# AAI595 Applied Machine Learning Final Project FINANCIAL PORTFOLIO MANAGEMENT USING MACHINE LEARNING

Pankajbharathi Sowmianarayanan
Department of Mathematical Science
Stevens Institute of Technology
Hoboken, NJ 07030
Email: psowmian@stevens.edu

Hemalatha Katta
Department of Mathematical Science
Stevens Institute of Technology
Hoboken, NJ 07030
Email: hkatta@stevens.edu

Neha Masurkar
Department of Mathematical Science
Stevens Institute of Technology
Hoboken, NJ 07030
Email: nmasurka1@stevens.edu

*Abstract*—The project, *Financial Portfolio Management Using Machine Learning*, addresses the challenges of predicting trends in stock market prices, a complex and dynamic system influenced by numerous factors. While traditional forecasting and diffusion modeling offer insights, they fall short in addressing the intricacies of market volatility and risk management. By integrating machine learning techniques, such as regression analysis, the project identifies relationships between independent features and dependent variables, such as stock prices, to enhance prediction accuracy. The system incorporates stock return predictions using advanced algorithms, sentiment analysis of financial news and social media, and risk-adjusted performance metrics. Through a comparative analysis of three machine learning algorithms, the project demonstrates the most effective model for prediction. Additionally, reinforcement learning enables dynamic portfolio rebalancing, adapting to market fluctuations and optimizing asset allocation. This innovative solution combines AI-driven insights with robust financial strategies to minimize risk and maximize returns, offering scalability and adaptability for diverse investment environments.

## I. Introduction

Financial portfolio management is a critical process aimed at optimizing asset allocation to achieve a balance between maximizing returns and minimizing risks. This project, titled *Financial Portfolio Management Using Machine Learning*, harnesses the power of data-driven techniques to create an intelligent and adaptive system for dynamic portfolio optimization. By integrating machine learning algorithms, natural language processing (NLP), and advanced financial metrics, the system analyzes historical data, market sentiment, and risk factors to inform investment decisions. The project aims to provide a modern, scalable solution that adapts to volatile market conditions and aligns with the diverse objectives of investors.

The cornerstone of this research is to explore innovative methodologies for predicting stock returns, conducting sentiment analysis, and optimizing portfolio allocation. The approach emphasizes balancing quantitative data-driven predictions with qualitative insights from financial news and social media. Key questions driving the project include identifying the most effective models for predicting stock performance, leveraging sentiment analysis to enhance decision-making, and using advanced optimization techniques to achieve an ideal risk-return trade-off. By addressing these challenges, the project aspires to offer transformative tools that revolutionize portfolio management strategies in the evolving financial landscape.

## II. Motivation

The motivation for this project stems from the increasing complexity and volatility of financial markets, which demand advanced tools for effective portfolio management. Traditional methods often fall short in adapting to dynamic market conditions and integrating vast, unstructured data sources like financial news and social media sentiment. By leveraging machine learning and natural language processing, this project seeks to bridge the gap between data-driven insights and practical investment strategies. The goal is to empower investors with a robust, adaptive system that not only maximizes returns but also effectively manages risks, providing a competitive edge in today's fast-paced and data-intensive financial environment.

## III. Related Work

In the past, significant research has been conducted on applying machine learning techniques to financial portfolio management. Early studies focused on using statistical methods like linear regression and time series models, such as ARIMA, for stock price prediction. However, these methods often struggled to capture the complex, non-linear relationships inherent in financial markets. The advent of machine learning brought forth more robust models, including Random Forest, Gradient Boosting Machines, and Support Vector Machines, which improved prediction accuracy by handling non-linear dependencies. Additionally, advancements in deep learning, particularly with Long Short-Term Memory (LSTM) networks, have enabled the modeling of sequential and time-dependent data, proving effective for predicting stock trends and returns.

Beyond stock prediction, sentiment analysis has gained traction in portfolio management. Researchers have explored the use of natural language processing (NLP) techniques, such as VADER and BERT, to analyze financial news and social media data for market sentiment. Studies have shown that incorporating sentiment analysis into decision-making improves portfolio performance by capturing qualitative market signals often overlooked by traditional models. Moreover, modern portfolio optimization techniques, like those leveraging the Sharpe Ratio, Mean-Variance Optimization, and reinforcement learning, have pushed the boundaries of dynamic asset allocation. These advancements highlight a growing trend towards integrating AI-driven solutions into financial strategies, forming the foundation upon which this project is built.

