Analysis & Prediction of S&P 500 Companies Stock Performance

Data Science Group 5, CFG Degree Spring 2025:
Wei Shi, Avril Childs, Bushra Abodher, Esther Nansubuga, Laura Davies, Laura Lloyd

Introduction

Project Aims and Objectives

The primary objective of this project is to analyse and predict the stock performance of S&P 500 companies by leveraging key financial indicators and macroeconomic data. We aim to:

- Identify the financial metrics most strongly correlated with stock outperformance

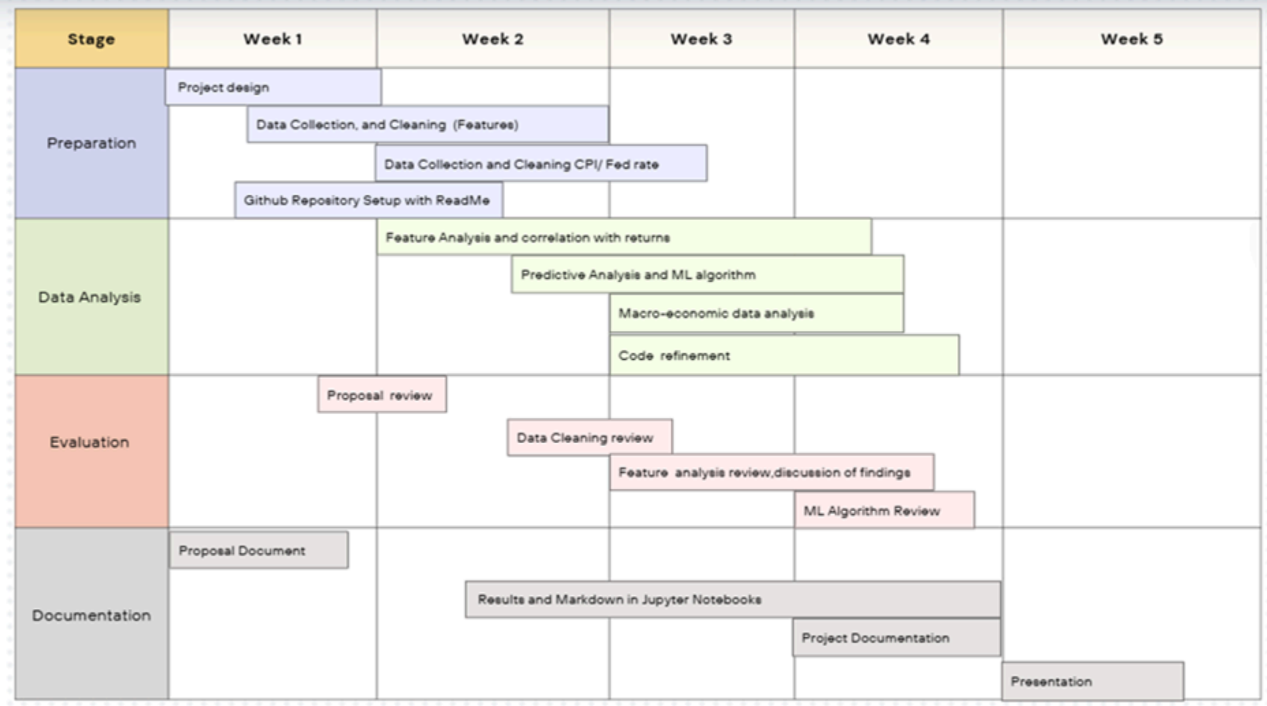
- Analyse macroeconomic indicators, namely Consumer Price Index (CPI) and Federal Reserve Interest Rates to assess their effect on the S&P 500 index performance.

- Develop a Machine learning model to select 10 stocks each month that are likely to outperform the market based on historical data.

Target Audience

Our primary target audience are individual investors and investment analysts and brokers with moderate financial and technical knowledge.

Project Roadmap



Stage	Week 1	Week 2	Week 3	Week 4	Week 5
Preparation	Project design				
	Data Collection, and Cleaning (Features)				
		Data Collection and Cleaning CPI/ Fed rate			
	Github Repository Setup with ReadMe				
Data Analysis		Feature Analysis and correlation with returns			
		Predictive Analysis and ML algorithm			
			Macro-economic data analysis		
			Code refinement		
Evaluation	Proposal review				
		Data Cleaning review			
			Feature analysis review,discussion of findings		
				ML Algorithm Review	
Documentation	Proposal Document				
		Results and Markdown in Jupyter Notebooks			
				Project Documentation	
					Presentation

Background

Individual investors often find it challenging to choose stocks that generate profitable returns while effectively managing risk. Stock performance is influenced by human buying and selling behaviours which are driven by how investors interpret a company's financial health, news about the market or economy and general sentiment. One of the biggest challenges is resisting the urge to make investment decisions based on news headlines and social media hype, rather than fundamental analysis. However, as Warren Buffett famously said, "You don't buy a stock, you invest in a company".

A mix of company-specific financial metrics and broader economic indicators are often analysed to make investment decisions. In this project we focus on **8 common financial features** commonly associated with stock performance, detailed in the following table:

Category	Metric	Description	Stock Outperformance
Valuation Ratios	Price-to-earnings P/E ratio	It measures a company's share price relative to its earnings per share	**Reasonable P/E**
	EV/EBITDA ratio	Compares enterprise value (EV) to operating earnings before interest and tax payment	**Low EV/EBITDA**
	PEG ratio	(price/earning to growth) Adjusts the P/E ratio for the company's earnings growth	**PEG < 1**
Return Ratios	Return on Assets (ROA)	Indicates how efficiently a company uses assets to generate profit	**High ROA**
	Return on Equity (ROE)	Reflects how effectively a company uses shareholder funds to generate profit	**High ROE**
Profitability Margins	EBITDA Margin	Assesses a company's profitability and financial performance	**High EBITDA margin**
	EPS Growth	(earnings per share) Tracks profit per share over time	**High EPS Growth**
Cash Flow	CFO Growth	(cash flow from operations growth) monitors cash from core operations	**Consistent CFO**

Table 1 : 8 Financial Features

Stock performance is also shaped by macroeconomic indicators, which influence investor confidence and spending behaviours. This project studies the following:

1. **Consumer Price Index (CPI):** This is a measure of inflation and rising CPI indicates increasing prices and reduced purchasing power which can reduce consumer spending and corporate profits.
2. **Federal Reserve interest rate (also known as the federal funds rate):** An increase in rates makes borrowing more expensive, which can slow economic growth and impact corporate earnings.

