8thFinal3 (1).pdf

by Aritra Ray

Submission date: 29-Apr-2025 01:01PM (UTC+0530)
Submission ID: 2532796959
File name: 8thFinal3_1_.pdf (1.02M)
Word count: 2896
Character count: 15945

Chapter 1

1.1 Introduction

This section provides an overview of the project, i.e., motivation, objectives, and scope. Optimizing portfolios has been one of the key focuses of financial decision-making, where having a trade-off between risk and return through diversification is of the utmost importance. For the project, the application of Modern Portfolio Theory (MPT) in real-world practice is used to create a optimized mutual fund portfolio system. Lastly, the current study is an improvement and extension of my previous project, which had created an initial framework based on standard MPT. Building further from that, the current project integrates more advanced time series forecasting models, i.e., seasonal ARIMA, in an attempt to make more accurate projections of future returns, thereby easing some of the limitations faced in the previous system.

Literature Survey A comprehensive survey of the available literature on portfolio optimization and financial modeling was achieved. Earlier research has examined an array of methods, such as heuristic methods, dynamic programming, and conventional optimization models, with special emphasis on the Markowitz model. For instance, research by Markowitz (1952) work discusses the foundations of portfolio selection, laying down the concept of achieving equilibrium between risk and return by means of diversification. Additionally, the work done by Elton and Gruber (1976) discusses the efficacy of portfolio selection in synthesizing both the Markowitz and Sharpe models. DeMiguel's (2009) study checks the empirical difficulties associated with portfolio diversification, highlighting the weaknesses of conventional models when applied to actual data. Collectively, these studies emphasize the need for a flexible, data-oriented approach that can be adapted to suit the needs of individual investors' interests—a task that my earlier project commenced and which is further developed in the present study.

Objective The main goal of this project is to create a mutual fund portfolio optimization system. The system seeks to balance return and risk as per user-specified constraints including risk tolerance, asset class, and investment size. By integrating sophisticated forecasting models to predict future returns with the conventional MPT paradigm, the project seeks to enhance the accuracy of diversified minimum-risk portfolio recommendations. This is an improvement on the original project that gives users actionable fund selection and asset allocation advice to facilitate them to make informed investment choices.

Scope This research maximizes the diversification of the mutual fund portfolios with the Markowitz MPT model by employing historical NAV and fund-specific data to precise scale-invariant risk-return estimation and ranking of the funds. A continuation of my earlier work, this research utilizes the most advanced machine learning algorithms to suggest a scalable data-driven portfolio recommendation platform for Indian mutual funds, with wider finance implications.

1.2 Methodology

The objective of this methodology is to optimize mutual fund portfolios using an enhanced version of Markowitz's Modern Portfolio Theory (MPT). This approach integrates historical Net Asset Value (NAV) data, machine learning techniques, and a genetic algorithm to construct a diversified portfolio tailored to the user's risk preference.

1.2.1 Data Preparation and Integration

1.2.1.1 Main Dataset

The primary dataset consist of 38 columns that include MF-related attributes such as risk measures and performance indicators. But the data is lacking in scheme codes. In order to meet this, a fuzzy logic search is used to infer and add scheme codes from a scheme code vs fund name data set, completeness assured.

Sample of Key Dataset Columns:

Column	Description
Scheme Code	Unique identifier inferred via fuzzy logic search
Funds	Official fund name
Fund Manager	Name of the managing entity or individual
Category	Categorical tag (e.g., EQ-L&M, DT-CR, GOLD-ETF)
AUM (in Rs. cr)	Assets Under Management, in crores
Expense Ratio (%)	Annual expense as a percentage of AUM
Return (%) 1 mo	Fund return over the past one month
Return (%) 1 yr	Fund return over the past one year
52 Week High (NAV)	Highest NAV in the last 52 weeks
52 Week Low (NAV)	Lowest NAV in the last 52 weeks
Inception Date	Date on which the fund was launched
Benchmark Index	Reference index for performance comparison
Fund Type	Open/Closed-ended designation

1.2.1.2 Correlation Matrix Dataset

A standalone CSV file presents MF schema code-related correlation coefficients. This dataset is later used in portfolio diversification and optimization.

1.2.2 User Input Parameters

To make the portfolio personalized, the user defines the following parameters:

- **n**: The number of mutual funds to be used in the portfolio.
- **Forecast Date**: A user-defined future date (in `YYYY-MM-DD` format) for which the NAV-based returns will be forecasted using time series modeling.
- **Risk Amount:** The investor's risk tendency (e.g., low, moderate, high), which is used to filter funds based on clustering criteria.

