Stock Forecasting and Portfolio Optimization Based on ARIMA-GARCH, Random Forest and Monte Carlo Models

Jingyuan San

Department of Software Engineering, South China Normal University, Huizhou, 516000, China
20222034035@m.scnu.edu.cn

Abstract. Stock forecasting has long been a popular issue in financial markets. Also comparing the effectiveness of different forecasting models is enormously important. In this paper, the prediction accuracy of the traditional Autoregressive Integrated Moving Average-Generalized Autoregressive Conditional Heteroskedasticity (ARIMA-GARCH) model and the machine learning model (Random Forest) was compared. 15 high-tech stocks were selected and predicted through 2 models. This paper then optimized the stock portfolio based on the prediction results. The collected data was pre-processed to generate the model predictions and then the portfolio was optimized using the Monte Carlo algorithm. The ARIMA model used the AIC criterion to select optimal parameters. Additionally, the GARCH model was utilized in conjunction with ARIMA model in order to eliminate the effect of heteroskedasticity. While in the Random Forest model, the training was done by partitioning the training set and the test dataset. The experimental results indicated that the random forest model in machine learning provided more accurate prediction results than the ARIMA-GARCH model. Additionally, the optimized portfolio showed very substantial expected returns.

Keywords: Stock Prediction, Machine Learning, Portfolio Optimization.

1 Introduction

In the realm of finance, stocks are a fundamental component, with investors acquiring shares issued by listed companies to gain shareholder status and associated rights. Concurrently, listed companies utilize stock issuance to entice investors and secure funds for business expansion, product innovation, and service enhancement. Investors seek to optimize returns by diversifying their investment portfolios, underscoring the significance of analyzing stock return forecasts. Despite the recent surge in interest in stock return forecasting, the unprecedented volatility of the stock market has rendered accurate predictions increasingly challenging. While traditional econometric models retain their merits, many have faltered in accurately predicting stock returns. Therefore, the exploration of new models through machine learning and other methodologies with superior performance deserves more attention.

The continuous advancement of artificial intelligence technology has led to the

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_61

emergence of powerful self-learning algorithms in the field of system science. Models like decision trees, random forests, neural networks, support vector machines, and time series models have been successfully applied across various fields, yielding substantial results. In the financial sector, machine learning can find application in numerous scenarios. For example, one article focusses on Artificial Neural Networks (ANN), which is a powerful nonparametric model with the ability to recognize complex patterns and handle intricate financial data. Many studies have demonstrated that ANN surpasses traditional statistical methods in various applications. ANN can effectively address the constraints of traditional methods when dealing with nonlinear relationships and high-dimensional data. Unlike traditional methods that rely on linear assumptions, ANN can capture intricate nonlinear relationships through a multi-layer network structure [1]. In the context of these scenarios, it is imperative to prioritize research aimed at optimizing investment portfolios based on projected stock returns. In [2], the author employed the Autoregressive Integrated Moving Average (ARIMA) model to forecast the price index of the Istanbul stock market. The study highlights that the ARIMA (3,1,5) model exhibits a strong fit in the prediction process. The predicted value closely mirrors the actual value, showcasing minimal prediction error, and thus, it is suitable for economic decision-making and investment strategy development. Another article examines the use of a linear Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model to predict daily closing prices in the Nigerian stock market. The findings indicate that the GARCH (1,2) model can effectively assess the volatility of financial data, offering investors a valuable risk management tool [3]. In the realm of stock return forecasting, traditional models such as the ARIMA model are frequently employed, whereas machine learning models are less prevalent. Therefore, this study seeks to compare the predictive accuracy of traditional models with machine learning models and employ the findings to enhance investment portfolio optimization.

In this paper, 15 representative high-tech stocks were selected using yfinance, and their daily logarithmic returns were calculated. The study utilized both the traditional ARIMA-GARCH model and the random forest model in machine learning to predict stock returns. A comparison was made using the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) values to assess the prediction results of the two models. Ultimately, the predicted results were used to optimize the investment portfolio, with the analysis demonstrating that the machine learning model outperformed the traditional model in predicting stock returns.

2 Methodology

2.1 Dataset Preparation

Data Collection. This paper analyzes 15 technology stocks such as, Apple (AAPL), Microsoft (MSFT), Google (GOOGL). Data is sourced from Yahoo Finance and encompasses the period from January 2016 to July 2023. The extensive dataset will be utilized to forecast the returns of each stock and develop covariance matrices to understand the interrelationships between these stocks. After this, the data will be employed to optimize portfolio allocations, ensuring that the investment strategy is effective in maximizing returns while also being cautious in managing risk. The daily closing price of each stock was designated as the target variable and used in the subsequent return calculations.