## IV. SOLUTIONS

### A. *Dataset Description*

The dataset for this project is focused on analyzing financial market data, specifically the daily adjusted closing prices of stocks from four major companies: Apple (AAPL), Microsoft (MSFT), Google (GOOG), and Amazon (AMZN). The data spans the period from December 13, 2023, to December 13, 2024, and was sourced using the y-finance library, a widely trusted tool for obtaining historical stock data. Adjusted closing prices were selected to account for factors like stock splits and dividends, ensuring accurate reflection of true value changes. These prices form the foundation for calculating meaningful metrics, such as daily percentage returns, which provide insight into stock performance trends. This data set offers a reliable and comprehensive basis for further analysis, laying the groundwork for informed investment decision-making.

The dataset for this project is focused on analyzing financial market data, specifically the daily adjusted closing prices of stocks from four major companies: Apple (AAPL), Microsoft (MSFT), Google (GOOG), and Amazon (AMZN). The data spans the period from December 13, 2023, to December 13, 2024, and was sourced using the y-finance library, a widely trusted tool for obtaining historical stock data. Adjusted closing prices were selected to account for factors like stock splits and dividends, ensuring accurate reflection of true value changes. These prices form the foundation for calculating meaningful metrics, such as daily percentage returns, which provide insight into stock performance trends. This data set offers a reliable and comprehensive basis for further analysis, laying the groundwork for informed investment decision-making.

To prepare the dataset for machine learning applications, several preprocessing steps were performed. First, daily returns were computed using the formula (Price Today−Price Yesterday)/Price Yesterday, utilizing the pct_change() function in Pandas. This transformation standardized the data by expressing price changes as percentages, making it scale-invariant and suitable for statistical and predictive modeling. Missing values were handled by removing rows with incomplete data, such as the first row, where no previous price exists for comparison. Subsequently, time-lagged features were introduced by shifting daily returns to align the data such that today's return serves as a feature for predicting the next day's return. The data was then split into training (80%) and testing (20%) subsets, ensuring effective model training and evaluation. These steps resulted in a well-structured dataset, ready for advanced modeling tasks.

The refined dataset was applied to a variety of financial analysis tasks, including stock return prediction, sentiment analysis, and portfolio optimization. By integrating machine learning techniques such as Random Forest Regression, the dataset enabled the development of models capable of predicting future returns based on historical trends. Sentiment analysis of financial news and social media added a qualitative dimension to the dataset, allowing for better-informed investment decisions. Moreover, portfolio optimization techniques like Mean-Variance Optimization utilized the dataset to create balanced portfolios that maximize returns while minimizing risks. Through a combination of robust data collection, meticulous preprocessing, and advanced modeling, this dataset serves as a powerful tool for dynamic financial portfolio management and decision-making.

### B. *Data Collection & Stock Prediction*

Data collection is the first step in building any stock prediction system and involves gathering reliable and accurate historical stock market data. The focus is on collecting adjusted closing prices for selected stocks, which account for stock splits and dividends to reflect true value changes. Sources like the Yahoo Finance API (y-finance library) are widely used for their reliability and ease of access. The data spans a specific period, such as a year, to capture meaningful trends and patterns. For example, daily stock data for companies like Apple (AAPL), Microsoft (MSFT), Google (GOOG), and Amazon (AMZN) can form the basis of analysis.

In the preprocessing phase, raw data is cleaned and prepared for further analysis. Key preprocessing steps include calculating daily returns using the formula:

Daily Return=Price Today−Price Yesterday/Price Yesterday

This transformation makes the data scale-invariant and suitable for predictive modeling. Missing values, such as the first row with no prior data, are handled by removing or inputting them. The outcome is a clean dataset of daily percentage changes, ready for feature and target extraction.

```
[*********************100%***********************]
Ticker          AAPL      AMZN      GOOG      MSFT
Date
2023-12-14  0.000758 -0.009540 -0.005748 -0.022545
2023-12-15 -0.002726  0.017298  0.004805  0.013117
2023-12-18 -0.008503  0.027339  0.025030  0.005179
2023-12-19  0.005360 -0.001817  0.006633  0.001637
2023-12-20 -0.010714 -0.010859  0.011296 -0.007073
```

