The Application of Ridge Regression, Random Forest and Mean-Variance Model in Portfolio Optimization

Tianyi Lu

Department of Economic and Finance, City University of Hong Kong, Hong Kong, 999077, China
tianyilu3-c@my.cityu.edu.hk

Abstract. In recent years, with the increasing volatility in global financial markets, which is driven by factors including geopolitical events and policy uncertainty, the need for more effective portfolio optimization techniques has highly intensified. This study explores the use of machine learning, specifically Ridge regression and Random Forest model, under the framework of the Mean-Variance Model to optimize portfolio returns. Historical stock between July 2014 and July 2023 was used. Key technical indicators such as moving averages (MA50, MA200), volatility, and volume-based metrics were utilized as input features. The specific models used in this study were Ridge Regression and Random Forest for comparing the performance of the linear and non-linear models. The predicted returns from both models' test sets were incorporated into the Mean-Variance portfolio optimization, aiming to maximize the Sharpe ratio. The back-testing results between July 2023 and July 2024 showed that both machine learning-enhanced portfolios outperformed the benchmark portfolio based on actual market returns, with the Random Forest portfolio achieving superior risk-adjusted returns and lower volatility. These findings represent the potential of utilizing machine learning models to enhance traditional portfolio optimization strategies.

Keywords: Portfolio Optimization, Ridge Regression, Random Forest, Mean-Variance Model

1 Introduction

In the current global economic landscape, with the characterization of increasing complexity and uncertainty of economics and finance, the volatility of the stock market will increase [1]. Geopolitical events, including the Russia-Ukraine conflict, have exacerbated economic uncertainty, by their significant disruption in global supply chains [2], as well as fluctuations in commodity prices, thus leaving substantial risks to investors. In that case, portfolio optimization has become crucial for maximizing investors' risk-adjusted returns. The Mean-Variance model, which was initially developed by Harry Markowitz in 1952, is a classic portfolio optimization technique, aiming at maximizing the Sharpe Ratio of a portfolio [3].

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_62

Given the heightened volatility in today's financial markets, the need for effective portfolio optimization strategies is more urgent than ever.

The application of advanced algorithms, particularly concerning Artificial Intelligence (AI), has led to revolutions in various areas, including medicine and physics. Techniques such as decision trees and regression models have been widely employed to enhance decision-making processes across sectors. For instance, according to a study conducted by Kinar et al. [4], machine learning can help evaluate a patient's risk for colon cancer by analyzing their medical record. In finance, these methods are also widely used but still have some potential. Angelini explored the application of neural networks in credit risk assessment [5], showing that classical feedforward neural networks can be effectively used to evaluate one's credit risk [6]. Regarding portfolio optimization, there are also plenty of studies conducted with the aid of AI techniques in this area. Day et al. found that the Long Short-Term Memory (LSTM) model is the best model to predict the stock market [7], and while incorporating it into a portfolio optimization model, it can outperform ETFs. Additionally, Mousavi created a dynamic portfolio trading system with the aid of multitree genetic programming [8], and the results show that the adjusted return outperforms other traditional dynamic portfolios. Andriyashin has also analyzed the application of decision trees, incorporated with implied volatility, which can be used to pick stocks from XETRA DAX, and the results display significant predictive power. Research conducted by Sohrab Mokhtari and Kang Yen shows that applying a linear regression model from the perspective of fundamental analysis to predict the market is feasible [9], but the accuracy is not good. Another model from a technical perspective fails to outperform the market, highlighting the need for further exploration of AI techniques in the finance industry.

Despite these advancements and constraints, there remains a critical need to refine portfolio optimization models to better adapt to the ever-changing market conditions. The practicality of optimizing investment portfolios to generate the best possible risk-reward ratio should be emphasized, especially in volatile markets. This paper will focus on the Mean-Variance model and examine whether the application of machine learning can improve portfolio returns. First, financial data will be sourced from Yahoo Finance. Next, Ridge regression and Random Forest will be applied separately to project future stock returns. Then, the predicted returns will be used as part of the inputs for the Mean-Variance Model. Last but not least, the resulting portfolio returns will be compared with those of a portfolio that uses purely past stock data.

2 Method

2.1 Dataset Preparation

For this study, historical stock data originates from Yahoo Finance, through Yfinance package in Python, and covers the period from July 1st, 2014 to July 1st, 2023. The raw dataset includes daily adjusted closing prices and volumes for the eight stocks this essay aims to analyze: Apple (AAPL), Microsoft (MSFT), Amazon (AMZN), Tesla (TSLA), AT&T (T), JPMorgan Chase (JPM), Procter & Gamble (PG), and Coca-Cola (KO). As for pre-processing, missing value is handled and daily return is calculated as the percentage change in adjusted closing price for each stock. The dataset is then divided into three parts: A training set covering the period between 2014.7.1 and 2019.7.1, a validation set covering the period between 2019.7.1 and 2021.7.1, and a test set covering the period between 2021.7.1 and 2023.7.1.

