Sustainable Portfolio Construction: Clustering, Forecasting, and Optimization of NASDAQ-100 Stocks with ESG Integration

By

Gonzalo González

July 2025.

Abstract

This document addresses the gap in the literature related to two emerging trends: sustainable portfolio construction and research on the technology sector. It aims to integrate ESG factors with financial performance using machine learning techniques applied to NASDAQ-100 technology stocks to create sustainable portfolios.

A three-stage framework is developed: First, the stocks are clustered using four different combinations of variables, where the approach with the best silhouette score is chosen. Then, for each cluster, machine learning models are fitted to the data to forecast return and volatility, selecting the one with the lowest RMSE. Finally, with the forecasted values, different ESG-aware portfolios are obtained and compared with portfolios created without considering the ESG score, by applying a Monte Carlo simulation and obtaining the portfolio metrics under uncertainty.

The results showed that a simple approach, using historical returns, volatility, and volume values, generates the most homogeneous and interpretable clusters, with two groups. In forecasting, the SVM model is the best in predicting both returns, while XGBoost and Random Forest excel in forecasting volatility. Finally, the ESG-aware portfolios can sacrifice a small portion of return and volatility to improve the ESG score by a greater amount.

These findings propose a new framework to create ESG-aware portfolios based on large market indexes using clustering, machine learning forecasting, and optimization, offering alternatives to investors where a sustainable portfolio can be obtained without sacrificing financial performance.

Keywords: Sustainable finance, ESG investing, technology stocks, portfolio optimization, machine learning, clustering

Contents

List of Figures

List of Tables

1 Introduction

According to Thomson Reuters (2024), investors have increased interest in environmental, social, and governance (ESG) initiatives. Companies view them as a competitive advantage and have begun to incorporate them into their corporate strategies to attract investors and enhance their brand image. Another recent trend is the investment in the technology sector, according to Atomico and Invest Europe (2024), where stocks tend to have high returns and volatility (Morris et al. (2024)). While both trends may appeal to investors, they are often contradictory.

Currently, the technology sector is facing some challenges in ESG that need to be taken into account. Data centers consumed 200 terawatt-hours of electricity in 2022, more than some countries, like Argentina (Plan Be Eco (2024)). On the social side of the ESG spectrum, USC Viterbi School of Engineering (2022) showed that artificial intelligence responses are not free of bias, finding that between 3.4% and 38.6% of the data showed signs of prejudice when asked about gender, religion, race, and profession. These environmental and social concerns are joined by governance risks. According to Infosecurity Magazine (2025), the Law Firm LA Piper recorded €1.2 billion of General Data Protection Regulation fines across Europe in 2024, and €2.9 billion in 2023.

All of these ESG risks can reduce the value of the company if appropriate decisions are not made to mitigate them. Therefore, risk-tolerant and ESG-focused investors need a framework to obtain an optimal portfolio with technology companies, such as the components of the NASDAQ-100 index, that incorporates ESG-related risks, to create a profitable and sustainable investment alternative.

The dominant theory in portfolio selection has been the Modern Portfolio Theory (MPT), developed by Markowitz (1952), which seeks maximum return for a given level of risk, or minimum risk for a given return. Even though ESG was not integrated, it laid the groundwork for integrating MPT with ESG factors. For example, Torricelli et al. (2022) included ESG scores to get optimal portfolios from the EURO STOXX Index with a high level of sustainability, without necessarily sacrificing the return on investment.

To improve the performance of the portfolio, forecasting models can be developed to predict returns and volatility (Allen et al. (2019)). The ARIMA, for returns, and GARCH,

for volatility, were the academic benchmark, using only lagged values as predictors (Hossain et al. (2015)), which can be limiting because they fail to account for external information. It has been shown how machine learning algorithms tend to outperform traditional models (Jin (2024)), given their ability to capture complex patterns in the data. Models like Neural Networks, XGBoost, and Random Forest were fitted to predict returns of stocks in the New York Stock Exchange (Chin et al. (2022)), as well as the volatility of stocks in the S&P500 (Zhang et al. (2024)).

A practical challenge in applying forecasting models across large portfolios is scalability, because for each stock, multiple models will be fitted to the stock's return and volatility. Training a separate model for each stock is time-consuming and computationally expensive. As shown by Li et al. (2023), clustering stocks and fitting the models to the cluster can increase the forecast accuracy. When using the K-means algorithm to cluster stocks, accuracy metrics, such as the Root Mean Squared Error (RMSE), are decreased by between 0.02 and 0.4 units compared to the forecast without clustering, which is equivalent to a range between 18% and 63% error decrease.

However, no study analyzes the NASDAQ-100 index components and the ESG-integrated portfolio. Existing research focuses on bigger indexes, like the S&P500, or regional indexes, such as EURO STOXX. As a result, the technology sector remains unexplored in the context of investment strategies with ESG factors. This creates a gap in understanding how ESG criteria affect portfolio construction and performance in a high-return and high-volatility sector, like the NASDAQ-100.

Additionally, no approach uses portfolio optimization, enhanced by return and volatility forecasting, to obtain an accurately predicted portfolio, which is improved by clustering to reduce computational costs. Studies tend to either use them individually, or forecast stock returns and volatility to optimize a portfolio, or cluster stocks to improve forecasting accuracy. Therefore, the possible interactions between these approaches have not been investigated.

These two gaps motivate this study to highlight the missing intersections in the literature between clustering, forecasting, and portfolio optimization techniques, with the NASDAQ-100 components data and the ESG factors. For ESG-focused investors, there is value in obtaining portfolios considering ESG scores from the NASDAQ-100 index. However, to get the best of this technique, the stocks should be clustered, and then their return and volatility should be forecasted for an accurate approach. This way, a highly sustainable portfolio, with desired levels of returns and volatility, can be constructed for investors interested in ESG.

From an academic perspective, bridging the gap between these areas of study contributes to the growing literature on sustainable finance. It proposes an integrated

methodology that combines clustering, forecasting, and optimization to construct portfolios based on extensive stock indexes, including ESG factors.

To address these gaps, the present study sets the following objective: To construct and evaluate investment portfolios composed of NASDAQ-100 technology stocks that achieve high ESG standards while also delivering strong financial performance in terms of return, volatility, and risk-adjusted return. Specifically, the study aims to cluster NASDAQ-100 technology stocks based on historic return, volatility, and volume values, as well as fundamental analysis indicators, and to analyze the financial behavior of each resulting group. It also seeks to forecast these stocks' rolling returns and volatility by fitting machine learning models that use lagged values and technical analysis indicators. Finally, the study aims to construct optimized portfolios using criteria such as return, volatility, Sharpe ratio, and ESG scores, and to compare their performance with benchmark portfolios, through descriptive statistics and financial metrics.

These objectives lead to the following research question:

> *Is it possible to predict returns and volatility using machine learning models, and then construct an investment portfolio with optimization techniques, based on technology sector stocks, using return, volatility, and ESG scores, that still outperforms the market?*

The research question addresses the two gaps identified in the literature in a unified framework: the lack of ESG-integrated portfolios focused on the NASDAQ-100 index, and the absence of research that combines clustering, forecasting, and portfolio optimization. By asking whether it is possible to construct a portfolio using return and volatility forecasts enhanced by clustering techniques, and constrained by ESG factors, the research responds to the practical and academic needs stated before.

Concretely, the stocks in the NASDAQ-100 index will be clustered using K-means constrained, with different combinations of fundamental analysis indicators, and return, volatility, and volume values. These clusters will be compared to a cluster made by the principal components, keeping at least 80% of the variance, and the best one in terms of silhouette score will be selected. Then, for each cluster, four different machine learning models will be fit to the return and volatility: Random Forest, XGBoost, Neural Networks, and Support Vector Machine, with technical analysis indicators, and lagged values of return, volatility, and volume used as predictors. The model with the smallest root mean squared error will be selected for each case, and its parameters will be used to fit the model with the complete data and to forecast each stock in the cluster. Finally, four portfolios will be created: maximum return, minimum variance, maximum Sharpe ratio, and minimum ESG risk score, which will be compared with portfolios that do not

consider ESG variables. The stability of the portfolio will be evaluated with a Monte Carlo simulation. The output will be the recommendations of optimal asset allocation for specific profiles of investors.

The results of this research contribute to the dominant theory in portfolio management by adding ESG factors. Traditionally, MPT does not consider this variable, neither in the objective function nor in the constraints. This will align with the growing demand for sustainable investment frameworks and modernize MPT to accommodate ethical issues.

In terms of methodology, this study offers a structure that combines clustering, machine learning time series forecasting, and portfolio optimization, an approach that is not found in the current literature. This integration can improve forecast accuracy, reduce computational cost, and create a tailored portfolio. By applying this to an extensive index, such as the NASDAQ-100, the need for scalable investment strategies in a high-volatility sector is addressed. This can be used for other indices, offering practical tools for ESG-interested investors and contributing to sustainable finance.

This study finds that it is possible to predict returns and volatility using machine learning models, and then construct an investment portfolio with optimization techniques, based on technology sector stocks, such as NASDAQ-100 stocks, using return, volatility, and ESG scores. Specifically, historical return, volatility, and volume values can create high-performing and interpretable clusters, such as a group with high return, volatility, and volume and a cluster with low values. Then, Support Vector Machine with a third-degree polynomial trend is the most accurate model to forecast both clusters' return, given its ability to capture non-linear patterns. At the same time, the tree-based models, such as Random Forest and XGBoost, are better for volatility when data is more structured and predictable. Finally, when the ESG risk score is integrated to construct a portfolio, the result is a portfolio that performs closely to the unconstrained one, sacrificing a small proportion of return and volatility, to improve its ESG performance.

2 Literature Review

The objective of this chapter is to review the existing literature related to the main steps proposed in this document. The first section focuses on stock clustering, discussing its significance in financial analysis, the algorithms used, and the relevant variables employed as predictors. The second section is about stock forecasting, highlighting its role in portfolio construction, the performance of ML models compared with statistical models, and the predictor variables. Then, the third section shows how portfolio optimization with ESG factors is a new and underdeveloped trend, optimizing portfolios with different goals and checking the stability of the optimal allocation. Finally, the last section relates to how the NASD-100 index is integrated with clustering, forecasting, and optimization.

2.1 Stock Clustering

Clustering techniques have emerged as a valuable tool in financial analysis, particularly for simplifying the forecasting of returns and volatilities in large markets. Given that indices such as the S&P500 include hundreds of companies, clustering can reduce the computational cost by grouping the datapoints in representative groups. Moreover, clustering can enhance forecasting accuracy, as shown by Li et al. (2023). Accuracy metrics, such as the Root Mean Squared Error (RMSE), are decreased by between 0.02 and 0.4 units compared to the forecast without clustering, which is equivalent to a range between 18% and 63% error decrease.

Hierarchical clustering (Miccichè et al. (2005)) was the first method used to group stocks. It builds a dendrogram where the datapoints start as a one-member cluster, and then they are grouped one by one, bringing together the closest two clusters in one set, until all the datapoints are in the same category. Mantegna (1999) showed how there is detectable value in clustering stocks from S&P500, by using this algorithm, finding three clusters, corresponding to the energy, raw materials, and consumer goods sectors.

K-Means is an iterative algorithm that assigns datapoints to the closest centroid until the within-cluster sum of squares is not improved (Arthur & Vassilvitskii (2006), or a

certain number of iterations is reached. This algorithm was applied by Ghosh (2025), aiming to cluster 49 stocks in the Indian market, identifying five clusters that are more characterized by their market movements than by their sectors.

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Huang et al. (2019)) groups datapoints that are part of dense regions, detecting noise by labelling low-density points as outliers. Rukmi et al. (2019), clustered companies in the Indonesian Stock Exchange using the DBSCAN algorithm, finding 24 clusters and 372 outliers.

Comparing these three algorithms, K-Means is widely used in the literature compared to DBSCAN and hierarchical clustering. Suganthi & Kamalakannan (2015) showed how this algorithm excels in its simplicity and velocity, making it more suitable for stock market data mining applications.

Regarding features used for clustering, fundamental analysis variables have been applied in many studies. These variables are obtained from the financial statements of a company and show how it performs in terms of profitability, liquidity, and solvency. Boloș et al. (2025) worked with the S&P500 index, utilizing profitability ratios (such as ROA, ROE, and EBITDA), liquidity measurements (Cash and Current ratios), and solvency indexes (Debt-to-EBITDA, Asset-to-Equity). The optimal number of clusters, given by a silhouette score of 0.9, was two groups.

In a more straightforward approach regarding the variables, He et al. (2023) used daily returns and volatility of the stocks of Ping An Bank. Safari-Monjeghtapeh & Esmaeilpour (2024) extended this approach by clustering companies listed on the Tehran Stock Exchange across multiple time windows, from weekly to annual. Their findings suggest that the longer the windows, the more stable and informative the clustering results.

However, when multiple variables are included in the analysis, the model becomes computationally expensive, and it is easier to have highly correlated inputs. Using dimensionality reduction techniques, like Principal Components Analysis (PCA), can be helpful to reduce the number of variables while keeping their variability (Jolliffe & Cadima (2016)). This algorithm uses linear algebra to transform the dataset with potentially correlated variables into a new set of uncorrelated predictors.Agarwal et al. (2021) clustered NYSE stocks based on daily price and used PCA to retain 99% of the total information before clustering.

To compare different clustering approaches, the silhouette method is the standard technique (Rousseeuw (1987)). It measures how likely it is for a record to be misclassified in its cluster, compared to the closest centroid. Ridwan et al. (2022) compared

K-means and average linkage clustering on the Indonesian market index IDX80. The quality of the clusters was evaluated using the silhouette score, and it was found that K-Means outperformed the average linkage in terms of the silhouette score.

In terms of clustering, empirical comparisons indicate that K-Means tends to be the best-performing algorithm for stocks. It is important to develop different variable approaches, with fundamental analysis indicators, stock characteristics, such as return, volatility, and volume, a mixture of both, and the use of PCA to reduce dimensions. All of these must be compared using the silhouette score, taking care that the average value is over 0.25, and the score of the datapoints should be positive in all cases.

2.2 Stock Forecasting

Forecasting expected returns and volatility is crucial for effective portfolio optimization, as shown by Allen et al. (2019). It is important to highlight the different types of models, such as statistical-based, machine learning-based, and hybrid approaches, as well as exogenous variables, like lagged returns, volatility, volume values, and technical analysis indicators. In this section, the models and variables mentioned above will be described to address the research question innovatively and consistently with the literature approach.

Models are compared using accuracy metrics, which can be scale-dependent or non-scale-dependent. In the first class, the Root Mean Squared Error (RMSE) returns a value on the same scale as the predicted variable, making it easier to interpret. On the other hand, the Mean Absolute Percentage Error (MAPE) is measured in percentage, relative to the actual values, which can distort the results when the scale is small (Shcherbakov et al. (2013)). As stated by Mallikarjuna & Rao (2019), the RMSE is more suitable when working with stock returns.

Additionally, to get the models with the best parameters, a grid and randomized search can be performed using time series cross-validation. The first search looks into every possible combination, which can be computationally expensive if the parameter space is large. For cases like this, a randomized search will look into a random combination of parameters. Zhou (2021) shows how it is possible to improve the accuracy of an ML model using this technique.

Naïve or dummy models are used as a benchmark in forecasting (Shmueli et al. (2019)), which is important with stocks, given their unpredictability. The most common values used as naïve are the mean, the first lag, or the seasonal lag. Beck et al. (2025) showed how even complex models, such as Long Short-Term Memory Neural Networks, fail to beat the benchmark when forecasting exchange rates, NASDAQ-100

stocks, and macroeconomic price indicators.

To mitigate the unpredictability of stock time series, the curve can be smoothed using rolling windows. Instead of forecasting the outcome variable in the next period, using the current one, rolling windows use the average of a time window (that can be a week, a month, or even a year) in a forward way. Feng et al. (2024) found that rolling and expanding windows improve the predictive power of models, making the volatility forecasts more accurate for the S&P500 index.

The Auto Regressive Integrated Moving Average (ARIMA) model (Nethaji et al. (2024)) has been widely used for stock forecasting. It takes three parameters: the number of lags used as predictors (AR component), the number of differences taken from the value and their lags (I component), and the number of residual lags (MA component). Li (2024), used ARIMA models to forecast a stable market index (S&P500) and a more unstable one (CSI300), showing that these types of models are suitable for stable time series, which translates into low volatility.

The General Auto Regressive Conditional Heteroskedasticity (GARCH) model (Celestin et al. (2025)) is the benchmark for volatility forecasting, often combined with ARIMA to model returns and volatility jointly. It uses the number of lagged conditional variances and the number of lagged squared residuals as parameters. Odah (2025) successfully forecasted Apple Inc. stock volatility using this model, specifically, with GARCH(0,1) being the most efficient in capturing the dynamic patterns of this asset.

Machine learning models have outperformed statistical models, such as ARIMA, for stock returns and volatility. Using Support Vector Machine (Drucker et al. (1996), a model that uses linear algebra to transform the data into a high-dimensional space to fit data with non-linear behaviour, Santos Bezerra & Melo Albuquerque (2019) showed how this model can have an RMSE two orders of magnitude smaller than GARCH models, when predicting the volatility of the S&P500 and the Brazilian Ibovespa. The predictors used were the first lag of the volatility.

Another ML technique that is used for stock forecasting is Neural Networks, a model inspired by the human brain, where neurons are connected, starting from the input neurons (predictors), until it gets to the output neurons (outcome) (Sazlı (2006)). Sun et al. (2025) suggests that even simple feed-forward Neural Networks can perform better than advanced approaches, such as Transformers and Long Short Term Memory Neural Networks, in various prediction tasks.

Random Forest is another effective algorithm frequently used for stock market prediction (Breiman et al. (2017)), using multiple decision trees, taking the averages of their

predictions, to get a more accurate outcome. Azman et al. (2025) used price, return, and volume to fit a Random Forest model to multiple indices. This algorithm was also built with technical analysis variables, such as SMA, EMA, RSI, and Bollinger bands, to forecast the performance of SPY (Deep et al. (2025)).

Similarly, XGBoost, another tree-based algorithm that takes weak predictors to recursively forecast the residuals (Friedman (2001)), can also be deployed for time series forecasting using lagged returns and volatilities as predictors (Gifty & Li (2024)), along with other ML models. When comparing approaches, XGBoost had a lower RMSE and MAPE, which makes it a more accurate model for forecasting the Google stock market price.

In summary, the literature suggests that the most effective approach for stock return and volatility forecasting is to use the rolling window variables in a group of ML models, such as Neural Networks, Random Forest, XGBoost, and SVM, use time series cross-validation with a grid or randomized search, compare the models using the RMSE, and utilize lagged return, volatility, and volume values, with technical analysis indicators as predictors.

2.3 Optimal Portfolio

Portfolio optimization is an important part of investment management, where the goal is to obtain the optimal allocation proportion, given an investment goal, commonly to minimize volatility, maximize return, or a combination of both (Khan & Tanwani (2024)). In most cases, it uses optimization techniques, with models that can be expressed as linear programming if the objective function and constraints are linear, or non-linear programming if any of these are non-linear. However, an approach that considers ESG factors has not been developed on a large scale.