Fig1. Daily Stock Returns for AAPL, AMZN, GOOG, and MSFT

Stock prediction uses machine learning models to estimate future stock prices or returns based on historical data. A popular approach involves time-lagged prediction models, where today's returns are used as features to predict tomorrow's returns. Random Forest Regression is often chosen due to its robustness to non-linear patterns, ability to handle diverse datasets, and ease of interpreting feature importance.

The workflow includes aligning features and targets by shifting daily returns by one day, splitting the data into training (80%) and testing (20%) subsets, and training the model on historical trends. The performance of the model is evaluated using metrics such as Mean Squared Error (MSE) or R-squared values. The predicted results provide insights into potential future stock movements, enabling better investment strategies.

This systematic approach to data collection and prediction forms the backbone of any intelligent portfolio management system, supporting informed decision-making under uncertainty while aiming to maximize returns.

Predicted Returns:

| | AAPL | MSFT | GOOG | AMZN |
|---|---|---|---|---|
| 0 | 0.004288 | 0.008756 | 0.010687 | 0.010114 |
| 1 | 0.000408 | 0.001570 | 0.005154 | 0.007640 |
| 2 | 0.005667 | 0.001091 | 0.001823 | 0.002296 |
| 3 | 0.002966 | 0.006747 | 0.008083 | 0.001250 |
| 4 | 0.002722 | 0.004104 | 0.002577 | -0.001060 |

Fig2. Predicted Stock Returns for AAPL, MSFT, GOOG, and AMZN

### *C. Random Forest Regression*

Random Forest Regression is an ensemble learning algorithm that combines the outputs of multiple decision trees to improve prediction accuracy. It works by constructing a collection of decision trees using bootstrapped subsets of the training data. In each decision tree, a random subset of features is considered for each split, which helps create diverse trees that are less likely to be overfit to the data. This diversity enhances the overall performance of the model by reducing variance and making it more robust to noise.

In regression tasks, Random Forest predicts an output by averaging the predictions from all individual trees in the forest. This process helps to smooth out the results and avoid overfitting that can occur with a single decision tree. The algorithm is well-suited for modeling complex, non-linear relationships between features and the target variable. Additionally, Random Forest provides important insights into feature importance, which can guide further analysis or feature selection in other models.

One of the key advantages of Random Forest Regression is its ability to handle large datasets with high dimensionality, making it applicable across various domains. It can also manage missing data effectively by using surrogate splits during the tree-building process. Despite these strengths, the algorithm can be computationally expensive, especially with large datasets or a high number of trees, and its black-box nature makes it harder to interpret compared to simpler models like linear regression.

Random Forest Regression is widely used in applications such as financial forecasting, where it can predict stock prices or returns based on historical data, medical research for predicting disease outcomes, and in real estate for estimating property values. Its ability to balance accuracy with complexity makes it a popular choice for tasks where prediction precision is critical.

### *D. Long Short-Term Memory Network*

**1. LSTM in the Project Context**: In the project focused on financial portfolio management, LSTM (Long Short-Term Memory) was introduced as an enhancement for stock price prediction. Stock prices follow sequential patterns, and traditional methods often fail to capture long-term dependencies in such data. LSTM, as an advanced recurrent neural network (RNN), was selected because of its ability to remember historical stock price patterns over extended periods. This feature is crucial for financial forecasting, where past performance often influences future trends, making LSTM an effective tool for predicting future stock returns in the context of dynamic portfolio management.

**2. LSTM Architecture for Stock Price Prediction:** In this project, LSTM was used to model the time-series nature of stock prices, which are influenced by various market factors over time. The LSTM network's architecture, with its memory cells and gating mechanisms, helped learn temporal dependencies between stock prices on different days. The forgetting gate allowed the model to discard irrelevant information, while the input and output gates controlled the flow of new and old data, helping the network focus on the most important trends. This ability to retain important information over time made LSTM a valuable tool for improving the accuracy of stock price predictions.