Data Processing. In this paper, the traditional ARIMA-GARCH model is compared with a machine learning model (i.e. the random forest), for which different data processing efforts are performed for the two prediction approaches.

For the ARIMA-GARCH model, after confirming that there are no missing values in the time series data, this paper cleans the data and adds frequencies to the returns. This step is critical for accurate model training and forecasting. The daily logarithmic return for each stock in the dataset was calculated after that. Logarithmic returns are preferred in financial modeling because they simplify the mathematics involved in continuous compounding calculations and provide a more accurate picture of investment performance over time.

For the random forest (machine learning) model, the corresponding features and target values for each stock are first defined. The features include 20-day moving average, 50-day moving average and volatility. The target value is the logarithmic return of each stock. Missing values in the dataset are processed through a function to ensure data integrity. Before data processing, the dataset is checked to see if it contains all the necessary features and target columns, and if there are any missing, the data processing for that stock is skipped. The feature matrix contains the technical indicators of the stock and the target variable is the corresponding logarithmic return. Finally, the dataset is divided into training and test sets for subsequent model training and performance evaluation.

2.2 Stock Return Prediction

This study utilized the ARIMA-GARCH model and the Random Forest model in machine learning for forecasting stock returns. It also provided commentary on the forecast results using RMSE and MAPE.

ARIMA-GARCH Model. The ARIMA model, widely recognized in finance for its forecasting capabilities, is a fundamental approach for analyzing time series data. It comprises three key components: Autoregressive (AR), Differencing (I), and Moving Average (MA) [4, 5]. The AR component examines how past values influence the present, while differencing helps to make a non-stationary series stationary by removing trends. The MA part focuses on the relationship between past forecast errors and current values. However, in scenarios where the data exhibits volatility clustering—where periods of high variance are followed by higher variance, and low variance by low variance—the ARIMA model alone may not perform adequately. Such a situation involves conditional variance, which can be captured by extending the ARIMA model with a Generalized Autoregressive Conditional Heteroskedasticity (GARCH) framework. This combination allows the model to account for changing variance over time, which is common in financial data. Therefore, in most cases, the GARCH model is used alongside the ARIMA model to capture data volatility and enhance model prediction accuracy.

In this paper, the time series data undergo a rigorous smoothness test to ensure their stability. Subsequently, optimal parameters are determined using the AIC quasi-test to construct the most appropriate model. The residuals obtained from the ARIMA model are then used as parameters in the GARCH model to capture the data's volatility.

Random Forest Model. Random forest modeling is an integrated learning method in machine learning principally used for classification and regression tasks. It is a model composed of multiple decision trees that enhance overall model performance and stability by combining the predictions of these decision trees. The Random Forest Model was proposed by Leo Breiman in 2001 [6]. The basic idea is to improve the generalization ability of the model by combining multiple weak learners (usually decision trees) to create one strong learner. The random forest model is founded on ensemble learning and the principle of 'crowdsourcing'. This method entails creating multiple decision tree models and merging their outcomes through voting (for classification) or averaging (for regression) to generate more dependable predictions. Furthermore, during the construction of each tree, the random forest algorithm randomly selects a subset of features at each node split instead of considering all of them, which serves to decrease the model's variance and mitigate overfitting [7, 8].

In this study, moving average and volatility features were generated through logarithmic returns calculations using stock data obtained from Yahoo Finance. The dataset was divided into a training set and a test set. Following this, a Random Forest model was trained and evaluated to predict stock returns.

2.3 Portfolio Optimization

The Monte Carlo Method (MCM) is a set of statistical techniques used for numerical computation through stochastic sampling [9, 10]. It is extensively applied in physics, finance, engineering, and computer science. The fundamental concept involves estimating the properties of a phenomenon or system by conducting a high volume of stochastic simulations, particularly when traditional analytical methods are impractical or ineffective due to the complexity of the problem.

This study aims to achieve the highest possible return for a given level of risk or to minimize the risk for a given level of expected return by selecting the best mix of financial assets through portfolio optimization and Monte Carlo methods. Initially, backtesting is used to evaluate the performance of the assets from July 2023 to July 2024. Despite occasional volatility and minor drawdowns, the strategy demonstrates relatively stable growth, indicating strong risk management capabilities and resilience to market fluctuations. Monte Carlo simulations are then employed to calculate the efficient frontier, resulting in the selection two representative portfolios: one with the highest Sharpe ratio and another with the lowest volatility.

