**Report for portfolio optimization and forecasting**

# Table of content

[Table of content 1](#_Toc184062060)

[1. Introduction 2](#_Toc184062061)

[2. Data Source 2](#_Toc184062062)

[1.1 Ethics and Privacy 2](#_Toc184062063)

[1.2 Reliability 2](#_Toc184062064)

[1.3 Integrity 3](#_Toc184062065)

[1.4 Validity 3](#_Toc184062066)

[1.1 Data Dates 3](#_Toc184062067)

[1.2 Nature of Data 3](#_Toc184062068)

[1.3 Data Structure 3](#_Toc184062069)

[4. Data Deficiencies 3](#_Toc184062070)

[5.Analytics approaches 3](#_Toc184062071)

[**5.1 Step 1**: Select the Top 20 Stocks by Market Capitalization 3](#_Toc184062072)

[**5.2 Step 2**: Constructing the Portfolio 4](#_Toc184062073)

[**5.3 Step 3**: Forecasting with LSTM and use predicted prize optimize portfolio 7](#_Toc184062074)

[**5.4 Step 4**: Identifying Trading Opportunities 12](#_Toc184062075)

[5. Conclusion 14](#_Toc184062076)

[6. Personal Reflection 15](#_Toc184062077)

[7. Reference 15](#_Toc184062078)

# 1. Introduction

Financial markets are always characterized by non-linear relationships and huge noise, which makes the analysis challenging. This project builds a portfolio of the top 20 stocks by market capitalization, optimizing the prediction and identifying buy and sell points using machine learning and technical indicators. In this project: data from Yahoo Finance and Wikipedia (Wikipedia,2024) are used to screen these stocks, and the mean-variance model (Markowitz, 1967) is used to construct the portfolio, balancing risk and return. The Long Short-Term Memory (LSTM) model (Hochreiter & Schmidhuber, 1997) was used to predict future stock prices, and then portfolio optimization was performed to maximize return and minimize risk by identifying trading points through technical indicators such as Moving Average Convergence / Divergence (MACD) (Appel,1979).

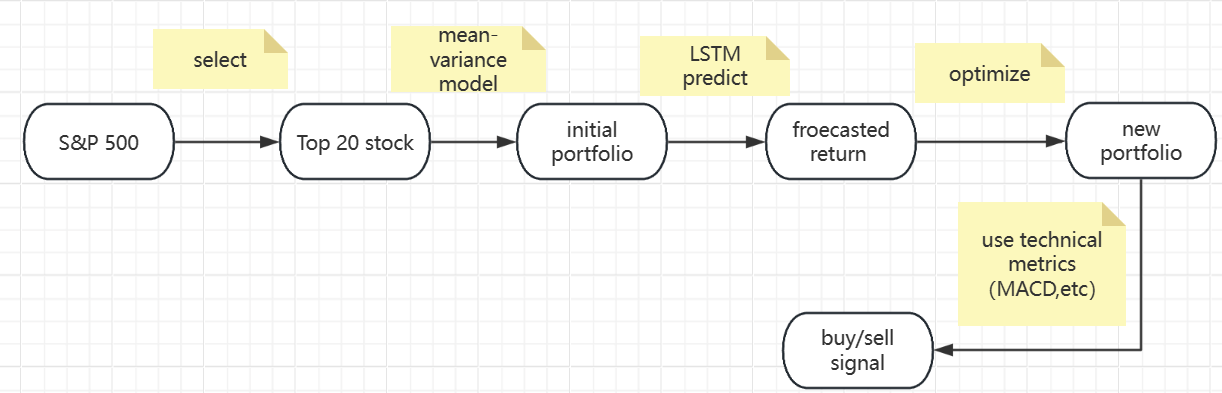


Fig 1. Flow chart of this project

# 2. Data Source

The data is sourced from Yahoo Finance (using the yfinance Python library) and downloaded from Wikipedia .



