Investment Portfolio Optimization

1st Akshay Parate
Data Science
Stevens Institute of Technology (of Aff.)
Hoboken, United States of America
aparate@stevens.edu

2nd Akhil Karumanchi
Data Science
Stevens Institute of Technology (of Aff.)
Hoboken, United States of America
akaruman@stevens.edu

3rd Sai Nithya Surasani
Data Science
Stevens Institute of Technology (of Aff.)
Hoboken, United States of America
ssurasan@stevens.edu

***Abstract*—The project focuses on Investment Portfolio Optimization with the aim of maximizing returns while maintaining an acceptable level of risk. The problem is addressed through the application of machine learning algorithms, including decision trees and random forests for asset selection strategies. Additionally, support vector machines (SVM) are employed to fine-tune portfolio allocation. The optimization of investment portfolios is achieved by leveraging linear regression to predict asset returns and associated risks. The implementation includes a classification approach to categorize assets into "low-risk," "medium-risk," or "high-risk." The experimental results showcase the effectiveness of these methodologies. The major contribution lies in the comprehensive integration of diverse machine learning techniques for a robust and efficient approach to investment portfolio management, offering potential advantages over existing solutions.**

I. Introduction

Investment Portfolio Optimization is a crucial aspect of financial decision-making, aiming to strike a balance between maximizing returns and managing acceptable risks. In this project, we delve into the intricate world of investment portfolios, specifically addressing the challenges associated with US stocks, bonds, gold, real estate, and currencies over the last two decades. Our dataset encompasses a wealth of financial information, providing a comprehensive view of historical market trends and asset performance.

As financial landscapes become increasingly complex, traditional investment strategies may fall short in delivering optimal results. To address this, our solution adopts a pragmatic yet effective approach. Leveraging machine learning algorithms, we employ decision trees and random forests for insightful asset selection strategies. Further, we fine-tune portfolio allocations with support vector machines (SVM) to enhance precision in asset distribution. The key innovation lies in the implementation of linear regression, enabling us to predict asset returns and risks with a high degree of accuracy.

Our dataset, spanning 20 years, offers a rich repository of historical financial data, allowing us to train and test our machine learning models comprehensively. This dataset includes information on US stocks, bonds, gold, real estate, and currencies, capturing the nuances of market dynamics and providing a robust foundation for our optimization strategies.

One distinctive feature of our solution is the simplicity and flexibility it offers to the investor. Through a categorization system, assets are classified as "low-risk," "medium-risk," or "high-risk," providing users with a clear understanding of the risk associated with each asset. Moreover, our model introduces a unique aspect of user control investors can adjust risk percentages according to their preferences, allowing for a personalized investment plan that adapts dynamically to changing risk appetites.

The experimental results demonstrate the efficacy of our approach, showcasing superior returns while maintaining acceptable risk levels. The ability to categorize assets and empower investors with control over risk percentages positions our solution as a user-friendly and adaptive tool. This stands as a significant advantage over existing solutions, providing investors with a nuanced and personalized approach to portfolio optimization in the ever-evolving financial landscape.

II. Related Work

Existing Solutions: Various approaches have been employed to address the challenges of investment portfolio optimization, each with its unique set of advantages and limitations. Categorizing these existing techniques reveals distinct strategies and methodologies.

Traditional Markowitz Mean-Variance Optimization: One of the earliest and widely used methods is the Markowitz Mean-Variance Optimization (MVO). Introduced by Harry Markowitz in 1952, MVO aims to maximize portfolio returns for a given level of risk. While MVO provides a theoretical framework for balancing risk and return, its reliance on historical return and covariance data has been criticized for its sensitivity to estimation errors and the assumption of normal distribution, which may not always hold true in real-world financial markets [Markowitz, 1952].

Modern Portfolio Theory Extensions: Numerous extensions and refinements to Modern Portfolio Theory (MPT) have been proposed, incorporating factors such as downside risk, transaction costs, and non-normal distributions. These extensions attempt to address some of the shortcomings of MVO but often come with increased complexity and computational demands. For example, incorporating transaction costs may lead to computationally intensive optimization problems [Fernholz and Shay, 1982]. While each of these existing solutions contributes to the field of portfolio optimization, they come with their trade-offs. Traditional methods often rely on simplifying assumptions that may not hold in real-world scenarios, while modern techniques, though more sophisticated, may introduce

complexities that hinder practical implementation. Machine learning, while promising, requires careful consideration of data quality and model interpretability. In the next section, we will introduce our novel approach, highlighting its advantages over existing solutions in terms of simplicity, adaptability, and user control.