1.2.3 Initial Clustering and Risk Profiling

To gain insights into the structure of the dataset and to facilitate risk-based filtering, two clustering techniques are applied:

1.2.3.1 Clustering Based on Structural Features

- **Features:** Assets Under Management (AUM) and Expense Ratio.
- **Observation:** In the PCA plot, the left-hand cluster is predominantly composed of debt funds (low risk), the right-hand cluster mainly includes equity funds (high risk), and the intermediate region corresponds to hybrid funds or fund-of-funds (moderate risk).

- **Interpretation of PCA Axes:** The PCA axes represent principal components derived from linear combinations of the input features. *PCA Component 1* largely captures the variance associated with return profiles and category distinctions (e.g., equity vs debt), while *PCA Component 2* explains secondary variations, including fund size (AUM) and Expense Ratio. Thus, the separation along these axes reflects differences in both return-based performance and structural characteristics of the funds.

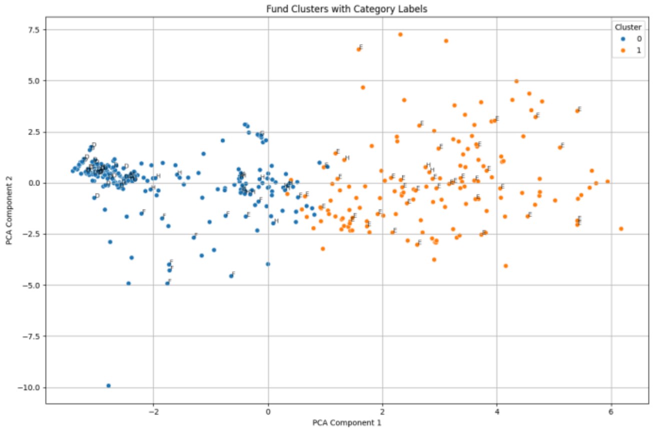

FIGURE 1.1: Equity-Debt Allocation Breakdown

1.2.3.2 Clustering Based on Historical Returns

Using historical returns (from 1 month to 10 years), the funds are clustered into three groups using t-SNE. The following table summarizes the cluster-wise risk profile based on average return and volatility:

Cluster	Avg Return (%)	Avg Volatility (%)	Risk Level
0	7.71	2.80	Low Risk
1	26.07	14.75	High Risk
2	14.63	8.13	Moderate Risk

Interpretation of t-SNE Axes: The t-SNE components represent nonlinear projections of the original return vectors into a two-dimensional space that preserves local structure. While the axes themselves do not have fixed, interpretable meanings like in PCA, funds that are close together in this plot share similar return trajectories across time horizons. *t-SNE Component 1 and Component 2* can thus be understood as abstract axes encoding patterns in historical performance — such as consistently high returns, stability, or volatility over time.

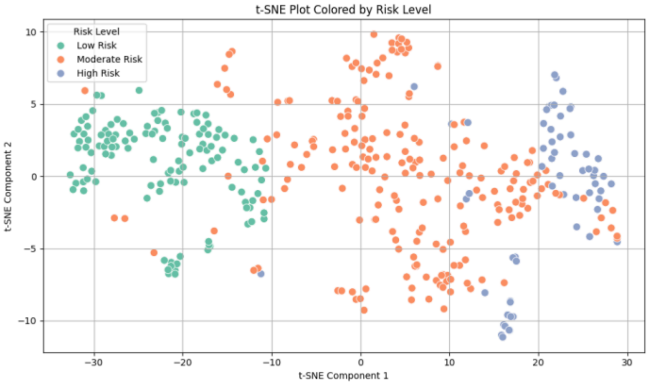

FIGURE 1.2: Return Based Clustering

Based on this clustering, the investor's input on risk preference (low, moderate, or high) is mapped to the corresponding cluster(s). For example, if the user selects "Moderate Risk", the portfolio optimization will only consider funds from Cluster 2.

1.2.4 Historical Data Retrieval and Time Series Modeling

For each selected fund, an API call (e.g., `https://api.mfapi.in/mf/147931`) is used to fetch historical Net Asset Value (NAV) data.

