A Comparative Analysis of Portfolio Optimization: LSTM vs. Random Forests

Jinyuan Sun

International School, Beijing University of Posts and Telecommunications, Beijing, China

jinyuan@bupt.edu.cn

Abstract. Portfolios represent a structured framework aiming to maximize returns while simultaneously minimizing risks, with applications spanning various domains, including stock investments, production optimization, and engineering models. Portfolio optimization remains a perpetual subject of interest within the realm of finance, and recent advancements in deep learning techniques provide a novel perspective for its exploration. This essay focuses on a selection of healthcare-related stocks, namely LLY, PGR, CI, and UNH, within the timeframe of May 10, 2019, to October 16, 2023. Initially, this study involves training Long Short-Term Memory (LSTM) and Random Forest models on data up to September 5th, employing a 30-day shifting window to capture stock trends. Subsequently, a portfolio strategy is employed for forecasting results, yielding both a model maximum Sharpe ratio and another minimizing risk along the efficient frontier of Monte Carlo simulations. By employing these two portfolio models and comparing their results with the performance of the NASDAQ over the same period, this essay concludes that both models outperform the NASDAQ and the LSTM models in this research demonstrate superior performance relative to Random Forest models.

Keywords: Mean-variance model; portfolio management; LSTM model; Random Forest.

1. Introduction

The selection of a project portfolio holds significant importance for numerous firms, and this significance extends to investors as well [1]. A well-diversified portfolio offers investors the potential for increased yields while maintaining an equivalent level of risk [2]. Markowitz's classical mean-variance (MV) theory, introduced in 1952, has long stood as the cornerstone of modern portfolio theory. Markowitz's pioneering work established a rational framework for making portfolio management decisions by quantifying the trade-off between risk and return. Over time, extensive research has been undertaken to develop model variations that better adapt to real-world conditions [3-5].

In the realm of applied portfolio research, diversification has long been regarded as a fundamental strategy for risk mitigation. Recent financial sector events, such as the GameStop short squeeze in early 2021 and the market turmoil triggered by the pandemic, have underscored the critical role of diversification in managing unforeseen risks and optimizing portfolio returns [6]. Diversifying across a range of assets and sectors has been a key tenet in traditional portfolio management.

However, in the contemporary financial landscape, the challenge goes beyond diversification alone. Accurately predicting future returns while managing risk is the quintessential task of portfolio optimization. This is where machine learning techniques have begun to shine. The advent of cutting-edge AI models, particularly the combination with LSTM networks, offers a promising avenue to tackle the ever-present challenges of unpredictability and volatility [7, 8].

Notably, a combination of the Mean-Variance (MV) framework with machine learning methodologies has shown impressive performance in contrast to traditional approaches [9]. Yet, despite the potential benefits, there remains a significant research gap in exploring the full capabilities of machine learning in portfolio optimization. While LSTM has demonstrated its effectiveness in stock price prediction [10-11], the application of LSTM and other machine learning models in the portfolio space has not received the attention it deserves.

One study that closely aligns with our research is the work of Fischer and Krauss, who employed LSTM networks to manage portfolios composed of S&P 500 constituents. This study included a comparative analysis with memoryless machine learning models such as logistic regression and random forests [12]. However, it's worth noting that the data and methodologies in their study are outdated, highlighting the need for more contemporary investigations in this evolving domain.

In essence, the financial landscape is evolving rapidly, and the time is ripe to bridge the gap between traditional portfolio management and advanced machine learning techniques. The potential for enhancing portfolio allocation strategies, optimizing risk-adjusted returns, and addressing the modern challenges of the financial world remains a frontier that warrants further exploration. Machine learning holds the key to this uncharted territory, and the research community must rise to meet this exciting challenge.

The main goal of this study is to ascertain the superior approach, whether it is more prudent to leverage the cutting-edge LSTM model or resort to the conventional Random Forest, a traditional machine-learning technique. The intention is to enhance the accuracy of predictions regarding future returns and covariance, thereby enabling more effective strategies for asset allocation.