Requirements

Technical Requirements:

- **Jupyter Notebook:** source code in Python and Markdown.
- **Microsoft Excel and CSV:** S&P 500 data storage.
- **Statistical Libraries (Python):** Numpy, Pandas, Scipy, Sci-kit Learn, Joblib.
- **Data visualisation libraries (Python):** Matplotlib, Seaborn, Plotly.
- **GitHub:** granted access control exclusively to our group members.

Non-technical Requirements:

- **Financial and Economic Literacy:** Basic understanding of company financial terminology and economic concepts.
- **Statistical Knowledge:** Familiarity with fundamental statistics, including summary metrics (mean, median, maximum, minimum), frequency distributions, and percentiles.
- **Purpose and Success Criteria:** Prioritizing prediction accuracy and usability.
- **User-Friendly Design:** Ensuring clear interpretability—delivering understandable insights, reports, and visualizations.
- **Accessibility:** Making results accessible to investors with non-technical backgrounds.

Data Collection

Data Requirements

- S&P 500 companies monthly stock prices and the 8 financial indicators *(see table 1)*
    - Jan 2015 - July 2024 for individual feature analysis and Machine Learning model training and testing
    - August 2024 to Dec 2024 for out-of-sample testing
- Macroeconomic data via APIs (CPI and Federal Reserve Interest Rate)

Data availability and sources

Monthly Stock Prices, Financial Indicators and Ratios:

Data sourced from S&P CapitalIQ, a third-party data vendor, compiled into Excel format. It includes monthly stock pricing data and financial ratios for the eight study features from 2015 to 2024.

Macro-economic data:

- Consumer Price Index (CPI) was obtained through the U.S. Bureau of Labor Statistics (BLS) Public Data API, which provided historical time series data in JSON format. The API allowed us to query multiple time series across various time periods, ensuring comprehensive analysis.

- Federal Interest Rate (EFFR) data was obtained via a RESTful API from the New York Fed website. Users had to provide inputs on the duration required to retrieve the correct endpoint. The remote API can be found here.

Data Cleaning and Transformation

Data cleaning strategies were tailored to each financial feature, addressing its unique challenges based on the analysis *(refer to Appendix 1).*

1. **Missing values and Not Meaningful (NM) values**

- Companies with substantial missing data (>50%) were dropped.
- Where relevant, NM/NaN and missing values were replaced with zeros, previous values (forward filling) or an aggregate value.

2. **Replacing missing/ zero values**

- Zero values were either replaced with either 0, previous values (forward filling) or an aggregate value depending on the feature.

- For features that can have zero values (e.g. ROA or ROE), long-run zero trends were analyzed to identify anomalies, which were then treated as NM values.

3. **Outlier management**

- This is feature-specific and exploratory analysis was first applied to determine threshold values (e.g. PEG value > 100 is not meaningful).
- Winsorization was applied to remove extreme values.

4. **Data transformation:** The feature and stock data are initially structured in a time-series wide format, with each feature stored in separate dataframes. To facilitate correlational analysis with returns, the data is converted into a long format.

Implementation and Execution

Development Approach

We followed a Scrum and Agile approach to structure our workflow efficiently:

- After conducting a **SWOT analysis** we allocated roles based on individual strengths, and adopted a **feature-driven development approach**, ensuring each team member took ownership of a specific financial feature for exploration, cleaning, and analysis.
- As project deadlines approached, we structured our workflow by breaking down tasks, and working in **sprints** to maintain efficiency.
- Team members independently analysed their features in separate Jupyter Notebook files while using GitHub branches for **version control** and to merge updates.
- Cleaned data for each feature was shared and used within the machine learning and predictive modelling.
- Regular **Stand-Up meetings** were held to share analyses, conduct peer code reviews, discuss progress, and address roadblocks—fostering collaboration, adaptability and continuous improvement.
- Using an **iterative approach**, we continuously refined our data analysis based on insights gained through exploratory analysis, adapting methods as needed. For example, we emphasized **refactoring** (e.g. encapsulating code within functions) to enhance readability and reusability.

Team Member Roles:

Member	Roles	Role Assignment	Data Analysis	Documentation
Wei	Technical Lead	Lead on stock market analysis, overall methodology, statistical analysis and model design	Returns, CFO Growth ML algorithm And testing	Proposal Project document
Avril	Project Lead	Facilitating meetings, oversee project progress, code, project milestones and requirements	PEG ratio ROC CFO Growth	Proposal Project document
Bushra	Task Coordinator	Delegating tasks and tracking team activities	P/E Ratio Fed Rate (API)	Presentation slides
Laura L	Submissions Lead	Collating, editing, and submitting of project assignment documents	EPS Growth CPI (API)	Proposal Project document
Laura D	GitHub Coordinator	Managing GitHub repository, resolving merge conflicts, source code file	ROE EBITDA Margin	README file Source code
Esther	Presentation Lead	Oversee delegation of presentation slides, script and time keeping.	EV/EBITDA	Presentation slides

Table 2: Team task assignments

Data Analysis

The following data analysis were undertaken using python libraries within Jupyter notebooks:

1. **Stock Performance**:

This was analysed using Monthly Returns (%) for feature analysis. For the ML model, classification of Sharpe Ratio (a measure used to evaluate the risk-adjusted return of an investment) was used:

Sharpe ratio >= 2 - "outperforming" (label = 1)
Sharpe ratio < 2 - "underperforming" (label = 0)

2. **Individual Feature and Macro-economic Analysis:**

The 8 key financial features *(refer to Appendix 2)* and two macroeconomic factors *(refer to Appendix 4*) were time-aligned and analysed:

Exploratory data analysis:

- **Distribution Plots**: Visualizing spread and skew to understand data distribution and outliers
- **Aggregate Study** : Calculation of mean and median to summarize key trends.
- **Interquartile Ranges and proportions**: Assessing data dispersion and

Correlational Analysis

- **Correlation Matrix**: Examining relationships between financial features, stock returns, and macroeconomic data with the S&P 500 index.
- **Linear correlation**: Used Pearson's coefficient via Pandas library for both company-level and aggregated data.
- **Data Visualization**: Represented through scatter plots, line plots and heatmaps to illustrate patterns and relationships.
- **Quartile Analysis**: Leveraging quartiles to assess data distribution and correlation strength.