The minimum volatility portfolio is a method of obtaining the optimal allocation, given that the investment return is more certain. It considers the returns covariance matrix of the stocks in the study, and optionally, a constraint that ensures a minimum expected return. On the other hand, the maximum return portfolio, as the name implies, returns an asset allocation that ensures the optimal expected return on investment. The objective function contains the portfolio return, and, in the constraints, a maximum allowed volatility can be set. Kumar et al. (2022) showed the importance of constructing an optimum return and volatility portfolio, dominating a simple equal-weight asset allocation in terms of mean and variance in most cases.

A hybrid solution can be reached if the Sharpe ratio is used. This index shows how much additional unit of expected return the portfolio gets for each unit of volatility,

compared to a risk-free asset, ensuring a more balanced result in terms of return and volatility (Goetzmann et al. (2002)). The objective function contains the maximization of this index, and a maximum allowed volatility, with a minimum expected return that can be set to bound the solution space. Qu & Zhang (2023) used this model to create a portfolio with stocks with the most significant market capitalization on a group of industries, with this approach obtaining better results than the benchmarks in return, volatility, and total Sharpe index.

However, these models do not consider any ESG factor when constructing portfolios. One of the most renowned scores is the ESG risk rating provided by Morningstar Sustainalytics (2024), measuring how vulnerable a company is to ESG risk, and how it mitigates these risks. It comprises five categories: from 0 to 10, a company has negligible risk; from 10 to 20, it is low; from 20 to 30 is medium; from 30 to 40 is high, and over 40 is severe. The rating is composed of corporate and stakeholder governance, material ESG issues, and systemic ESG issues.

Incorporating ESG variables into portfolio optimization is a recent trend in academia. For example, Torricelli et al. (2022) explores what happens when the ESG score is placed as a constraint, having a minimum value allowed. They use a set of European stocks and ESG data from Refinitiv and show how, with this approach, portfolios tend to avoid investing in fossil fuel sectors at the expense of a slight drop in returns and Sharpe ratios. On the other hand, the study by Varmaz et al. (2024) focuses on integrating ESG directly into the objective function. This approach is more flexible, indicating how portfolios with moderate ESG scores can still perform competitively.

To evaluate portfolios, a Monte Carlo simulation can be conducted, where the returns follow a multivariate normal distribution, with a mean equal to the mean return, and a variance equal to the covariance matrix (Pedersen (2014)). This approach can assess how the portfolios will behave under uncertainty. Measures like Value at Risk and probability of loss (also in the literature as Limited Expected Loss) are used to compare the simulations. Gambrah & Pirvu (2014) found that the probability of loss is a more conservative metric when compared to VaR, generating very cautious portfolios.

The inclusion of ESG factors in portfolio optimization is relatively recent; therefore, more research needs to be done. A correct approach for this problem is to optimize a portfolio using four different models: maximum return, minimum volatility, maximum Sharpe Ratio, and minimum ESG score. This will ensure the availability of options for different investor profiles. The stability of the portfolio should be evaluated using a Monte Carlo simulation to ensure that, in a real-life scenario, the solution behaves as expected.

2.4 NASDAQ-100 studies

According to the latest NASDAQ report (NASDAQ, Inc. (2025)), the returns of the NASDAQ100 tend to be higher than the S&P500 returns, an index that is usually used to represent the market. This is accompanied by a higher volatility, which makes these stocks ideal for risk-tolerant investors seeking for investment opportunities with high returns. However, there are some concerns about the ESG performance of these companies. Cohen (2023) demonstrated that there are social issues in the NASDAQ-100 components that need to be mitigated to maximize their firm value.

On the other hand, the use of ML models with NASDAQ-100 data has been growing in recent years. Bulani et al. (2025) clustered stocks in the NASDAQ100 index using returns and volatility and applying the K-Medoids algorithm, similar to K-means, but using a real datapoint as a centroid. They obtained three groups, with an average silhouette score of 0.34. However, there were two distinct groups: the stocks with high return/ high volatility, and the stocks with low return/ low volatility.

For forecasting, Zhao (2025) built a Random Forest regressor model to predict returns and volatility of the NASDAQ-100 index, using fundamental analysis indicators, plus lagged return, volatility, and volume values, obtaining an R^2 value of 0.78 and 0.73 on the test set, and a mean squared error of 0.015 and 0.028, respectively for returns and volatilities. The authors remark that the error of the return model is low, and the error of the volatility model is reasonably intermediate, being this last value more influenced by external variables, such as macroeconomic events.

Abraham et al. (2004) compared the performance of the SVM against multiple neural network-based models on the NASDAQ-100 index, using past observations as predictors. The results demonstrate how the NN models outperform SVM on the training set, yet SVM obtains lower RMSE on the test data partition. This implies a possible overfitting of the NN models to the data, with the model following the noise on the training set, causing it to fail in the predictions on the test set.

Another study compared multiple tree boosting methods, such as LightGBM, XGBoost, CatBoost, and a weighted fusion model, combining the output of these tree algorithms (Huang (2024)). The findings suggest that the ensemble model outperforms the boosting methods, with a mean average error of 5.195, against the 5.197 of LightGBM, 5.199 of the CatBoost, and 5.202 of the XGBoost. However, since the scores are very close to each other, the author suggests that these models are suitable for NASDAQ-100 predictions.

Wu (2024) constructed a portfolio based on NASDAQ-100 stocks with a mixed objective of maximizing the return and minimizing the volatility, obtaining a portfolio com-

posed of 10 stocks, all of them with similar weights, where companies like Amazon, Google, Adobe, and Airbnb were part of it, and other companies, such as Microsoft and Apple were omitted.

Despite the growing literature on NASDAQ-100 modelling, notable gaps remain, the most important one being the lack of ESG factors in NASDAQ-100-based portfolio optimization. This can prevent investors who are interested in profitable and sustainable investments from investing in a high-return market. Furthermore, fundamental and technical analysis indicators have not been included as predictors for clustering and forecasting these stocks. The inclusion of these variables can improve the forecasting accuracy, impacting the precision of the portfolio.

2.5 Summary

In summary, the literature demonstrates the opportunity that investors have in a high-return market like the NASDAQ-100; however, there are concerns related to the ESG performance of its components. There are no studies that include the ESG factors in portfolio optimization, showing a gap in the literature that needs to be addressed.

When evaluating models in other markets, it has been shown how clustering can improve forecasting accuracy and reduce its complexity, using fundamental analysis indicators and stock descriptive statistics to group companies using the K-Means algorithm. With the outcomes of clustering, machine learning models can be fit to them, and these models can forecast returns and volatility of each stock inside the cluster, employing technical analysis indicators, lagged return, volatility, and volume values. With the predicted values, a portfolio optimization can be performed, using the ESG score from Morningstar Sustainalytics, maximizing returns and the Sharpe ratio, and minimizing volatility and ESG score. Each one of these steps should be under the CRISP-DM framework to ensure a structured, repeatable, and business-focused approach.

3 Methodology

The structure of this study is based on the CRISP-DM Methodology (Chapman et al. (1999)), and it will be used for the three stages: Clustering, Forecasting, and Optimization.

3.1 Clustering

3.1.1 Business understanding

Developing forecasting models for each stock on the NASDAQ-100 index can be computationally expensive and inefficient since, as stated before, clustering can enhance forecasting. The clustering section of this study aims to find the best way to group stocks based on fundamental analysis indicators and market indexes, such as return, volatility, and volume.

3.1.2 Data understanding

The data is obtained from the yfinance package in Python, an API that accesses the Yahoo Finance website to retrieve information about stocks, indexes, currencies, and other types of financial assets (Aroussi (2025)). The variables used in the analysis are the following:

- **Profitability ratios**
  - **Return on Asset (ROA)**: It measures how efficiently a company uses its assets to generate revenue. The bigger the ratio, the more efficient the company. It is computed in the following way:

$$ROA = \frac{\text{Net Income}}{\text{Total Assets}} \tag{3.1}$$

  - **Return on Equity (ROE)**: Ratio that shows how efficiently a company uses its equity to generate revenue. It is obtained as follows:

$$ROE = \frac{\text{Net Income}}{\text{Shareholder's Equity}} \tag{3.2}$$

  - **Operating Margin**: Measures the proportion of the operating income relative to the revenue:

$$\text{Operating Margin} = \frac{\text{Operating Income}}{\text{Revenue}} \tag{3.3}$$

  - **Earnings before interest, taxes, depreciation and amortization (EBITDA) margin**: Shows the proportion of earnings from the total revenue. It is computed as follows:

$$EBITDA\ Margin = \frac{\text{EBITDA}}{\text{Revenue}} \tag{3.4}$$

- **Liquidity ratios**
  - **Current Ratio**: Measures the ability of a company to cover short-term debt with current assets. The bigger the ratio, the more liquidity the company has. The ratio is computed as follows:

$$Current\ Ratio = \frac{\text{Current Assets}}{\text{Current Liabilities}} \tag{3.5}$$

  - **Quick Ratio**: Similar to Current Ratio, but stricter, given that the inventories are excluded from the current assets:

$$Quick\ Ratio = \frac{\text{Current Assets - Inventories}}{\text{Current Liabilities}} \tag{3.6}$$

- **Solvency ratio**
  - **Debt-to-Equity Ratio**: Shows how a company can cover its total debt with its equity. The lower the ratio, the more solvent the company is. It is obtained by the following equation:

$$Debt - to - equity\ Ratio = \frac{\text{Total Debt}}{Shareholder's Equity} \tag{3.7}$$

- **Daily stock prices and trading volumes**: These values are used to compute the historic return, volatility, and volume, from January 1st, 2015, to April 30th, 2025.

3.1.3 Data preparation

The daily stock prices are transformed using the logarithmic return equation:

$$R_t = ln(\frac{P_t}{P_{t-1}}) = ln(P_t) - ln(P_{t-1}) \tag{3.8}$$

P_t and R_t are the stock price and return on period t. Using this, two more variables are obtained:

- **Average daily return**: Measures the average rate of change in the stock price. It is computed using the following equation:

$$E(R_t) = \frac{\sum_{i=1}^{n} R_{t+1-i}}{n} \tag{3.9}$$

  Where n is the number of data points.

- **Historic volatility:** Is a measure of the spread of the returns, obtained as the standard deviation:

$$\sigma(R_t) = \sqrt{\frac{\sum_{i=1}^{n}(R_t - E(R_t))^2}{n-1}} \tag{3.10}$$

  Where σ_t is the volatility of the financial asset.

Using the daily trading volume values, the following variable is obtained:

- **Average daily trading volume**: Measures the average amount of daily transactions produced involving each stock:

$$\text{Average Volume} = \frac{\sum_{i=1}^{n} Volume}{n} \tag{3.11}$$

These variables will be scaled before fitting the clustering model, given that K-Means is a distance-based algorithm, and the different scales can influence the result.

3.1.4 Modeling

Four approaches will be compared in terms of input variables:

- Profitability, liquidity, and solvency ratios, plus overall return, volatility, and volume.
- Profitability, liquidity, and solvency ratios.
- Overall return, volatility, and volume.

- All the variables, using PCA, keeping 80% of the variance.

To avoid single-member and oversized clusters, the 'KMeansConstrained' function will be used (Levy-Kramer (2018)). The number of clusters will range from 2 to 10, with a minimum member size of 5 datapoints, a maximum of 50, fixing a seed for reproducibility, and comparing the silhouette scores.

3.1.5 Evaluation

The different approaches will be compared using the silhouette score, and the one that has the highest average value will be used to cluster the stocks. The silhouette score ranges from -1 (totally misclassified) to 1 (well-clustered). It is computed as follows:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{3.12}$$

Where $a(i)$ is the average distance from point i to all other points in the same cluster and $b(i)$ is the smallest average distance from point i to any other cluster.

3.1.6 Deployment

Each stock and the cluster number will be saved in a CSV, which will be used in the forecasting step. The chosen approach will give the optimal number of clusters and their belonging to fit a forecasting model for each cluster's return and volatility.

3.2 Forecasting

3.2.1 Business Understanding

As stated in previous sections, a forecasting-based portfolio optimization performs more accurately when using the overall mean returns and volatility. To achieve this, it is important to develop an unbiased model that effectively predicts the characteristics of the stocks. The forecasting section aims to develop models that can accurately forecast the return and volatility of each cluster and predict these values for each stock inside the clusters.

3.2.2 Data understanding

There are two sources of input data for forecasting:

- **Daily stock prices and trading volume:** In the same way as in the clustering section, stock prices and trading volumes are used as an input to be transformed into rolling statistics and technical analysis indicators.
- **Cluster label:** The clustering section output will be used to filter the dataset and fit the model to each cluster.

3.2.3 Data preparation

After filtering the dataset to obtain the stock prices and volumes for each cluster, the following features are computed:

- **Technical analysis indicators**
  - **Simple Moving Average (SMA)**: The SMA smoothes the stock price over a time window. It is interpreted as the tendency of the series. It is obtained as the price average over that window:

$$SMA_t = \frac{1}{N} \cdot \sum_{t=0}^{N-1} P_{t-i} \tag{3.13}$$

    Where N is the window size and P_t is the asset's price at time t.

  - **Exponential Moving Average (EMA)**: Unlike SMA, where all past values have the same weight, EMA emphasizes recent values. It is computed recursively:

$$EMA_t = \alpha \cdot P_t + (1 - \alpha) \cdot EMA_{t-1} \tag{3.14}$$

    Where α is the smoothing factor, obtained as $\frac{2}{N+1}$

  - **Bollinger bands**: Measures the series' volatility; the greater the distance between the bands and the SMA, the more volatile the series is. The equations are as follows:

$$UBB_t = SMA_t + 2 \cdot \sigma_t \tag{3.15}$$

$$LBB_t = SMA_t - 2 \cdot \sigma_t \tag{3.16}$$

    Where UBB_t and LBB_t are the upper and lower bands, and σ_t is the standard deviation of the prices over the N periods.

– **Relative Strength Index (RSI)**: Measures momentum, how fast the price of a stock changes, ranging from 0, where the price has only been going down, to 100, with the stock price going only up. The equation is as follows:

$$RSI_t = 100 - \frac{100}{1 + \frac{avg\_gains_t}{avg\_loss_t}} \tag{3.17}$$

– **Stochastic Oscillator**: It measures the position of the current closing price relative to its price range over a recent period:

$$\%K_t = \frac{P_t - L_t}{H_t - L_t} \tag{3.18}$$

Where H_t and L_t are the highest and lowest values over a desired window.

– **On-Balance Volume (OBV)**: A transaction volume-based metric shows the cumulative volume flow. It is computed as follows:

$$OBV_t = OBV_{t-1} + sgn(P_t - P_{t-1}) \cdot V_t \tag{3.19}$$

V_t is the trading volume at time t. The function $sgn(x)$ takes the follwoing values:

$$sgn(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

All these variables will be computed with a window size of 21 days, equivalent to 1 trading month, and will be lagged to avoid look-ahead bias.

- **Lagged average monthly return:** Obtained using Equation 3.9 and lagged to be used as predictor. It is calculated as a rolling average of the 21-day window.
- **Lagged monthly volatility:** Obtained using Equation 3.10 and lagged to be used as a predictor using a 21-day window rolling standard deviation.
- **Lagged average monthly trading volume:** Computed using Equation 3.11 and lagged to be used as a predictor using a 21-day window rolling average.

The outputs of the models will be:

- **Forecasted average monthly return:** Is the forecasted average return for the next 21 working days.

- **Forecasted monthly volatility**: Is the forecasted volatility for the next 21 working days.

These values should be standardized for each stock, and the cluster average will be obtained, given that prices can be on different scales, influencing technical indicator variables.

3.2.4 Modeling

A naive model, plus four machine learning models, will be fit for each cluster's return and volatility. For this, a time series cross-validation of 5 folds (Scikit-Learn Developers (2025*c*)) and a randomized search will be performed to obtain the best parameters for each model based on the RMSE (Scikit-Learn Developers (2025*b*)):

- **Random Forest** (Scikit-Learn Developers (2025*a*)):
  - **Estimators:** Number of trees used to estimate the output. Ranging from 10 to 100, by 10.
  - **Maximum depth:** Controls the depth of the splits. Ranging from 3 to 10.
  - **Maximum features:** Limits the number of features used in each tree. Ranging from 3 to 15, by 3.
  - **Minimum samples in a leaf:** Set the minimum number of datapoints that must be in a final leaf. Ranging from 1 to 10.
- **XGBoost** (XGBoost Contributors (2025)):
  - **Estimators:** Number of trees used sequentially. Ranging from 100 to 1000, by 50.
  - **Maximum depth:** Controls the depth of the splits. Ranging from 1 to 10.
  - **Learning rate:** Scales the contribution of each tree. Ranging from 0.01 to 0.1 by 0.01.
  - **Regulation alpha:** Adds a penalty for large leaf weights, which can drive weights towards zero. It can be 0, 0.1, 1, or 10.
  - **Regulation lambda:** Adds a penalty for large leaf weights, using squared values. It can be 1, 5, 10, or 50.
  - **Minimum child weight:** Controls how much data a tree needs before splitting. It can be 1, 5, or 10.
- **Neural Networks** (Scikit-Learn Developers (2025*d*)):

- **Hidden layer sizes:** Number of neurons and layers in the network architecture. It can be (6),(3),(4), (4, 3), (6, 3), (5, 2), (7, 2), (7, 5, 3), (8, 5, 2), (5, 4, 3), or (4, 3, 2).
- **Learning rate:** Controls how the learning rate changes during training. It can be constant (fixed learning rate), invscalling (gradually decreasing), or adaptive (keeps the rate constant and reduces it when the model stops improving).
- **Learning rate initialization:** Is the starting value of the learning rate. It can be 0.01, 0.001, or 0.0001.
- **Activation:** Transforms each neuron output, adding non-linearity. It can be identity (linear), logistic (outputs between 0 and 1), tanh (outputs between -1 and 1), or relu (identity for positive values, 0 for negative).
- **Alpha:** Adds penalty for large weights. It can be 0.0001, 0.001, 0.01, 1, or 10.

- **Support Vector Machine** (Scikit-Learn Developers (2025*e*)):
  - **Kernel:** Model transformation to detect non-linear patterns. It can be linear (does not detect non-linear patterns), RBF (for non-linear relationships), or poly (uses polynomial relationships),
  - **C:** Controls errors penalization. It can be 0.1, 1, 10, or 100.
  - **Gamma:** Defines how far the influence of a single datapoint extends. It can be scale (based on the variance of the features), auto (based on the number of features), 0.01, or 0.1
  - **Epsilon:** Tolerance to errors for predictions. It can be 0.01, 0.1, or 1.

3.2.5 Evaluation

Since the values of return and volatility are close to 0, the models will be evaluated in a scale-aware metric, such as the RMSE. The model with the lowest value will be the one that will be selected to forecast the return and volatility for each cluster. The RMSE equation is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{3.20}$$

Where e_i is the forecast error in period i.

In addition, the histogram of residuals will be used to check whether the chosen model is free of bias and is reliable for forecasting.

3.2.6 Deployment

The selected models with their hyperparameters will fit their corresponding stocks using the whole dataset. This prediction is used as an input for the portfolio optimization section.

3.3 Optimization

3.3.1 Business understanding

The business purpose of the study is to give actual recommendations to investors about how much to invest on each company in the NASDAQ-100 technology sector stocks, which is achieved through mathematical optimization.