Fig2. yfinance Python library

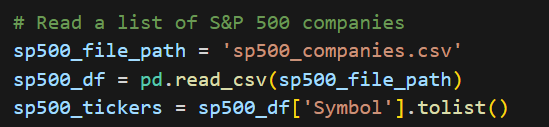


Fig 3. Dataset downloaded from Wikipedia

1.1 Ethics and Privacy: The data is publicly available, related to market information, and does not contain Personally Identifiable Information (PII) (Schwartz,2011). It is used responsibly for research purposes only to create optimized portfolios. No privacy concerns or anonymization are required.

1.2 Reliability: Yahoo Finance data is from official exchanges and financial providers, generally reliable for major stocks and indices. The analysis uses historical data, so delays are not an issue.

1.3 Integrity: Missing values in the historical data are handled by dropping NaN entries during data cleaning. After the top 20 stocks are selected th Front and Back Fill is used.

1.4 Validity: Financial data is affected by market volatility, which impacts forecast validity. Machine learning is used for time series forecasting instead of traditional methods.

**3.** Data Characteristics

1.1 Data Dates: The data spans from 1 January 2013 to 31 December 2023, covering different market conditions. Five years of data is used for neural network forecasting.

文本

描述已自动生成

Fig 4. Data date

1.2 Nature of Data: The data is time series and consists of adjusted closing prices for multiple stocks. Daily returns are calculated as input features for optimization and forecasting.

1.3 Data Structure: The data is stored in a pandas DataFrame with stock symbols as columns and dates as rows.

# 4. Data Deficiencies

Missing values are handled using forward and backward filling methods, and removing rows/columns with all null values. Stocks with less than 1000 trading days are filtered out. However, These processes may cause problems due to market instability.

# 5.Analytics approaches

## **5.1 Step 1**: Select the Top 20 Stocks by Market Capitalization

The process involved selecting the top 20 companies by market capitalization from the S&P 500 component stocks. Specifically, the selection criteria included:

Market Capitalization: Each company's market capitalization was calculated to determine the largest firms by total value.

Data Collection: Historical stock data for each company was downloaded from Yahoo Finance for the period between January 1, 2013, and December 31, 2023. A retry mechanism was implemented to handle data retrieval issues, such as "Too Many Requests" errors.

Minimum Data Requirement: Companies with insufficient historical data (fewer than 1000 records) were excluded to ensure data quality.

Top 20 Stocks by Market Cap: After calculating the market capitalization, the top 20 companies were selected for visualization.

Libraries used:

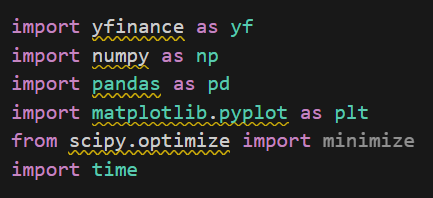


Fig 5. Python library used in this step

Visualization:

The under visualization showed the top 20 stocks in the S&P 500. It also shows the stock market capitalization in descending order to provide a pool of stocks for later portfolio building using Markowitz's model(mean-variance).

图表, 直方图

描述已自动生成

Fig 6. Top 20 Stock by market capitalization

## **5.2 Step 2**: Constructing the Portfolio

A portfolio of selected stocks was constructed using a simple mean-variance optimization technique to allocate weights to different stocks. The use of optimization (such as minimizing variance while maximizing return) ensures a balanced risk-reward ratio. By selecting weights that maximize the Sharpe Ratio (Sharpe, 1994) or minimize risk, the constructed portfolio aims to achieve optimal performance.

Stock Selection and Data Collection: A list of top 20 stock tickers was provided. The historical price data for these stocks was downloaded using the Yahoo Finance API (yfinance). The data was collected for the period from January 2013 to December 2023, focusing on the adjusted closing prices.