The retrieved data undergoes several preprocessing steps:

- Dates and NAV values are parsed and standardized.
- The NAV series is resampled to month-end values to maintain consistency.
- Monthly **log-returns** are computed to stabilize variance.

Based on the user-specified forecast date, the number of months ahead to forecast is calculated relative to the last available NAV data point.

A Seasonal ARIMA (SARIMA) model with configuration $(1, 1, 1) \times (1, 1, 1, 12)$ is then fitted to the historical log-returns to capture both short-term dependencies and yearly seasonality.

Using the fitted model:

- Future monthly log-returns are forecasted.
- The cumulative product of forecasted returns is computed to obtain the expected percentage return over the forecast horizon.

The final output is a vector of predicted returns, denoted as μ_{pred}, which serves as an input for the portfolio optimization phase.

1.2.5 Optimization Formulation

Objective Function:

$$\min_{x,w} \quad f(x, w) = \lambda_1 \cdot \text{AvgCorr}(x) + \lambda_2 \cdot w^T C w \tag{1.1}$$

Subject to:

$$\sum_{i=1}^{N} x_i = n_{\text{select}}, \quad x_i \in \{0,1\}, \quad \forall i \in \{1,\ldots,N\} \quad \text{(Selection constraint)} \tag{1.2}$$

$$\sum_{i=1}^{N} w_i = 1, \quad w_i \geq 0, \quad \forall i \in \{1,\ldots,N\} \quad \text{(Weight constraint)} \tag{1.3}$$

$$x_i = 0 \Rightarrow w_i = 0, \quad \forall i \in \{1,\ldots,N\} \quad \text{(Non-selection weight constraint)} \tag{1.4}$$

$$w^T \mu_{\text{pred}} \geq R_{\text{target}} \quad \text{(Return constraint)} \tag{1.5}$$

Average Pairwise Correlation:

$$\text{AvgCorr}(x) = \frac{\sum_{i \neq j} C_{ij} x_i x_j}{\sum_{i \neq j} x_i x_j} \tag{1.6}$$

where:

- N is the total number of funds.
- n_{select} is the number of funds to be selected.
- x_i is a binary variable indicating whether fund i is selected $(x_i = 1)$ or not $(x_i = 0)$.
- w_i represents the weight of fund i in the portfolio.
- C is the covariance matrix of the funds.
- μ_{pred} is the predicted return vector according to the SARIMA model.
- R_{target} is the minimum required portfolio return.
- λ_1, λ_2 are weight parameters for average correlation and variance, respectively.

1.2.5.1 Genetic Algorithm Implementation

The portfolio optimization problem was solved using a Genetic Algorithm (GA) with chromosomes comprising fund selection vectors and allocation weights. The fitness function minimized a weighted sum of portfolio variance and average correlation, with weights λ_1 and λ_2 based on the risk profile. Constraints ensured the selection of exactly n_{select} funds, weight normalization, and a minimum expected return R_{target}. The GA was run for 10,000 generations with a population size of 30, using single-point crossover and swap mutation to evolve solutions.

1.3 Results

This section details the outcomes of our optimization framework across three investor risk profiles—Low, Moderate, and High. We begin with outlining the key parameters—namely the target return (R_{target}) and the objective weights (λ_1, λ_2)—that guide the Genetic Algorithm. We then showcase the selected funds and their exact allocation weights in tabular form, followed by two intuitive visualizations: a pie chart illustrating individual fund weightings and another showing the overall equity versus debt breakdown. Finally, we conclude with key allocation insights that tie together performance, diversification, and risk-return trade-offs.

1.3.1 Parameter Selection and Allocation Rationale

To tailor the optimization to different risk appetites, we derived key parameters as follows:

Target Return R_{target} Based on clustering by historical returns (1 mo–10 yr), we obtained cluster average 1yr returns:

Cluster	**Avg Return (%)**	**Risk Level**
0	7.71	Low Risk
1	26.07	High Risk
2	14.63	Moderate Risk

We then set

$$R_{\text{target}} = \begin{cases} 0.0771, & \text{Low Risk,} \\ 0.1463, & \text{Moderate Risk,} \\ 0.2607, & \text{High Risk.} \end{cases}$$