2.2 Linear Regression

Linear regression is a fundamental predictive model that assumes a linear relationship between input features and the target variable. In this study, a Ridge regression model was used to prevent overfitting by adding an L2 regularization term controlled by the hyperparameter alpha. The target variable for this linear regression model is daily return. As for input features, include the adjusted closing price as well as several technical factors including the 50-day and 200-day moving averages (MA50, MA200), 50-day volatility, the Relative Strength Index (RSI), and volume-based metrics, i.e. 50-day moving average of volume. The Ridge model is trained using the training set and further processed using the validation dataset. Model evaluation will be conducted concerning the performance of the validation set and test set. The evaluation metrics include Mean Squared Error (MSE) and R-squared.

2.3 Random Forest

Random Forest is a widely used ensemble learning method. In this study, Random Forest model is implemented using the RandomForestRegressor class from the sci-kit-learn library, with the number of decision trees (n_estimators) set to 100. As for feature selection, the model used the same technical indicators as the Ridge regression model for consistency, i.e. 50-day and 200-day moving averages (MA50 and MA200), 50-day volatility, Relative Strength Index (RSI), and 50-day moving average of volume. The model is trained and validated by the same dataset as Ridge Regression has used. Also, the model evaluation will be based on the result of the validation set and test set, where key evaluation metrics include MSE and R-square.

2.4 Portfolio Optimization

In this section, Mean-Variance Model (MVM) is used to construct an optimized portfolio with the primary objective of maximizing the Sharpe ratio [10], which reflects the portfolio's risk-adjusted return.

In this optimization process, three distinct sets of return data were used as inputs. The first two datasets were derived from the test set predictions of the machine learning models (Ridge regression and Random Forest) for the period between July 1, 2021, and July 1, 2023. These predicted returns were combined with real stock return data spanning from July 1, 2014, to August 31, 2021, to form two separate input datasets. The third dataset consisted solely of real stock returns for the entire period from July 1, 2014, to July 1, 2023, which aims to create a benchmark portfolio to evaluate the performance of the previous two portfolios.

To optimize the portfolio, there were additional constraints on the weights of each stock: the sum of all stock weights was required to equal 1, and individual stock weights were restricted to fall between 5% and 30%. The optimization process utilized the Scipy package to solve for the weight distribution that maximized the Sharpe ratio. Key portfolio statistics, including the expected portfolio return and portfolio volatility, were calculated post-optimization to evaluate the efficiency of the resulting portfolios.

The optimized mean-variance portfolios based on the predicted returns from the Ridge regression and Random Forest models were evaluated against a mean-variance portfolio optimized using actual stock returns from the test set. The evaluation involved a comparison of key portfolio metrics, including expected returns, portfolio volatility, and Sharpe ratios.

Furthermore, each of the optimized portfolios was simulated for a future period spanning from July 2023 to July 2024. During this simulation, both cumulative returns and excess returns (the difference between portfolio returns and actual market portfolio returns) were calculated for the Ridge and Random Forest portfolios. This comparison aims to provide insight into how effectively the machine learning models' predicted returns outperform the real market. Apart from that other evaluation metrics i.e. volatility, and Sharpe ratio will also be studied to compare the effectiveness of using machine learning to enhance the practicality of Mean-Variance Model.

3 Results and Discussion

3.1 Performance of Ridge Regression and Random Forest

Table 1 compares the performance of Ridge Regression and Random Forest, showing that Ridge Regression outperformed Random Forest. Ridge Regression achieved a lower MSE of 0.000037, indicating greater prediction accuracy compared to Random Forest's MSE of 0.00014. Additionally, Ridge Regression had a higher R^2 value of 0.7, indicating that 70% of the variance in the data could be explained by this model, whereas Random Forest only explained 52% with an R^2 of 0.52.

Table 1. The performance of the models.

Model	MSE	R-square
Ridge Regression	0.000037	0.7
Random Forest	0.00014	0.52

3.2 Portfolio Comparison

Fig. 1 displays the cumulative return generated by the three portfolios, and it is obvious that Ridge Regression and Random Forest Portfolio outperform True Market

Portfolio, generating a higher return. Fig. 2 shows the excess return that Ridge and Random Forest generate over time, which is always above zero, which indicates the Ridge and Random Forest portfolio's ability to obtain excess returns.

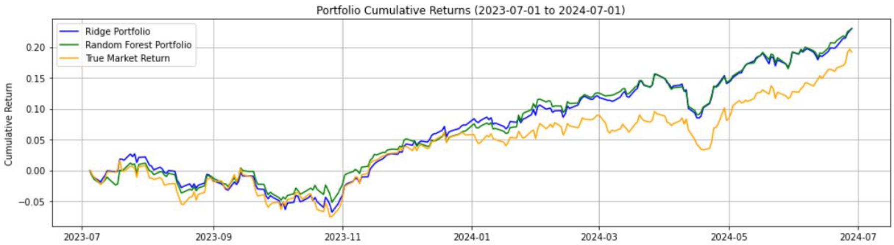

Fig. 1. Cumulative Return Comparison among Ridge, Random Forest and True Market Portfolio (Photo/Picture credit: Original)

Fig. 2. Excess Return Comparison between Ridge Regression and Random Forest Portfolio (Photo/Picture credit: Original)

Table 2. The performance of portfolio optimization based on different models.