To achieve this objective, the study initiates by selecting a diverse set of 100 sector-specific stocks from the latest constituents of the S&P 500 index. Subsequently, historical stock price data spanning 1117 days is employed to train both the LSTM neural network and the Random Forest model, with a shifting window of 30 days for consistent training. These models are then equipped to forecast the stock prices for the forthcoming day, utilizing data from the preceding 30 days. The projected values are converted into percentage changes, representing daily returns, and a shrinkage method is applied to compute the covariance matrix. The Mean-Variance optimization method is then used to find the ideal portfolio weights for every day. The study iterates this process for the following 30 days, continuously adjusting the portfolio weights based on the most recent predictions. Ultimately, after the 30-day period elapses, the cumulative portfolio returns are computed and compared against the performance of the NASDAQ index, serving as a benchmark.

The subsequent sections of this study are organized as follows: In section 2 the data utilized for this study are presented, the stock selection procedure is explained, and a detailed summary of the selected stocks is given. In Section 3, a comprehensive exposition of the methodologies applied in this research is presented, encompassing LSTM and random forests method. Section 4 involves a comparative analysis of the effectiveness of these two approaches, juxtaposed with a benchmark asset. Finally, Section 5 offers the concluding remarks for this paper and explores possible avenues for further investigation.

2. Data and pre-processing

The daily stock data utilized in this research has been sourced from Kaggle, a well-regarded data repository (https://www.kaggle.com/datasets/svaningelgem/nyse-100-daily-stock-prices) and is provided by the New York Stock Exchange (NYSE). The process of stock selection is grounded in the assessment of average daily returns during the period spanning from May 10, 2019, to October 16, 2023. The selection criteria involve ranking stocks in descending order based on their performance,

and the top 100 performing stocks are chosen for training the LSTM and random forest models. Finally, the paper ultimately narrows its focus to four stocks within the healthcare and insurance sectors, specifically represented by LLY, PGR, CI, and UNH, as depicted in Table 1. Besides, the distribution of returns for the four stocks is shown in Figure 1.

Table 1. Selected stocks

Stock Symbol	Company
LLY	Eli Lilly and Company
PGR	The Progressive Corporation
CI	The Cigna Group
UNH	UnitedHealth Group Incorporated