Objective Weights λ_1 and λ_2 The objective

$$\min_{x,w} \lambda_1 \,\mathrm{AvgCorr}(x) \; + \; \lambda_2 \, w^\top C w$$

balances *diversification* (AvgCorr) against *variance* (risk). We chose:

$$(\lambda_1, \lambda_2) = \begin{cases} (0.1,\, 0.9), & \text{Low Risk (minimize variance)}, \\ (0.5,\, 0.5), & \text{Moderate Risk (balanced)}, \\ (0.9,\, 0.1), & \text{High Risk (maximize diversification)}. \end{cases}$$

1.3.2 Low Risk Profile

1.3.2.1 Selected Funds and Allocations

Index	Funds	Category	Weight
1	HSBC Banking and PSU Debt Fund-Reg(G)	DT-B&PSU	0.2278
2	Invesco India Contra Fund(G)	HY-ARB	0.0319
3	JM Arbitrage Fund(G)	HY-ARB	0.2906
4	Kotak Global Emerging Mkt Fund(G)	FOF-OVR	0.1329
5	SBI Liquid Fund-Reg(G)	DT-LIQ	0.3167

Table 1.1: Selected Funds for Low Risk Profile

- *Expected Return* : 8.66%
- *Portfolio Variance* : 0.898

1.3.2.2 Weight Allocation Pie Chart

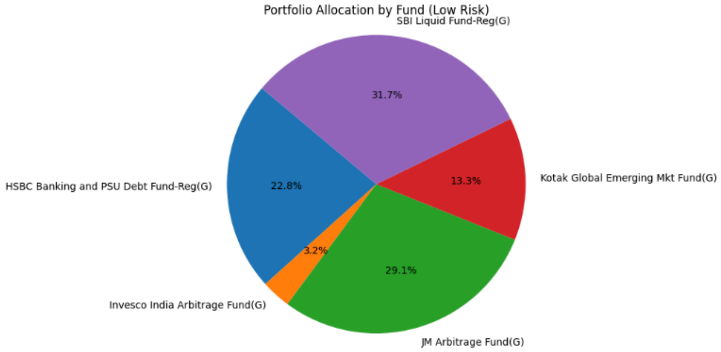

Figure 1.3: Weight Allocation among Selected Funds (Low Risk)

1.3.2.3 Equity vs. Debt Allocation

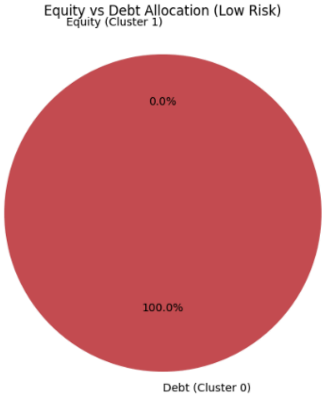

Figure 1.4: Equity vs Debt Asset Allocation (Low Risk)

1.3.3 Moderate Risk Profile

1.3.3.1 Selected Funds and Allocations

Index	Funds	Category	Weight
1	Axis Gold ETF	GOLD-ETF	0.0686
2	DSP Credit Risk Fund-Reg(G)	DT-CR	0.4448
3	Invesco India Contra Fund(G)	EQ-VAL	0.1112
4	Invesco India Financial Services Fund(G)	EQ-FIN	0.1450
5	Quantum Gold Fund ETF	GOLD-ETF	0.2304

Table 1.2: Selected Funds for Moderate Risk Profile

- *Expected Return* : 22.26%
- *Portfolio Variance* : 0.434

1.3.3.2 Weight Allocation Pie Chart

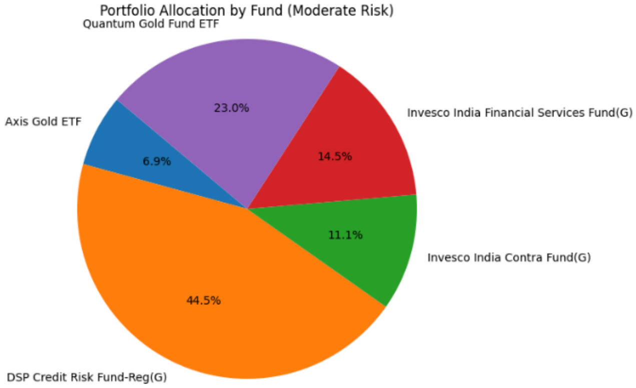

Figure 1.5: Weight Allocation among Selected Funds (Moderate Risk)

