Optimizing Investment Strategies: A Random Forest Approach to Stock Return Prediction and Portfolio Management

Jingxuan Bian[1, *] and Jiaxin Lin[2]

[1]Computer Science, Hebei University, 071002, Hebei, China
[2]Finance and Computer Science, Boston College, 140 Commonwealth Ave Chestnut Hill MA 02467, USA
*20221202028@stumail.hbu.edu.cn

Abstract. Quantitative finance is becoming an increasingly useful tool in modern financial markets by utilizing computational power to optimize investment strategies. The idea of Quantitative finance originated from early theories like the Efficient Market Hypothesis and Random Walk Theory proposed by Louis Bachelier in 1900. Today, machine learning has revolutionized how financial data is analyzed, with models such as Random Forest providing valuable insights for stock price prediction and portfolio management. This study focuses on employing the Random Forest model for predicting quarterly stock returns and volatilities by using financial data from the NASDAQ 100 Index. Some of the key corporate characteristics that are used in this study are net profit, return on equity (ROE), and total liabilities. The model aims to identify key factors that influence stock performance through analyzing the key characteristics. The predictions generated by the model are then used to construct optimized investment portfolios, which are tested against benchmark portfolios, the NASDAQ 100 index, in a back testing framework. The results demonstrate that the Random Forest model effectively captures patterns in stock performance and enhances decision-making in portfolio management. The results of this study also highlight the potential of machine learning in improving stock selection and asset allocation strategies. In addition, this approach is contributing to more informed decision-making in long-term investment portfolios.

Keywords: Random Forest, Return Prediction, Machine Learning.

1 Introduction

The earliest stock market in human history can be traced back to the 13th century, when Venetian merchants in Italy started trading government securities for the first time. The first company that issued stock was called The Dutch East India Company back in 1602, where people could buy shares from the company and receive dividends when the company became profitable. However, when the stock market kept on expanding and became increasingly efficient, traditional finance theories became hard

© The Author(s) 2025
J. Lu (ed.), *Proceedings of the International Workshop on Navigating the Digital Business Frontier for Sustainable Financial Innovation (ICDEBA 2024)*, Advances in Economics, Business and Management Research 315,
https://doi.org/10.2991/978-94-6463-652-9_70

to predict and the idea of Quantitative Finance came into play. It is a method that uses mathematical models, statistical techniques, and computational tools for investment decisions in the financial market. The idea of Quantitative Finance first came up in 1900, when Louis Bachelier came up with the idea of Efficient Market Hypothesis and the Random Walk Theory. Nowadays, there are more advanced computers that can do hundreds of millions of calculations for machine learning techniques and Artificial Intelligence (AI) models that can help people to make predictions and improve their trading strategies.

Random forest, one of the machine learning methods that is often being used for forecasting various features in many fields. As AI became more and more affordable and easy to access for all human beings, fields such as Biology, Chemistry, and Medicine started applying machine learning to help forecast or simulate future events, especially using the random forest method. For example, Wang et al. used random forests to predict the hot spots in protein in protein interfaces that provides crucial information for the research on protein [1]. Another team who predicts the rate constants of semivolatile organic compounds with hydroxyl radicals and ozone in indoor air uses multiple models for forecasting, including QSAR, LASSO, and random forests, and shows random forest was one of the models that were the most effective statistical models [2]. When machine learning comes to the financial market, machine learning is very useful for forecasting stock price movements. Pushpendu Ghosh's team was able to employ random forests and long short-term memory networks as their machine learning methods for predicting directional movements of stock prices for intraday trading [3]. Research shows that random forests are very successful and efficient methods in various machine learning methods. Therefore, our team decided to use random forest as our method to predict future stock return.

In finance and investment, the ability to predict quarterly returns and effectively select assets is paramount to achieving successful portfolio management. Anthony et al. has undertaken a comprehensive project aimed at leveraging machine learning techniques to enhance the prediction of quarterly returns and inform asset selection decisions. The project is structured into four core modules: Data, Machine Learning, Portfolio Construction, and Back testing. By utilizing data from the NASDAQ 100 Index and incorporating corporate characteristics, particularly financial data. They employ the Random Forest machine learning model to predict quarterly returns and volatility. Subsequently, these predictions are used to construct portfolios, which are then compared to benchmarks during the Back testing phase. This paper delves into the methodologies employed, the findings derived from the analysis, and the implications of the results for portfolio management and asset selection strategies.