3 Results and Discussion

3.1 Stock Forecast

As illustrated by Table 1, the Random Forest model surpasses the Arima-Garch model with significantly smaller RMSE and MAPE values, indicating lower error rates. The Random Forest model excels at capturing intricate nonlinear relationships in the data, making it more suitable for the highly nonlinear stock market. Moreover, it automatically performs feature selection during training, enhancing accuracy by identifying the most impactful variables. In contrast, the ARIMA-GARCH model is better suited for linear time series and may not effectively capture nonlinear features. It also requires manual selection of lag terms and volatility deconstruction, potentially limiting its effectiveness in stock forecasting. Additionally, the Random Forest model demonstrates greater resilience to noise and overfitting through the combination of multiple decision trees, while the ARIMA-GARCH model is sensitive to outliers and model settings, raising the risk of inaccurate predictions.

Table 1. The performance of two models in stock prediction

Model	Arima-Garch		Random forest	
Forecast	RMSE	MAPE	RMSE	MAPE
AAPL	0.0229	4.7508	0.0198	2.3976
ADBE	0.0296	5.2943	0.0179	3.1211
AMD	0.0464	75.7199	0.0181	2.8197
AMZN	0.0255	216.4061	0.0211	3.2721
CRM	0.0315	6.2891	0.0367	2.4004
CSCO	0.0204	85.8538	0.0377	2.835
GOOGL	0.0238	3.6431	0.0294	3.6989
IBM	0.0201	6.0165	0.0181	1.8276
INTC	0.0303	27.1481	0.0224	2.3514
MSFT	0.0213	9.9739	0.0199	1.6011
NVDA	0.0421	27.7751	0.0207	2.9589
ORCM	0.0275	68.8097	0.0212	2.6064
QCOM	0.0336	6.5784	0.0184	2.2943
TSLA	0.0522	7.5313	0.0261	1.882
TXN	0.0232	8.2783	0.0203	1.8768

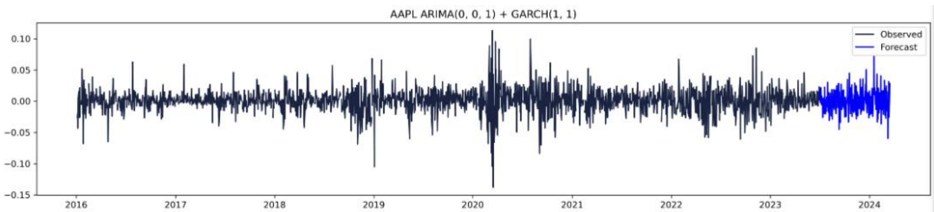

Fig. 1. Prediction of Apple with Arima-Garch model (Photo/Picture credit: Original)

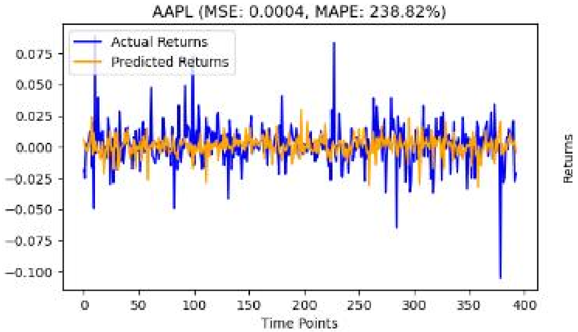

Fig. 2. Prediction of Apple with Random Forest model (Photo/Picture credit: Original)

The charts illustrate the prediction results for Apple using the Arima-Garch model and the Random Forest model in Fig. 1 and Fig. 2, respectively. The RMSE and MAPE values suggest that the Random Forest model outperforms the Arima-Garch model. This could be due to information loss or outliers during data preprocessing and differencing in the time series, or it may be attributed to suboptimal parameter selection. Although the Random Forest model provides better predictions in terms of RMSE and MAPE, it still struggles to accurately capture the volatility of the original data. Factors such as macroeconomic policies, unforeseen events, and other external market influences are challenging to predict and significantly contribute to stock market volatility, and capturing these volatilities is exactly what the random forest model lacks.