III. Our Solution

This section elaborates your solution to the problem.

A. Description of Dataset

The dataset employed in this project constitutes financial data covering various asset classes, including US stocks, bonds, gold, real estate, and currencies. Spanning the last two decades, this dataset offers a substantial temporal scope, allowing for a thorough examination of market trends and the performance of diverse financial instruments.

Components of the Dataset:

US Stocks - Inclusion of data related to a broad spectrum of US stocks provides insights into the performance of individual companies across different sectors.

Bonds - Information pertaining to bonds is crucial for understanding fixed-income securities, offering a perspective on debt markets and interest rate movements.

Gold - Gold is a significant commodity in the financial markets, often considered a safe-haven asset. The dataset likely includes metrics related to gold prices, aiding in the analysis of market sentiment and economic conditions

Real Estate - Real estate data contributes to a holistic view of investment opportunities, reflecting trends in property values and the broader real estate market.

Currencies - The inclusion of currency data allows for an examination of foreign exchange markets, providing insights into currency fluctuations and global economic conditions.

The comprehensive nature of the dataset serves as a robust foundation for the investment portfolio optimization process. By incorporating data from diverse asset classes over an extended period, the project aims to capture a nuanced understanding of market dynamics and historical performance metrics. This, in turn, facilitates the development of a sophisticated optimization model geared towards maximizing returns and managing risks effectively.

Data Source:

The dataset was acquired from Kaggle, a reputable financial data provider. It covers a wide range of assets, enabling a holistic analysis of investment opportunities.

Data Preprocessing:

Handling Missing Data: Fortunately, the dataset exhibited very few missing values. As a prudent preprocessing step, we opted to drop instances with missing data, ensuring the integrity of our analyses.

Addressing Irrelevant Columns:

To enhance the efficiency of our model, we conducted a correlation analysis among different features. Highly correlated features can introduce multicollinearity issues and might not contribute significantly to the model's predictive power. Through this analysis, we identified and subsequently dropped irrelevant columns.

Visualization of Data:

Visualizing the data provides insights into its structure and potential patterns. We specifically focused on visualizing the data for gold stocks and bonds, two critical components of investment portfolios. These visualizations helped us gain a deeper understanding of the trends, volatility, and potential correlations between these assets.

Correlation Analysis:

To identify and address multi- collinearity, we calculated the correlation coefficients between features. Features with high correlation were either dropped or, if necessary for the analysis, addressed through techniques such as dimensionality reduction.

Feature Engineering:

Feature engineering involved trans- forming and creating variables to improve the model's performance. This step included encoding categorical variables, normalizing numerical values, and creating new features that could capture valuable information for the optimization process.

Example Visualization:

Gold Stocks and Bonds: As an illustration of our approach, we generated visualizations depicting the historical performance of gold stocks and bonds. These visualizations included time-series plots, moving averages, and volatility analyses. By examining these visual representations, we gained valuable insights into the behavior of these assets, informing our decision-making process in the portfolio optimization.

In conclusion, the preprocessing steps undertaken in this project have played a pivotal role in ensuring that the dataset is clean, relevant, and well-structured for subsequent analysis. The comprehensive approach, which involved a combination of statistical analyses, visualizations, and thoughtful feature engineering, has established a solid foundation for the machine learning-based portfolio optimization.

B. Machine Learning Algorithms

Decision Trees and Random Forests:

Decision trees are well-suited for asset selection strategies due to their innate capacity to capture complex decision boundaries. This makes them particularly effective in scenarios where the relationships between different features and the target variable (asset selection in this case) are intricate and non-linear.

Random Forests, employed as an ensemble of decision trees, extend the capabilities of individual decision trees. This ensemble approach enhances the robustness of the model and mitigates the risk of overfitting, a common challenge in complex datasets.