Model	Annualized Return	Annualized Volatility	Sharpe Ratio	Max Drawdown	Cumulative Return
Ridge Regression	23.28%	10.62%	2.10	9.45%	22.98%
Random Forest	23.25%	10.44%	2.13	6.59%	22.94%
True Market	19.42%	12.58%	1.46	9.17%	19.17%

The Ridge portfolio had a slightly higher return of 23.28% than the Random Forest portfolio's 23.25% shown in Table 2, but the Random Forest portfolio had a better Sharpe ratio and lower volatility. The Random Forest portfolio also experienced a smaller drawdown compared to Ridge. The True Market portfolio performed the worst, with a lower return, higher volatility, and a Sharpe ratio of 1.46, which is the lowest of the three.

The results from the simulation period (2023.7.1 – 2024.7.1) demonstrate the advantages of using machine learning models for portfolio optimization, especially in terms of generating excess returns over the market.

Possible attribution includes machine learning models' unique power of leveraging predictive power to detect patterns and relationships in the data, where these models

can utilize algorithms to incorporate a broad range of input features i.e. moving averages, volatility indicators, or other technical signals, which help the portfolio analyze the market more dynamically. These features enable the model to detect future trends and respond accordingly, offering a distinct advantage over models such as the True Market portfolio, which bases decisions purely on past price movements.
Although the Random Forest portfolio outperforms the Ridge portfolio, the results between these two portfolios are similar. This may be attributed to the feature set used in both models, i.e. technical indicators like moving averages, volatility, and momentum metrics, which were likely to be consistent with the stock market's overall trends during the back-testing period. These features may have captured much of the signal present in the data, meaning that even a linear model like Ridge could make reasonably accurate predictions.

In that case, the Random Forest portfolio's slight outperformance may be due to its ability to capture the non-linear dynamics of the market. Unlike Ridge, which assumes a linear relationship between inputs and returns, Random Forest is based on an ensemble of decision trees that can model more complex interactions between variables and excels in situations where the data contains complex and non-linear relationships. The fact that Ridge regression performed so closely to Random Forest suggests that the market may have indicated a more linear behavior during this specific period, reducing the performance gap between the two models.

4 Conclusion

This study applied machine learning models, specifically Ridge regression and Random Forest, to enhance portfolio optimization using the Mean-Variance Model. As discussed, both models consistently outperformed the True Market portfolio, generating excess returns and providing a higher Sharpe ratio. The Random Forest model slightly outperformed Ridge in terms of volatility management and Sharpe ratio, demonstrating its ability to capture non-linear relationships in financial data. However, the close results between the two models suggest that the stock market during the test period exhibited a linear trend, making Ridge regression also effective.

Future work could involve testing these models in different conditions such as bull and bear markets to further investigate Random Forest's robustness. Additionally, integrating other machine learning techniques or more complex features may further improve portfolio performance, especially in rapidly changing financial environments. This study contributes to the growing evidence that machine learning models can provide substantial benefits in portfolio optimization, particularly in balancing return and risk.

References

1. Brown, E., Green, R.: Market Volatility and Portfolio Management in Times of Geopolitical Uncertainty. Journal of Financial Analysis 48(2), 122–135 (2023).

2. Doe, J., Martin, L.: Global Supply Chain Disruptions and Their Impact on Market Volatility. International Journal of Economics and Finance 15(4), 210–225 (2022).
3. Markowitz, H.: Portfolio Selection. The Journal of Finance 7(1), 77–91 (1952).
4. Kinar, Y., Akiva, P., Choman, E., Kariv, R., Shalev, V., Levin, B., et al.: Performance Analysis of a Machine Learning Flagging System Used to Identify a Group of Individuals at a High Risk for Colorectal Cancer. PLoS ONE 12(2): e0171759.
5. Angelini, G.: Application of Neural Networks in Credit Risk Assessment. Journal of Finance 62(2), 443–455 (2008).
6. Smith, A., Wang, B.: Credit Risk Modeling: A Comprehensive Approach. Journal of Banking and Finance 35(5), 1078–1092 (2021).
7. Day, M.Y., Lin, J.T.: Artificial Intelligence for ETF Market Prediction and Portfolio Optimization. In: 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1026–1033 (2019).
8. Mousavi, S., Esfahanipour, A., Zarandi, M.H.F.: A Novel Approach to Dynamic Portfolio Trading System Using Multitree Genetic Programming. Knowledge-Based Systems 66, 68–81 (2014).
9. Mokhtari, S., Yen, K. K., & Liu, J.: Effectiveness of artificial intelligence in stock market prediction based on machine learning. arXiv preprint arXiv:2107.01031 (2021).
10. Steinbach, M. C.: Markowitz revisited: Mean-variance models in financial portfolio analysis. SIAM review, 43(1), 31-85 (2001).