3.2 Portfolio Optimization

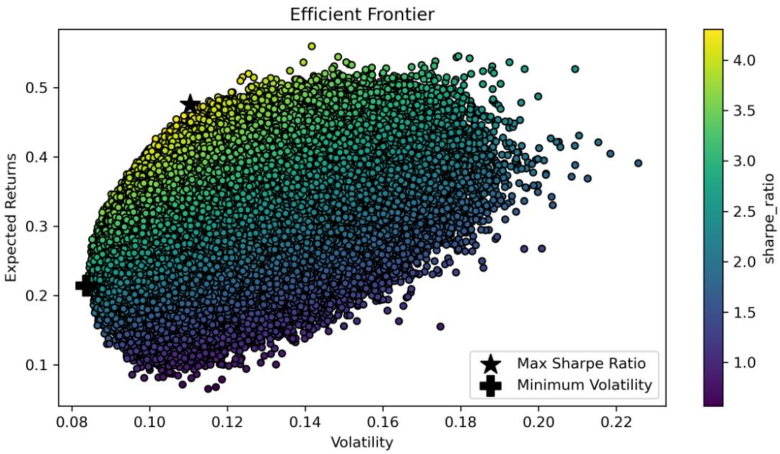

Fig. 3. Monte Carlo simulation (Photo/Picture credit: Original)

Through Monte Carlo simulation, 10 million portfolio weights are generated and visualized in Fig. 3. The leftmost boundary of the shape is the efficient frontier, representing optimal portfolios when considering about sharp ratio and volatility. This study picks two representative portfolios —— maximum Sharpe ratios and minimum volatilit .y

Table 2. Largest sharp ratio outcome

Time period	
Start date	2023/7/3
End date	2024/7/15
Total months	12
Backtest	
Annual return	57.54%
Cumulative returns	59.83%
Annual volatility	10.87%
Sharpe ratio	4.24
Calmar ratio	12.67
Stability	0.88
Max drawdown	-4.54%
Omega ratio	1.95
Sortino ratio	7.48
Skew	0
Kurtosis	-0.14
Tail ratio	1.31
Daily value at risk	-1.19%

Backtesting the portfolio with the highest Sharpe ratio showed a cumulative return of 59.83% shown in Table 2, a Sharpe ratio of 4.24, and an annual volatility of 10.87%. It is remarkable that sharpe ratio is decreasing constantly, which may imply that the portfolio is becoming less effective.

Fig. 4. Rolling Sharpe ratio(6-month) (Photo/Picture credit: Original)

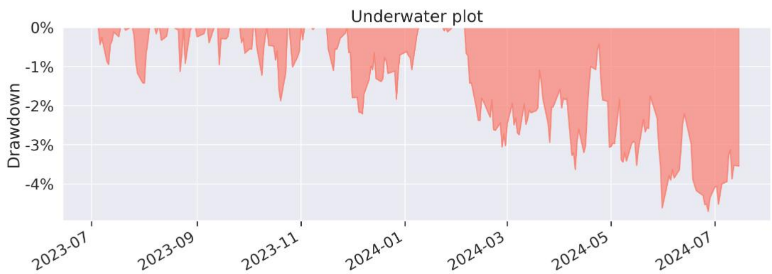

Fig. 5. Underwater plot (Photo/Picture credit: Original)

Through backtesting the portfolio with the lowest volatility, this study found that the annual volatility slightly dropped to 8.239%, decreased by approximately 2.5%, but the cumulative return and Sharpe ratio both decrease by half. From the previous graph and the outcome at this page, it can be inferred that there is a marginal effect: Sacrificing returns does not achieve the expected low volatility. Also, Fig. 4 and Fig. 5 show that the Sharpe ratio is continuously decreasing, supporting the conclusion from the previous paragraph.

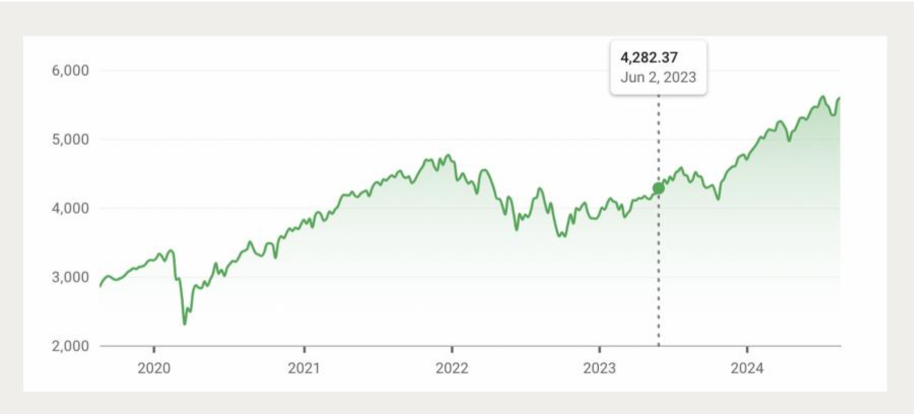

Fig. 6 S&P Return: Jun 2, 2023 (Photo/Picture credit: Original)