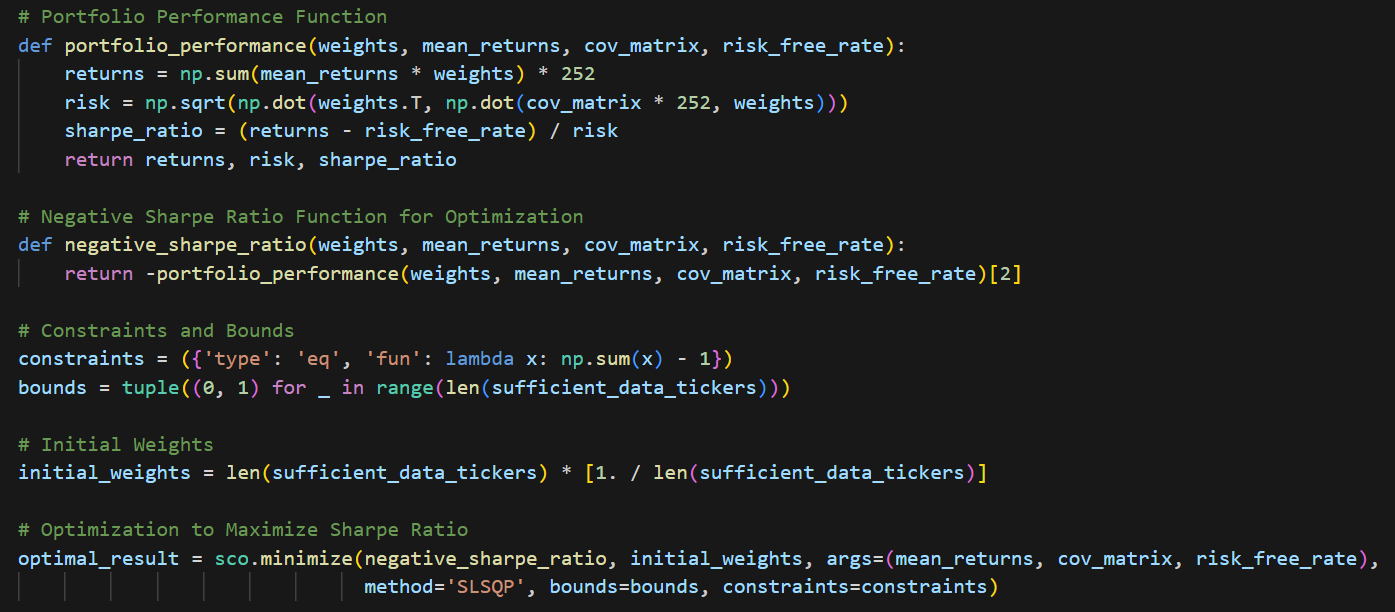
Data Cleaning and Filtering: Stocks with insufficient trading data (less than 1,000 trading days) were filtered out to ensure the reliability of the dataset. This filtering step ensured that only those stocks with significant trading history were included in the portfolio.

Mean and Covariance Calculation: This step calculates the mean return and covariance matrix of a stock after screening its daily returns. It is then input into the mean-variance model for use.

Mean-variance optimization: The mean-variance method is used to optimize the portfolio to maximize the Sharpe ratio . The Sharpe ratio is a measure of risk-adjusted returns and is calculated by dividing the difference between the portfolio return and the risk-free rate (set at 2% in this project) by the volatility of the portfolio. Later, to find the optimal portfolio weights that maximize the Sharpe ratio, the Sequential Least Squares Programming (SLSQP) method is used.

Portfolio Allocation: Using the optimized weights, stocks with a small percentage are filtered out. The final portfolio obtained is represented as a pie chart with the weights allocated to each stock. The goal of this method is to achieve a balanced risk-return ratio by investing in a portfolio of stocks with the best trade-off between expected return and risk.

The use of mean-variance optimization techniques, such as maximizing the Sharpe ratio or minimizing risk, ensures a balanced risk-return profile. With the objective of maximizing returns at a given level of risk, investments are allocated among the top 20 stocks by market capitalization, resulting in an optimal portfolio.