1.3.3.3 Equity vs. Debt Allocation

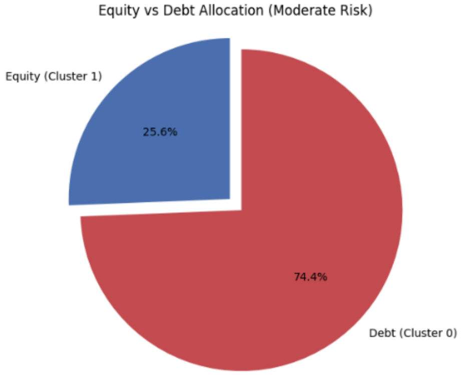

FIGURE 1.6: Equity vs Debt Asset Allocation (Moderate Risk)

1.3.4 High Risk Profile

1.3.4.1 Selected Funds and Allocations

Index	Funds	Category	Weight
1	DSP Small Cap Fund-Reg(G)	EQ-SML	0.0554
2	Nippon India Retirement Fund–Wealth Creation(G)	HY-SOL	0.0475
3	Quant Large & Mid Cap Fund(G)	EQ-L&M	0.3812
4	SBI Magnum Midcap Fund-Reg(G)	EQ-MID	0.2569
5	SBI PSU Fund-Reg(G)	EQ-THEM	0.2589

TABLE 1.3: Selected Funds for High Risk Profile

- *Expected Return* : 33.97%
- *Portfolio Variance* : 0.951

1.3.4.2 Weight Allocation Pie Chart

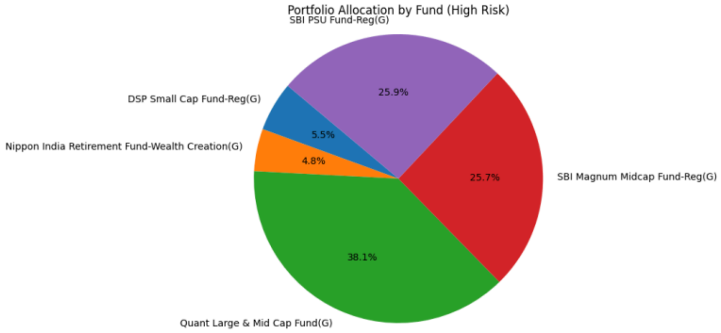

Figure 1.7: Weight Allocation among Selected Funds (High Risk)

1.3.4.3 Equity vs. Debt Allocation

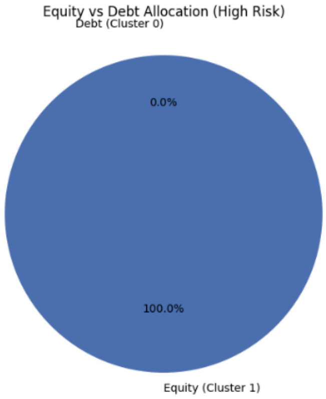

Figure 1.8: Equity vs Debt Asset Allocation (High Risk)

1.3.5 Key Allocation Insights

Across all three risk profiles, the optimized portfolios exhibit expected behavior: the Low Risk portfolio achieves a modest return with low portfolio variance, ensuring capital preservation; the Moderate Risk portfolio generates a higher return while maintaining relatively low variance through diversified asset allocation; and the High Risk portfolio delivers the highest return, accompanied by the highest variance, as expected from a fully equity-oriented strategy. Importantly, in all cases, the realized expected returns exceed their respective target returns R_{target}, thereby validating the effectiveness of our optimization framework in aligning with investor objectives.

After reviewing the optimized portfolios for all three risk profiles, we observe clear, intuitive asset-class tilts:

- **Low Risk**: Allocated exclusively to debt and arbitrage funds (100% debt), ensuring capital preservation and minimal volatility.
- **Moderate Risk**: Split between 74.4% debt (credit-risk and arbitrage) and 25.6% equity, balancing stable income with growth potential.
- **High Risk**: Invested entirely in equity funds (100% equity), maximizing exposure to higher-volatility, higher-return assets.

These allocation patterns demonstrate that our Genetic Algorithm—driven by the profile-specific target returns R_{target} and weights (λ_1, λ_2)—effectively:

- Achieves or surpasses each cluster's historical average return.
- Balances diversification and risk according to the chosen (λ_1, λ_2) trade-offs.
- Produces portfolios whose asset-class compositions align precisely with the investor's stated risk preference.