**3. Benefits of LSTM for Financial Forecasting:** LSTM's primary advantage in this project is its ability to handle and predict stock prices based on historical data. Financial markets are volatile, and predicting stock returns requires understanding of long-term trends. Unlike traditional models that might fail to capture such patterns, LSTM networks maintain an internal memory, making them adept at predicting future returns based on past data. By incorporating LSTM into the stock return prediction workflow, the model can adapt to various market conditions and adjust predictions, accordingly, thus improving the robustness of portfolio optimization strategies.

**4. LSTM for Portfolio Management:** LSTM was used to enhance portfolio optimization by providing more accurate predictions of stock returns, which are essential for determining the best asset allocation. By predicting the future returns of stocks like Apple, Microsoft, Google, and Amazon, the LSTM model helped inform decisions about which assets to invest in and how much to allocate to each. The dynamic nature of the LSTM model also made it suitable for real-time adjustments in the portfolio, enabling better risk management and maximizing returns. This integration of LSTM into the portfolio management system allowed for more adaptive and data-driven investment decisions.

**5. Future Potential of LSTM in the Project:** While LSTM improved the stock return prediction accuracy, future enhancements could involve incorporating more complex variations of LSTM, such as bidirectional LSTMs or stacked LSTMs, to further enhance performance. Additionally, combining LSTM with other techniques, such as reinforcement learning for dynamic portfolio rebalancing, could provide a more comprehensive solution. By leveraging advanced techniques like these, the project could evolve to handle even more complex financial data, offering a robust, scalable, and adaptive system for managing investment portfolios in fluctuating market environments.

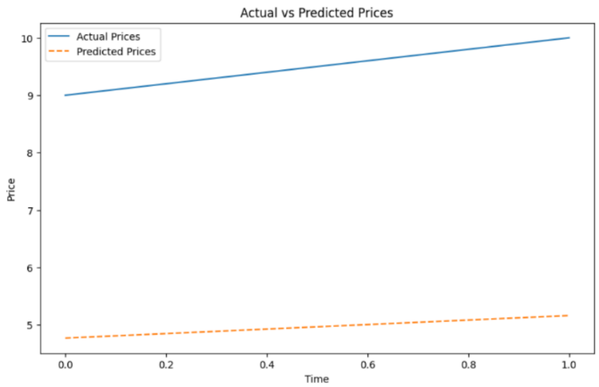



Fig 3. Line graph for One stock price using LSTM

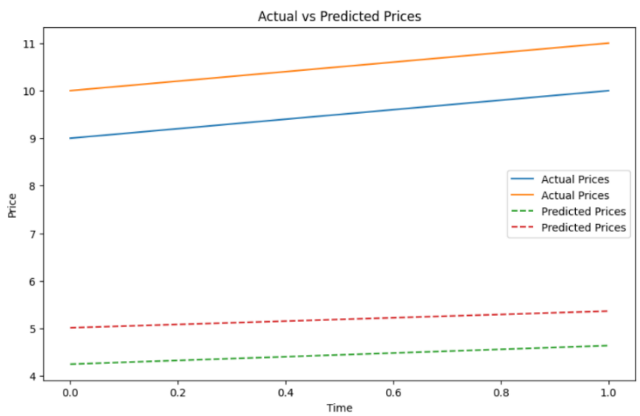



Fig 4. Line Graph for Multiple Stock Using LSTM

*E. Sentiment Analysis*

Sentiment analysis in the context of financial portfolio management is a crucial aspect of understanding market behavior and predicting asset movements. It involves analyzing financial news, social media, and other textual data sources to gauge public sentiment towards specific stocks or the market in general. By examining the tone of news articles, social media posts, or investor sentiment reports, sentiment analysis tools can help predict market trends that may not be immediately apparent from historical data alone. This allows investors to make more informed decisions, potentially giving them an edge in a competitive market.

In this project, sentiment analysis uses two prominent techniques: VADER and BERT. VADER, or Valence Aware Dictionary and Sentiment Reasoner, is a highly efficient tool for analyzing short, informal texts such as tweets or headlines. It assigns sentiment scores—positive, neutral, or negative—and provides a compound score to reflect the overall sentiment. On the other hand, BERT, a more advanced Natural Language Processing (NLP) model, can understand complex contexts in longer and more nuanced texts. It allows for a deeper analysis of sentiment, providing more accurate insights into market sentiment from financial reports, news, and social media.