Fig 7.The main part portfolio allocation coding

Libraries /tools used:

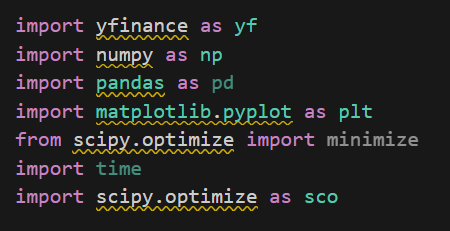


Fig 8. Python library used in this step

Visualization:

This visualization showed the allocation of the portfolio of top 20 stock that is selected by market capitalization. will be used for later comparisons with portfolios constructed after LSTM prediction.

图表, 饼图

描述已自动生成

Fig 9.Optimal Portfolio Weight Allocation

## **5.3 Step 3**: Forecasting with LSTM and use predicted prize to optimize portfolio

The LSTM model is chosen because it is suitable for time series data, which is crucial in financial forecasting’s has the advantage of processing on long-term dependencies data (Yong,2019) and therefore can effectively capture market trends and volatility’s is better than traditional machine learning models in handling continuous data and better than other time series models (ARIMA) can handle non-linear models better, which is why LSTM is widely used in financial forecasting.

Analyses conducted:

The analysis involves collecting historical stock data, preprocessing it, training the LSTM model and using it for forecasting. The portfolio is then optimized based on the predicted returns. The following are the key components of the analysis:

Data Acquisition and Preprocessing: Historical stock data from and for the six stocks in the portfolio in the previous step, from 1 January 2013 to 20 November 2024, was collected using yfinance. Missing values were processed using forward and backward padding. The data was then scaled using MinMaxScaler to normalize the values and thus stabilize the LSTM learning process.

Data Conversion: The scaled data is converted into a sequence suitable for LSTM input with a 50-day lookback window. This window size was tuned to capture enough temporal information to make accurate predictions while maintaining computational efficiency (to calculate the predictions for the six stocks individually, a window of 100 would be too long and would affect the efficiency of the actual use).

Model Architecture and Training: The LSTM model used has two layers, each with 128 units. To prevent overfitting, a culling layer with a ratio of 0.2 was added to the model. The model was compiled using the Adam optimizer and trained with a 20% validation ratio. To avoid overfitting and ensure efficient learning, callbacks such as ReduceLROnPlateau and EarlyStopping were used.

Evaluation and prediction: The model was evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-squared (R²). If the evaluation metrics are satisfactory (MSE < 10,000, MAE < 50, R² > 0.7), the model is used to forecast stock prices for 2024. Each forecast is generated iteratively using the last 50 days of training data as input.

Portfolio optimization: Expected returns are calculated based on the predicted prices and portfolio optimization is performed using scipy.optimize.minimize to maximize the Sharpe ratio. In order to even out the risk of the portfolio to prevent the possibility of having only a single stock, the boundaries of each stock weight are set between 0.05 and 0.4.

Libraries Used:

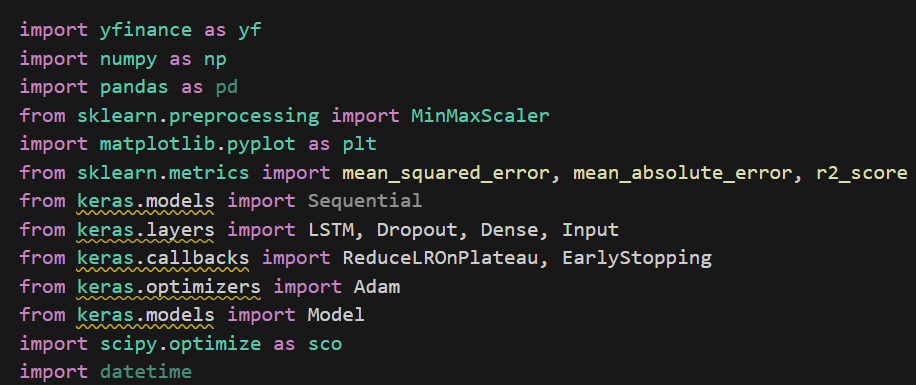


Fig 10. Python library used in this step

Answering Research Questions Using the Models:

The LSTM model was used to predict future stock prices, and the predicted prices were then used to estimate expected returns. This allowed us to answer questions related to the expected performance of each stock. The optimization step provided insights into how the portfolio could be adjusted to maximize returns while minimizing risk, thereby addressing questions related to optimal investment strategies.

Assumptions and Dataset Modifications:

Assumptions: It was assumed that the historical price data could adequately represent future market conditions, which is common in financial modeling. Additionally, a look-back period of 50 days was assumed to be sufficient to capture relevant trends without overfitting.

Dataset Modifications: The data was normalized using MinMaxScaler to ensure that all features were on a comparable scale, aiding in the stability of LSTM training. Missing values were handled by forward and backward filling to ensure continuity in the time series, which is crucial for LSTM models.

文本

描述已自动生成

Fig 11.The main part of LSTM forecasting

Visualization:

This visualization, Shows a comparison between the predicted (Orange Line) and actual performance (Blue Line) of each stock in the portfolio that I have selected and optimized for mean-variance over the course of 24 years, with indices such as R-squared MAE and the images showing that the predictions are better allowing for the optimization of portfolio and the use of predictive models for post-November '24 trading opportunities.

图形用户界面, 图表

描述已自动生成

Fig 12. Price forecasts for individual stocks in the original portfolio

All six stocks basically met our previously mentioned requirements (MSE < 10,000, MAE < 50, R² > 0.7), except for one stock that was off by 3 points, which could be due to excessive volatility, so the forecasts can be used for portfolio optimization.





Fig 13.The evaluation of forecasting in six stocks

After the forecasting of LSTM Based on the projected future returns, we optimized the portfolio again, using the same mean-variance model with the following weightings for each stock in the portfolio. This visualization, showing the optimized portfolio weight shares.

图表, 饼图

描述已自动生成

Fig 14. Optimal portfolio weight allocation

In order to prove that the optimized portfolio (Orange Bar) has significant advantages over the first portfolio (Blue Bar) based on mean-variance, I measured the Sharpe ratio return and volatility of this portfolio, and the data shows that the optimized portfolio based on LSTM prediction has higher return and Sharpe ratio, and less volatility, and the visualization shows the comparison between the two.

图表, 条形图

描述已自动生成

Fig 15. Comparison of portfolio in step2 and step3

## **5.4 Step 4**: Identifying Trading Opportunities

The analysis uses Simple Moving Averages (SMA), Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD) and Bollinger Bands to identify price trends and generate buy and sell signals. In order to capture future price fluctuations, crossover signals of long and short term averages are used along with the LSTM model to predict future prices.

SMAs of different periods (10 and 50) are calculated to identify crossover points such as golden crosses and death crosses. MACD is calculated based on the difference between the 12-period and 26-period Exponential Moving Averages (EMAs) as well as the 9-period signal line. Bollinger Bands are also used to detect overbought and oversold conditions based on standard deviation channels. Buy and sell signals are determined based on the above indicators.

Libraries Used:

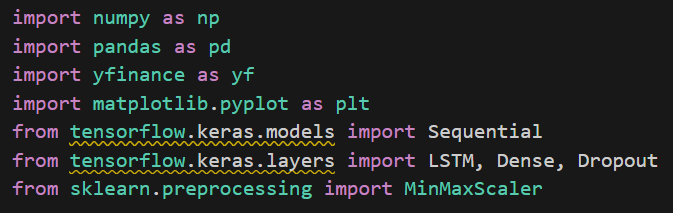


Fig 16. Python library used in this step

Questions answered :

When to Buy?:

Golden Cross (SMA 10 and SMA 50): Buy signals are generated whenever the shorter-term SMA 10 crosses above the longer-term SMA 50.