Predictive Algorithm and Machine Learning

Methodology:

- Three models— Logistic Regression with LASSO regularization, Random Forest, and Neural Networks— to compare their performance

- Target variable is categorical, representing performance outcomes: 0 for underperforming and 1 for outperforming.

- The features used in the models include the EV/EBITDA ratio, P/E ratio, PEG ratio, ROA, ROE, EBITDA margin, EPS growth and CFO growth.

- The models were trained on data between January 2015 and January 2022 and tested on data after January 2022.

- Grid Search was used for hyperparameter tuning, identifying the best parameter combinations for the models. *(see Appendix 4)*

Model Evaluation and Testing :

The model performances were evaluated using the following methods (*see Appendix 4*) to determine the best performing model:

- Confusion Matrix (which included Precision, recall and F1 Score)
- Receiver Operating Characteristic (ROC) curve
- Real-life Application Testing (Out-of-Sample) testing was applied using data between August 2024 and December 2024

Implementation process and challenges

Implementation Achievements:

- **Data Processing Success** – Effectively cleaned, transformed, and structured large datasets, ensuring high-quality input for machine learning.
- **Model Implementation** – Successfully developed and deployed machine learning models.
- **Agile & Iterative Approach** – Continuously adapted methods to align with the project's overarching goal of building a machine learning model.
- **Strong Team Coordination** – Regular stand-up meetings fostered a problem-solving collaboration and a cohesive team.
- **Clear documentation** – Well-structured code comments, README files, and commit messages enhanced understanding and ensured seamless team alignment.

Implementation Challenges:

- **Data Complexity & Cleaning** – Managing and analysing a large dataset while maintaining data integrity posed challenges. For example, our API was limited to 10 years , while the financial dataset extended further.. To resolve this, we aligned all data to a common timeframe.
- **Financial and statistical knowledge -** The team had varying levels of financial and statistical expertise, which was addressed through knowledge sharing and research and appointing a Technical Lead with a background in stock trading and statistical knowledge to provide key guidance.
- **Time Constraints** – Tight deadlines required balancing project tasks with daily responsibilities, family life, and time zone differences. To mitigate this, we appointed a Project Lead to keep the team on track and held regular meetings (2 - 3 times a week) to provide opportunities for discussion. Team members supported each other by stepping in or reallocating tasks to keep the project on track.
- **Code Integration and standardisation** – Team members utilized different development environments , occasionally resulting in compatibility issues. We addressed this by standardising file formats (eg. cleaned data files from `.xls` to `.csv`.) , implementing peer code reviews, aligning cleaning and analytical strategies where possible. A Github Coordinator was appointed to manage merge conflicts and maintain code organisation.

Conclusions

Summary of Findings

Financial Features : Analysis of the 8 financial features individually did not reveal any strong relationship with monthly stock price returns at individual company data level for shorter term returns. P/E ratio had the highest correlation but was still scored low at 0.1, suggesting no single feature on its own significantly correlates statistically in the short-term returns.

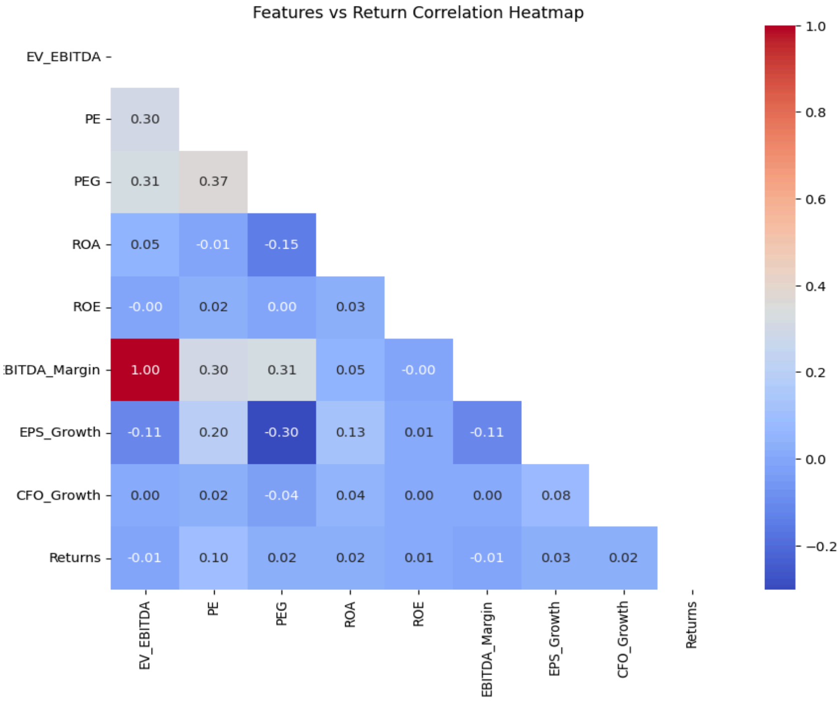


Fig 1. Heatmap of correlations between each financial feature and monthly price return

Analysis using lagged returns and rolling aggregates however showed potential correlation between some features and returns, and signal that there is potential influence in longer term stock returns *(see Appendix 2).*

Predictive analysis using machine learning identified EPS growth and ROA are the top 2 features of importance. However, it must be noted that scores were closely matched between all the features.