3.3.2 Data understanding

The input data for this section is the following:

- **Forecasted monthly average returns and volatility:** Output from the previous section, used to obtain portfolio return and volatility.
- **ESG score:** Measures of the ESG risk, provided by Morningstar Sustainalytics, ranging from 0 (negligible risk) to 40+ (severe).
- **S&P500 daily index level:** Obtained by the use of yfinance package. It will be used to set threshold values for the return and volatility of the portfolio.
- **Average of the U.S. one-year t-bill:** The value in April will be used as a risk-free asset. It needs to be transformed into a daily equivalent. It is obtained from Board of Governors of the Federal Reserve System (US) (2025).
- **NASDAQ-100 components daily prices**: Used to compute the correlation matrix.

3.3.3 Data preparation

To fit the optimization models, the following data transformations need to be done:

- **Daily equivalent of the U.S. one-year t-bill:** Since the original value is in an annual horizon, it needs to be transformed to a daily horizon to be comparable to the forecasted returns. It is obtained with the following equation.

$$r_{daily} = (1 + r_{annual})^{1/252} - 1 \tag{3.21}$$

Where r_{daily} and r_{annual} are the daily and annual risk-free rates, respectively.

- **Return and volatility threshold:** These values are used as bounds for the portfolio. They are obtained using Equation 3.9 and Equation 3.10 on the S&P500 daily index level, and the results are multiplied by 1.5, to constrain even more the return, and to allow a slack on the volatility.
- **ESG score threshold:** Set to 15, the middle number for a low-risk ESG score.
- **Covariance matrix:** It is obtained to compute the portfolio volatility. First, the correlation matrix is obtained using the NASDAQ-100 components' returns in the following way.

$$corr_{i,j} = \frac{\sum_{t=0}^{T}(r_t^i - R_i) \cdot (r_t^j - R_j)}{\sqrt{\sum_{t=0}^{T}(r_t^i - R_i)^2 \cdot \sum_{t=0}^{T}(r_t^j - R_j)^2}} \tag{3.22}$$

Where $corr_{i,j}$ is the correlation between the returns of assets i and j, r_t^i and r_t^j are the returns of assets i and j in the period t and R_i and R_j are the average returns of assets i and j.

Then the correlation matrix and the forecasted volatilities are used to compute the covariance matrix:

$$\sigma_{i,j} = \sigma_i \cdot \sigma_j \cdot cov_{i,j} \tag{3.23}$$

Where $\sigma_{i,j}$ is the covariance between assets i and j and σ_i and σ_j are the forecasted volatilities.

The outputs of the model are the following:

- **Weights:** Proportion of capital to be invested in each stock. In ranges from 0 to 1, and the sum of all weights must be equal to 0.
- **Portfolio return:** Average return of the portfolio, computed as follows:

$$R_p = \sum_{\forall i \in I} \omega_i \cdot R_i \tag{3.24}$$

Where R_p is the portfolio return, and ω_i and R_i are the weight and return of asset i.

- **Portfolio volatility:** Measures the overall uncertainty of the portfolio return, ob-

tained as:

$$\sigma_p^2 = \sum_{\forall i \in I} \sum_{\forall j \in I} \omega_i \cdot \omega_j \cdot \sigma_{i,j} \tag{3.25}$$

Where σ_p^2 is the squared value of the portfolio volatility and $\sigma_{i,j}$ is the covariance between asset i and j.

- **Portfolio Sharpe ratio:** Measures the extra unit of return per unit of risk. It is computed as follows:

$$Sharpe_p = \frac{R_p - r_f}{\sigma_p} \tag{3.26}$$

Where $Sharpe_p$ is the Sharpe ratio of the portfolio.

- **Portfolio ESG score:** Is the average ESG score of the portfolio, obtained in the following way:

$$ESG_p = \sum_{\forall i \in I} \omega_i \cdot ESG_i \tag{3.27}$$

ESG_p is the average ESG score of the portfolio and ESG_i is the ESG score of asset i.

3.3.4 Modeling

Four optimization models are deployed to obtain four different optimal portfolios:

- **Maximize return, with a maximum volatility and ESG score as constraints:** The model formulation is as follows:

Objective Function:

$$Maximize \quad R_p \tag{3.28}$$

Constraints:

$$\sigma_p^2 \leq \sigma_{threshold}^2 \tag{3.29}$$

$$ESG_p \leq ESG_{threshold} \tag{3.30}$$

$$\sum_{\forall i \in I} \omega_i = 1 \tag{3.31}$$

$$\omega_i \geq 0, \quad \forall i \in I \tag{3.32}$$

Where Equation 3.29 constraints the portfolio volatility using the variance threshold $\sigma^2_{threshold}$, Equation 3.30 sets a maximum average value for the ESG score as $ESG_{threshold}$, Equation 3.31 sets the sum of the weights equal to 1, and Equation 3.32 is the non-negativity constraint.

- **Minimize volatility, with a minimum return and maximum ESG score as constraints:** The model formulation is the following:

  Objective Function:

$$Minimize \quad \sigma_p^2 \tag{3.33}$$

  Constraints:

$$R_p \geq R_{threshold} \tag{3.34}$$

$$ESG_p \leq ESG_{threshold} \tag{3.35}$$

$$\sum_{\forall i \in I} \omega_i = 1 \tag{3.36}$$

$$\omega_i \geq 0, \quad \forall i \in I \tag{3.37}$$

  Where Equation 3.34 constraints the portfolio return using the threshold $R_{threshold}$.

- **Maximize the Sharpe ratio, with a minimum return, and maximum volatility and ESG score as constraints:** The model formulation is the following:

  Objective Function:

$$Maximize \quad Sharpe = \frac{R_p - r_f}{\sigma_p} \tag{3.38}$$

  Constraints:

$$R_p \geq R_{threshold} \tag{3.39}$$

$$\sigma_p^2 \leq \sigma^2_{threshold} \tag{3.40}$$

$$ESG_p \leq ESG_{threshold} \tag{3.41}$$

$$\sum_{\forall i \in I} \omega_i = 1 \tag{3.42}$$

$$\omega_i \geq 0, \quad \forall i \in I \tag{3.43}$$

- **Minimize the average ESG score, with a minimum return and maximum volatility as constraints:** The model formulation is the following:

  Objective Function:

  $$Minimize \quad ESG_p \tag{3.44}$$

  Constraints:

  $$R_p \geq R_{threshold} \tag{3.45}$$

  $$\sigma_p^2 \leq \sigma_{threshold}^2 \tag{3.46}$$

  $$\sum_{\forall i \in I} \omega_i = 1 \tag{3.47}$$

  $$\omega_i \geq 0, \quad \forall i \in I \tag{3.48}$$

These models are compared with a maximum return, minimum volatility, and maximum Sharpe ratio portfolios without the ESG considerations of the constraint in Equation 3.30. The amplpy package is used to solve the optimization problem, applying the Bonmin solver for non-linear problems (AMPL Optimization Inc. (2025)).

3.3.5 Evaluation

A Monte Carlo simulation approach with 1000 cases will be used to evaluate the stability of the portfolios. The forecasted returns will serve as the multivariate normal distribution's mean, and the volatility is leveraged to compute the covariance matrix by multiplying them by the correlation matrix. After this, the portfolios will be compared using central tendency, spread metrics, and financial indicators in terms of their return, volatility, ESG score, and Sharpe ratio. Two additional metrics will be included:

- **Value at Risk (VaR):** It is the maximum loss that will not be exceeded with a

given confidence level. It is obtained as follows:

$$VaR_{\alpha} = Quantile_{1-\alpha}(R_p) \tag{3.49}$$

Where α is the confidence level, which is set to 95%

- **Probability of loss:** Is the probability of obtaining negative returns in the portfolio. It is computed in the following way:

$$P_{loss} = \frac{\#\{R_p < 0\}}{M} \tag{3.50}$$

Where $\#\{R_p < 0\}$ is the number of cases among the simulation where the portfolio return was negative, and M is the number of simulation runs.

3.3.6 Deployment

The deployment of the model is the final recommendation on how much to invest in each stock, according to the investor's investment profile.

4 Data Description

4.1 Clustering variables

The summary statistics of the variables used for clustering can be seen in Table 4.1. There are 75 records, corresponding to the 75 stocks under study. The companies can be seen in Table A1.1. In the profitability ratios, the EBITDA margin is the metric with the least variation, with a standard deviation equal to half the mean. On the other hand, the Operating margin is the variable with the most variation, caused by Strategy Incorporated (MSTR), a company with a score equal to -53.314. Both liquidity ratios have similar values in all descriptive statistics, given the similarity in their formulae.

Variable	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum
ROA	0.0894	0.0851	-0.1883	0.0441	0.0741	0.1303	0.5324
ROE	0.3377	0.6387	-1.1476	0.1024	0.2116	0.4007	4.6822
EBITDA margin	0.3010	0.1414	0	0.1959	0.3097	0.3943	0.7465
Operating margin	-0.4913	6.1833	-53.3144	0.1220	0.2298	0.3073	0.4992
Quick Ratio	1.477	1.380	0.136	0.685	0.899	1.797	6.987
Current Ratio	1.9788	1.5778	0.3630	0.9490	1.3370	2.5670	8.1550
Debt-to-Equity	132.3567	377.5375	1.1480	30.8520	57.3290	117.5660	3178.8370
Historic Return	0.0609%	0.0394%	-0.0532%	0.0384%	0.0566%	0.0846%	0.2087%
Historic Volatility	2.1453%	0.6420%	1.1935%	1.6645%	2.0008%	2.5030%	4.4174%
Historic Volume	17.5M	56.2M	0.5M	2.0M	3.8M	9.1M	462.2M

Table 4.1: Clustering variables descriptive statistics

For clustering, the correlation between variables is important. A high correlation may indicate multicollinearity, meaning redundant information is stored more than once. In Figure 4.1, a heatmap of the correlation matrix is shown, where the most important values are: Current ratio and Quick ratio, with a value of 0.96, given that both are liquidity ratios; ROA and Historic Volume, with a value of 0.6, that might be caused on the influence of the ROA value in the intention to buy on the investor; and ROE and Debt to Equity, due to the use of equity as a financial leverage.

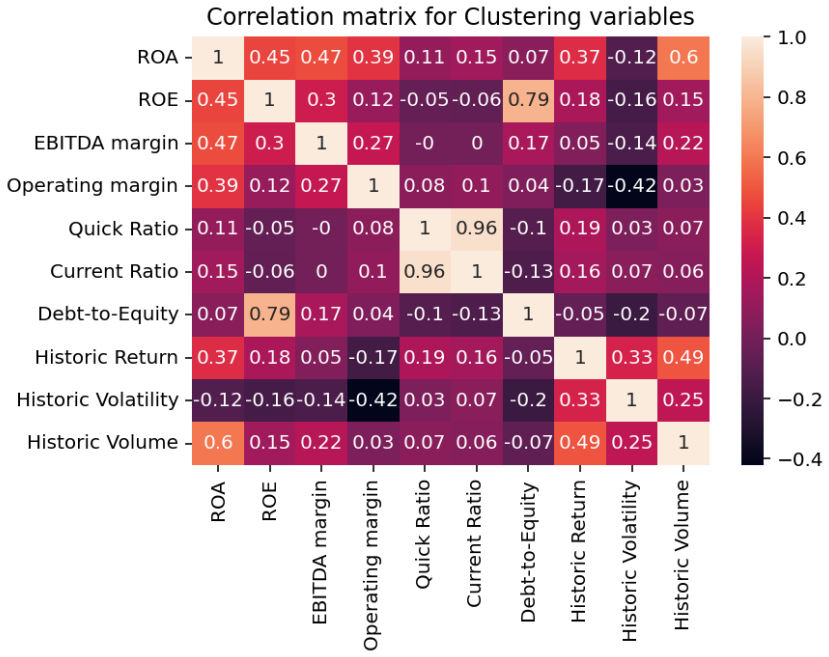
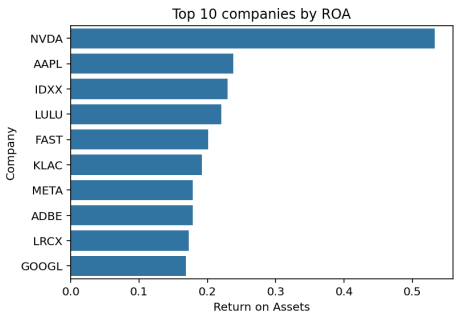
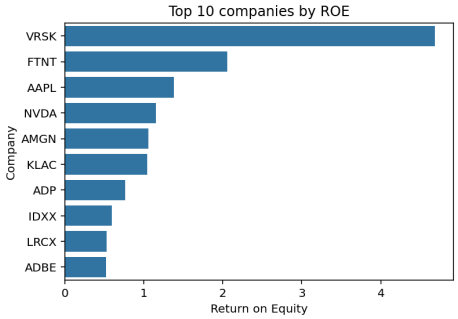


Figure 4.1: Correlation of Clustering variables

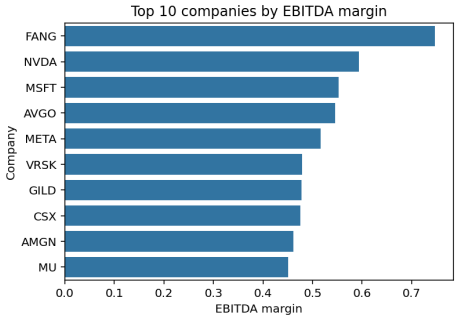
In Figure 4.2, the top-performing companies in profitability ratios can be seen. NVIDIA (NVDA) is the only stock that appears consistently in all four plots, showing its efficiency in using assets and equity, while having strong operational profitability. Meta (META), Verisk Analytics (VRSK), and KLA Corporation (KLAC) each rank in the top 10 for three of the four ratios: for Meta in ROA, EBITDA margin, and Operating margin; for Verisk in ROE, EBITDA margin, and Operating margin; and KLA in ROA, ROE, and Operating margin. Lastly, Diamond Energy (FANG) is the best company in terms of EBITDA and Operating margin, utilizing operational profitability efficiently but inefficiently using assets and equity, compared to other companies in the index.



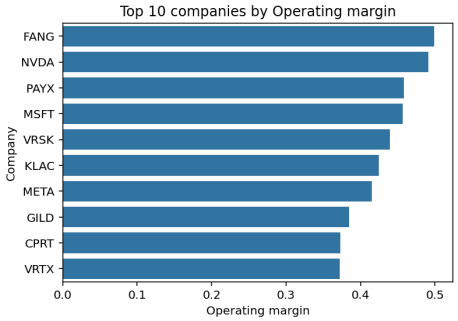
(a) Top performing companies in ROA



(b) Top performing companies in ROE



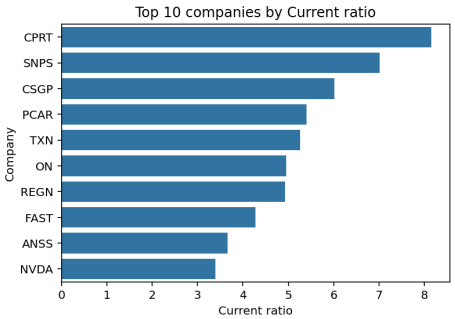
(c) Top performing companies in EBITDA Margin



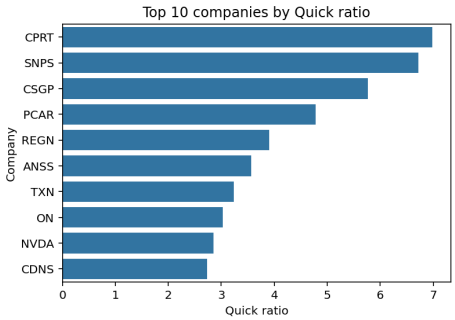
(d) Top performing companies in Operating Margin

Figure 4.2: Top performing companies in profitability ratios

Regarding liquidity ratios, the top performing companies in the Current and Quick ratio are displayed on Figure 4.3. Both measures share the same companies in the top 3, being Copart (CPRT), Synopsys (SNPS), and CoStar Group (CSGP), which are the companies that can solve their short-term debt in a better way. Looking at the solvency plot in Figure 4.4, NVIDIA makes the top 10 in all criteria.

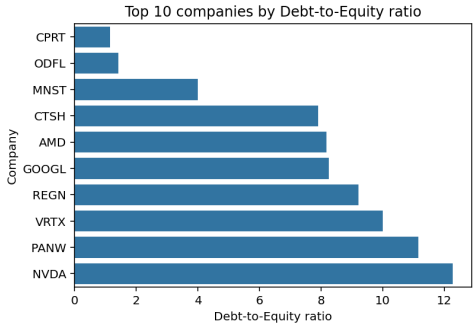


(a) Top performing companies in Current Ratio



(b) Top performing companies in Quick Ratio

Figure 4.3: Top performing companies in liquidity ratios

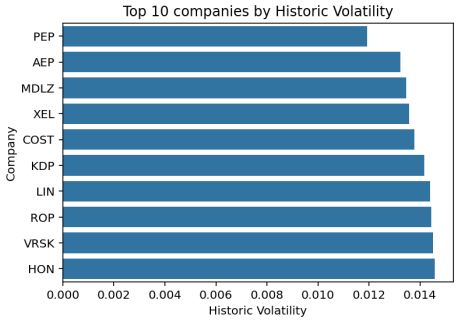


(a) Top performing companies in Debt to Equity Ratio

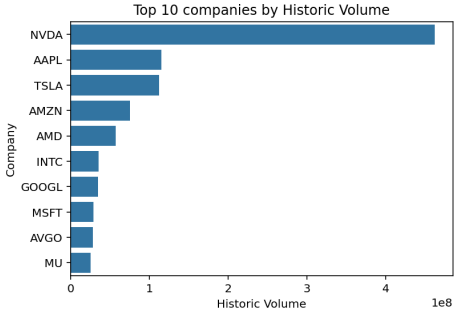
Figure 4.4: Top performing companies in solvency ratios

In terms of historic return, volatility, and volume, it is noticeable in Figure 4.5 that no company that appears in the return top appears in the volatility top. These two variables are usually in conflict; it is difficult for a stock to get high returns while keeping low volatility. On the other hand, the return plot shares four companies with the volume plot, NVIDIA, Advanced Micro Devices (AMD), Broadcom (AVGO) and Tesla (TSLA), indicating that investors tend to invest in high return stocks instead of low volatility companies, for the NASDAQ-100 index, given that these stocks overperform the market in return, but underperform in volatility.

(a) Top performing companies in Return



(b) Top performing companies in Volatility



(c) Top performing companies in Volume

Figure 4.5: Top performing companies in Return, Volatility and Volume

4.2 Forecasting variables

In Table 4.2, the descriptive statistics for the averaged forecasting variables are displayed. There are 191625 records, one for each stock and day. It can be noticed that the SMA and EMA follow a similar distribution of values, which is explained by both being moving averages that are computed differently. The RSI and the Stochastic Oscillator have a small proportion of variance.

Variable	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum
SMA_lag	132.576	132.336	17.433	47.466	92.576	170.527	814.199
EMA_lag	132.569	132.332	17.434	47.458	92.573	170.516	814.199
BB_upper_lag	225.655	242.566	26.915	74.891	157.210	287.996	1657.056
BB_lower_lag	39.498	45.741	-49.972	10.274	32.187	50.996	193.300
RSI_lag	53.073	1.413	48.547	52.392	52.977	54.043	56.285
Stoch_K_lag	56.661	3.365	45.256	55.084	56.608	58.813	63.179
OBV_lag	1,446.7M	6,840.4M	-297.4M	107.9M	246.0M	547.7M	58,806.1M
Lagged_Return	0.0627%	0.0382%	-0.0411%	0.0421%	0.0584%	0.0853%	0.2147%
Lagged_Volatility	1.9178%	0.5682%	1.0193%	1.5003%	1.8249%	2.2309%	3.5706%
Lagged_Volume	17.5M	56.5M	0.5M	2.0M	3.8M	9.1M	465.2M

Table 4.2: Forecasting variables descriptive statistics

Looking into the correlation heatmap in Figure 4.6, a correlation equal to 1 between SMA and EMA is noticed, which agrees with the information shown on Table 4.2. For the endogenous variables, only the volatility is correlated with its lag, with a value of 0.59.