2 Method

2.1 Dataset Preparation

By utilizing data from the NASDAQ 100 Index and incorporating corporate characteristics, particularly financial data, we employ the Random Forest machine learning model to predict quarterly returns and volatilities. Subsequently, these predictions are used to construct portfolios, which are then compared to benchmarks during the Back testing phase.

We obtained the financial data of Nasdaq 100 index stocks from 2010 to 2020 through the yfinance interface, including securities such as Apple, Airbnb, Adobe, and others. We selected financial indicators for each security, including net profit, operating income, ROE, total liabilities, total assets, operating profit, return on investment, volatility of income growth rate, net profit growth rate, and operating margin, across every year and quarter.

Subsequently, we divided the dataset into two parts: a training set and a testing set. The training set comprised data from 2010 to 2018, while the testing set included data from 2019 to 2020. The latter was utilized for back testing and optimizing our model, as well as refining our investment portfolio strategies. Our focus was on predicting two key variables: the expected return rate and expected volatility for the subsequent quarter.

2.2 Machine Learning-based Stock Price Prediction

Random Forest. Random Forest, an ensemble learning technique that leverages the power of multiple decision trees, has emerged as a highly effective and versatile tool for various predictive modeling tasks [4-6]. By constructing a forest of decision trees, each trained on a subset of features and samples drawn randomly from the original dataset through bootstrap sampling, Random Forest achieves remarkable performance gains over individual decision trees. This approach not only enhances the overall accuracy of predictions but also imparts robustness against noise and outliers in the data. So this led our team to choose this model as the solution for the project.

We have chosen random forest as the prediction model, which has the following characteristics: 1) Integration advantage: Random forests effectively reduce the risk of overfitting that may exist in a single decision tree by constructing multiple decision trees and integrating their prediction results, thereby improving the accuracy and stability of overall predictions. 2) Feature importance assessment: Random forests can evaluate the importance of various input features (such as bond yields, debt ratios, etc.) on the prediction results, which helps identify key factors that affect bond yields. 3) High dimensional data processing: The bond market involves a large amount of data, including historical yields, macroeconomic indicators, market sentiment, etc. Random forests can effectively process these high-dimensional data and extract useful information for prediction.

Adjusting Parameters. In the task of prediction, we have utilized the Random Forest model and fine-tuned its crucial parameters to optimize prediction performance. Specifically, we enhanced the model's robustness and accuracy by increasing the number of trees (n_estimators), allowing it to examine the data from multiple perspectives through the lens of distinct decision trees. In the meantime, we set the maximum depth of each tree (max_depth) to prevent overfitting by limiting their excessive growth. By capping the maximum depth, we ensured that the model maintained its ability to handle complex data while preserving good generalization capabilities. This strategic parameter optimization has empowered the Random Forest model to exhibit heightened precision and stability in bond yield predictions, providing a more reliable data foundation for investment decision-making.

2.3 Portfolio Optimization

Portfolio A. Portfolio A is designed to capitalize on the principles of Modern Portfolio Theory. The core objective is to construct a portfolio that balances risk and return, maximizing expected returns while maintaining a level of volatility acceptable to the investor. To achieve this, we utilize Monte Carlo simulation, a powerful stochastic method that involves repeatedly sampling from a probability distribution to estimate the likelihood of various outcomes.

Monte Carlo simulation is employed in Portfolio A to assess the potential performance of various investment portfolios under varying market conditions. By simulating thousands or even millions of scenarios based on historical data and assumptions about future market behavior, we can generate a distribution of potential outcomes for the portfolio's return and volatility. This distribution allows us to estimate the probability of achieving specific return targets and to quantify the portfolio's risk profile.

From this subset of stocks, we use Monte Carlo simulation to draw multiple portfolios, each comprising the top 10 stocks with the highest predicted returns. By simulating the performance of these portfolios over a range of market conditions, we can identify the optimal combination that balances risk and return most effectively. This optimization process leverages the vast number of potential portfolio configurations generated by Monte Carlo simulation [7-10], ensuring that our final portfolio selection is based on a thorough and systematic analysis.

Portfolio B. In contrast to Portfolio A's active approach, Portfolio B adopts a more passive strategy that aims to replicate the performance of the Nasdaq 100 index. This benchmark-tracking strategy provides a point of comparison for evaluating the performance of Portfolio A and allows us to assess the added value of our active investment strategy.