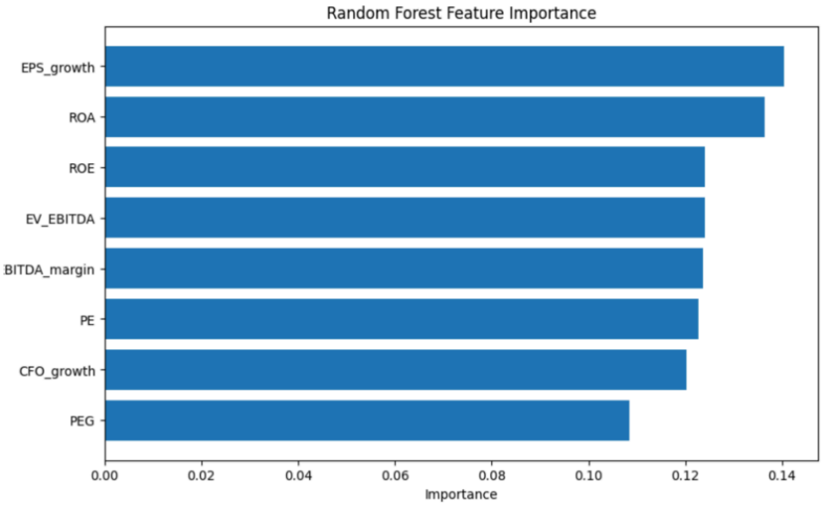


Fig.2 Feature Importance using Random Forest

Consumer Price Index (CPI) Federal Reserve Interest Rates

Relationship Between CPI & Stock Prices Inflation impacts stock performance differently depending on economic policies and investor sentiment. Periods of high inflation (like 2021-2022) saw stock market corrections but also sector-specific gains (tech & AI). Lower inflation years (2015, 2024 expectations) have often fueled stronger stock rallies due to Federal Reserve rate cuts & economic optimism. (*See Appendix 3, Fig A3.1*).

Federal Reserve Interest Rates (EFFR)

In 2015–2018 there was a gradual rate hike with steady stock growth due to strong economic sentiment. When rates slashed to near-zero due to COVID, the market rebounded quickly. In 2022–2023, there was a sharp EFFR increase to fight inflation, which lead to a decrease in stock prices. However, in 2024, though rates stabilized at high levels and yet the stock market surged on optimism *(see Appendix 3, fig. A3.2)*

Higher EFFR generally leads to lower stock growth due to costlier borrowing, and more cautious investing and lower rates lead to higher stock growth (cheaper loans, more spending). However, the stock market reacts not just to rates, but also to Federal Reserve signals and general economic outlook.

Machine Learning Model: The project also successfully established a functional stock selection model *(see Appendix 4 for more details)*. However, out-of-sample testing did not conclusively demonstrate market outperformance, potentially due to limitations in data scope, feature selection, or model calibration.

Data analysis Strategy Review

Our project leveraged robust team collaboration to conduct thorough data cleaning, gaining deep insights into the target variable, eight key financial features and macroeconomic factors. We analysed their distributions trends, correlations, and outliers, laying a strong foundation for feature engineering. The derived features proved instrumental in enhancing the predictive power of our machine learning models.

We developed and evaluated three machine learning models, Logistic Regression with LASSO regularization, Random Forest, and Neural Networks. After rigorous comparison, the Random Forest model demonstrated superior performance and was selected as the optimal approach for this project.

Recommendations for Future Study

In-depth study on external factors and macro-economic influences

Further exploratory research is needed:

- Deeper investigation into the causes of non-meaningful values and outliers.
- Additional analysis using longer-term stock returns is required to assess the impact of financial features on long-term investments, taking into consideration lagged reaction to the release of these indicators.
- Further study should also account for external events and macroeconomic factors as confounding factors.

To enhance the model's precision and predictive capability, we also propose the following improvements:

- Extending the training data period to capture a broader range of market conditions.
- Refining feature selection by incorporating domain-specific expertise.
- Integrating the Fama-French 5-Factor model to account for additional risk factors.
- Experimenting with alternative definitions of "outperform," such as adjusting the Sharpe Ratio threshold (currently set at >2).

Document Authors: Avril Childs, Laura Lloyd and Wei Shi

APPENDIX 1 - Examples of Data Cleaning Strategies

1. Analysis of data for Zeros and Data anomalies

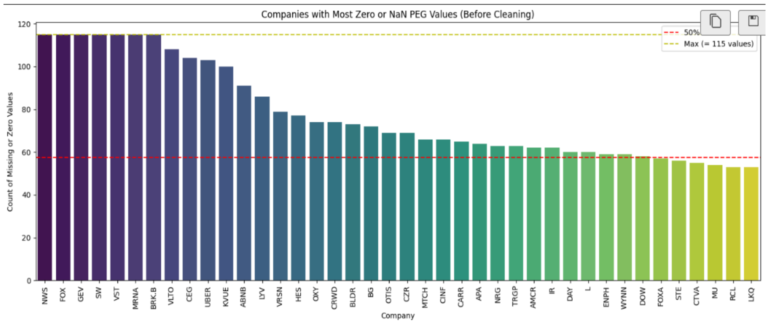


Fig A1.1 Visualising Zero Counts in PEG Ratio data

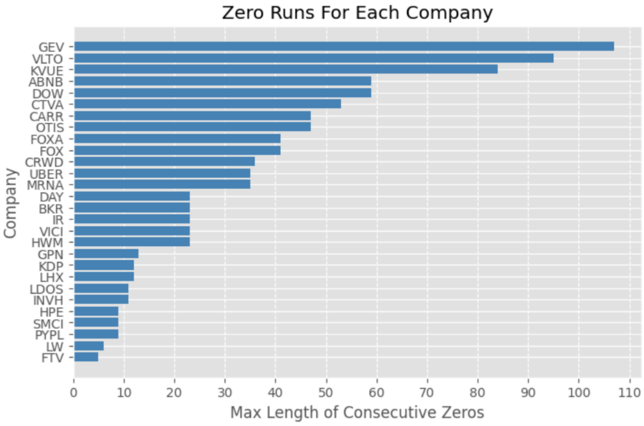


Fig. A1.2 Visualising Zero runs in ROA

Zeros were dealt with dependent on the financial feature. For example, PEG ratios do not have zero values and so if the number of zeros is more than 50% of the column, the columns are dropped. However in ROA where zeros can be valid, zeros values are examined for long runs which can indicate data errors.