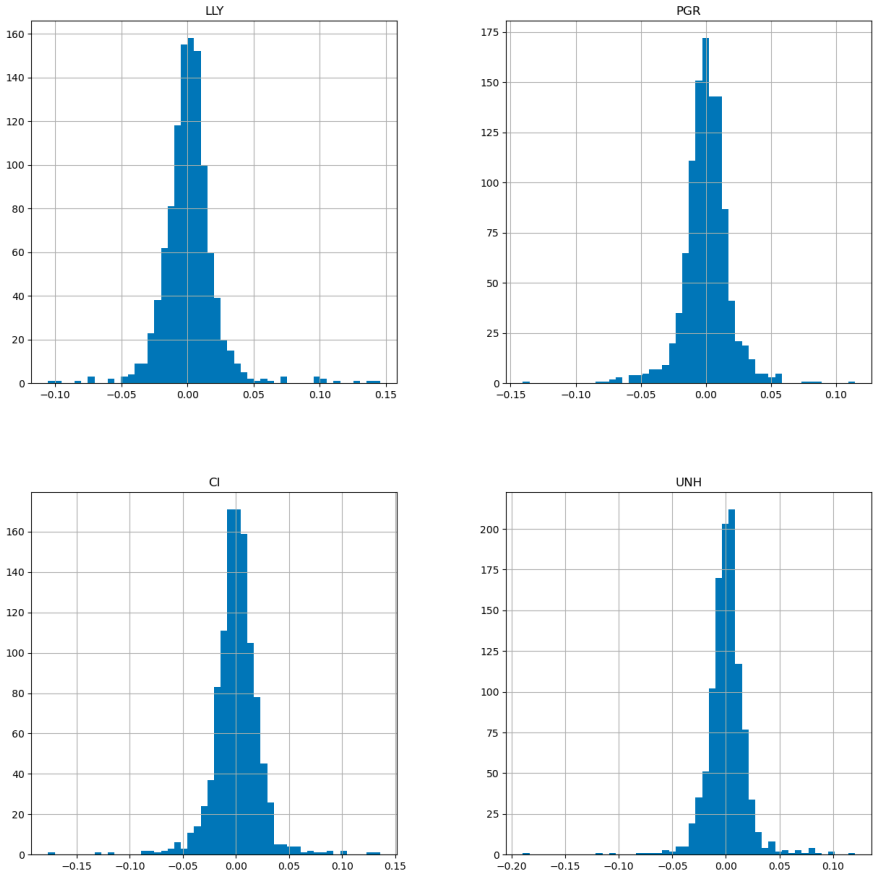

Figure 1. Distribution of returns for the four stocks

This study examines a period of 1147 market days, from May 10, 2019, to October 16, 2023. The model is trained using the first 1117 days of this period, from May 10, 2019, to September 1, 2023. The portfolio's performance is then evaluated during the remaining 30 market days, from September 1, 2023, to October 16, 202. This article opts for 'Adj Close' as the daily price and computes the computes the daily returns of the i-th particular asset R_{it} on day t as follows.

$$R_{it} = \log\left(\frac{V_{t+1}^{i}}{V_{t}^{i}}\right) \tag{1}$$

3. Method

The research methodology comprises four distinct phases. To commence, the study undertakes the selection of the top 100 stocks from all stocks in the New York Stock Exchange (NYSE). Subsequently, the research involves the training of a LSTM network and a traditional algorithm based

on random forest. This training employs historical stock price data spanning 1117 days, enabling the models to forecast the following day's stock prices with a shifting window of 30 days. Concurrently, the study employs the shrinkage method to estimate the covariance among the selected assets. In the third stage, the study calculates the ideal portfolio weights on a daily basis using the mean-variance optimization approach. Subsequently, the study proceeds to contrast portfolio returns through two distinct approaches for predicting stock prices. These returns are juxtaposed with the performance of a NASDAQ benchmark.

3.1. Monte Carlo model

The Monte Carlo model, a powerful and versatile computational technique, has found wide-ranging applications across diverse fields, with particular significance in the realm of finance [13]. Named after the renowned Monte Carlo Casino, this method is predicated on random sampling and statistical analysis to solve complex problems. In finance, it plays a pivotal role in risk assessment, option pricing, and portfolio optimization. By simulating numerous possible outcomes based on probabilistic inputs, the Monte Carlo model provides insights into uncertain financial scenarios, enabling investors and decision-makers to make informed choices in the face of uncertainty.

3.2. Mean-Variance Model

The Modern Portfolio Theory (MPT), pioneered by Harry Markowitz in 1952, represents a seminal framework in the domain of portfolio optimization. Central to this approach is the concept of Mean-Variance Optimization (MVO), a method designed to assist investors in fine-tuning their portfolios by striking an equilibrium between risk and return. MVO serves as the fundamental tool within MPT, guiding investors in making strategic asset allocation decisions that can maximize their investment returns while prudently managing the associated risks.

Let w_i be the weight of the i-th asset such that $\sum_i w_i = 1$ and μ_i be the expected return of the i-th asset. The portfolio's expected returns can be calculated by the following equation.

$$\mu_p = \sum_i w_i \mu_i \tag{2}$$

Within the Mean-Variance Optimization (MVO) framework, risk is conventionally symbolized by the variance of the portfolio's returns. This variance, in turn, encapsulates the collective impact of two crucial components. Firstly, it encompasses the weighted variances of each individual asset held within the portfolio. Secondly, it factors in the weighted covariances between every possible pair of assets in the portfolio. This comprehensive evaluation of variance serves as a fundamental risk metric in MVO, enabling investors to gauge and manage the overall risk exposure in their portfolios by considering the interplay of individual asset volatilities and their correlations.

$$Variance: \sigma_P^2 = \text{var}\left(\sum_i w_i r_i\right) = \sum_{ij} w_i w_j \text{cov}(r_i r_j) \tag{3}$$

Within the Mean-Variance (MV) model, the focus is directed towards two compelling portfolio constructs: the portfolios with the lowest volatility and the highest Sharpe ratio [13]. The minimum volatility portfolio, characterized by q=0, represents an investment choice reflective of a staunch aversion to risk. On the other hand, the Sharpe ratio, a widely embraced metric for appraising risk-adjusted returns, assumes a pivotal role in this model. It quantifies the trade-off between investment risk and potential returns, offering a crucial gauge for investors and analysts in assessing portfolio performance and optimization. This essay delves into the dynamics of these two intriguing portfolios within the framework of the MV model, elucidating their significance in the context of financial decision-making.

$$Sharpe\ ratio = \frac{R_p - R_f}{\sigma_p} \tag{4}$$

Where the current risk-free rate is denoted as R_f. Within this research, the focal point is the maximum Sharpe ratio portfolio, chosen as the target portfolio for in-depth analysis.