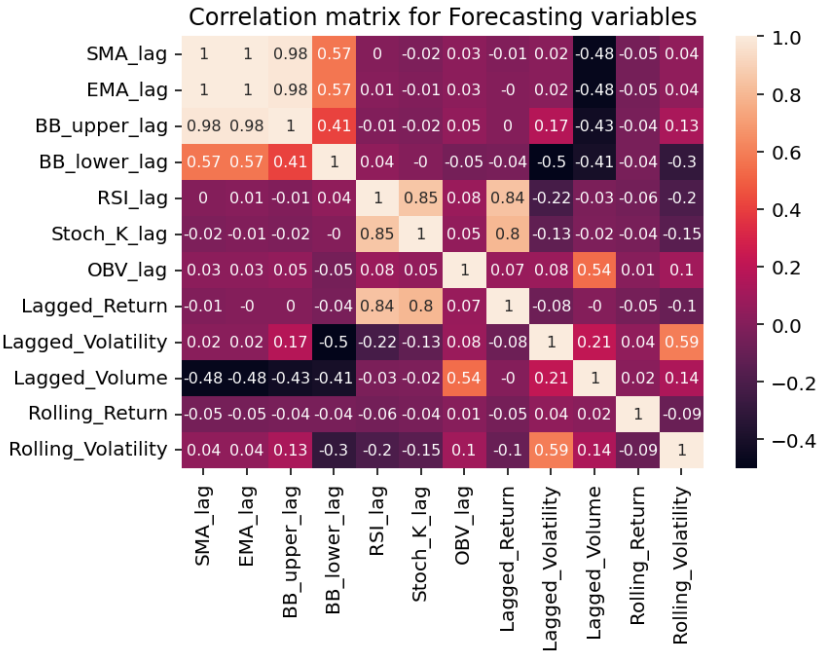


Figure 4.6: Correlation of Forecasting variables

In Figure 4.7, a time series plot can be seen for the six stocks with higher average return. These companies are, from left to right, top to bottom, Advanced Micro Devices, Broadcom, Axon Enterprise, Strategy Incorporated, NVIDIA, and Tesla. These returns appear to be centered around zero, with periods of higher volatility, such as with MSTR between 2022 and 2024, and lower volatility, like AVGO between 2016 and 2018.

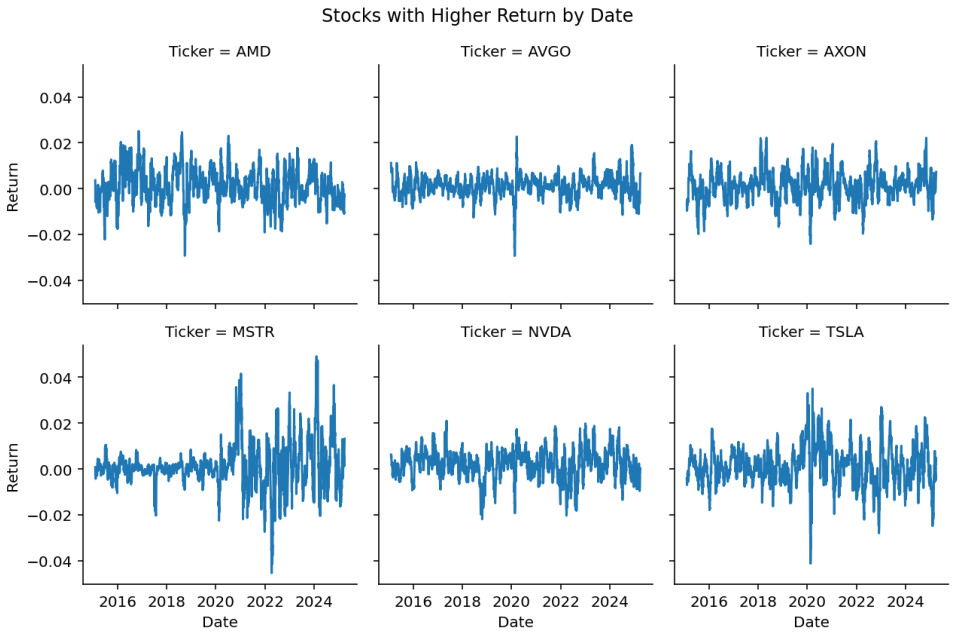


Figure 4.7: Stocks with higher Return by Date

The time series plot for volatility can be seen in Figure 4.8 for the six stocks with lower average values. These companies are, from left to right, top to bottom, American Electric Power Company (AEP), Costco Wholesale Corporation (COST), Keurig Dr Pepper (KDP), Mondelez International (MDLZ), Pepsi (PEP), and Xcel Energy (XEL). Most values are around 0% and 0.02%, except for a period in 2020 where there is a spike of high volatility, most likely due to the pandemic of that year.

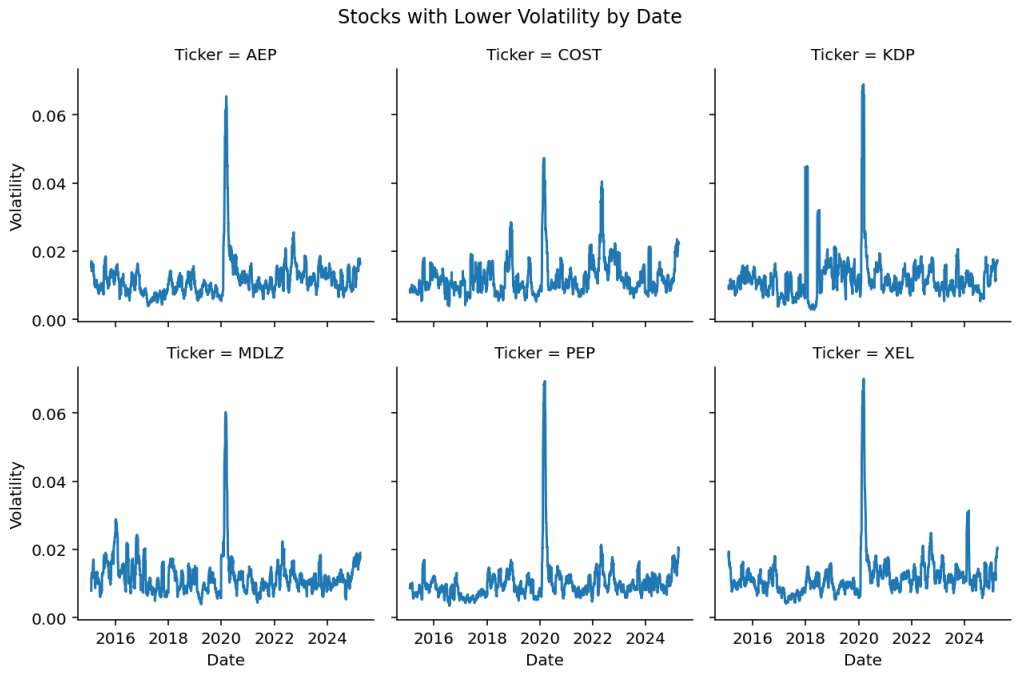


Figure 4.8: Stocks with lower Volatility by Date

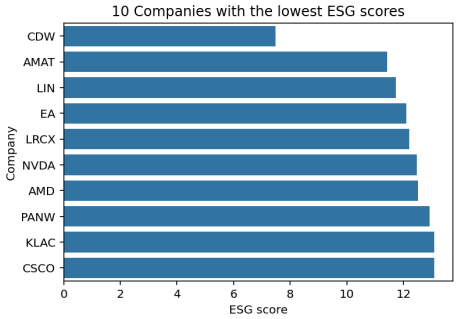
4.3 Optimization variables

The summary statistics of the ESG score are shown in Table 4.3. There are 75 records, corresponding to the stocks in the study. This variable ranges from 7.49 (negligible) to 37.31 (high), and the mean and median are close to the "medium" category, which indicates that half of the values are either in the negligible and low categories, and the other half in the medium and high categories.

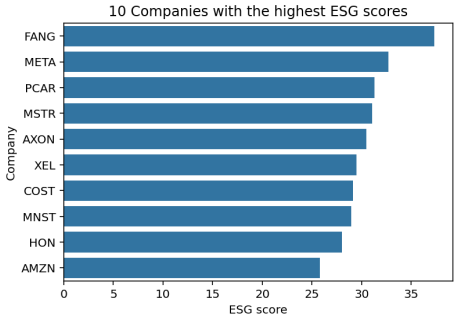
Variable	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum
ESG score	19.50373	5.793303	7.49	15.305	18.81	22.675	37.31

Table 4.3: ESG score descriptive statistics

Looking into the best and worst performing companies in Figure 4.9, only CDW Corporation (CDW) has a negligible risk. On the other hand, Diamond Energy, META, Paccar (PCAR), Strategy Incorporated, and Axon Enterprise (AXON) have a high risk. It can be deduced that most of the values are in the low and middle categories.



(a) Top 10 companies with the lowest ESG score



(b) Top 10 companies with the highest ESG score

Figure 4.9: Companies by ESG scores

The S&P500 returns are used to set a portfolio return and volatility threshold. There are 2596 records, one per day. The summary statistics are displayed in Table 4.4. Comparing the mean value, which used to compute the return threshold, with the historic return in Table 4.1, it can be inferred that the S&P500 returns fall inside the first quantile of the historic return, meaning that, to set an applicable threshold, the returns of the index should be multiplied by a factor greater than one. The same can be said about the volatility, where the standard deviation of the returns of the S&P500 index is less than the minimum value for historic volatility.

Variable	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum
S&P500 returns	0.0383%	1.1524%	-12.7652%	-0.3848%	0.0645%	0.5792%	9.0895%

Table 4.4: S&P500 returns descriptive statistics

The correlation matrix of the stocks is used as an approximation for the covariance matrix when multiplied by the forecasted volatility. In Figure 4.10, a heatmap of the stocks with at least one absolute correlation value higher than 0.8 can be seen. There are no high negative correlations, which will make diversification difficult. This may happen due to the stocks having similar movement directions simultaneously. The highest correlation is between Lam Research Corporation (LCRX) and Applied Materials (AMAT) with a value of 0.88, followed by Cadence Design Systems (CDNS) and Synopsys (SNPS).

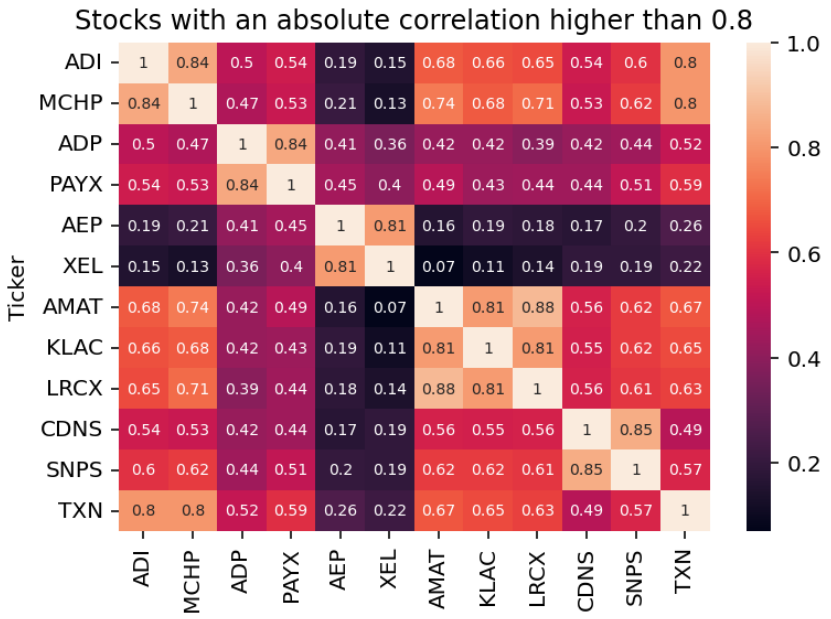


Figure 4.10: Stocks with an absolute correlation higher than 0.8

5 Results

5.1 Clustering

5.1.1 Using all variables

When clustering with all the variables, the best number of clusters is 2, with an average silhouette score close to 0.25, as shown in Table 5.1. In general, all the values are between 0.11 and 0.25, having a close range with low values for a silhouette score.

Number of clusters	Silhouette score
2	0.2226
3	0.2115
4	0.2527
5	0.1267
6	0.1619
7	0.1601
8	0.1262
9	0.1152
10	0.1261

Table 5.1: Silhouette scores obtained by clustering with all variables

Zooming in on the 4 clusters configuration in Figure 5.1, 2 clusters appear to be correctly grouped, being clusters 0 and 3, both with an average score around 0.35, and a size of 49 and 8, respectively. On the other hand, the five members of cluster 1 are likely misclassified, with a score of -0.27. The 13 members of cluster 2 are ambiguously assigned and have a score of -0.01.

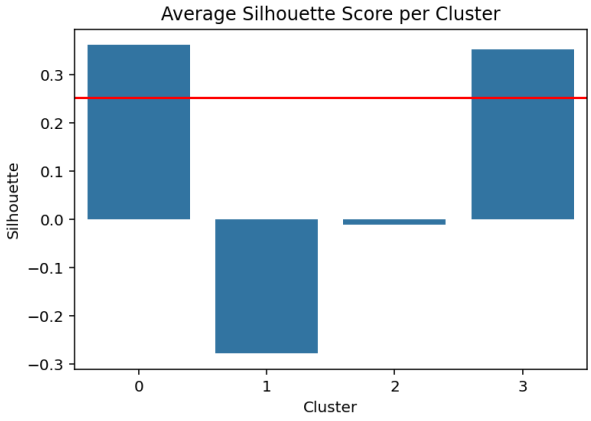


Figure 5.1: Average Silhouette score for every cluster with all variables

5.1.2 Using fundamental analysis indicators

For this approach, the iteration with 2 clusters has a better silhouette score, with a value of 0.29. As shown in Table 5.2, these values range from 0.17 to 0.29, with the optimal configuration improving significantly the silhouette value.

Number of clusters	Silhouette score
2	0.2951
3	0.2847
4	0.2539
5	0.2450
6	0.2428
7	0.2235
8	0.1978
9	0.1656
10	0.1750

Table 5.2: Silhouette scores obtained by clustering with fundamental analysis indicators

With the optimal number of clusters being two, the average silhouette score for each group is shown in Figure 5.2. Cluster 0 is the biggest one with 50 members, the maximum allowed in this configuration of K-means constrained, and its score is 0.45. Cluster 1 has 25 members and a score of -0.01, showing ambiguity on the final labels of the clusters, potentially meaning that an optimal clustering can be reached without the constraints.

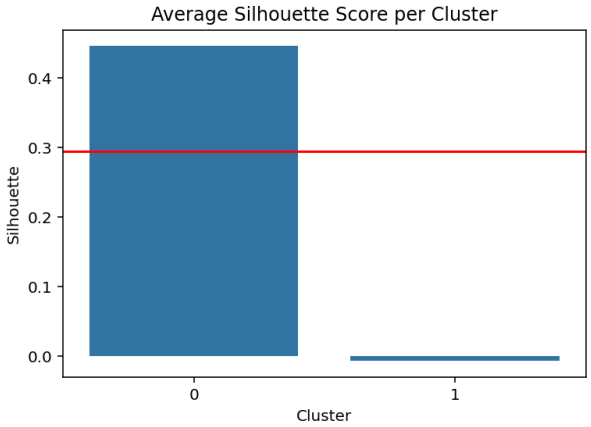


Figure 5.2: Average Silhouette score for every cluster with fundamental analysis indicators

5.1.3 Using return, volatility, and volume values

Using a simple approach, with historic return, volatility, and volume values, the results improve. The best average silhouette score occurs when the number of clusters is set to 2, with a value of 0.39. All the values displayed in Table 5.3 tend to outperform other approaches, ranging from 0.29 to 0.39.

Number of clusters	Silhouette score
2	0.3902
3	0.3599
4	0.3636
5	0.2943
6	0.3018
7	0.3267
8	0.3152
9	0.3070
10	0.2929

Table 5.3: Silhouette scores obtained by clustering with return, volatility, and volume

In Figure 5.3, a plot of the average silhouette score per cluster can be seen. Cluster 0 is the dominant group with 50 members, the maximum allowed, and a score of 0.54. Cluster 1 comprises 25 companies with an average score of 0.09, showing signs of an ambiguous classification.

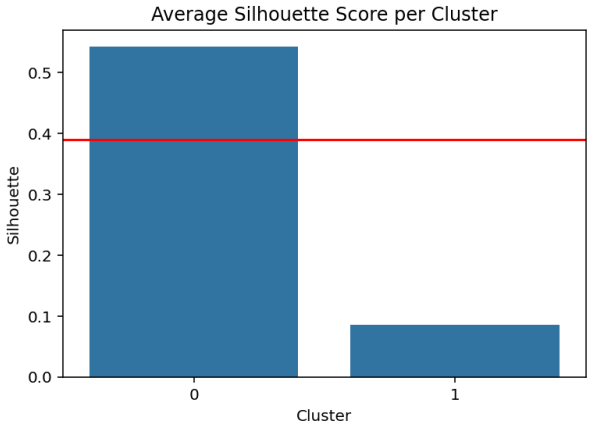


Figure 5.3: Average Silhouette score for every cluster with return, volatility, and volume

5.1.4 Using Principal Components Analysis

For the approach with PCA, it is important to decide how many components to use. In Figure 5.4, the proportion of variance by the number of PCA components can be seen. Setting a value of 0.8 as a threshold, represented by the red line, five components are needed to reach this value.

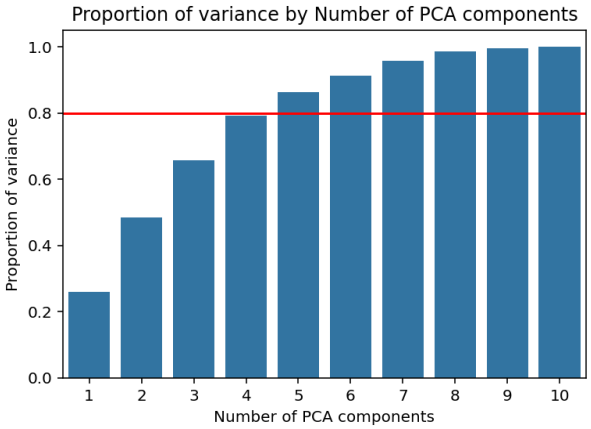


Figure 5.4: Proportion of variance by PCA components

With the 5 PCA components, the best number of clusters is four, with a value of 0.31, improving the score obtained using all variables. The silhouette scores in Table 5.4 range from 0.12 to 0.31.

Number of clusters	Silhouette score
2	0.2658
3	0.2982
4	0.3075
5	0.2275
6	0.1945
7	0.1692
8	0.1285
9	0.1189
10	0.1225

Table 5.4: Silhouette scores obtained by clustering with PCA

The average silhouette scores for the 4 clusters configuration can be seen in Figure 5.5. Clusters 0 and 3 are the only ones with a significant positive value, with a score of 0.20 and 0.43, and a size of 15 and 50, respectively. Cluster 1 is highly likely to be misclassified, with a score of -0.26 and 5 members. Finally, cluster 2 scores near 0, meaning that the five members are classified ambiguously.