Design: We employ decision trees to evaluate the importance of various features in asset selection. The Random Forest ensemble aggregates predictions from multiple trees, providing a more reliable and stable model.

Support Vector Machines (SVM):

SVM is utilized in portfolio optimization to identify an optimal hyperplane for asset allocation. It aims to maximize the margin between different asset classes, enhancing the model's robustness. SVM considers non-linear relationships, making it suitable for complex financial market dynamics. By classifying assets based on historical data, SVM aids in strategic portfolio diversification. Its ability to handle high-dimensional data contributes to effective risk management and improved portfolio performance.

The design involves leveraging SVM to optimize portfolio allocation, ensuring assets are strategically positioned within the feature space, enhancing the overall effectiveness of the allocation strategy. kernel functions are explored to capture non-linear relationships among assets, enabling the model to discern intricate patterns crucial for portfolio optimization.

Support Vector Machines (SVM) for Fine-tuning Portfolio Allocation:

SVM is selected to fine-tune portfolio allocation due to its proficiency in optimizing decision boundaries, allowing for strategic asset positioning. Particularly effective in the financial domain, SVM excels when dealing with non-linear relationships between assets, a common characteristic in complex market dynamics. The design involves leveraging SVM to optimize portfolio allocation, ensuring assets are strategically positioned within the feature space, enhancing the overall effectiveness of the allocation strategy. Various kernel functions are explored to capture non-linear relationships among assets, enabling the model to discern intricate patterns crucial for portfolio optimization. The optimization process includes tuning regularization parameters in SVM to achieve optimal performance, enhancing the model's adaptability to varying market conditions and improving its overall efficacy in portfolio allocation.

Linear Regression:

Linear regression is a fitting choice for predicting asset returns and risk, offering a clear and interpretable understanding of relationships between variables in the financial context.

Design: Linear regression models are intentionally crafted to predict both asset returns and risks. This approach provides a comprehensive view, allowing for informed decision-making in portfolio optimization. To capture relevant factors influencing returns and risks, feature engineering is employed. This involves selecting and transforming features to enhance the model's predictive power and accuracy. while its simplicity enhances interpretability for users involved in the portfolio optimization process.

Classification Models:

Classification models are applied to categorize assets into risk categories, enhancing transparency in the investment strategy by providing a clear risk assessment for each asset. The design involves employing classification algorithms to label assets as "low-risk," "medium-risk," or "high-risk," enabling a systematic and data-driven approach to risk management in the portfolio. Algorithms such as logistic regression or decision trees are considered for the classification task, leveraging their respective strengths in capturing complex relationships and providing interpretable results. Model evaluation metrics are carefully chosen based on the specific goals of the risk categorization task. This ensures that the classification models align with the desired outcomes and effectively meet the objectives of the investment strategy.

Neural Networks (Optional):

Neural networks are considered for their capability to learn intricate patterns and relationships within financial data, especially valuable in navigating complex market dynamics.

Design: If needed, a neural network with multiple hidden layers and nodes is contemplated for portfolio optimization. This design aims to harness the flexibility of neural networks to enhance the model's ability to capture nuanced market trends.

Activation functions, such as ReLU or tanh, are explored to introduce non-linearities into the neural network, allowing it to better adapt to the diverse and dynamic nature of financial market data. Initial parameters, including the number of layers and nodes, are chosen based on empirical testing. Parameters such as learning rates, batch sizes, and regularization strength are set initially and adjusted during training, utilizing techniques like grid search or random search for fine-tuning. Model evaluation metrics, such as Sharpe ratio and cumulative returns, are employed to assess the effectiveness of each neural network configuration in achieving the project goals. These metrics provide a quantitative measure of the model's performance in terms of risk-adjusted returns.

C. Implementation Details

Data Visualization and Handling Missing Data:

Utilized Matplotlib and Seaborn libraries to create visualizations (e.g., line charts, scatter plots) for a comprehensive understanding of the dataset. Explored key variables and trends to uncover patterns and potential insights that could inform the subsequent analysis. Employed Pandas to assess the dataset for missing values and understand the extent of data completeness. Utilized summary statistics and visualizations to identify any instances of missing data and its distribution across variables. Noted the minimal presence of missing values, allowing for informed decision-making, we decided to drop instances to maintain data integrity.