2. Outlier analysis and management

Winsorization involves transforming extreme values to reduce their impact without entirely removing them. For instance, values below the 5th percentile may be set to the highest value within that percentile, and those above the 95th percentile may be set to the lowest value within that percentile.

Histograms were used to analyse frequency and distribution to determine outliers and examine effectiveness of winsorization and outlier management. For example, after winsorization, number of outliers remained high. PEG > 100 is deemed not meaningful, so a values > 100 were dropped - which is visualised in the change in the x-axis limit.

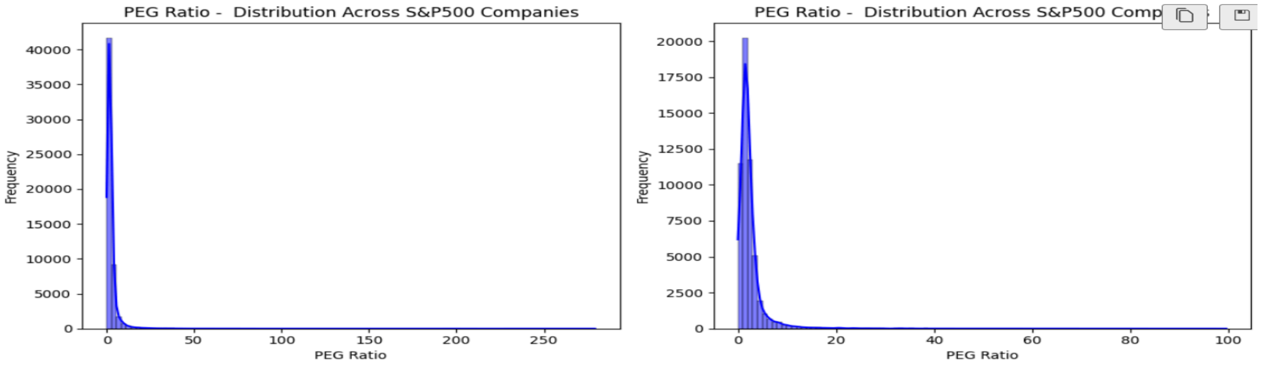
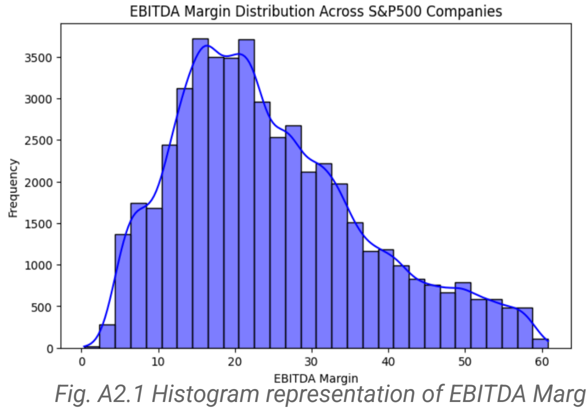


Fig A1.3 Histogram of PEG ratio before and after winsorisation and filtering PEG < 100

APPENDIX 2 - Examples of Data Analysis - Financial Features/Indicators

1. Exploratory analysis - Frequencies and distribution



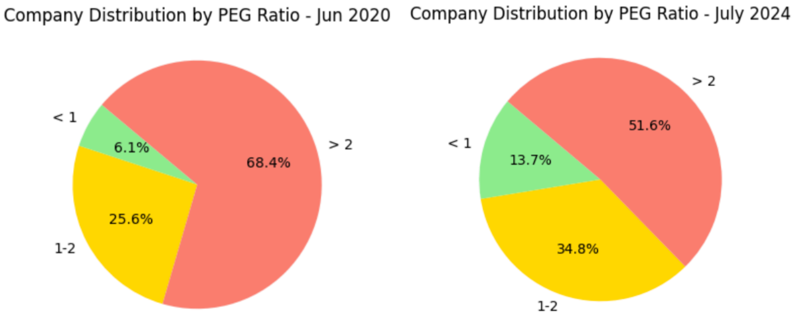
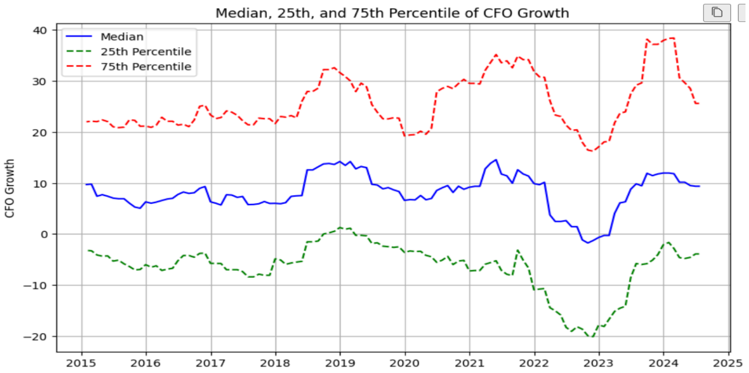


Fig. A2.1 Histogram representation of EBITDA Margin

Fig A2.2 Pie chart of PEG Ratio year 2020 vs 2024

Each feature was explored using descriptive analysis such as frequency, proportions, and aggregates. For example, the pie chart in the PEG analysis was used to observe the difference between Jun 2020 during the COVID-19 pandemic and now.

2. Time Series - Median graphs



Median over time was plotted in time-series line graphs to provide identifying cyclical patterns or significant events affecting stock prices. Interquartile ranges were used to establish if the trend is similar between lower and upper quartile ranges.

Majority of the features showed a significant change in 2020 during the pandemic. CFO Growth was an exception and the downtrend in 2022 was attributed to the increase in the Federal Interest Rate.

Fig A2.3 Median CFO Growth values over Time

Feature vs Returns

3. Scatter Plots

Scatter plots were also used to visualise correlation between the features and returns. In this instance P/E ratio appear to congregate along the zero line, confirming the correlation coefficient value of 0.0947 of this feature.