RSI: Buy signals are generated when the RSI drops below 40, indicating oversold conditions.

MACD: Buy signals are generated when the MACD line crosses above the Signal Line.

Bollinger Bands: Buy signals are triggered when the predicted price moves below the lower Bollinger Band, indicating the asset is oversold.

When to Sell?:

Death Cross (SMA 10 and SMA 50): Sell signals occur when SMA 10 crosses below SMA 50.

RSI: Sell signals are generated when the RSI rises above 60, indicating overbought conditions.

MACD: Sell signals are generated when the MACD line crosses below the Signal Line.

Bollinger Bands: Sell signals occur when the predicted price moves above the upper Bollinger Band, indicating overbought conditions.

Assumptions and Modifications:

It was assumed that historical price trends and the relationships captured through SMAs would continue to be indicative of future price movements.

Market efficiency was partially relaxed, assuming that technical indicators could identify short-term opportunities not immediately reflected in the price.

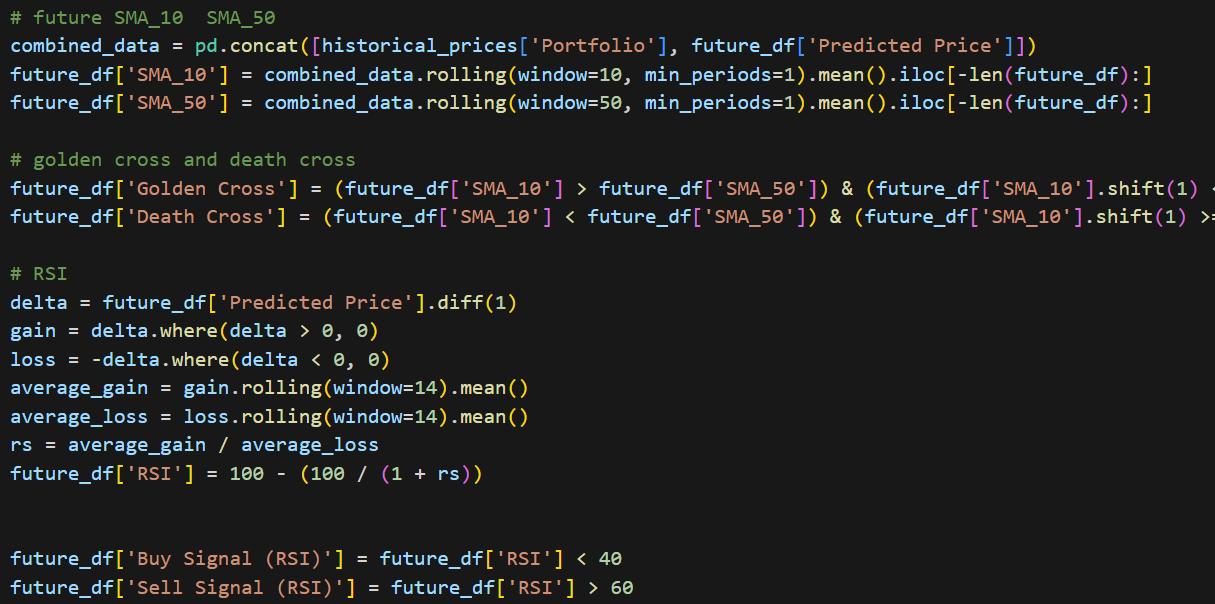


Fig 17. The part of the prediction of signal for transaction

Visualization

The predicted price is shown by the blue line.

The 10-day and 50-day SMAs are shown by orange and purple dashed lines, respectively.

Also, different signals for buy and sell from different indicators are shown.