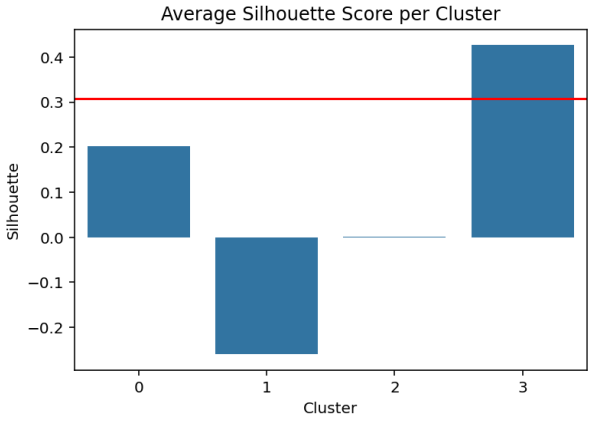


Figure 5.5: Average Silhouette score for every cluster with PCA variables

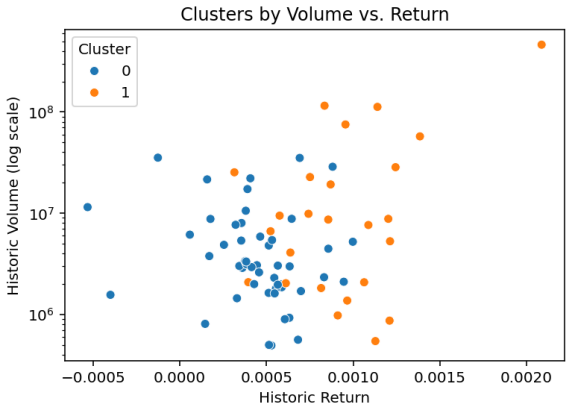
5.1.5 Chosen clustering approach

According to what was shown before, given its high overall silhouette score and the absence of negative average scores on every cluster, the approach that uses only return, volatility, and volume values is chosen. This, being a simple approach, makes clustering more interpretable. In Table 5.5, the values of the centroids can be seen. There is a clear distinction between a group with high values for all three variables (cluster 1) and a group with lower values (cluster 0). In Table A1.2 and Table A1.3, the clusters' membership can be seen, with cluster 0 having Google, Microsoft, and Adobe as members; meanwhile, cluster 1 contains Apple, Amazon, Meta, and NVIDIA.

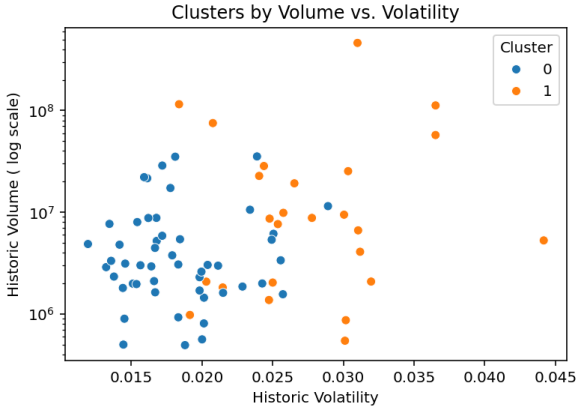
Cluster	Return	Volatility	Volume
0	0.0443%	1.8349%	6,396,245
1	0.0941%	2.766%	39,598,277

Table 5.5: Centroid of the best performing cluster

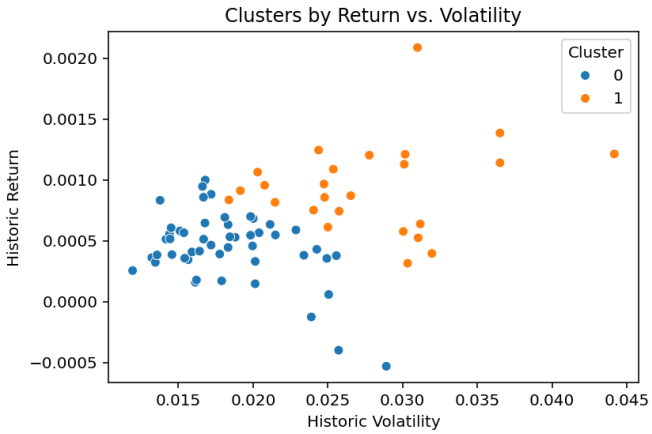
The differences can be noticed when the variables are plotted, as in Figure 5.6. Some leaks in the plots use volume; however, in the Return vs. Volatility plot, the distinction is clear, with no visible leakage between the clusters.



(a) Return vs. Volume



(b) Volume vs. Volatility



(c) Return vs. Volatility

Figure 5.6: Cluster analysis scatterplots

5.2 Forecasting

5.2.1 Cluster 0 return

For cluster 0 return, the results of the parameter randomized search for the Random Forest model can be seen in Table 5.6. The best RMSE approach uses 40 estimators,

a maximum depth of 6 branches, three maximum features, and at least nine samples in a leaf.

RMSE	Estimators	Max Feaures	Max Depth	Min Samples Leaf
1.197538	40	6	3	9

Table 5.6: Cross-validation values for Cluster 0 Random Forest model for return

For Neural Networks, it uses an architecture of 2 hidden layers with 4 and 3 neurons each, with a starting learning rate of 0.01 that gradually decreases, with a Tanh function as activation, which sets outputs between -1 and 1, and an alpha value of 10.

RMSE	Hidden Layer Sizes	Learning Rate	Learning Rate Initialization	Activation	Alpha
0.598017	(4, 3)	Invscaling	0.01	Tanh	10

Table 5.7: Cross-validation values for Cluster 0 Neural Networks model for return

An even better cross-validation score is obtained by SVM, as shown in Table 5.8. This is achieved using a third-degree polynomial kernel, an error penalization of 10, a gamma value of 0.01, and a tolerance of errors of 0.1.

RMSE	Kernel	C	Gamma	Epsilon
0.591918	Polynomial	10	0.01	0.1

Table 5.8: Cross-validation values for Cluster 0 SVM model for return

The XGBoost cross-validation score is shown in Table 5.9. The best model uses 100 estimators, a maximum depth of 9, a learning rate of 0.01, a penalty of large weights of 1, a squared penalty of 5, and a minimum of 1 datapoint to allow the tree to split.

RMSE	Estimators	Max Depth	Learning Rate	Alpha	Lambda	Weight
0.820264	100	9	0.01	1	5	1

Table 5.9: Cross-validation values for Cluster 0 XGBoost model for return

When the performance of the models in both the train and test sets is compared, it is clear that SVM achieves the lowest score on the test set, making it the best approach to model cluster 0 return. The Naive model has a low score, even outperforming Random Forest and XGBoost, which can be overfitting the training set.

Model	RMSE Train	RMSE Test
Naive	0.618219	0.41399
Random Forest	0.53599	0.700979
XGBoost	0.399514	0.489936
Neural Networks	0.597003	0.412211
SVM	0.594932	0.388442

Table 5.10: RMSE scores for train and test sets for Cluster 0 return

Checking the histogram of residuals of the chosen model in Figure 5.7, the residuals are centered around 0, showing no sign of bias. The curve is slightly skewed to the left, but without extreme outliers, then this model is considered reliable to forecast cluster 0 return.

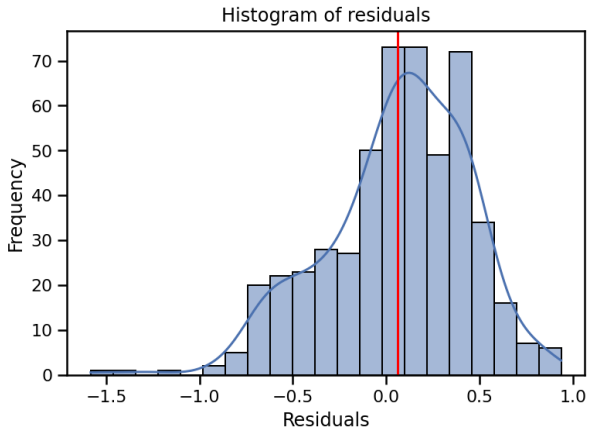


Figure 5.7: SVM residuals diagnosis for cluster 0 return

5.2.2 Cluster 0 volatility

The results of the randomized search for cluster 0 volatility using the Random Forest algorithm are displayed in Table 5.11. It uses 70 estimators, three maximum features, a maximum depth of 3, and a minimum of 8 samples in a leaf to allow a split.

RMSE	Estimators	Max Feaures	Max Depth	Min Samples Leaf
1.513073	70	3	3	8

Table 5.11: Cross-validation values for Cluster 0 Random Forest model for volatility

The best performing Neural Network is obtained with three hidden layers with 5,4, and 3 neurons, respectively, a constant learning rate of 0.001, a ReLu activation function (equal to the identity function for positive values, and 0 for negative), and a regularization factor of 10. These results are summarized in Table 5.12

RMSE	Hidden Layer Sizes	Learning Rate	Learning Rate Initialization	Activation	Alpha
0.661808	(5, 4, 3)	Constant	0.001	ReLU	10

Table 5.12: Cross-validation values for Cluster 0 Neural Networks model for volatility

The best performing SVM model uses a Radial Basis Function that suits non-linear relationships, a penalty of 0.1 for significant errors, a gamma value of 0.01, and a tolerance of 0.01.

RMSE	Kernel	C	Gamma	Epsilon
0.67741	Radial Basis Function	0.1	0.01	0.01

Table 5.13: Cross-validation values for Cluster 0 SVM model for volatility

For XGBoost, the best number of estimators is 100, 9 is set as a max depth, the learning rate is equal to 0.01, a regularization alpha equal to 1, a regularization lambda equal to 5, and a minimum of 1 datapoint to allow the tree to split.

RMSE	Estimators	Max Depth	Learning Rate	Alpha	Lambda	Weight
1.136739	100	9	0.01	1	5	1

Table 5.14: Cross-validation values for Cluster 0 XGBoost model for volatility

In Table 5.15, the train and test RMSE scores can be seen for the different models. All but Random Forest outperformed Naive, with XGBoost obtaining the smallest value and being selected as the forecasting model for cluster 0 volatility, given that there is no sign of overfitting.

Model	RMSE Train	RMSE Test
Naive	0.800851	0.412252
Random Forest	0.554526	0.420961
XGBoost	0.444358	0.405031
Neural Networks	0.800884	0.409893
SVM	0.658324	0.406366

Table 5.15: RMSE scores for train and test sets for Cluster 0 volatility

The histogram of residuals in Figure 5.8 shows the residuals centered around 0, indicating the absence of bias. The right tail is longer than the left one; however, no extreme outliers exist. Therefore, XGBoost is considered reliable for forecasting cluster 0 volatility.

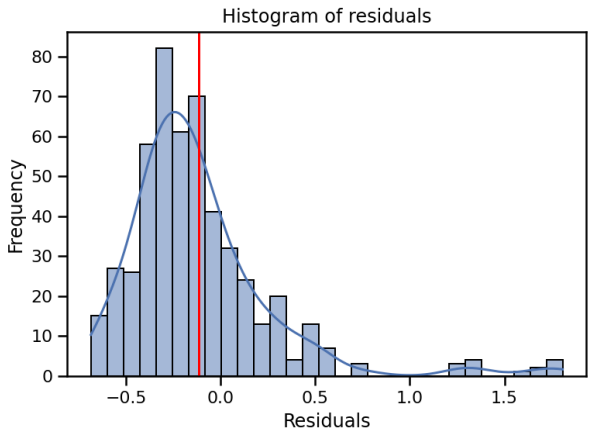


Figure 5.8: XGBoost residuals diagnosis for cluster 0 volatility

5.2.3 Cluster 1 return

For cluster 1 return, the results of the parameter randomized search for the Random Forest model can be seen in Table 5.16. The best RMSE approach uses 40 estimators, a maximum depth of 3 branches, three maximum features, and at least five samples in a leaf.

RMSE	Estimators	Max Feaures	Max Depth	Min Samples Leaf
0.767927	40	3	3	5

Table 5.16: Cross-validation values for Cluster 1 Random Forest model for return

For Neural Networks, the results are displayed in Table 5.17, where it is shown that the best approach is to use two hidden layers with 4 and 3 neurons, respectively, a constant learning rate of 0.001, a ReLU activation, and a regularization parameter of 10.

RMSE	Hidden Layer Sizes	Learning Rate	Learning Rate Initialization	Activation	Alpha
0.670708	(4, 3)	Constant	0.001	ReLU	10

Table 5.17: Cross-validation values for Cluster 1 Neural Networks model for return

An even better cross-validation score is obtained by SVM, as shown in Table 5.18. This is obtained by using a polynomial kernel of third degree, an error penalization of 0.1, a gamma value of 0.01, and a tolerance of errors of 0.01.

RMSE	Kernel	C	Gamma	Epsilon
0.656164	Polynomial	0.1	0.01	0.01

Table 5.18: Cross-validation values for Cluster 1 SVM model for return

For XGBoost, the best parameters are 100 estimators, a maximum depth of 8 splits, a learning rate of 0.09, a regularization alpha of 10, a regularization lambda of 1, and a minimum of 1 datapoint to allow the tree to split.

RMSE	Estimators	Max Depth	Learning Rate	Alpha	Lambda	Weight
0.666463	100	8	0.09	10	1	1

Table 5.19: Cross-validation values for Cluster 1 XGBoost model for return

When comparing the RMSE values of the train and test sets in Table 5.20, only SVM can outperform Naive. XGBoost shows signs of overfitting, with a very low RMSE in the training set and a high value on the test set. SVM is chosen as the forecasting model for cluster 1 return.

Model	RMSE Train	RMSE Test
Naive	0.656755	0.573853
Random Forest	0.555444	0.718896
XGBoost	0.261903	0.604739
Neural Networks	0.638579	0.642512
SVM	0.66002	0.553961

Table 5.20: RMSE scores for train and test sets for Cluster 1 volatility

In Figure 5.9, the histograms of residuals for the SVM model can be seen. The mean is centered near zero, and the curve resembles a normal distribution. Then, the SVM model is free of bias and is reliable for forecasting cluster 1 return.

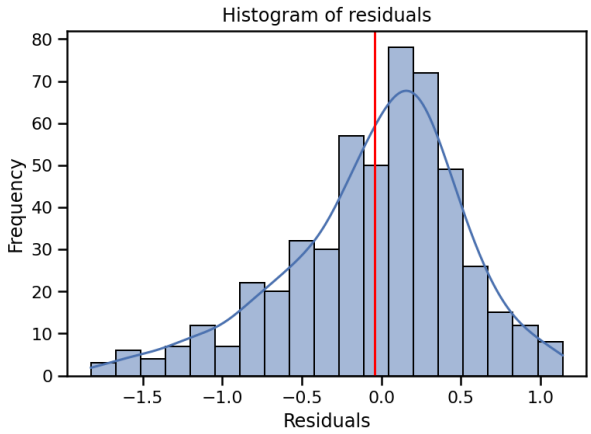


Figure 5.9: SVM residuals diagnosis for cluster 1 return

5.2.4 Cluster 1 volatility

The results of the randomized search for cluster 1 volatility using the Random Forest algorithm are displayed in Table 5.21. It uses 80 estimators, three maximum features, a maximum depth of 3, and a minimum of 8 samples in a leaf to allow a split.

RMSE	Estimators	Max Feaures	Max Depth	Min Samples Leaf
0.834593	80	3	3	8

Table 5.21: Cross-validation values for Cluster 1 Random Forest model for volatility

The Neural Network score is displayed in Table 5.22, where the best RMSE is obtained when an architecture of 3 hidden layers with 8, 5, and 2 neurons is used, with a constant learning rate of 0.001, an identity activation function, and a regularization parameter equal to 10.

RMSE	Hidden Layer Sizes	Learning Rate	Learning Rate Initialization	Activation	Alpha
0.595965	(8, 5, 2)	Constant	0.001	Identity	10

Table 5.22: Cross-validation values for Cluster 1 Neural Networks model for volatility

The best performing SVM model uses a Radial Basis Function that suits non-linear relationships, a penalty of 0.1 for significant errors, a gamma value of 0.01, and a tolerance of 0.01.

RMSE	Kernel	C	Gamma	Epsilon
0.647801	Radial Basis Function	0.1	0.01	0.01

Table 5.23: Cross-validation values for Cluster 1 SVM model for volatility

For XGBoost, the best number of estimators is 850, 1 is set as a max depth, with all estimators being weak learners, the learning rate is equal to 0.05, a regularization alpha equal to 0, without penalizing large weights, a regularization lambda equal to 5, and a minimum of 10 datapoint to allow the tree to split.

RMSE	Estimators	Max Depth	Learning Rate	Alpha	Lambda	Weight
0.682238	850	1	0.05	0	5	10

Table 5.24: Cross-validation values for Cluster 1 XGBoost model for volatility

Comparing all models in terms of RMSE in the training and test set, Random Forest is the only algorithm able to outperform Naive, making it the chosen model to forecast cluster 1 volatility. The values are displayed in Table 5.25.

Model	RMSE Train	RMSE Test
Naive	0.730189	0.55796
Random Forest	0.451029	0.544576
XGBoost	0.428185	0.629463
Neural Networks	0.57191	0.671624
SVM	0.56937	0.598105

Table 5.25: RMSE scores for train and test sets for Cluster 1 volatility

The residual diagnosis in Figure 5.10 shows a curve centered near 0, showing no signs of bias. However, it is positively skewed, but with no extreme outliers. Then, Random Forest is a reliable model to forecast cluster 1 volatility.