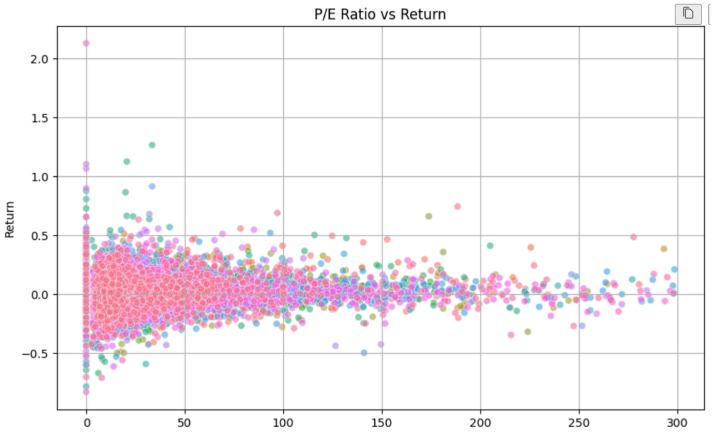


Fig. A2.4 Scatter plot of P/E Ratio vs Returns

4. Using aggregated averages:

Medians were compared over time. A time-matched correlation was not obvious, but when the graphs were smoothed with rolling medians, trends emerged suggesting longer term correlations.

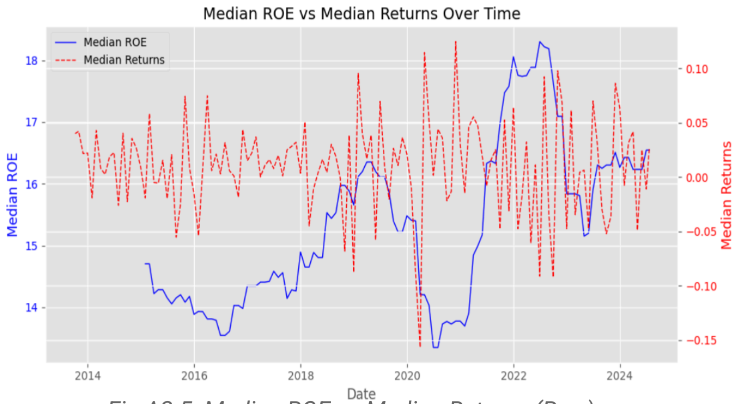


Fig A2.5 Median ROE vs Median Returns (Raw)

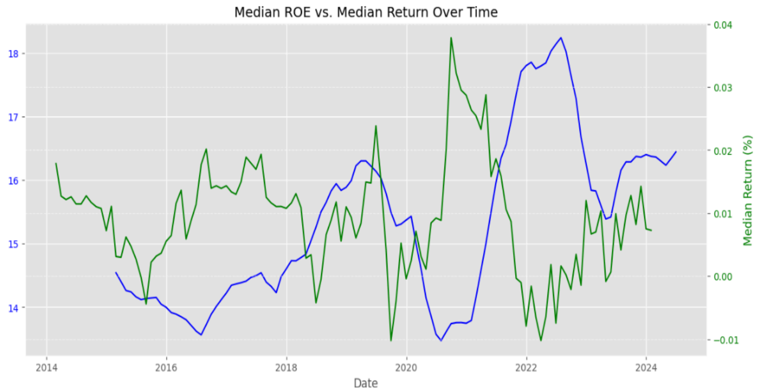


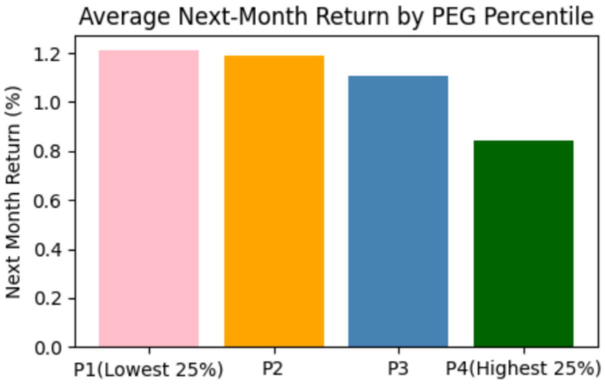
Fig A2.6 - Smoothed Rolling Medians of ROE vs Returns

5. Analysing Quartiles with next-month returns

Quartile bar plots are used to visualise how each percentile performs in the next month stock returns and is a useful way to see if there is possible lagged relationship between the financial feature and returns

This graph suggests that on average the companies with lowest percentile PEG ratios experience the highest next month returns.