The sentiment scores produced by VADER and BERT are then used to enhance stock return prediction models. Sentimental data provides additional features that can be integrated into machine learning models, helping to forecast stock prices more accurately. By incorporating real-time sentiment data, these models can adapt quickly to market shifts caused by breaking news or public opinion. For example, positive sentiment surrounding a company's earnings report may influence its stock price, and sentiment analysis allows these changes to be captured early, improving the accuracy of predictions.

Visualization tools, such as word clouds, further enhance sentiment analysis by visually representing the most impactful terms within financial news articles or social media posts. These visualizations help investors and analysts quickly grasp the key themes that are driving market sentiment. In this project, for example, frequent terms such as "stock," "growth," and "earnings" were identified, providing insight into market trends and sentiments surrounding companies or industries. This form of analysis helps investors stay updated on market moods and align their strategies accordingly.

Overall, sentiment analysis serves as a powerful tool in financial portfolio management by providing a more comprehensive view of the market. While traditional financial metrics such as price-to-earnings ratios and historical returns are essential, sentiment analysis adds an additional layer of insight by factoring in the psychological and emotional factors influencing market behavior. By integrating sentiment data into portfolio optimization and risk management strategies, investors can improve their decision-making processes, making portfolios more adaptable to ever-changing market conditions and potentially increasing returns while managing risk.

| | title | sentiment |
|---|---|---|
| 0 | Scott Boras Takes Shot At New York Yankees At ... | 0.068182 |
| 1 | Broadcom Shares Jump After Chipmaker Predicts ... | -0.212500 |
| 2 | Appleが2025年に自社製ネットワークチップ搭載のApple TVとHomePod mi... | 0.368182 |
| 3 | Australia’s media movers and shakers on who su... | 0.037037 |
| 4 | 350-Powered ’30 Ford Model A Coupe Street Rod | 0.000000 |

Fig 5. News Article Sentiment Analysis


Fig 6. Word Cloud of Financial News

*F. Q-Learning (Reinforcement Learning)*

It is a model-free reinforcement learning algorithm used to find the optimal action-selection policy for an agent interacting with an environment. It works by assigning a value, called the Q-value, to each state-action pair, which represents the expected future reward for taking a particular action in a specific state. The agent iteratively updates the Q-values based on the rewards received after each action. This allows the agent to learn the most effective strategy for maximizing cumulative rewards over time, even without knowing the environment's underlying dynamics.

The core of Q-Learning is the Q-value update rule, which combines immediate rewards with the maximum expected future rewards.

The Q-value for a state-action pair is updated using the formula

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[ r + \gamma \cdot \max_{a'} Q(s',a') - Q(s,a) \right]$$

where the learning rate (α\alphaα) controls how much new information overrides the old, and the discount factor (γ\gammaγ) determines the importance of future rewards. This process continues until the Q-values converge, indicating the optimal policy.

In financial portfolio management, Q-Learning can be applied to optimize asset allocation by learning the best decisions (e.g., buy, sell, or hold) based on market conditions. The agent, using historical data as the environment, learns which actions yield the best returns. Over time, it refines its strategy to maximize profits while managing risk. For example, the agent learns when to buy or sell stocks to achieve the best possible portfolio returns, adjusting its actions according to the evolving market environment.

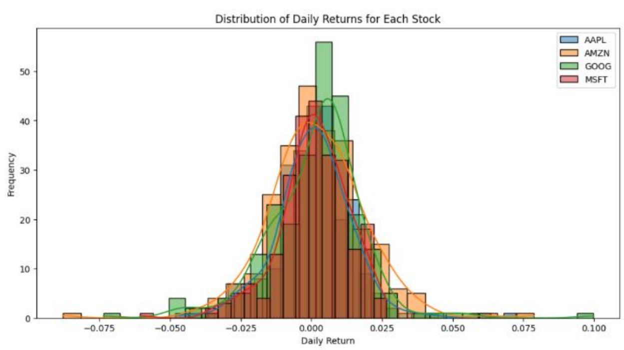



Fig. 7 Distribution of Daily Returns for Each Stock

## V. FUTURE RESEARCH DIRECTION

**Expanded Asset Classes**: Extend the model to handle a broader range of asset classes such as bonds, cryptocurrencies, or real estate, improving its versatility and applicability to various investment portfolios.