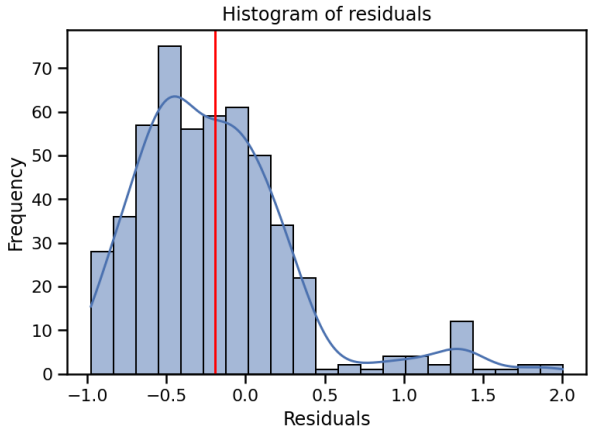
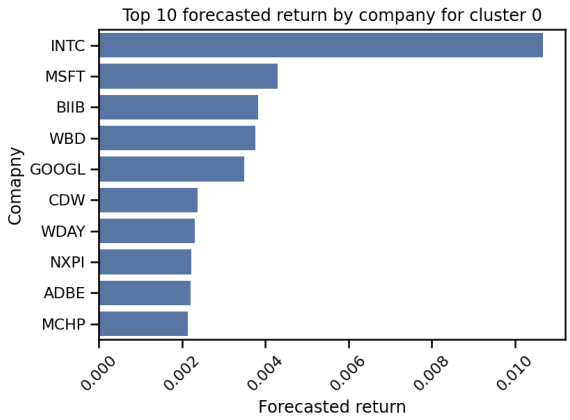


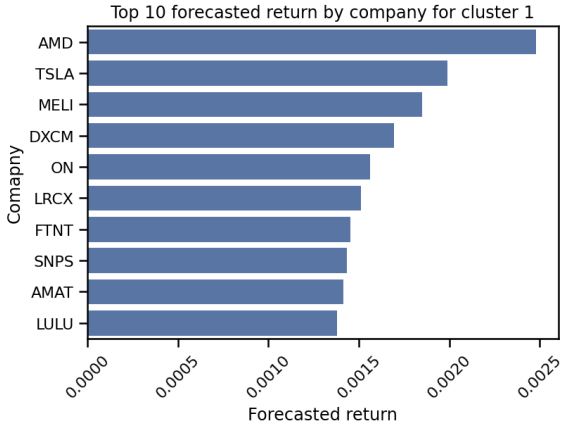
Figure 5.10: Random Forest residuals diagnosis for cluster 1 volatility

5.2.5 Chosen forecasting approach

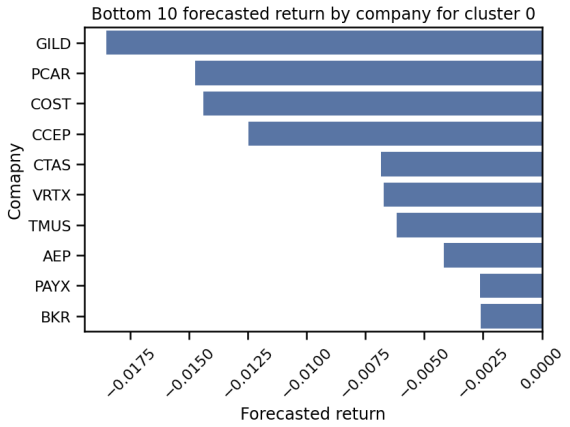
As mentioned, an SVM with a third-degree polynomial kernel is chosen for cluster 0 and cluster 1 returns. The forecasted values for each stock can be seen in Table A1.4 and Table A1.6. The smallest forecasted returns for cluster 0 belong to Gilead Sciences (GILD), PACCAR (PCAR), Costco, and Coca-Cola Europacific Partners, all with a value less than -1%. Cluster 1 contains Strategy Incorporated, NVIDIA, Broadcom (AVGO), Axon, and Meta as the companies with lower returns. On the other hand, Intel, Microsoft, and Biogen are the best-performing companies in terms of return from cluster 0. At the same time, Advanced Micro Devices (AMD), Tesla, and MercadoLibre (MELI) are the top performers of cluster 1.



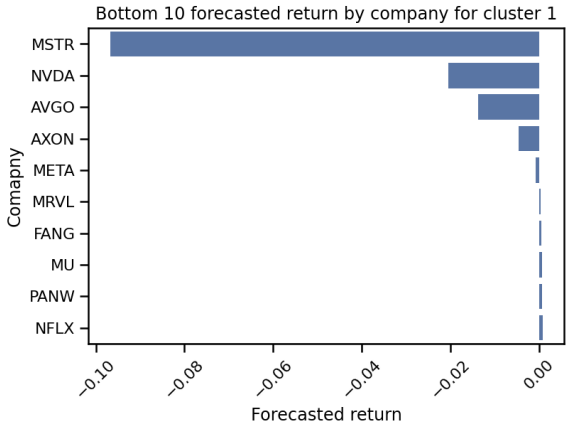
(a) Top 10 returns for cluster 0



(b) Top 10 returns for cluster 1



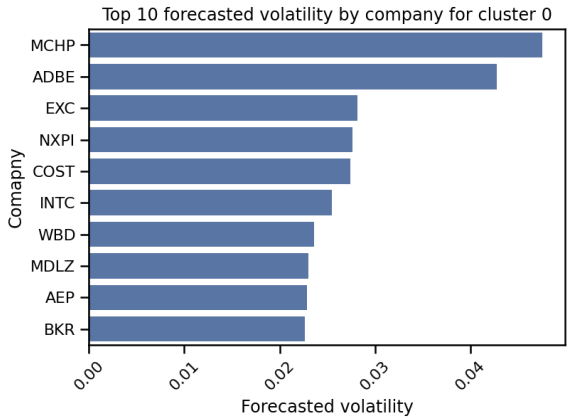
(c) Bottom 10 returns for cluster 0



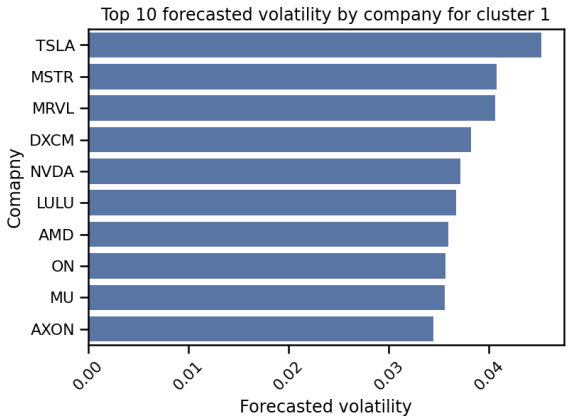
(d) Bottom 10 returns for cluster 1

Figure 5.11: Return forecasting results

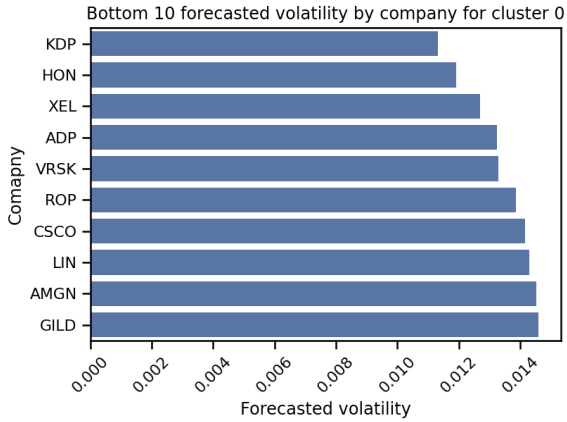
The tree-based models are the most accurate for volatility, with an XGBoost model chosen for cluster 0 and a Random Forest for cluster 1. The forecasted volatilities are present in Table A1.5 and Table A1.7. The smallest volatilities for cluster 0 are KDP, Honeywell International (HON), and Xcel Energy; meanwhile, for cluster 1, Cadence Design Systems (CDS), Take-Two Interactive Software, and Apple are the best-performing companies for volatility. The most volatile stocks belong to Microchip Technology Incorporated (MCHP), Adobe, and Exelon Corporation, in cluster 0, and Tesla, Strategy Incorporated, and Marvell Technology in cluster 1.



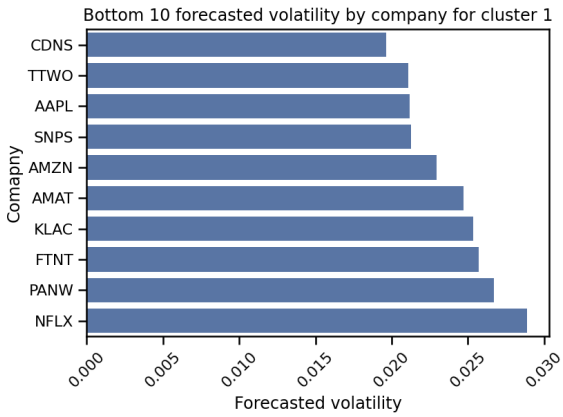
(a) Top 10 volatilities for cluster 0



(b) Top 10 volatilities for cluster 1



(c) Bottom 10 volatilities for cluster 0



(d) Bottom 10 volatilities for cluster 1

Figure 5.12: Volatility forecasting results

5.3 Optimization

5.3.1 Without including ESG scores

After solving the optimization problem for a maximum return portfolio without including ESG scores, the weights are shown in Figure 5.13. The majority of the investment has to be put into Intel Corp (INTC), while Microsoft (MSFT) and Biogen (BIIB) share a fifth of the investment. Warner Bros Discovery (WBD) has around 4.4% of the investment.

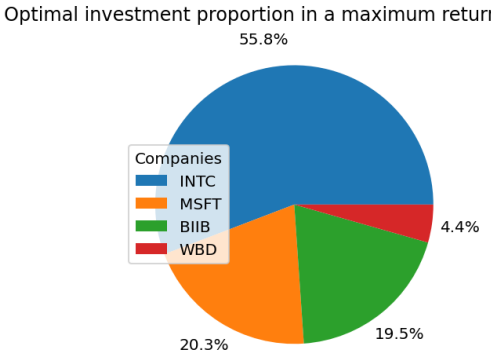


Figure 5.13: Proportion of investment for a maximum return portfolio

The minimum volatility portfolio is highly diversified, with 13 stocks netting around 95% of the investment amount. Keurig Dr Pepper (KDP) and Xcel Energy (XEL) each obtain a fourth of the amount, followed by Biogen, Take-Two Interactive Software (TTWO), and Verisk Analytics (VRSK).

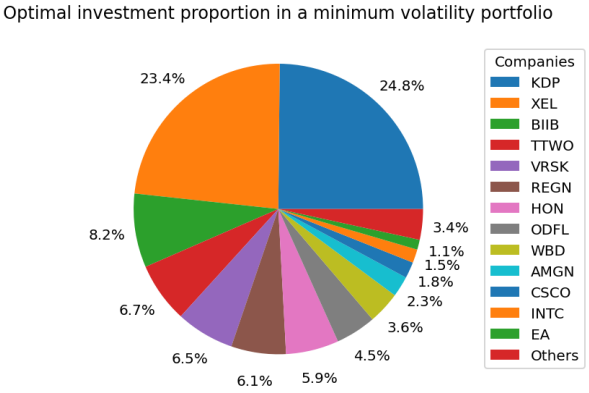


Figure 5.14: Proportion of investment for a minimum volatility portfolio

The composition of the maximum Sharpe ratio portfolio, as displayed in Figure 5.15, is almost identical to the maximum return portfolio. The difference lies in a lesser amount invested in Intel Corp, and a slightly higher amount put on Microsoft, Biogen, and Warner.

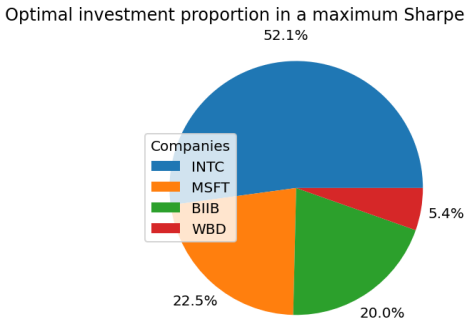


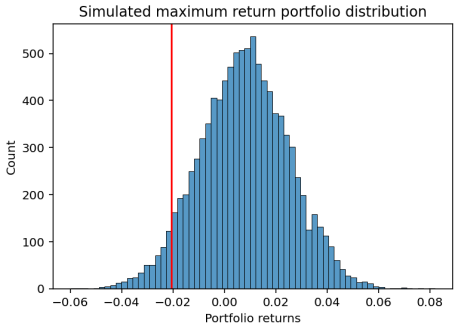
Figure 5.15: Proportion of investment for a maximum Sharpe ratio portfolio

In Table 5.26, the parameters of the optimized models can be seen. The maximum return and maximum Sharpe ratio portfolio have similar behaviours, being both more volatile but with less ESG risk than the minimum volatility portfolio. The ESG score values are close to the medium risk category.

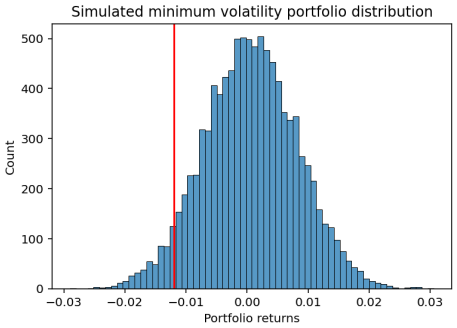
Portfolio	**Return**	**Volatility**	**Sharpe ratio**	**ESG score**
Maximum return	0.774%	1.729%	0.439	18.601
Minimum Volatility	0.058%	0.763%	0.056	22.435
Maximum Sharpe ratio	0.750%	1.673%	0.439	18.521

Table 5.26: Results of the portfolio optimization without including ESG scores

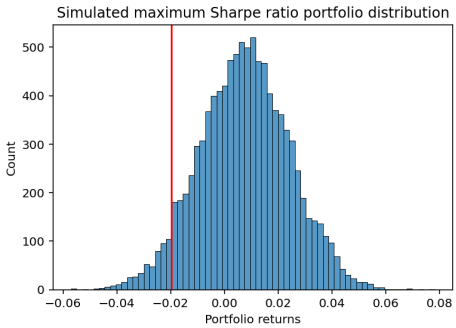
The distributions of the simulated portfolios and the Value at Risk, represented by a red line, can be seen in Figure 5.16. Both the maximum return and maximum Sharpe ratio have a VaR around -2%, which can be interpreted as there is a 5% probability that the portfolio will lose at least 2% of its value. The minimum volatility portfolio has a VaR around -1%, making it a more secure option compared with the rest of the portfolios.



(a) Returns distribution for maximum return portfolio



(b) Returns distribution for minimum volatility portfolio



(c) Returns distribution for maximum Sharpe ratio portfolio

Figure 5.16: Return distributions and VaR for the simulated portfolios without including ESG score

Looking at the different metrics obtained when portfolios are simulated, in Table 5.27, it can be concluded that the return, volatility, Sharpe ratio, and ESG score have minor changes from the optimization step. If the focus is on the probability of loss, the maximum return portfolio has a 32% probability of having a negative return, making it the smallest value.

Portfolio	Return	Volatility	Sharpe ratio	ESG score	Probability of loss	VaR
Maximum return	0.774%	1.733%	0.438	18.601	32.430%	-2.060%
Minimum Volatility	0.057%	0.766%	0.055	22.435	47.200%	-1.193%
Maximum Sharpe ratio	0.750%	1.677%	0.439	18.521	32.530%	-1.969%

Table 5.27: Results of the portfolio simulation without including ESG scores

5.3.2 Including ESG scores

When ESG scores are added as a constraint, CDW Corporation (CDW), the company with the lowest ESG score, is added to the portfolio along with the companies that were present in the maximum return portfolio without including ESG score, such as INTC, Warner, Microsoft, and Biogen.

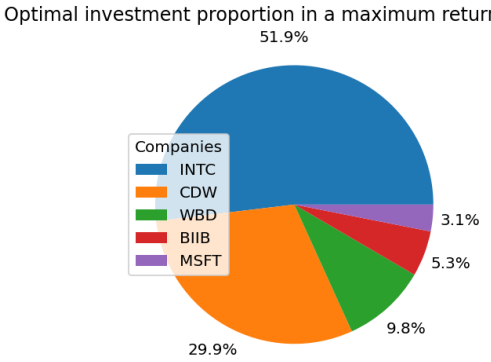


Figure 5.17: Proportion of investment for a maximum return portfolio

A more diversified portfolio is obtained when minimum volatility is set as the goal, as displayed in Figure 5.18. Electronic Arts (EA), CDW Corporation, Keurig Dr Pepper, and Cisco Systems (CSCO) hold between 10 and 14% each.

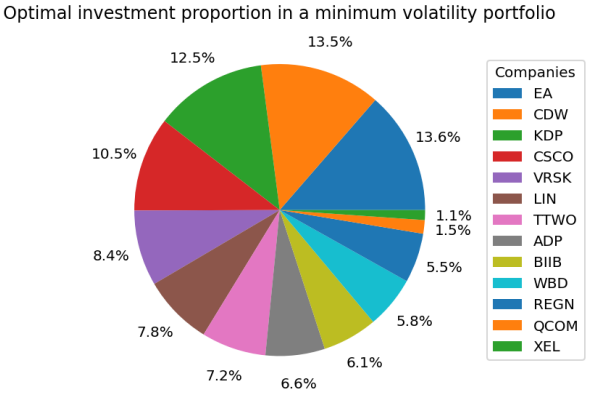


Figure 5.18: Proportion of investment for a maximum return portfolio

The optimal Sharpe ratio portfolio is almost identical to the maximum return portfolio. Both invest in Intel, CDW, Warner, Microsoft, and Biogen, with the last three holding a slightly bigger proportion of the investments.

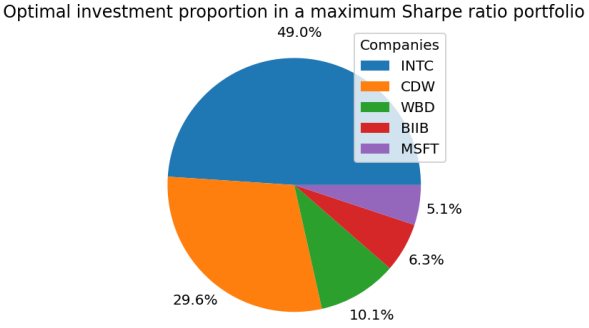


Figure 5.19: Proportion of investment for a maximum Sharpe ratio portfolio

When the ESG score is minimized, a single-stock portfolio is obtained, with CDW holding 100% of the investment amount.

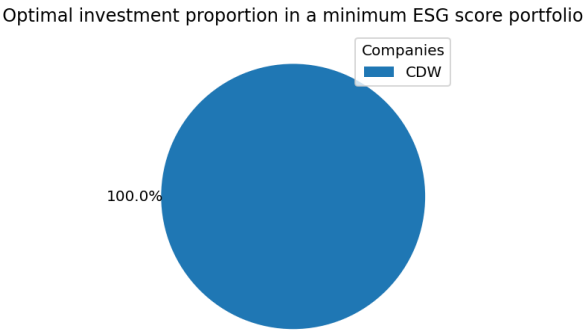


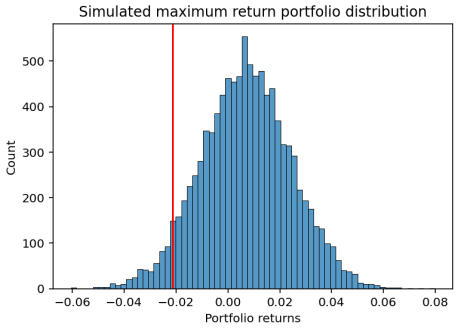
Figure 5.20: Proportion of investment for a minimum ESG score portfolio

In Table 5.28, the different portfolio metrics can be seen. In the same way as in the case when the ESG scores are not included, the maximum return and maximum Sharpe ratio portfolio have similar values, with high return, volatility, and Sharpe ratio, and an ESG score of 15. The minimum ESG score portfolio has a score of 7.5; however, its volatility is high, and its return is between the maximum return and minimum volatility portfolios.