1.4 Conclusion

This project bridges the gap between theoretical Modern Portfolio Theory and practical fund selection by integrating advanced time-series forecasting, clustering analytics, and a Genetic Algorithm into a unified system. By distilling the mutual fund universe into risk buckets—using structural (AUM and expense ratio) and performance (1 month–10 year returns) clustering—concrete, data-backed targets for returns and trade-off weights were established. Seasonal ARIMA modeling provided forward-looking return estimates, enriching the Markowitz framework with predictive power. Finally, a PyGAD-powered Genetic Algorithm generated bespoke portfolios aligned with different risk profiles, demonstrating intuitive asset-class allocations: 100% debt for conservative investors, a 74.4/25.6 debt-equity mix for moderate investors, and 100% equity for aggressive investors.

Beyond optimization, the system emphasizes transparency and interpretability. Investors are clearly informed about fund selection logic, weight distribution, and diversification trade-offs through intuitive visualizations such as pie charts and summary tables.

Future extensions could include dynamic user preferences (e.g., ESG criteria, sectoral views), advanced forecasting models (e.g., Prophet, LSTM), and real-time re-optimization. Ultimately, the system can be fully deployed as a web-based platform, offering users an interactive and personalized mutual fund advisory experience.

Appendix A

Appendix

A.1 SARIMAX Forecasting Model Parameters

The key parameters used in SARIMAX modeling of monthly NAVs were tuned separately for each fund to capture seasonality and trend. General parameter ranges explored:

- **p, d, q**: Autoregressive, differencing, and moving average orders.
- **P, D, Q, s**: Seasonal counterparts with a fixed seasonality $s = 12$ (monthly data).
- **Exogenous Variables**: None included; purely univariate models.

Final model selection was based on minimizing the AIC (Akaike Information Criterion) while ensuring residuals approximate white noise.

A.2 PyGAD Genetic Algorithm Settings

The Genetic Algorithm was implemented using the `pygad` Python library with the following core settings:

- **Population Size**: 300 chromosomes
- **Gene Space**:
  - Binary selection (x_i) for fund inclusion.
  - Continuous allocation weights (w_i) between 0 and 1.
- **Selection Method**: Steady-State Selection
- **Crossover**: Single-Point Crossover with probability 0.8
- **Mutation**: Random Mutation with probability 0.05
- **Stopping Criteria**: 200 generations or fitness convergence.

A.3 Data Preprocessing Steps

The following preprocessing was applied to raw mutual fund data:

- **Missing Data**: Rows with missing critical NAV or return fields were removed.
- **Scaling**: Features like AUM and Expense Ratio were standardized.
- **Outlier Treatment**: Funds with abnormal NAV fluctuations were excluded from forecasting models.

A.4 Potential Website Deployment Plan

To enhance accessibility and usability, the project framework can be extended into a web-based platform. The preliminary plan would involve:

- **Frontend**: Developed using ReactJS with visualization libraries (e.g., Chart.js) for portfolio charts and fund comparisons.
- **Backend**: A lightweight Flask or Express.js server to handle user inputs (risk appetite, horizon) and dynamically generate optimized portfolios.
- **Hosting**: Cloud deployment via platforms like Heroku, AWS, or Vercel for scalability.

This deployment would allow end-users to interactively generate and visualize customized mutual fund portfolios in real-time.

Bibliography

Chopra, V. K. and Ziemba, W. T. (1993). The effect of errors in the input data on optimization: A multi-portfolio comparison. *The Journal of Portfolio Management*, 19(2):6–11.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How efficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953.

Elton, E. J. and Gruber, M. J. (1976). Efficient portfolio selection: A synthesis of the markowitz and sharpe models. *Journal of Finance*, 31(5):1345–1365.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Merton, R. C. (1972). An analytic derivation of the efficient portfolio frontier. *The Journal of Financial and Quantitative Analysis*, 7(4):1851–1872.

8thFinal3 (1).pdf

ORIGINALITY REPORT

6%	4%	2%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	1lib.to Internet Source	2%
2	Submitted to Greenwood High Student Paper	1%
3	Submitted to Cranfield University Student Paper	<1%
4	Submitted to Great Lakes Institute of Management Student Paper	<1%
5	files.ele-math.com Internet Source	<1%
6	www.financialexpress.com Internet Source	<1%
7	Submitted to Erasmus University of Rotterdam Student Paper	<1%
8	masterstrokeonline.com Internet Source	<1%
9	iiste.org Internet Source	<1%
10	mdpi-res.com Internet Source	<1%
11	icmrr.org Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		