**Advanced Reinforcement Learning Techniques**: Implement more advanced RL algorithms like Deep Q-Networks (DQN) or Actor-Critic methods to further improve portfolio rebalancing and strategy optimization.

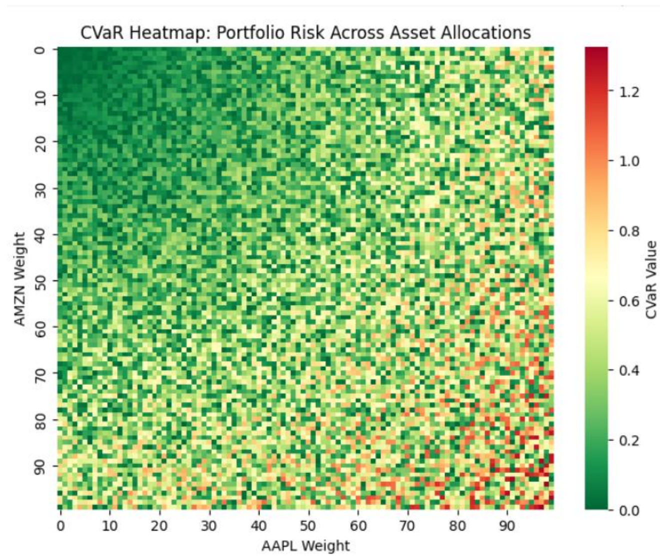
**Quantum Computing for Portfolio Optimization**: Explore quantum computing algorithms for optimizing portfolio allocations, leveraging quantum speedups to solve complex problems faster than classical methods.

**Decentralized Finance (DeFi) Integration**: Integrate with DeFi platforms to enable portfolio management that includes crypto assets, yield farming, or decentralized lending markets, offering a new frontier in portfolio diversification.

## VI. CONCLUSION

In conclusion, the integration of artificial intelligence into portfolio management is revolutionizing the way investors approach the market. Machine learning significantly enhances return prediction and risk management, enabling more accurate forecasting and data-driven decision-making. Sentiment analysis adds an extra layer of insight, capturing real-time market sentiment and news trends to further refine stock price predictions and portfolio strategies. Reinforcement learning takes this further by enabling dynamic and adaptive rebalancing, ensuring portfolios remain aligned with constantly evolving market conditions.

This AI-driven approach provides a powerful, modern framework for both individual investors and institutions, allowing for more efficient, scalable, and adaptable strategies. By combining these techniques, investors can maximize returns, reduce risks, and navigate the complexities of today's financial landscape with greater confidence and precision. The result is a sophisticated portfolio management system that not only enhances profitability but also fosters long-term sustainability in a volatile market environment.




HeatMap for Risk Evaluation

## VII. CONTRIBUTION

1) Pankajbharathi – Coded the stock prediction, backtest portfolio, LSTM & Sentiment Analysis, prepared the presentation slides, and did the final project report.
2) Hemalatha – Did Data Collection & Data preprocessing of the data.
3) Neha – Did Optimization Portfoilio & Evaluation.

## VIII. REFERENCES

1) https://www.researchgate.net/publication/382028685_FINANCIAL_PORTFOLIO_OPTIMIZATION_USING_MACHINE_LEARNING_AND_DATA_ANALYTICS
2) https://www.researchgate.net/publication/362660222_Portfolio_Optimization_using_Artificial_Intelligence_A_Systematic_Literature_Review
3) https://dspace.mit.edu/bitstream/handle/1721.1/157186/masuda-jmasuda-meng-eecs-2024-thesis.pdf?sequence=1&isAllowed=y
4) https://www.sciencedirect.com/science/article/pii/S2405918821000155
5) https://www.sciencedirect.com/science/article/abs/pii/S0957417423001410
6) https://iaeme.com/MasterAdmin/Journal_uploads/IJCA/VOLUME_3_ISSUE_1/IJCA_03_01_002.pdf
7) https://link.springer.com/article/10.1007/s10614-024-10604-6