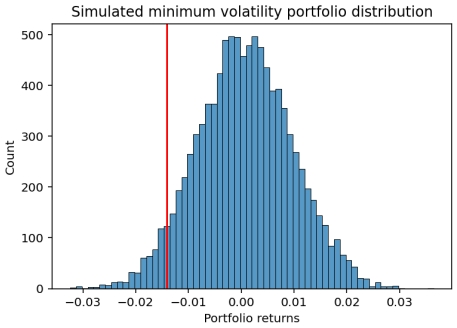
Portfolio	Return	Volatility	Sharpe ratio	ESG score
Maximum return	0.695%	1.729%	0.394	15
Minimum Volatility	0.058%	0.901%	0.047	15
Maximum Sharpe ratio	0.677%	1.681%	0.394	15
Minimum ESG score	0.238%	1.722%	0.129	7.490

Table 5.28: Results of the portfolio optimization including ESG scores

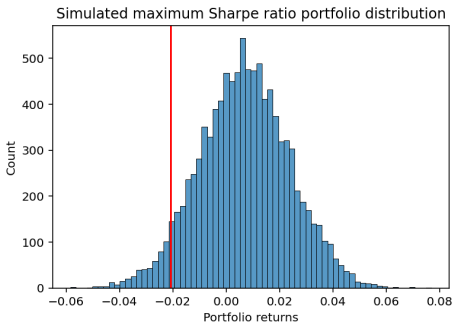
The distributions of the simulated portfolios and the Value at Risk, represented by a red line, can be seen in Figure 5.21. The minimum ESG score portfolio has the highest absolute VaR, which is caused by the absence of diversification in its portfolio. This value can be interpreted as there is a 5% probability that the portfolio will lose at least 2.5% of its value. Similar to the portfolios without considering ESG scores, the maximum return and maximum Sharpe ratio portfolios are close to the -2% VaR, while the minimum volatility portfolio has the lowest value around -1.5%



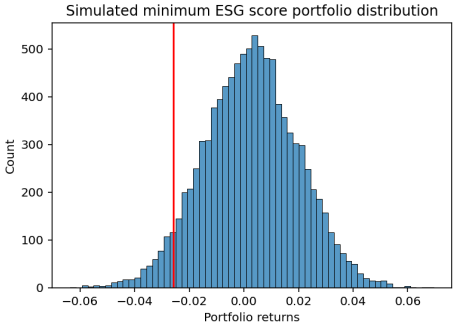
(a) Returns distribution for maximum return portfolio



(b) Returns distribution for minimum volatility portfolio



(c) Returns distribution for maximum Sharpe ratio portfolio



(d) Returns distribution for minimum ESG score portfolio

Figure 5.21: Return distributions and VaR for the simulated portfolios including ESG score

Similar to the case when the ESG score is not included, when the portfolios are simulated, the values of the return, volatility, Sharpe ratio, and ESG score are almost identical to the optimization output. Again, the maximum return portfolio is the one that has a lower probability of loss, with 34.1%. The maximum ESG score portfolio has a high probability of loss, which makes it a risky option and not recommended unless the interest of investors in the ESG factors is greater than the financial performance of their investments.

Portfolio	Return	Volatility	Sharpe ratio	ESG score	Probability of loss	VaR
Maximum return	0.699%	1.735%	0.394	15	34.100%	-2.139%
Minimum Volatility	0.067%	0.901%	0.058	15	47.380%	-1.411%
Maximum Sharpe ratio	0.681%	1.686%	0.395	15	34.100%	-2.079%
Minimum ESG score	0.254%	1.724%	0.139	7.490	43.950%	-2.579%

Table 5.29: Results of the portfolio simulation including ESG scores

5.3.3 Comparing the two approaches

The inclusion of the ESG score as a goal only generated a low-return and high-risk single-stock portfolio. This is not recommended unless there is a greater interest in

ESG factors than in financial performance.

However, when ESG was set as a constraint, the sacrifice in return and volatility is rewarded with a lower ESG score, allowing the creation of a more sustainable portfolio. On Table 5.30, the percentage of variation of ESG portfolio metrics compared to the baseline is shown. For the maximum return and maximum Sharpe ratio portfolios, there is a decrease of approximately 10% in the return and Sharpe ratio, and an increase of 0.1% and 0.5% in volatility. This sacrifice allowed a decrease of 20% of the ESG score, significantly improving its sustainability. On the other hand, the minimum volatility portfolio increased the return and volatility by 17%, making it more profitable, but uncertain; meanwhile, the Sharpe ratio increased 5%, meaning that this portfolio is more profitable per unit of volatility. Its ESG score decreased by 33%, a significant improvement in sustainability.

The values of VaR and the probability of loss are similar; then, under uncertainty, ESG-aware portfolios perform almost as well as the baseline.

Portfolio	Return	Volatility	Sharpe ratio	ESG score
Maximum return	-9.654%	0.115%	-9.946%	-19.358%
Minimum Volatility	17.453%	17.656%	5.029%	-33.141%
Maximum Sharpe ratio	-9.286%	0.553%	-9.970%	-19.010%

Table 5.30: Percentage of variation of ESG portfolio metrics compared to the baseline

As an actual recommendation, it is suggested that, for an aggressive investor profile, the maximum return strategy be followed, mainly investing in Intel Corp, followed by CDW Corporation, Warner, Biogen, and Microsoft. This approach will provide a high-return portfolio, tolerating high volatility. For a risk-adverse investor, it is recommended to follow the minimum volatility portfolio. Investing primarily in Electronic Arts, CDW Corporation, Keurig Dr Pepper, and Cisco Systems. This way, a low-risk portfolio can be obtained at the expense of the return. A balanced approach is achieved with the maximum Sharpe ratio portfolio, modifying the maximum return portfolio to invest a slightly lower amount in Intel Corp, and increasing the weight on the rest of the stocks.

6 Conclusions

Reflecting on the results shown in the previous section, it can be concluded that it is possible to predict returns and volatility using machine learning models, and then construct an investment portfolio with optimization techniques, based on technology sector stocks, such as the NASDAQ-100 index, using return, volatility, and ESG scores. These portfolios outperform the market (S&P500) in terms of return, but not in volatility. Additionally, the portfolios obtained with ESG variables had a performance close to the portfolios obtained without the ESG score, meaning that it is possible to sacrifice a portion of return and volatility to create a more sustainable portfolio.

Specifically for the clustering section of this study, companies were clustered using a simple approach, with historic return, volatility, and volume values. This method outperformed the other 3 in terms of silhouette score: the approach using fundamental analysis variables, the one with all the variables, and the method where PCA is used to reduce the dimensions of the input. Two groups resulted from the selected clustering method: A high return, volatility, and volume group and a low return, volatility, and volume cluster.

In terms of forecasting, fundamental analysis variables, with lagged values of average monthly return, volatility, and volume, were used as predictors with four models: Random Forest, XGBoost, Neural Networks, and SVM. For cluster 0 return, an SVM with a third-degree polynomial kernel was the approach with a lower RMSE, outperforming the Naive model. The volatility of cluster 0 is forecasted using an XGBoost with 100 estimators, being the best-performing model. Similar to cluster 0, cluster 1 return is forecasted with an SVM with a third-degree polynomial kernel. Finally, a Random Forest regressor with 80 estimators was used for cluster 1 volatility. With the returns being highly unpredictable, a non-linear method, such as the SVM with a polynomial kernel, outperforms the tree-based approaches, which can be simple for returns, and the Neural Networks, which can overfit. On the other hand, volatilities are better forecasted by tree-based models, where the data is more structured and predictable.

Then, in the portfolio optimization section, ESG-aware portfolios were constructed using four methods: maximum return, constraining volatility and ESG score; minimum volatility, subject to a minimum return and a maximum ESG score; maximum Sharpe

ratio, constraining return, volatility, and ESG score; and minimum ESG score, with a minimum return and a maximum volatility. These were compared with portfolios that do not use ESG variables, where the ESG-aware portfolios perform similarly to the non-aware portfolios, obtaining similar results in VaR and probability of loss.

The final recommendation to investor varies based on their profile. For an aggressive investor profile, it is suggested that the maximum return strategy be followed, investing in Intel Corp, CDW Corporation, Warner, Biogen, and Microsoft to achieve a high return portfolio and tolerate high volatility. For the risk-adverse investor, it is recommended to invest in multiple companies, such as Electronic Arts, CDW Corporation, Keurig Dr Pepper, and Cisco Systems, as in the minimum volatility portfolio, to achieve a low-risk portfolio, at the expense of the return. If a balanced approach is needed, the maximum Sharpe ratio portfolio is the suggested alternative, obtained by investing in the same companies as in the maximum return portfolio, but modifying the weights, investing a lower amount in Intel Corp, and increasing the investment in the rest of the companies.

The present research contributes to the literature by merging two trends in portfolio management, sustainable finance in the context of a high-return and high-volatility market, as the technology companies in the NASDAQ-100 index. It offers practical insights, such as clustering with simple variables (historic return, volatility, and volume values), which creates more homogenous and interpretable groups than using fundamental analysis variables. Plus, for highly unpredictable variables, like stock returns, SVM can outperform simple methods, such as Naive forecasting, and more complex models, like Neural Networks. Finally, when including ESG scores in portfolio optimization, a small proportion of return and volatility can be sacrificed to obtain a more sustainable portfolio.

Additionally, a new framework is developed, using clustering as a pre-processing step for forecasting, avoiding the computational cost of fitting ML models to each stock. By grouping companies with similar market characteristics, the framework allows the user to fit cluster-level ML models, providing a scalable approach, suitable for large portfolios and realistic applications.

Contrasting with previous studies, the results of the clustering section of this paper agree with Safari-Monjeghtapeh & Esmaeilpour (2024) in the use of long time windows to get stable and informative clusters. The groups formed by fundamental analysis variables, as done by Boloș et al. (2025), lacked homogeneity, especially on cluster 1, with a silhouette score near 0. The same can be said about the PCA clustering done by Jolliffe & Cadima (2016), where the empirical results of this study demonstrated that dimension reduction does not improve the silhouette scores.

In terms of forecasting, this study agrees with Santos Bezerra & Melo Albuquerque (2019) in the use of Support Vector Machine, tending to obtain low RMSE values for return prediction, as well as with Azman et al. (2025) and Gifty & Li (2024) with Random Forest and XGBoost, respectively, for volatility forecasting. On the other hand, the Neural Networks model was not chosen for any clusters return or volatility, making room for improvement by the use of advanced models, such as Long-Short Term Memory Neural Networks and Transformers, contrary to the statement made by Sun et al. (2025).

For portfolio optimization, the use of the ESG score in the constraints generates results where small proportions of the return and volatility are sacrificed to obtain a more sustainable portfolio, agreeing with Torricelli et al. (2022). When the score is placed as part of the objective function, a single-stock portfolio is obtained, selecting the company with the lowest value, without considering the return or volatility, contrary to the study made by Varmaz et al. (2024), where they showed how these portfolios can still perform competitively.

This study is limited by the need for a multi-period portfolio. In this case, transaction costs, such as commissions, slippage, and bid-ask spread, must be added. These costs will allow the portfolio to be rebalanced, changing the investment weights for stocks between periods. In the same vein, ESG scores are dynamic and updated periodically. For simplicity, this paper treats the score as a static variable.

In addition, there are types of variables that could have been included in the analysis, such as calendar-based variables (day, month, and year), macroeconomic indicators (Gross Domestic Product, Inflation, and Interest Rate), and sentiment data (news sentiment score, social media mentions, and insider trading data). These variables could have improved the performance of clustering and forecasting, capturing patterns that were not detected by the variables in the study.

The inclusion of a utility function for portfolio optimization can create more tailored portfolios. This document focuses only on maximum return, minimum volatility, maximum Sharpe ratio, and minimum ESG score, which correspond to specific profiles of investors. The use of the utility function will cover a broader range of investors in the spectrum, including, at the same time, return, volatility, and ESG score in the objective function, with the preference weights that reflect each investor's unique priority.

Finally, the models used in this paper are hardly interpretable. SVM, Random Forest, XGBoost, and Neural Networks work as black boxes, receiving an input and delivering an output without explaining how to get from one step to the other. An interpretable model will retrieve insights about the weight of each variable for the final answer, in the form of coefficients or rules, that can explain how the algorithm transforms the input

into the output.

For future research, it is recommended that the limiting variables be added to the study. A multi-period portfolio with transaction costs and dynamic ESG scores will allow a more realistic simulation of multi-period portfolio strategies, and the addition of calendar-based variables, macroeconomic indicators, and sentiment data can enhance the performance of clustering and forecasting models.

The addition of interpretable models, such as Linear Regression, Decision Trees, Generalized Additive Models, and RuleFit, will improve the model interpretation and can even beat the black box models in accuracy in some cases. With this, models with a set of coefficients or rules can be interpreted for insights about feature importance. In case these models are not able to outperform the black box algorithms, model-agnostic interpretability tools can be applied, like Shapley Additive Explanation (SHAP) or Local Interpretable Model-Agnostic Explanations (LIME). This will show the feature importance of complex models, like Neural Networks, and give them an interpretation.

Beyond this, the study can be replicated in other markets, such as the MSCI World Information Technology Index, including stocks from the United States, European Union, and Japan, which is helpful for international diversification. This would allow to test this framework across different market structures, helping to assess its robustness and generalization on a global scale.

In terms of the ESG score, a breakdown of the components can be included, optimizing for Environmental, Social, and Governance, depending on the investor's interest. At the same time, other scores can be included, either individually or as a weighted average with the Morningstar Sustainalytics score, obtaining a more reliable score with reduced bias. This will allow for targeted portfolio designs that align with specific ESG priorities, though it can introduce complexity in score integration and interpretation.

In summary, this study proposes a new framework for ESG-aware portfolio construction based on a high-return and high-volatility market exemplified by the NASDAQ-100 index, by combining clustering, forecasting, and optimization. It contributes to two emerging topics in the literature, which are sustainable finance methodologies and the growing research on the technology sector. With this approach, ESG-focused investors can obtain sustainable portfolios with minimal sacrifice of return and volatility. Future research that can be built upon this may focus on different technology markets, other ESG scores, or the inclusion of variables, models, and tools that are interpretable, while improving forecast accuracy. The findings of this paper open the door to more granular analysis and adaptable methods that align financial performance with evolving ESG standards.

Bibliography

Abraham, A., Philip, N. S. & Saratchandran, P. (2004), 'Modeling chaotic behavior of stock indices using intelligent paradigms', *arXiv preprint cs/0405018* . Available at https://doi.org/10.48550/arXiv.cs/0405018.

Agarwal, T., Quelle, H. & Ryan, C. (2021), 'Principal Component Analysis for Clustering Stock Portfolios', *Arizona Journal of Interdisciplinary Studies* **7**(2021), 64–75. Available at https://journals.librarypublishing.arizona.edu/azjis/article/id/2384.

Allen, D., Lizieri, C. & Satchell, S. (2019), 'In defense of portfolio optimization: What if we can forecast?', *Financial Analysts Journal* **75**(3), 20–38. https://doi.org/10.1080/0015198X.2019.1600958.

AMPL Optimization Inc. (2025), *AMPLPy Quick Start*, AMPL. "Quick start" section of the AMPLPy Python API documentation, accessed June 2025, https://amplpy.ampl.com/en/latest/quick-start.html.

Aroussi, R. (2025), 'yfinance: Download market data from Yahoo! Finance API', https://pypi.org/project/yfinance/. Apache Software License; latest release as of June 8, 2025.

Arthur, D. & Vassilvitskii, S. (2006), k-means++: The advantages of careful seeding, Technical report, Stanford. Available at http://ilpubs.stanford.edu:8090/778/?ref=https://githubhelp.com.

Atomico and Invest Europe (2024), 'State of European Tech 2024: A Decade of Progress and the Road Ahead'. Available at https://www.investeurope.eu/news/newsroom/state-of-european-tech-2024-a-decade-of-progress-and-the-road-ahead/.

Azman, S., Pathmanathan, D. & Balakrishnan, V. (2025), 'A two-stage forecasting model using random forest subset-based feature selection and BiGRU with attention mechanism: Application to stock indices', *PLoS One* **20**(5), e0323015. https://doi.org/10.1371/journal.pone.0323015.

Beck, N., Dovern, J. & Vogl, S. (2025), 'Mind the naive forecast! a rigorous evaluation

of forecasting models for time series with low predictability', *Applied Intelligence* **55**(6), 395. https://doi.org/10.1007/s10489-025-06268-w.

Board of Governors of the Federal Reserve System (US) (2025), '1-year treasury bill: Secondary market rate', https://fred.stlouisfed.org/series/TB1YR. Accessed June 12, 2025. Monthly average value for May 2025: 3.92%.

Bolos, M.-I., Rusu, , Leordeanu, M., Sabău-Popa, C. D., Perțicaș, D. C. & Crișan, M.-I. (2025), 'K-Means Clustering for Portfolio Optimization: Symmetry in Risk–Return Tradeoff, Liquidity, Profitability, and Solvency', *Symmetry* **17**(6), 847. https://doi.org/10.3390/sym17060847.

Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. J. (2017), 'Classification and regression trees'. https://doi.org/10.1201/9781315139470.

Bulani, V., Bezbradica, M. & Crane, M. (2025), 'Improving Portfolio Management Using Clustering and Particle Swarm Optimisation', *Mathematics* **13**(10), 1623. https://doi.org/10.3390/math13101623.

Celestin, M., Vasuki, M., Kumar, A. D. & Asamoah, P. J. (2025), 'Applications of GARCH Models for Volatility Forecasting in High-Frequency Trading Environments', *Zenodo* **10**, 12–21. https://doi.org/10.5281/zenodo.14904200.

Chapman, P., Kerber, R., Clinton, J., Khabaza, T., Reinartz, T. & Wirth, R. (1999), 'The CRISP-DM Process Model', *no. C* . Available at https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf.

Chin, J. T., Lin, H. & Mei, Y. (2022), 'Machine Learning and the Cross-Section of Stock Returns', *Available at SSRN 4282614* . http://dx.doi.org/10.2139/ssrn.4282614.

Cohen, G. (2023), 'The impact of ESG risks on corporate value', *Review of Quantitative Finance and Accounting* **60**(4), 1451–1468. https://doi.org/10.1007/s11156-023-01135-6.

Deep, A., Shirvani, A., Monico, C., Rachev, S. & Fabozzi, F. (2025), 'Risk-Adjusted Performance of Random Forest Models in High-Frequency Trading', *Journal of Risk and Financial Management* **18**(3), 142. https://doi.org/10.3390/jrfm18030142.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. & Vapnik, V. (1996), Support vector regression machines, Vol. 9. Available at https://papers.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf.

Feng, Y., Zhang, Y. & Wang, Y. (2024), 'Out-of-sample volatility prediction: Rolling window, expanding window, or both?', *Journal of Forecasting* **43**(3), 567–582. https://doi.org/10.1002/for.3046.

Friedman, J. H. (2001), 'Greedy function approximation: a gradient boosting machine', *Annals of statistics* pp. 1189–1232. https://doi.org/10.1214/aos/1013203451.

Gambrah, P. S. N. & Pirvu, T. A. (2014), 'Risk measures and portfolio optimization', *Journal of Risk and Financial Management* **7**(3), 113–129. https://doi.org/10.3390/jrfm7030113.

Ghosh, D. (2025), 'Polyspectral mean based time series clustering of Indian stock market', *Discover Data* **3**(1), 10. https://doi.org/10.1007/s44248-025-00030-w.

Gifty, A. & Li, Y. (2024), 'A Comparative Analysis of LSTM, ARIMA, XGBoost Algorithms in Predicting Stock Price Direction', *Engineering and Technology Journal* **9**(8), 4978–4986. https://doi.org/10.47191/etj/v9i08.50.

Goetzmann, W. N., Ingersoll Jr, J. E., Spiegel, M. & Welch, I. (2002), 'Sharpening sharpe ratios'. https://doi.org/10.3386/w9116.

He, B., Gong, E., Li, L. & Yang, Y. (2023), 'A Stock Price Prediction Method Based on LSTM and K-Means', *Frontiers in Science and Engineering* **3**(6), 44–57. https://doi.org/10.54691/fse.v3i6.5121.

Hossain, A., Kamruzzaman, M. & Ali, M. A. (2015), 'ARIMA with GARCH family modeling and projection on share volume of DSE', *Economics* **3**(7-8), 171–184. Available at https://www.researchgate.net/publication/282731075_ARIMA_with_GARCH_Family_Modeling_and_Projection_on_Share_Volume_of_DSE.