To construct Portfolio B, we first identify all stocks within the Nasdaq 100 index that had positive annual returns from 2010 to 2017. This historical performance criterion ensures that the stocks selected have demonstrated a consistent ability to generate positive returns over an extended period.

3 Results and Discussion

3.1 Parameter Adjustment

At the beginning, we obtained values like the one in the Table 1, However, the model's performance can be further optimized through parameter tuning. For example, by adjusting the n_estimators and max_depth parameters from (n_estimators=100, max_depth=10) to (n_estimators=135, max_depth=20).

Table 1. The performance of the model.

	Old Return Rate	**Old volatility**	**New Return Rate**	**New volatility**
MSE	0.4792	0.0085	0.415	0.0075
R^2	0.7144	0.7557	0.75268	0.7851

Next, we discussed model evaluation and parameter tuning. To assess the model's performance, we typically use Mean Squared Error (MSE) and the coefficient of determination (R^2). First, Adjusting the maximum depth (x_depth) of the tree, which can prevent overfitting of the model. We try to set an appropriate value to balance the complexity and generalization ability of the model. And second, we increase the number of trees (n_estimators) until the performance improvement of the random forest reaches a stable point. The most suitable numerical value for this model. We tested various combinations and ultimately determined that n_estimators=135 and max_depth=20 was the optimal parameter combination. This tuning process further reduced the MSE and increased the R^2 values, making the model more reliable. These results demonstrate the model's capability in predicting financial indicators.

3.2 Portfolio A

Portfolio A achieved strong annualized returns shown in Table 2, far exceeding the benchmark's. However, this high return comes with higher risk. The annualized volatility and the maximum drawdown are higher than that of benchmark's. When the market fluctuates, these stocks have a greater impact. It may be due to the concentration in only the top 10 stocks. We can also see from the cumulative return curve that Portfolio A is more volatile, especially during market downturns. To see if the high risk taken by the strategy is worth, The Sharpe ratio of 0.937 indicates better risk-adjusted returns than the benchmark's 0.769. However, considering the 27% annualized return, the sharp ratio is underperforming. The strategy's stability under high risk needs further enhancement.

Table 2. Portfolio A vs. Benchmark

	Cumulative Return	**Annualized Return**	**Annualized Volatility**	**Max Drawdown**	**Sharpe Ratio**
Portfolio A	1.600851	0.271400	0.264696	-0.281110	0.937301
NASDAQ 100ETF(benchmark)	1.363956	0.175492	0.197897	-0.227967	0.769049
Portfolio	1.564705	0.243344	0.192110	0-.199835	1.145404

Our investment portfolio A is based on Markowitz investment theory and uses Monte Carlo simulation to draw curves and find the optimal investment portfolio. We use machine learning to predict the return and volatility of 100 stocks in the index. We choose stocks with an average volatility of less than 18% based on Smith's investment preferences, and invest in the top 10 stocks with the highest returns shown in Table 3.

Table 3. The top 10 return stocks

Stock code	Name	Average return	Average volatility
ADSK	Autodesk	11.423%	17.922%
FTNT	Fortinet	10.467%	17.016%
LULU	lululemon athletica inc	9.322%	16.517%
PANW	Palo Alto Networks, Inc.	8.993%	16.343%
MU	Micron Technology, Inc.	8.821%	16.597%
FANG	Diamondback Energy Inc	8.657%	17.731%
EA	Electronic Arts	7.853%	15.211%
TTWO	Take-Two Interactive Software Inc	7.529%	16.751%
ILMN	Illumina Inc	7.453%	15.327%
MRVL	Marvell Technology, Inc.	7.371%	16.679%

3.3 Portfolio B

In contrast to Portfolio A, we have also created another portfolio that is more like a passive strategy for the Nasdaq 100 index, in which case it serves as our benchmark. We have decided to identify all stocks with positive annual returns from 2010 to 2017, form an investment organization, and conduct back testing using data from 2018 to 2019. In addition, we calculate the average cumulative return by dividing the cumulative return by the number of quarters for each stock. After completing the previous steps, we only have 90 stocks left in our investment portfolio. Using these 90 stocks, we have decided to form an investment portfolio where 80% of the weight is occupied by all 90 stocks. We selected the top 10 stocks with the highest average return rate in the dataset to form the remaining 20% of the weight shown in Table 4.