3.3. LSTM

LSTM, short for Long Short-Term Memory, represents a cutting-edge prediction method in the realm of artificial intelligence. It operates as a specialized type of artificial neural network, frequently employed in the domain of deep learning to process and analyze sequential data. The LSTM's adeptness in handling long-term dependencies, coupled with its enhanced performance relative to conventional Recurrent Neural Networks (RNNs), has propelled its widespread adoption across diverse domains. Three essential parts—the input gate, the forget gate, and the output gate—are what make LSTMs so successful. These gates are an integral part of the architecture of the LSTM; they act as gatekeepers, carefully controlling the information flow through the memory cell. The introduction of these gates empowers the LSTM model to make selective decisions regarding information retention or dismissal within the memory cell. This selectiveness hinges on a dynamic interplay between the model's inputs and the current state of the cell. Consequently, the LSTM model is equipped to retain and recall valuable information while discarding less relevant data, contributing to its superior performance in capturing and processing sequential patterns. In Figure 2, it can be observed that the architectural layout of the LSTM model and its associated modules. Notably, every cell is essential to the sequential transmission of both the hidden state and the cell state from one cell to the next.

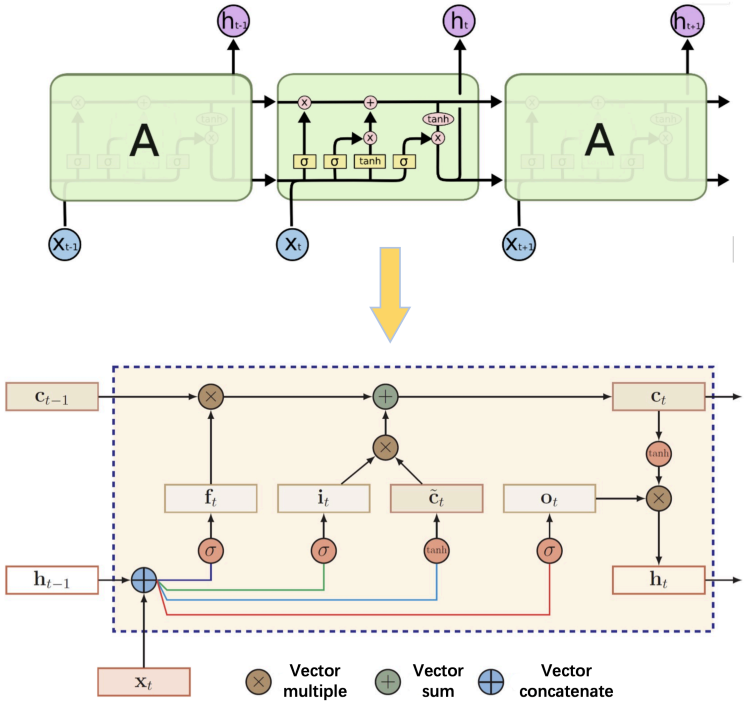

Figure 2. Structure of LSTM [8]

The machine learning method employed in this paper leverages the TensorFlow Sequential API. It consists of two LSTM layers, each comprising 100 units, and three dense layers. These dense layers are configured with 100, 50, and 1 connection, respectively. The inclusion of two LSTM layers is instrumental in facilitating the model's proficiency in capturing latent long-term relationships inherent in the data. Concurrently, the dense layers are responsible for processing non-time-series attributes within the dataset and mapping the outputs obtained from the LSTM layers into the desired output format. This holistic architecture ensures the model's ability to effectively capture both temporal dependencies and non-time-series characteristics within the data, enhancing its predictive capacity.

