analysis. Management science, 9(2), 277-293.

[2] Kris Boudt (2020). Machine Learning for Asset Managers, Quantitative Finance, 20(11), 1761-1762.

[3] Pinelis, M., & Ruppert, D. (2022). Machine learning portfolio allocation. The Journal of Finance and Data Science, 8, 35-54.

[4] Mueller-Glissmann, C., & Ferrario, A. (2024). Dynamic Asset Allocation Using Machine Learning: Seeing the Forest for the Trees. Journal of Portfolio Management, 50(5).

[5] Wijaya, A. Y., Fatichah, C., & Saikhu, A. (2023, November). Prediction of stock trend using random forest optimization. In 2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA) (pp. 1-6). IEEE.

[6] Tratkowski, G. (2020, May). Construction of Investment Strategies for WIG20, DAX and Stoxx600 with Random Forest Algorithm. In Contemporary Trends and Challenges in Finance: Proceedings from the 5th Wroclaw International Conference in Finance (pp. 179-188). Cham: Springer International Publishing.

[7] Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. The Review of Financial Studies, 28(3), 650-705.

[8] Moyoweshumba, E., & Seitshiro, M. (2025). Leveraging Markowitz, random forest, and XGBoost for optimal diversification of South African stock portfolios. Data Science in Finance and Economics, 5(2), 205-233.

[9] Gupta, S., & Yan, H. (2025). Using Large Language Models to Estimate Novel Risk: Impact on Volatility. Journal of Portfolio Management, 51(7).

[10] Ferreira, J. A. (2022). Models under which random forests perform badly; consequences for applications. Computational Statistics, 37(4), 1839-1854.

[11] Li J, Zhang H, Wang Y (2025).A Hybrid LSTM-GNN Model for Stock Price Prediction: Integrating Time-Series and Graph-Based Analysis.https: //arxiv.org/pdf/2502.15813.pdf