Huang, J. (2024), Prediction of closing prices for nasdaq listed stocks: A comparative study based on gradient boosting models, *in* 'Proceedings of the Highlights in Science, Engineering and Technology, SDPIT 2024', Vol. 92, DRPress, Tianjin, China, pp. 171–177. https://doi.org/10.54097/01rvrr58.

Huang, M., Bao, Q., Zhang, Y. & Feng, W. (2019), 'A hybrid algorithm for forecasting financial time series data based on DBSCAN and SVR', *Information* **10**(3), 103. https://doi.org/10.3390/info10030103.

Infosecurity Magazine (2025), 'Deterring Data Privacy Violations in Big Tech: Why Fines Aren't Enough'. Accessed June 3, 2025.

URL: *https://www.infosecurity-magazine.com/news-features/data-privacy-violations-big-tech/*

Jin, S. (2024), 'A Comparative Analysis of Traditional and Machine Learning Methods in Forecasting the Stock Markets of China and the US.', *International Journal of Advanced Computer Science & Applications* **15**(4). https://doi.org/10.14569/IJACSA.2024.0150401.

Jolliffe, I. T. & Cadima, J. (2016), 'Principal component analysis: a review and recent developments', *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* **374**(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.

Khan, A. A. R. & Tanwani, L. K. (2024), 'Portfolio optimization: Theory, methods, and applications', *ResearchGate* .
URL: *Available at https://www.researchgate.net/publication/381458899_Portfolio_Optimization_Theory_Methods_and_Applications*

Kumar, R. R., Stauvermann, P. J. & Samitas, A. (2022), 'An application of portfolio mean-variance and semi-variance optimization techniques: A case of fiji', *Journal of Risk and Financial Management* **15**(5), 190. https://doi.org/10.3390/jrfm15050190.

Levy-Kramer, J. (2018), 'k-means-constrained'.
URL: *https://github.com/joshlk/k-means-constrained*

Li, J. (2024), Study on the Fitness of ARIMA Model in Stock Forecasting, *in* 'SHS Web of Conferences', Vol. 208, EDP Sciences, p. 01028. https://doi.org/10.1051/shsconf/202420801028.

Li, M., Zhu, Y., Shen, Y. & Angelova, M. (2023), 'Clustering-enhanced stock price prediction using deep learning', *World Wide Web* **26**(1), 207–232. https://doi.org/10.1007/s11280-021-01003-0.

Mallikarjuna, M. & Rao, R. P. (2019), 'Evaluation of forecasting methods from selected stock market returns', *Financial Innovation* **5**(1), 40. https://doi.org/10.1186/s40854-019-0157-x.

Mantegna, R. N. (1999), 'Hierarchical structure in financial markets', *The European Physical Journal B-Condensed Matter and Complex Systems* **11**(1), 193–197. https://doi.org/10.1007/s100510050929.

Markowitz, H. (1952), 'Portfolio selection', *The Journal of Finance* **7**(1), 77–91. https://doi.org/10.2307/2975974.

Miccichè, S., Lillo, F. & Mantegna, R. N. (2005), Correlation Based Hierarchical Clustering in Financial Time Series, *in* 'Complexity, Metastability and Nonextensivity', World Scientific, pp. 327–335. https://doi.org/10.1142/9789812701558_0037.

Morningstar Sustainalytics (2024), *ESG Risk Ratings Brochure*. Accessed June 2025. https://www.sustainalytics.com/esg-data.

Morris, D., Hegarty, P., Nichols, V. & Glynn, D. (2024), 'The technology sector – An update after the summer volatility'. BNP Paribas Asset Management, accessed May 7, 2025.
URL: *https://www.bnpparibas-am.com/en-offshore/2024/09/12/the-technology-sector-an-update-after-the-summer-volatility/*

NASDAQ, Inc. (2025), 'When performance matters: Nasdaq-100® vs. s&p 500'. Available at https://www.nasdaq.com/articles/when-performance-matters-nasdaq-100-vs-s-and-p-500-q1-2025#:~:text=The%20Nasdaq-100%20is%20trailing%20the%20S%26P%20500%20year-to-date,return%20of%20436%25%20on%20a%20total%20return%20basis.

Nethaji, O., Premila, S. et al. (2024), 'Forecasting the stock market using ARIMA modeling and foresight.', *Journal of Computational Analysis & Applications* **33**(7). Available at https://www.researchgate.net/publication/384841942_Forecasting_the_stock_market_using_ARIMA_modeling_and_foresight.

Odah, S. T. (2025), 'Predicting stock price volatility using arch-garch models', *International Journal on Economics, Finance and Sustainable Development* **7**(4), 198–204. Available at https://www.researchgate.net/publication/392006497.

Pedersen, M. (2014), 'Portfolio optimization and Monte Carlo simulation', *Available at SSRN 2438121* . https://doi.org/10.2139/ssrn.2438121.

Plan Be Eco (2024), 'How Tech Companies Approach ESG Reporting'. Accessed June 3, 2025.
URL: *https://planbe.eco/en/blog/how-tech-companies-approach-esg-reporting/*

Qu, J. & Zhang, L. (2023), 'Application of maximum sharpe ratio and minimum variance portfolio optimization for industries', *Highlights in Business, Economics and Management* **5**, 205–223. https://doi.org/10.54097/hbem.v5i.5077.

Ridwan, A., Napitupulu, H. & Sukono, S. (2022), 'Decision-making in formation of mean-var optimal portfolio by selecting stocks using k-means and average link-

age clustering', *Decis. Sci. Lett* **11**, 431–442. https://doi.org/10.5267/j.dsl.2022.7.002.

Rousseeuw, P. J. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics* **20**, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Rukmi, A. M., Wahid, A. et al. (2019), Role of clustering based on density to detect patterns of stock trading deviation, *in* 'Journal of Physics: Conference Series', Vol. 1218, IOP Publishing, p. 012053. https://doi.org/10.1088/1742-6596/1218/1/012053.

Safari-Monjeghtapeh, L. & Esmaeilpour, M. (2024), 'Clustering of listed stock exchange companies active in the cement using the FPC clustering algorithm', *Data-Centric Engineering* **5**, e23. Available at https://www.cambridge.org/core/journals/data-centric-engineering/article/clustering-of-listed-stock-exchange-companies-active-in-the-cement-using-the-fpc-clustering-algorithm/A62C8843EACEB1A9CA38B32CF76DB783.

Santos Bezerra, P. C. & Melo Albuquerque, P. H. (2019), 'VOLATILITY FORECASTING: THE SUPPORT VECTOR REGRESSION CAN BEAT THE RANDOM WALK.', *Economic Computation & Economic Cybernetics Studies & Research* **53**(4). https://doi.org/10.24818/18423264/53.4.19.07.

Sazlı, M. H. (2006), 'A brief review of feed-forward neural networks', *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering* **50**(01). https://doi.org/10.1501/commua1-2_0000000026.

Scikit-Learn Developers (2025*a*), 'sklearn.ensemble.RandomForestRegressor', https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html. Accessed June 2025.

Scikit-Learn Developers (2025*b*), 'sklearn.model_selection.RandomizedSearchCV', https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html. Accessed June 2025.

Scikit-Learn Developers (2025*c*), 'sklearn.model_selection.TimeSeriesSplit', https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html. Accessed June 2025.

Scikit-Learn Developers (2025*d*), 'sklearn.neural_network.MLPRegressor', https:

//scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html. Accessed June 2025.

Scikit-Learn Developers (2025*e*), 'sklearn.svm.SVR', https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html. Accessed June 2025.

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., Kamaev, V. A. et al. (2013), 'A survey of forecast error measures', *World applied sciences journal* **24**(24), 171–176. Available at https://www.researchgate.net/publication/281718517_A_survey_of_forecast_error_measures.

Shmueli, G., Bruce, P. C., Gedeck, P. & Patel, N. R. (2019), *Data mining for business analytics: concepts, techniques and applications in Python*, John Wiley & Sons. Available at https://books.google.ie/books?hl=en&lr=&id=ZEewDwAAQBAJ&oi=fnd&pg=PR19&dq=Data+mining+for+business+analytics:+concepts,+techniques,+and+applications+in+%7BR%7D&ots=CKONLsElmP&sig=BwA7Ix6Vdp2qCO0TOIkUK8-uXG4&redir_esc=y#v=onepage&q=Data%20mining%20for%20business%20analytics%3A%20concepts%2C%20techniques%2C%20and%20applications%20in%20%7BR%7D&f=false.

Suganthi, R. & Kamalakannan, P. (2015), 'Analyzing stock market data using clustering algorithm', *International Journal of Future Computer and Communication* **4**(2), 108. https://doi.org/10.7763/IJFCC.2015.V4.366.

Sun, F.-K., Wu, Y.-C. & Boning, D. S. (2025), 'Simple Feedfoward Neural Networks are Almost All You Need for Time Series Forecasting', *arXiv preprint arXiv:2503.23621* . https://doi.org/10.48550/arXiv.2503.23621.

Thomson Reuters (2024), '2024 State of Corporate ESG: Navigating New Frontiers of Regulation'. Accessed June 3, 2025.
URL: *https://www.thomsonreuters.com/en-us/posts/wp-content/uploads/sites/20/2024/10/2024-State-of-Corporate-ESG-Report.pdf*

Torricelli, C., Bertelli, B. et al. (2022), 'ESG compliant optimal portfolios: The impact of ESG constraints on portfolio optimization in a sample of European stocks', *CEFIN WORKING PAPERS* . DOI: https://dx.doi.org/10.25431/11380_1291994.

USC Viterbi School of Engineering (2022), ''That's just common sense': USC researchers find bias in up to 38.6% of facts used by AI'. Accessed June 3, 2025.
URL: *https://viterbischool.usc.edu/news/2022/05/thats-just-*

common-sense-usc-researchers-find-bias-in-up-to-38-6-of-facts-used-by-ai/

Varmaz, A., Fieberg, C. & Poddig, T. (2024), 'Portfolio optimization for sustainable investments', *Annals of Operations Research* **341**(2), 1151–1176. https://doi.org/10.1007/s10479-024-06189-w.

Wu, Y. (2024), 'Risk Measurement and Portfolio Optimization Based on NASDAQ 100', *Advances in Economics, Management and Political Sciences* **85**, 281–291. https://doi.org/10.54254/2754-1169/85/20240928.

XGBoost Contributors (2025), 'XGBoost Parameters', https://xgboost.readthedocs.io/en/stable/parameter.html. Accessed June 2025.

Zhang, C., Zhang, Y., Cucuringu, M. & Qian, Z. (2024), 'Volatility forecasting with machine learning and intraday commonality', *Journal of Financial Econometrics* **22**(2), 492–530. https://doi.org/10.1093/jjfinec/nbad005.

Zhao, H. (2025), Predicting Stock Prices and Optimizing Portfolios: A Random Forest and Monte Carlo-Based Approach Using NASDAQ-100, *in* 'International Workshop on Navigating the Digital Business Frontier for Sustainable Financial Innovation (ICDEBA 2024)', Atlantis Press, pp. 883–892. https://doi.org/10.2991/978-94-6463-652-9_95.

Zhou, F. (2021), 'Cross-validation research based on rbf-svr model for stock index prediction', *Data Sci. Financ. Econ* **1**(1), 1–20. https://doi.org/10.3934/DSFE.2021001.

A1 Apendix

A1.1 Code

Code repository on GitHub

A1.2 Companies

Ticker	Company name
AAPL	Apple Inc.
ADBE	Adobe Inc.
ADI	Analog Devices, Inc.
ADP	Automatic Data Processing, Inc.
ADSK	Autodesk, Inc.
AEP	American Electric Power Company, Inc.
AMAT	Applied Materials, Inc.
AMD	Advanced Micro Devices, Inc.
AMGN	Amgen Inc.
AMZN	Amazon.com, Inc.
ANSS	ANSYS, Inc.
AVGO	Broadcom Inc.
AXON	Axon Enterprise, Inc.
BIIB	Biogen Inc.
BKR	Baker Hughes Company
CCEP	Coca-Cola Europacific Partners PLC
CDNS	Cadence Design Systems, Inc.
CDW	CDW Corporation
CHTR	Charter Communications, Inc.
CMCSA	Comcast Corporation
COST	Costco Wholesale Corporation
CPRT	Copart, Inc.
CSCO	Cisco Systems, Inc.
CSGP	CoStar Group, Inc.
CSX	CSX Corporation
CTAS	Cintas Corporation
CTSH	Cognizant Technology Solutions Corporation
DXCM	DexCom, Inc.
EA	Electronic Arts Inc.
EXC	Exelon Corporation
FANG	Diamondback Energy, Inc.
FAST	Fastenal Company
FTNT	Fortinet, Inc.
GILD	Gilead Sciences, Inc.
GOOGL	Alphabet Inc.
HON	Honeywell International Inc.
IDXX	IDEXX Laboratories, Inc.
INTC	Intel Corporation
KDP	Keurig Dr Pepper Inc.
KLAC	KLA Corporation
LIN	Linde plc
LRCX	Lam Research Corporation
LULU	Lululemon Athletica Inc.
MCHP	Microchip Technology Incorporated
MDLZ	Mondelez International, Inc.
MELI	MercadoLibre, Inc.
META	Meta Platforms, Inc.
MNST	Monster Beverage Corporation
MRVL	Marvell Technology, Inc.
MSFT	Microsoft Corporation
MSTR	Strategy Incorporated
MU	Micron Technology, Inc.
NFLX	Netflix, Inc.
NVDA	NVIDIA Corporation
NXPI	NXP Semiconductors N.V.
ODFL	Old Dominion Freight Line, Inc.
ON	ON Semiconductor Corporation
PANW	Palo Alto Networks, Inc.
PAYX	Paychex, Inc.
PCAR	PACCAR Inc
PEP	PepsiCo, Inc.
QCOM	QUALCOMM Incorporated
REGN	Regeneron Pharmaceuticals, Inc.
ROP	Roper Technologies, Inc.
ROST	Ross Stores, Inc.
SNPS	Synopsys, Inc.
TMUS	T-Mobile US, Inc.
TSLA	Tesla, Inc.
TTWO	Take-Two Interactive Software, Inc.
TXN	Texas Instruments Incorporated
VRSK	Verisk Analytics, Inc.
VRTX	Vertex Pharmaceuticals Incorporated
WBD	Warner Bros. Discovery, Inc.
WDAY	Workday, Inc.
XEL	Xcel Energy Inc.

Table A1.1: Companies included in the study

A1.3 Clustering results

Cluster 0 companies	
ADBE	HON
ADI	IDXX
ADP	INTC
ADSK	KDP
AEP	LIN
AMGN	MCHP
ANSS	MDLZ
BIIB	MNST
BKR	MSFT
CCEP	NXPI
CDW	ODFL
CHTR	PAYX
CMCSA	PCAR
COST	PEP
CPRT	QCOM
CSCO	REGN
CSGP	ROP
CSX	ROST
CTAS	TMUS
CTSH	TXN
EA	VRSK
EXC	VRTX
FAST	WBD
GILD	WDAY
GOOGL	XEL

Table A1.2: Cluster 0 companies

Cluster 1 companies
AAPL
AMAT
AMD
AMZN
AVGO
AXON
CDNS
DXCM
FANG
FTNT
KLAC
LRCX
LULU
MELI
META
MRVL
MSTR
MU
NFLX
NVDA
ON
PANW
SNPS
TSLA
TTWO

Table A1.3: Cluster 1 companies

A1.4 Forecasting results

Company	Return Forecast	Company	Return Forecast
ADBE	0.220%	HON	0.012%
ADI	0.017%	IDXX	0.136%
ADP	-0.103%	INTC	1.067%
ADSK	0.146%	KDP	-0.012%
AEP	-0.416%	LIN	-0.150%
AMGN	-0.226%	MCHP	0.214%
ANSS	0.168%	MDLZ	-0.169%
BIIB	0.383%	MNST	-0.251%
BKR	-0.260%	MSFT	0.429%
CCEP	-1.249%	NXPI	0.222%
CDW	0.238%	ODFL	0.173%
CHTR	0.065%	PAYX	-0.264%
CMCSA	0.043%	PCAR	-1.474%
COST	-1.440%	PEP	-0.016%
CPRT	-0.094%	QCOM	0.136%
CSCO	0.011%	REGN	0.083%
CSGP	0.052%	ROP	-0.063%
CSX	0.139%	ROST	-0.013%
CTAS	-0.685%	TMUS	-0.617%
CTSH	0.041%	TXN	0.079%
EA	-0.040%	VRSK	-0.126%
EXC	-0.145%	VRTX	-0.674%
FAST	-0.221%	WBD	0.376%
GILD	-1.853%	WDAY	0.230%
GOOGL	0.349%	XEL	0.019%

Table A1.4: Cluster 0 return forecast

Company	Volatility forecast	Company	Volatility forecast
ADBE	4.278%	HON	1.193%
ADI	2.202%	IDXX	1.730%
ADP	1.323%	INTC	2.546%
ADSK	2.075%	KDP	1.132%
AEP	2.286%	LIN	1.430%
AMGN	1.453%	MCHP	4.757%
ANSS	1.514%	MDLZ	2.299%
BIIB	2.060%	MNST	1.538%
BKR	2.267%	MSFT	1.533%
CCEP	1.547%	NXPI	2.762%
CDW	1.722%	ODFL	1.805%
CHTR	1.668%	PAYX	1.523%
CMCSA	1.557%	PCAR	1.793%
COST	2.744%	PEP	1.515%
CPRT	1.607%	QCOM	2.092%
CSCO	1.416%	REGN	2.225%
CSGP	1.877%	ROP	1.387%
CSX	1.622%	ROST	2.148%
CTAS	1.487%	TMUS	1.620%
CTSH	1.564%	TXN	1.890%
EA	1.687%	VRSK	1.329%
EXC	2.814%	VRTX	2.172%
FAST	1.558%	WBD	2.360%
GILD	1.460%	WDAY	2.241%
GOOGL	1.730%	XEL	1.269%

Table A1.5: Cluster 0 volatility forecast

Company	Return forecast
AAPL	0.106%
AMAT	0.141%
AMD	0.248%
AMZN	0.131%
AVGO	-1.367%
AXON	-0.462%
CDNS	0.130%
DXCM	0.169%
FANG	0.060%
FTNT	0.145%
KLAC	0.132%
LRCX	0.151%
LULU	0.138%
MELI	0.185%
META	-0.078%
MRVL	0.038%
MSTR	-9.677%
MU	0.070%
NFLX	0.078%
NVDA	-2.041%
ON	0.156%
PANW	0.070%
SNPS	0.144%
TSLA	0.199%
TTWO	0.118%

Table A1.6: Cluster 1 return forecast

Company	Volatility forecast
AAPL	2.120%
AMAT	2.471%
AMD	3.594%
AMZN	2.295%
AVGO	3.046%
AXON	3.445%
CDNS	1.962%
DXCM	3.825%
FANG	3.099%
FTNT	2.569%
KLAC	2.535%
LRCX	2.954%
LULU	3.670%
MELI	3.009%
META	3.061%
MRVL	4.060%
MSTR	4.081%
MU	3.561%
NFLX	2.888%
NVDA	3.716%
ON	3.568%
PANW	2.670%
SNPS	2.126%
TSLA	4.522%
TTWO	2.107%

Table A1.7: Cluster 1 volatility forecast