3.4. Random Forest

For classification and regression applications, Random Forest is a potent ensemble learning technique that is often used in data science and machine learning. It derives its strength from the combination of multiple decision trees, providing robustness and reducing overfitting. The algorithm's core idea is to aggregate the predictions from numerous individual decision trees to enhance the overall accuracy and generalization performance. The decision trees in a Random Forest are constructed using a process known as bootstrapped sampling, which involves randomly selecting subsets of the training data with replacement. Each tree is trained independently, leading to a diverse set of classifiers. Additionally, Random Forest introduces feature selection by randomly considering only a subset of features at each node split, which contributes to the model's robustness against noisy or irrelevant attributes. For problems involving classification, the Random Forest's final prediction is decided by a majority vote; for tasks involving regression, it is determined by an average. Mathematically, this can be represented as follows for a regression task.

$$\hat{y}(x) = \frac{1}{n}\sum_{i=1}^{n} Ti(x) \tag{5}$$

Where $\hat{y}(x)$ represents the predicted output, and Ti(x) represents the output of the i-th decision tree for input data point x. Random Forest's robustness, scalability, and capability to handle high-dimensional datasets make it a popular choice for various real-world applications, such as classification, regression, and feature importance ranking.

4. Results

In the pursuit of optimizing the LSTM model, an exhaustive exploration of various hyperparameters and network settings was conducted. The rigorous investigation unveiled the configuration that yielded the model's optimal performance. Notably, this superior performance was achieved when Time_Step was set to 8, LSTM_units to 50, Batch_Size to 64, and the model underwent training for a duration of 40 epochs.

Conversely, when it comes to the Random Forest algorithm, it was meticulously trained with a total of 1000 estimators. This approach involved the utilization of the preceding 30 days' closing prices as input to predict the subsequent day's closing price. Subsequently, the predicted values were employed to compute the daily return rate. The cumulative return over a 30-day span was then derived through the aggregation of these daily return rates. The outcomes of these analyses are presented below in Figure 3, shedding light on the comparative performance of these models.

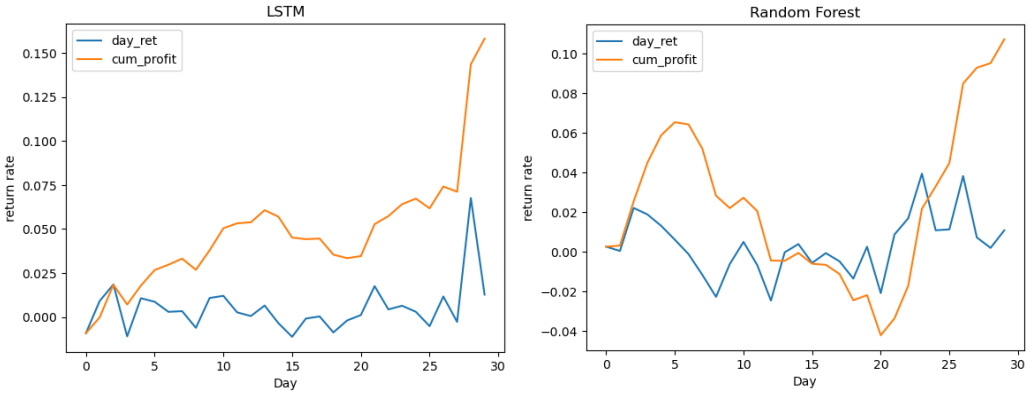

Figure 3. The daily return and cumulative return of Random Forest against LSTM.

The research uses the historical returns of the NASDAQ index as a benchmark during the test period to evaluate each model's performance. Subsequently, the research conducts an ex-post analysis to discern the actual returns generated by each portfolio. Encapsulating this evaluation is an annualized "tear sheet," a comprehensive summary displayed in the following graph. This visual representation delineates the cumulative returns achieved through two distinct methods, alongside the

performance of the NASDAQ index, offering a clear and illustrative comparison of the portfolio outcomes (See Figure 4).