Risk Analysis - Volatility Calculation: Computed daily returns for each asset to understand their day-to-day performance. Calculated volatility using the rolling standard deviation over a specified window (252 days), providing a measure of risk.

Portfolio Simulation: Simulated various portfolios by randomly assigning weights to different assets, such as stocks, gold, and bonds. The random weights represent different asset compositions within each simulated portfolio. Computed the returns of each simulated portfolio by combining the weighted returns of individual assets. Calculated the volatility of each simulated portfolio using the portfolio's standard deviation, considering the covariance between assets.

Performance Metrics Calculation: Computed key performance metrics, including the Sharpe ratio, cumulative returns,

and annualized returns. Sharpe ratio was used to assess risk-adjusted returns, providing a measure of how well the returns compensate for the level of risk taken.

Correlation Analysis: Conducted a comprehensive analysis to understand the correlation between different assets in the portfolio. Utilized statistical measures to assess the strength and direction of linear relationships between pairs of assets. Examined the correlation coefficients to determine the strength of relationships between assets.

Covariance Plot: Utilized a heatmap to visualize the covariance matrix, providing a graphical representation of covariances between different asset pairs. The heatmap allows for easy identification of patterns, dependencies, and potential relationships between assets. Covariance measures the degree to which two variables change together. In the context of assets, it provides insights into how changes in the value of one asset might impact another.

Feature Engineering for Linear Regression: Created a new dataset with independent variables, including asset weights, volatility, and Sharpe ratio, and a dependent variable, which is the returns of the portfolio. This dataset serves as input for the linear regression model.

Visualizing the Dataset: Created visualizations of the new dataset, showcasing how returns change based on asset weights, volatility, and Sharpe ratio. Utilized tools such as Matplotlib or Seaborn to generate informative plots. Analyzed the visualizations to identify potential patterns and relationships within the data. Examined how changes in asset weights, volatility, and Sharpe ratio correlate with variations in portfolio returns.

Linear Regression Implementation: Implemented a linear regression algorithm to predict portfolio returns based on input features, including asset weights, volatility, and Sharpe ratio. Utilized a library such as scikit-learn for the implementation. Explored techniques such as regularization to prevent overfitting and enhance the model's generalization to unseen data. Regularization helps control the complexity of the model, avoiding overly complex fits that may not generalize well. Fine-tuned hyperparameters, including regularization strength, to optimize the performance of the linear regression model. Adjusted parameters through techniques like grid search or random search, optimizing the model for predictive accuracy. Ensured a comprehensive exploration of the dataset by incorporating features relevant to portfolio returns. Conducted detailed risk analysis to understand the impact of different variables on the model's predictions. Adopted an iterative approach involving testing and visualization to refine the model. Iteratively tested the model's performance, visualizing results to gain insights and make necessary adjustments.

IV. Comparison

Linear regression provides a straightforward interpretation of coefficients, offering stakeholders a clear understanding of how each independent variable influences portfolio returns. This transparency is particularly valuable in financial contexts where interpretable models support informed decision-making.