Fig. A2.7 Quartile graph for PEG Ratio



APPENDIX 3 - Data Analysis - CPI/FED Rate

Consumer Price index (CPI) vs S&P 500 Returns

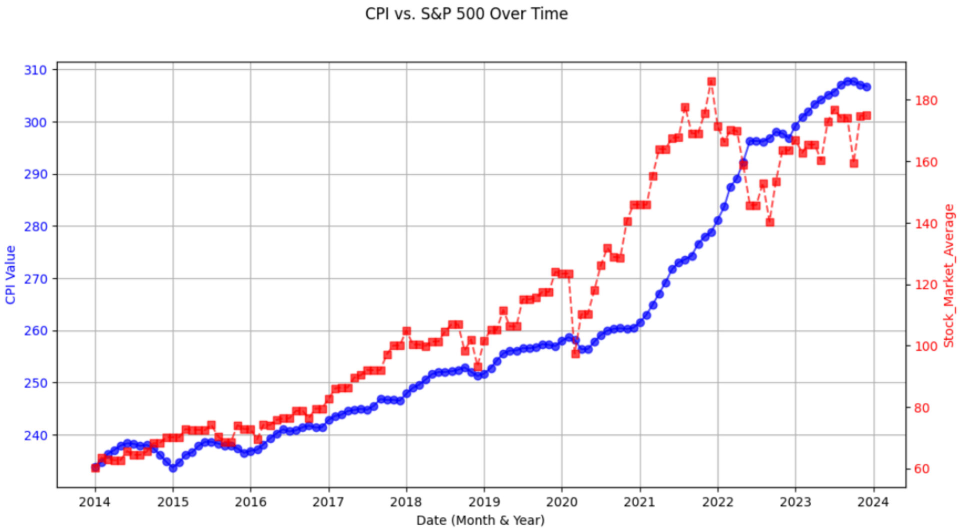


Fig. 3A.1 Graph of CPI vs S&P stock returns over time

Federal Reserve Interest Rates vs S&P 500 Returns

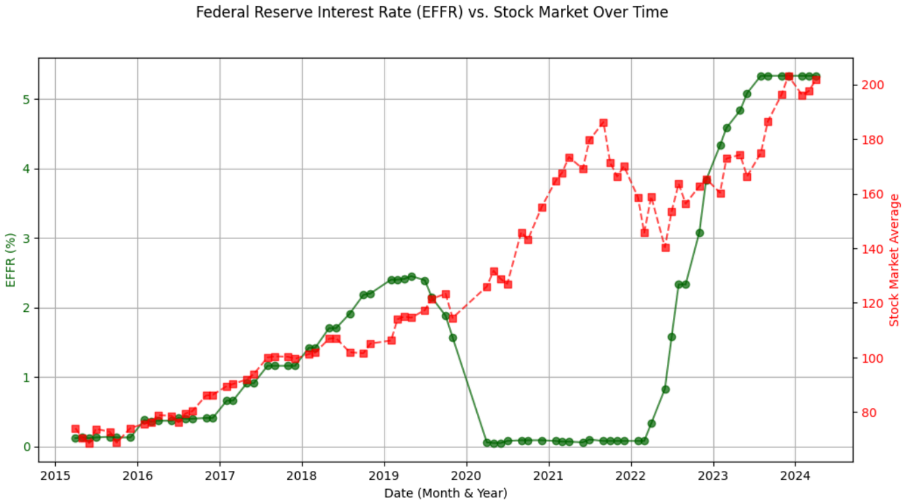
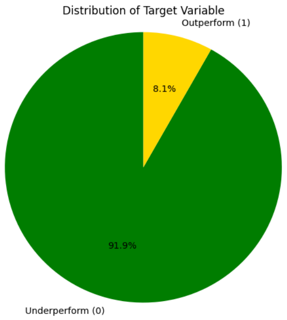


Fig. 3A.1 Graph of Federal Reserve Interest Rates vs S&P stock returns over time

APPENDIX 4 - Predictive Modelling and Machine Learning

The Imbalance in the distribution of target variable



The dataset is imbalanced, with over 90% of samples belonging to class 0 (underperforming) and only around 8% to class 1 (outperforming).
To address this during model training, we use the parameter class_weight='balanced'. This instructs the model to automatically assign weights to each class inversely proportional to their frequency, so that the minority class receives higher importance and the model does not become biased toward the majority class.

Fig 4A.1 Pie chart of Target Variable - Underperforming vs Outperforming

Model Evaluation

Using ROC curves

- Logistic Regression
  - Precision (1): 0.08
  - Recall (1): 0.56
  - F1-score: 0.14
- Random Forest
  - Precision (1): 0.12
  - Recall (1): 0.46
  - F1-score: 0.20
- Neural Networks
  - Precision (1): 0.14
  - Recall (1): 0.26
  - F1-score: 0.18

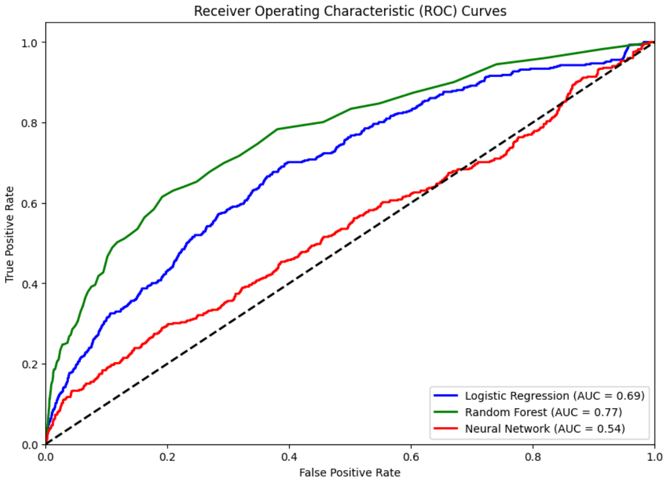


Fig 4A.2 ROC Graph for Model Valuation

The ROC curve indicates that the **Random Forest model** performed the best. Therefore, we selected Random Forest for our final analysis.