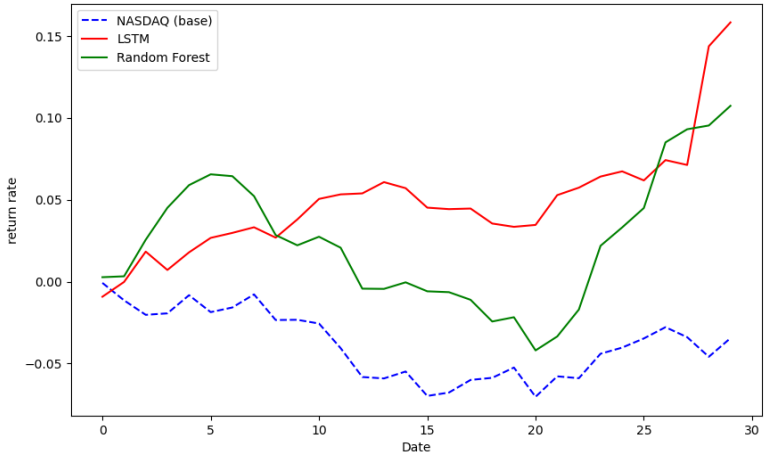

Figure 4. The cumulative return rates of NASDAQ, LSTM, and Random Forest.

5. Conclusion

Portfolio optimization represents both a well-established and widely embraced model in finance, driven by the dual objectives of maximizing returns while concurrently minimizing risks. This study specifically focuses on four prominent healthcare-related stocks—LLY, PGR, CI, and UNH—during the period spanning May 10, 2019, to October 16, 2023.

The research methodology commences with the utilization of the LSTM model and the algorithm of random forests to ingest and analyze the initial 1117 days of historical data, subsequently generating forecasts for the subsequent 30 days. Employing a shifting window of 30 days, these predictive models facilitate a comprehensive understanding of stock price trends. Subsequently, by applying portfolio strategies to these forecasting outcomes, the paper derives two critical models: one emphasizing maximum returns and the other prioritizing risk minimization, both located on the effective frontier within the context of Monte Carlo simulations.

Ultimately, these two models are integrated into portfolio construction, and their outcomes are contrasted with respect to the performance of the NASDAQ index within the same timeframe. This comprehensive analysis concludes that the LSTM model demonstrates superior predictive accuracy compared to the random forest model. Furthermore, both models outperform the NASDAQ, making them valuable tools for informed investment decisions within the healthcare-related stocks context.

References

[1] Archer, N. P., & Ghasemzadeh, F. (1996). Project portfolio selection techniques: a review and a suggested integrated approach. Available at: https://macsphere.mcmaster.ca/handle/11375/5415

[2] ASLAN, Y., & ÖZKAN, Ö. (Eds.). (2022). Innovative Approaches To Accounting, Finance And Auditing-4. Efe Akademi Yayınları.

[3] Kalayci, C. B., Ertenlice, O., & Akbay, M. A. (2019). A comprehensive review of deterministic models and applications for mean-variance portfolio optimization. Expert Systems with Applications, 125, 345–368.

[4] Chaweewanchon, A., & Chaysiri, R. (2022). Markowitz Mean-Variance Portfolio Optimization with Predictive Stock Selection Using Machine Learning. International Journal of Financial Studies, 10(3), 64.

[5] Jensen, M. C. (1978). Some anomalous evidence regarding market efficiency. Journal of Financial Economics, 6(2), 95–101.

[6] Zhang, D., Hu, M., & Ji, Q. (2020). Financial markets under the global pandemic of COVID-19.

Finance Research Letters, 36, 101528.
[7] Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. The North American Journal of Economics and Finance, 47, 552–567.
[8] Dixon, Matthew F., Igor Halperin, and Paul Bilokon. (2020). Machine Learning in Finance. Berlin and Heidelberg: Springer International Publishing.
[9] Chen, W., Zhang, H., Mehlawat, M. K., & Jia, L. (2021). Mean–variance portfolio optimization using machine learning-based stock price prediction. Applied Soft Computing, 100, 106943.
[10] Roondiwala, M., Patel, H., & Varma, S. (2017). Predicting stock prices using LSTM. International Journal of Science and Research (IJSR), 6(4), 1754-1756.
[11] Ma, Y., Han, R., & Wang, W. (2021). Portfolio optimization with return prediction using deep learning and machine learning. Expert Systems with Applications, 165, 113973.
[12] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654–669.
[13] Prahl, S. A. (1989, January). A Monte Carlo model of light propagation in tissue. In Dosimetry of laser radiation in medicine and biology (Vol. 10305, pp. 105-114). SPIE.