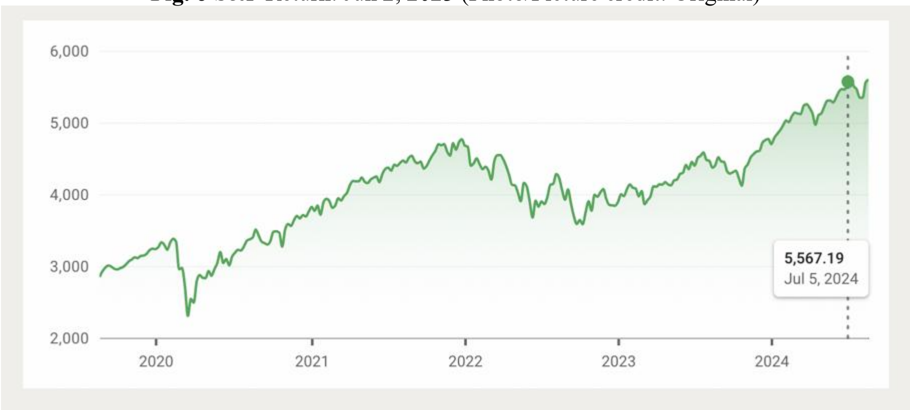

Fig. 7 S&P Return: Jul 5, 2024 (Photo/Picture credit: Original)

Based on calculations through Fig. 6 and Fig. 7, the S&P 500 index has increased by 30%. It is 8% higher than the returns of the lowest volatility portfolios, but nearly just half of that of the largest Sharpe ratio portfolios. This is due to the marginal effect mentioned earlier. This leads to the conclusion that optimal portfolios have relatively considerable returns.

4 Conclusion

This paper utilizes an ARIMA-GARCH model and a machine learning model to predict stock returns and optimize a stock portfolio based on the predictions. The ARIMA-GARCH model is tested by using the residuals from the ARIMA model as parameters for the GARCH model to complete the fitting. The Random Forest model is chosen for prediction in the machine learning model, and its prediction accuracy is found to be higher than that of the ARIMA-GARCH model based on the RMSE and MAPE tests. Lastly, the Monte Carlo algorithm is employed to optimize the stock portfolio, resulting in a portfolio with a significantly higher expected return.

However, the study has several limitations, including constraints in model selection and reliance on data. While the ARIMA-GARCH model is suitable for linear and conditionally heteroskedastic data, it may not perform well for nonlinear relationships. Similarly, the Random Forest model, reliant on feature selection, fails to capture the dynamic nature of time series effectively. Therefore, it would be prudent to consider integrating more advanced machine learning models such as LSTM and XGBoost to improve prediction accuracy in future studies.

References

1. Min, Q.: Statistical handbook: Statistical methods in finance, Chapter 18. Publisher, Location (2003).
2. Isenahd, G. M., Olubusoye, O. E.: Forecasting Nigerian stock market returns using ARIMA and artificial neural network models. CBN Journal of Applied Statistics 5(2), 25–48 (2014).
3. Arowolo, W. B.: Predicting stock prices returns using GARCH model. The International Journal of Engineering and Science 2, 32–37 (2013).
4. Shumway, R. H., Stoffer, D. S.: ARIMA models. In: Time series analysis and its applications: with R examples, pp. 75–163. Springer, New York (2017).
5. Kalpakis, K., Gada, D., Puttagunta, V.: Distance measures for effective clustering of ARIMA time-series. In: Proceedings 2001 IEEE International Conference on Data Mining, pp. 273–280. IEEE, Los Alamitos (2001).
6. Leo, B.: Random forests. Machine Learning 45, 5–32 (2001).
7. Rigatti, S. J.: Random forest. Journal of Insurance Medicine 47(1), 31–39 (2017).
8. Biau, G., Scornet, E.: A random forest guided tour. Test 25, 197–227 (2016).
9. Kroese, D. P., Brereton, T., Taimre, T., Botev, Z. I.: Why the Monte Carlo method is so important today. Wiley Interdisciplinary Reviews: Computational Statistics 6(6), 386–392 (2014).
10. Rubinstein, R. Y., Kroese, D. P.: Simulation and the Monte Carlo method. John Wiley & Sons (2016).