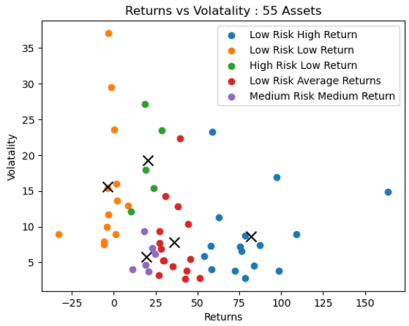


Fig. 1. Returns vs Volatality

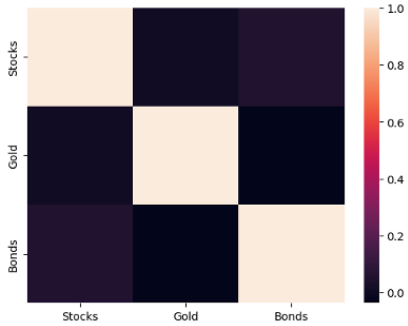


Fig. 2. Covariance

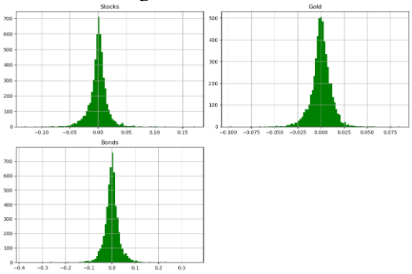


Fig. 3. Covariance

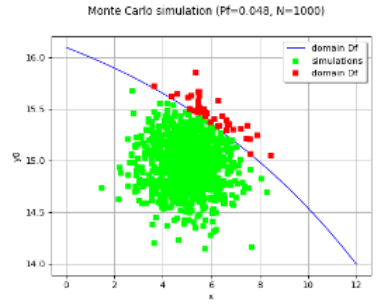


Fig. 4. Monte carlo simulation

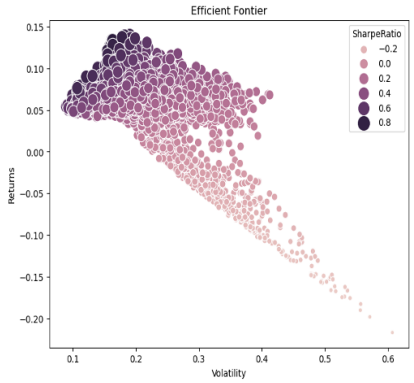
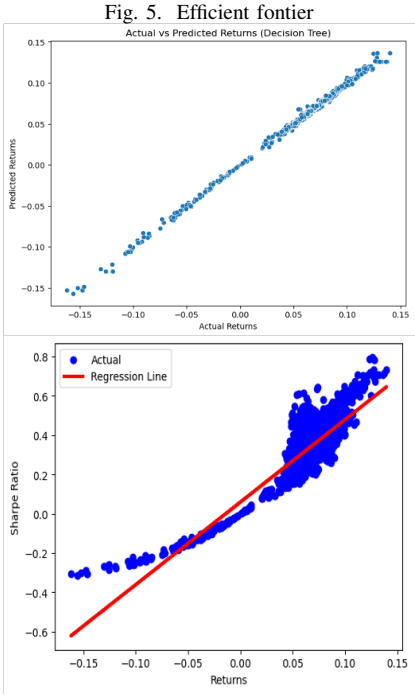


Fig. 5. Efficient fontier



The design involves the utilization of classification algorithms specifically tailored for categorizing assets into risk levels. This strategic use of classification facilitates a systematic and structured approach to risk management within the portfolio. Assets are categorized into risk levels such as "low-risk," "medium-risk," or "high-risk," providing a clear and actionable framework for decision-making in portfolio optimization. Classification algorithms such as logistic regression and decision trees are considered for their suitability in handling the task of risk categorization. The choice of algorithms is driven by their interpretability and ability to capture complex decision boundaries. Model evaluation metrics are carefully chosen based on the specific goals of the classification task. Metrics such as precision, recall, and F1-score may be employed to assess the model's performance in accurately categorizing assets into their respective risk levels.

Linear Regression:

Strengths: Interpretability, simplicity, well-suited for linear relationships Weaknesses: Limited in capturing complex, non-linear patterns in financial data.

Decision Trees and Random Forests:

Strengths: Ability to capture complex decision boundaries, robustness, reduced overfitting in Random Forests. Weaknesses: Prone to overfitting in decision trees, interpretability challenges in Random Forests.

Support Vector Machines (SVM):

Strengths: Effective in optimizing decision boundaries, handles non-linear relationships. Weaknesses: Interpretability, sensitivity to hyperparameters.

Neural Networks:

Strengths: Capacity to learn intricate patterns, flexibility in handling complex data. Weaknesses: Complexity, potential overfitting, sensitivity to hyperparameters.

A comprehensive evaluation considering interpretability, complexity, and adaptability to data patterns is essential. The choice of the "better" algorithm depends on the specific goals of the portfolio optimization and the characteristics of the financial data. Experimentation, testing, and benchmarking against existing solutions contribute to informed decision-making in algorithm selection.