Confusion Matrices - with and without hyperparameter tuning

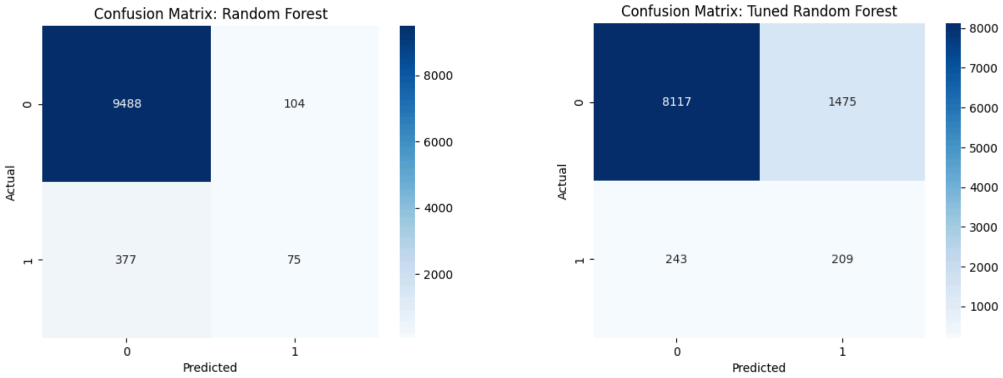
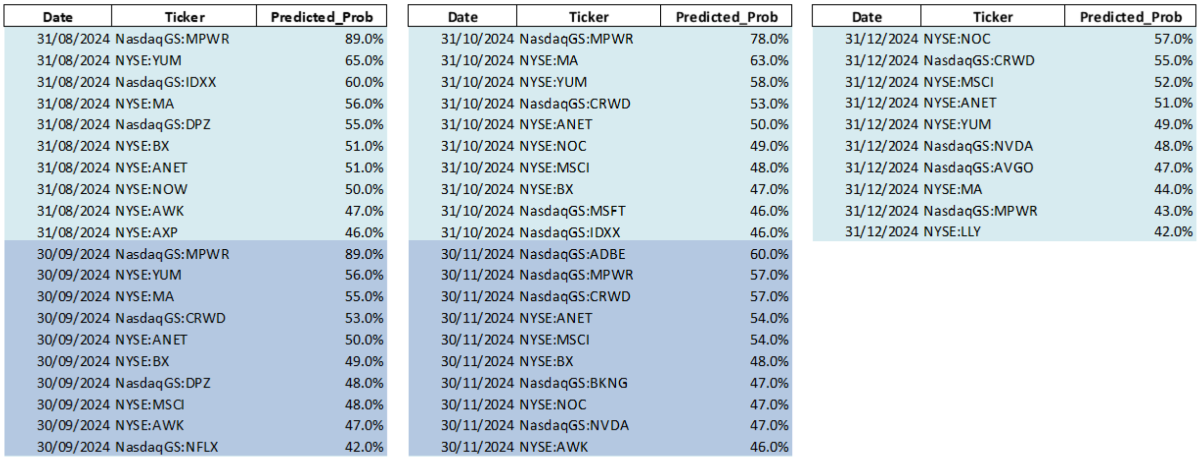


Fig A4.2 Confusion matrices of Random Forest before and after hyperparameter tuning

3. Real-life Application Testing

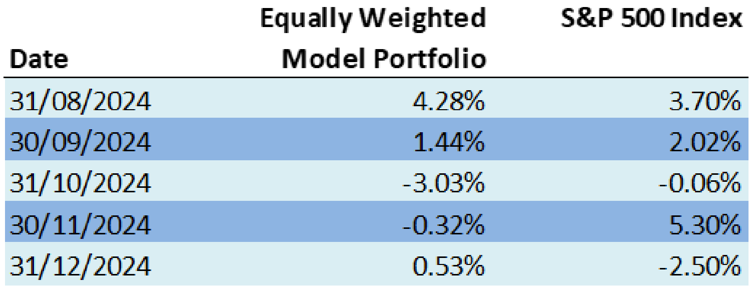
We applied the model in a real-world setting using data from September to December 2024. Each month, the model generated a list of stocks with their predicted probabilities of outperforming the market.



Date	Ticker	Predicted_Prob
31/08/2024	NasdaqGS:MPWR	89.0%
31/08/2024	NYSE:YUM	65.0%
31/08/2024	NasdaqGS:IDXX	60.0%
31/08/2024	NYSE:MA	56.0%
31/08/2024	NasdaqGS:DPZ	55.0%
31/08/2024	NYSE:BX	51.0%
31/08/2024	NYSE:ANET	51.0%
31/08/2024	NYSE:NOW	50.0%
31/08/2024	NYSE:AWK	47.0%
31/08/2024	NYSE:AXP	46.0%
30/09/2024	NasdaqGS:MPWR	89.0%
30/09/2024	NYSE:YUM	56.0%
30/09/2024	NYSE:MA	55.0%
30/09/2024	NasdaqGS:CRWD	53.0%
30/09/2024	NYSE:ANET	50.0%
30/09/2024	NYSE:BX	49.0%
30/09/2024	NasdaqGS:DPZ	48.0%
30/09/2024	NYSE:MSCI	48.0%
30/09/2024	NYSE:AWK	47.0%
30/09/2024	NasdaqGS:NFLX	42.0%

Date	Ticker	Predicted_Prob
31/10/2024	NasdaqGS:MPWR	78.0%
31/10/2024	NYSE:MA	63.0%
31/10/2024	NYSE:YUM	58.0%
31/10/2024	NasdaqGS:CRWD	53.0%
31/10/2024	NYSE:ANET	50.0%
31/10/2024	NYSE:NOC	49.0%
31/10/2024	NYSE:MSCI	48.0%
31/10/2024	NYSE:BX	47.0%
31/10/2024	NasdaqGS:MSFT	46.0%
31/10/2024	NasdaqGS:IDXX	46.0%
30/11/2024	NasdaqGS:ADBE	60.0%
30/11/2024	NasdaqGS:MPWR	57.0%
30/11/2024	NasdaqGS:CRWD	57.0%
30/11/2024	NYSE:ANET	54.0%
30/11/2024	NYSE:MSCI	54.0%
30/11/2024	NYSE:BX	48.0%
30/11/2024	NasdaqGS:BKNG	47.0%
30/11/2024	NYSE:NOC	47.0%
30/11/2024	NasdaqGS:NVDA	47.0%
30/11/2024	NYSE:AWK	46.0%

Date	Ticker	Predicted_Prob
31/12/2024	NYSE:NOC	57.0%
31/12/2024	NasdaqGS:CRWD	55.0%
31/12/2024	NYSE:MSCI	52.0%
31/12/2024	NYSE:ANET	51.0%
31/12/2024	NYSE:YUM	49.0%
31/12/2024	NasdaqGS:NVDA	48.0%
31/12/2024	NasdaqGS:AVGO	47.0%
31/12/2024	NYSE:MA	44.0%
31/12/2024	NasdaqGS:MPWR	43.0%
31/12/2024	NYSE:LLY	42.0%



Date	Equally Weighted Model Portfolio	S&P 500 Index
31/08/2024	4.28%	3.70%
30/09/2024	1.44%	2.02%
31/10/2024	-3.03%	-0.06%
30/11/2024	-0.32%	5.30%
31/12/2024	0.53%	-2.50%

The results show that the model-selected portfolio outperformed the S&P 500 Index in 3 months out of 5 (Aug, Nov, Dec 2024).
It did not show clear evidence of consistently outperforming the market

Fig 4A.3 Results of real-life test for data Aug 2024 - Dec 2024