Table 4. The top 10 return stocks weights allocated

Season	ADSK	EA	FANG	FTNT	ILMN	LULU	MRVL	MU	PANW	TTWO
2018 season 1	30.80%	12.44%	1.97%	6.10%	1.89%	9.76%	4.53%	2.24%	27.53%	2.75%
2018 season 2	1.40%	7.53%	2.17%	0.35%	13.30%	8.36%	28.88%	24.20%	1.42%	12.40%
2018 season 3	5.50%	26.70%	6.93%	20.02%	3.59%	31.48%	0.65%	0.99%	1.35%	2.78%
2018 season 4	23.27%	2.61%	1.52%	2.26%	4.04%	15.44%	2.20%	11.44%	1.90%	35.31%
2019 season 1	36.52%	3.95%	2.72%	0.23%	2.25%	0.76%	20.80%	6.01%	18.81%	7.94%
2019 season 2	20.31%	0.57%	1.66%	16.79%	1.14%	25.50%	5.14%	26.35%	2.06%	0.48%
2019 season3	2.13%	0.17%	12.46%	27.87%	12.90%	26.58%	0.88%	1.06%	6.65%	9.29%
2019 season 4	2.97%	1.53%	24.73%	26.80%	2.62%	3.28%	0.96%	3.02%	6.05%	28.04%

The top 10 stocks with average returns that we found in our investment portfolio are TTWO TTD、TSLA、TEAM、REGN、NXPI、NVDA、NFLX、FANG、AVGO。Except for REGN and FANG, which are pharmaceutical and energy companies, the other eight companies are technology companies. As we predicted, all these companies have been able to maintain good performance over the years. Even if we bought these stocks in 2018 and hold them until today, we can still earn at least 400% return from each stock.

4 Conclusion

In conclusion, the performance of both Portfolio A and B during the back testing period is in line with the objectives we set when constructing them. Portfolio A employs a Monte Carlo simulation and adjusts the weights of the portfolio on a quarterly basis. The dynamic adjustment mechanism provides it with better resilience in the face of elusive market conditions. However, the back testing results also show that Portfolio A is less resilient to market volatility and the balance of strategy between risk and return is not ideal. The impact of individual stocks on the overall portfolio is magnified by the strategy, especially during times of high market volatility. Portfolio B performs excellent during the back testing period, particularly in terms of risk control. However, a notable feature of Portfolio B is that it does not adjust its portfolio weights. Its performance may be challenged against high uncertainty of future market conditions. The approach predicts the future based on past performance; an assumption that does not always hold. To enhance the flexibility and adaptability of the strategy, a dynamic adjustment mechanism may need to be introduced in the future to cope with changes in the market environment.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

1. Wang, L., Liu, Z. P., Zhang, X. S., & Chen, L.: Prediction of hot spots in protein interfaces using a random forest model with hybrid features. Protein Engineering, Design & Selection 25(3), 119-126 (2012).
2. Wei, W., Sivanantham, S., Malingre, L., Ramalho, O., & Mandin, C.: Predicting the rate constants of semivolatile organic compounds with hydroxyl radicals and ozone in indoor air. Environmental Pollution 266, 115050 (2020).
3. Sarker, P., Sayed, A., Apu, A. S., Tasnim, S. A., & Mahmud, R.: A comparative review on stock market prediction using artificial intelligence. Malaysian Journal of Science and Advanced Technology, 383-404 (2024).
4. Rigatti, S. J.: Random forest. Journal of Insurance Medicine 47(1), 31-39 (2017).
5. Biau, G., & Scornet, E.: A random forest guided tour. Test 25, 197-227 (2016).
6. Breiman, L.: Random forests. Machine Learning 45, 5-32 (2001).
7. Mooney, C. Z.: Monte Carlo simulation (No. 116). Sage (1997).

8. Harrison, R. L.: Introduction to Monte Carlo simulation. In: AIP Conference Proceedings 1204, p. 17. NIH Public Access (2010).
9. Mahadevan, S.: Monte Carlo simulation. Mechanical Engineering-New York and Basel-Marcel Dekker, 123-146 (1997).
10. Bonate, P. L.: A brief introduction to Monte Carlo simulation. Clinical Pharmacokinetics 40, 15-22 (2001).

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.