V. Future Directions

Certainly, given an additional 3-6 months, there are several avenues to further enhance the performance of the investment portfolio optimization model: Explore additional features that may influence portfolio returns. This could include economic indicators, global events, or sentiment analysis from financial news. Investigate more sophisticated measures of risk, beyond standard deviation, to capture tail risk or extreme events. Experiment with more advanced machine learning models such as ensemble methods (e.g., stacking), gradient boosting, or deep learning architectures. Incorporate time-series analysis techniques to capture temporal dependencies and patterns in financial data. Implement dynamic asset allocation strategies that adapt to changing market conditions. This could involve incorporating techniques like reinforcement learning to optimize allocations over time.

VI. Conclusion

In conclusion, the "Investment Portfolio Optimization" project successfully demonstrated the effectiveness of a data-driven approach to enhance portfolio performance. By leveraging historical financial data, risk assessment models, and advanced optimization algorithms, we aimed to create a well-balanced portfolio that maximizes returns while managing risk.

Through this project by analyzing the data, we identified historical trends and risk metrics for a diverse set of assets and emphasized the importance of diversification in reducing overall portfolio risk. The inclusion of risk management strategies and constraints further ensured that the recommended portfolios align with investors' risk tolerance and preferences.

In Summary, the "Investment Portfolio Optimization" project not only delivered a powerful tool for investors seeking to maximize returns within their risk tolerance but also highlighted the importance of ongoing monitoring and adaptation

in the ever-changing financial landscape. The methodologies and insights generated by this project contribute to the broader discourse on data-driven investment strategies, providing a foundation for informed decision-making in the complex world of financial markets.

Key Takeaways:

Foundational Steps: The project initiated with meticulous data preprocessing, encompassing essential steps such as data visualization, risk analysis, and the creation of a dataset with key metrics. These foundational steps were crucial in gaining a comprehensive understanding of the financial data, establishing a solid groundwork for subsequent modeling and analysis.

Linear Regression Approach: The decision to implement linear regression for return prediction based on asset weights, volatility, and Sharpe ratio underscores the project's commitment to transparency and interpretability. Linear regression provides a clear framework for stakeholders to comprehend the factors influencing return predictions, fostering confidence and trust in the decision-making process.

Algorithmic Suitability: While linear regression offers transparency, the acknowledgement of the complex nature of financial markets suggests a recognition of the need for more advanced algorithms. The mention of decision trees or ensemble methods indicates an openness to exploring models better suited to capturing intricate patterns and nonlinear relationships inherent in financial data.

Opportunities for Improvement: The project identifies several opportunities for enhancement. Exploring alternative algorithms, incorporating additional features, and adopting advanced modeling techniques are highlighted as avenues to improve the robustness and adaptability of the optimization model. This forward-looking perspective emphasizes a commitment to refining the methodology for optimal performance.

Looking Forward: The iterative nature of model development is emphasized as a key takeaway. The project recognizes that continuous refinement and adaptation are integral to staying relevant in the dynamic world of finance. Considerations for feature engineering, dynamic asset allocation, risk management, and external data integration highlight a forward-thinking approach to further enhance the model's predictive power and applicability.

Limitations and Considerations: Acknowledging the limitations of linear regression in capturing complex market dynamics demonstrates a realistic understanding of the model's constraints. The project's focus on a subset of asset classes suggests a deliberate choice and points towards potential future expansions to include a broader array of assets for a more comprehensive and diversified perspective.

Continuous Iteration: The recognition of the dynamic nature of the financial landscape is a key takeaway. The project's conclusion is framed as a milestone rather than an endpoint, emphasizing the necessity for continuous iteration and adaptation of models to align with evolving market conditions and investor preferences. This iterative mindset positions the project as an ongoing exploration rather than a finite endeavor.

In essence, the project's key takeaways collectively contribute valuable insights to the field of investment portfolio optimization. By laying the groundwork for future enhancements, exploring advanced modeling techniques, and maintaining a forward-looking perspective, the project establishes itself as a significant contribution to the ongoing journey of refinement and adaptation in the ever-changing landscape of financial markets.

References-

https://www.kaggle.com/code/kashnitsky/a4-demo-linear-regression-as-optimization
https://smartasset.com/investing/guide-portfolio-optimization-strategies