图表, 折线图

描述已自动生成

Fig 18. buy and sell signals

# 5. Conclusion

In general, the findings demonstrated that combining traditional financial models (mean-variance) with machine learning techniques (LSTM) effectively achieved the aim of maximizing returns while minimizing risks. The constructed portfolio was optimized to handle different market conditions, and the forecasting and trading signals provided actionable insights for decision-making.

However, there are several limitations on my project:

Missing Data: Using forward or backward filling assumes past or future values are accurate estimates, which can be problematic in volatile markets. This approach risks introducing biases, making forecasts less reliable. For example, during significant market events like the 2020 pandemic, using simple imputation may lead to incorrect trend estimations. Exploring other methods, such as using LSTM to predict and fill missing values, could provide better accuracy.

Model Complexity and Overfitting: LSTM's complexity increases the risk of overfitting, particularly with noisy financial data. Although dropout layers were used, high training accuracy does not always translate well to real-world performance, especially under volatile market conditions. For example, when applying the model during a sudden market downturn, overfitting to stable historical trends may fail to adapt adequately.

Mean-Variance Optimization Limitations: The mean-variance approach assumes that returns are normally distributed and covariances are stable, but this is overly desirable in financial markets where fat-tailed effects and volatility clusters often occur. Under extreme market conditions, these assumptions no longer correspond to the status quo and optimised portfolios may not be well protected against significant losses.

Technical Indicators and Market Efficiency: SMA and MACD are based on historical patterns, conflicting with the efficient market hypothesis, which claims all information is already priced in. For example, using MACD signals may not consistently lead to profits in highly efficient markets where new information is instantly reflected in stock prices.

Bias from Assumptions: Assumptions like a 50-day look-back period and stable market conditions influenced results, limiting their applicability. For instance, a different look-back period might yield vastly different outcomes, affecting the reliability of portfolio optimization decisions during different market phases.

When managers or retail investors are in use of the above problems may lead to poor prediction ability, affecting the specific use. And because the above needs to be experimented with frequently based on updated data, it does not apply to the general public. Also may have poor interpretation.

# 6. Personal Reflection

One of the main challenges, during the training process of the LSTM model, I encountered the problem of overfitting, especially when using noisy data. To solve this problem, I added culling layers and tried different learning rates. While this was a time-consuming process, this iterative tweaking led me to better results.

I initially tried using LSTM, and since the results were not accurate, I added Convolutional Neural Networks (CNN) in the hope of improving the accuracy of the predictions. However, due to the increased complexity of the model and the risk of overfitting, I decided to simplify the model and focus on optimizing the LSTM by modifying the number of layers and the learning rate. This taught me the value of balancing the complexity of a model with its utility and the need to consider interpretability outside of realistic usage scenarios. In future projects, I plan to explore other portfolio optimization methods such as Support Vector Regression (SVR), and consider incorporating bonds or futures to diversify investment options.

# 7. Reference

1. Wikipedia contributors. (n.d.). List of S&P 500 companies. In Wikipedia. Retrieved November 21, 2024, from <https://en.wikipedia.org/wiki/List_of_S%26P_500_companies>
2. Markowitz, H. M. (1967). Portfolio selection: Efficient diversification of investments. Wiley.
3. Hochreiter, S. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
4. Appel, G.H. (1979). The stock option and no-load switch fund scalper's manual. Windsor Books.
5. Schwartz, P. M., & Solove, D. J. (2011). The PII problem: Privacy and a new concept of personally identifiable information. NYUL rev., 86, 1814.
6. Sharpe, W. F. (1994). The Sharpe ratio. Journal of Portfolio Management, 21(1), 49–58. <https://doi.org/10.3905/jpm.1994.409501>
7. Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural Computation, 31(7), 1235–1270. <https://doi.org/10.1162/neco_a_01199>

Code in GitHub:[YiyunXia926/topic\_in (github.com)](https://github.com/YiyunXia926/topic_in)