Machine Learning

Semester Project

Spyridoula Spyrakou (mtn2321) - Christos Georgios Foukanelis (mtn2324) - Panagiotis Tsilimidos (mtn2322)

Project Report

February 2024

Table of Contents

Problem Explanation

We selected stocks as a subject for this assignment.

Each stock (ticker is the name of the stock as it is registered in the stock exchange) is a company (from now on, company and share will be synonymous and mean the same thing) that is listed on a stock exchange and publicly traded. A company is required to share two things to stay public:

- The value (stock price - market capitalization) of the company
- Its financials must be published on a quarterly basis

Many models focus on time-series machine learning, i.e. based on the share price of previous moments/days to try to predict the next moment/day's price for quick gains. Something like this is called trading, it is a whole branch in finance now due to computers and is even used by large funds.

But trading is not investing in the classical sense of the term. Our model focuses on the investment side of stocks which has a longer-term horizon (months/years). Essentially, trading only takes into account one of the two elements that are required to be public, the value of the company, while for pure investing we mainly used the financials of the company to do fundamental investing.

Specifically, each company at the end of its fiscal quarter publishes three statements:

- **Income statement:** It contains the company's revenue, it is the accumulated revenue during that entire quarter.
- **Balance sheet:** Contains the company's assets and liabilities (debts, etc...). The balance sheet is a snapshot of the company at the end of the quarter.
- **Cash flow statement:** Contains the cash flow. How much money is flowing into the company, its source and where it is outflowing. Similar to the income statement, it contains the total cash flow during the quarter.

Data Preparation

Dataset and Model Goal

Using the data from these three statements we created rows and columns with the following rationale:

1. In each row we have consolidated the statements for one company for one year
2. Each row is a data point for training the model
3. We remove the date and the company name at the end to avoid bias
4. The columns are all the rows of the three statements that have now been transposed into columns
5. The columns contain the numbers for each company and year, for the entire year.

Data Cleanup – Curation

Dataframes transformation X

We pulled the data frames from Morningstar (https://www.morningstar.com) for the training data - features (X). The data frames were in this format for all three statements (excel files):

	SKYT_balance-sheet_Annual_As_Originally_Reported	2019	2020	2021	2022
0	Total Assets	190435000.0	263209000.0	263598000.0	305764000.0
1	Total Current Assets	84517000.0	76572000.0	74397000.0	116551000.0
2	Cash, Cash Equivalents and Short Term ...	4605000.0	7436000.0	12917000.0	30025000.0
3	Cash and Cash Equivalents	4605000.0	7436000.0	12917000.0	30025000.0
4	Inventories	15994000.0	27169000.0	17500000.0	13397000.0
...	...	...	...	...	...
118	Total Contractual Obligations due in year 3	0.0	6354000.0	6125000.0	3296000.0
119	Total Contractual Obligations due in year 4	0.0	6125000.0	32492000.0	2756000.0
120	Total Contractual Obligations due in year 5	0.0	38572000.0	6411000.0	2363000.0
121	Total Contractual Obligations due Beyond	0.0	122884000.0	116477000.0	42082000.0
122	Total Contractual Obligations - Interests ...	0.0	-92000000.0	-86559000.0	-14162000.0

123 rows × 5 columns

Some of them, per company, had different years, so we had to keep the common years for each trio of statements before we brought them into the format we wanted. In the end it came down to this format:

SKYT_balance-sheet_Annual_As_Originally_Reported	Total Assets	Total Current Assets	Cash, Cash Equivalents and Short Term Investments	Cash and Cash Equivalents	Inventories	Raw Materials, Consumables and Supplies	Work-in-Process	Trade and Other Receivables, Current	Trade/Accounts Receivable, Current	Taxes Receivable, Current	...	Total Lease Liability - Beyond
2019	190435000.0	84517000.0	4605000.0	4605000.0	15994000.0	6342000.0	9652000.0	61968000.0	60991000.0	521000.0	...	0.0
2020	263209000.0	76572000.0	7436000.0	7436000.0	27169000.0	7450000.0	19719000.0	29995000.0	21357000.0	0.0	...	89350000.0
2021	263598000.0	74397000.0	12917000.0	12917000.0	17500000.0	10161000.0	7339000.0	40126000.0	23022000.0	745000.0	...	84116000.0
2022	305764000.0	116551000.0	30025000.0	30025000.0	13397000.0	13038000.0	359000.0	62839000.0	29683000.0	169000.0	...	10943000.0

4 rows × 123 columns

Then for each company we joined these three statements side by side, with the years as the base (like an inner join in SQL based on year - first column). Finally, we append all these grouped dataframes underneath each other.

But this created the most difficulties, as we discovered that not all dataframes have the same column-features. Therefore we had to merge only for the common columns and since we counted around 450 features, we ended up with 28. We considered them a small number and even some basic financial data were missing. Therefore, after we identified some of them, we threw two stocks (dataframes) out as they were the only ones that did not have one or two features that we wanted and by those stocks alone being removed, we were ultimately up around a dozen features. The final dataset for X ended up like this:

	Year	Total Assets	Total Current Assets	Cash, Cash Equivalents and Short Term Investments	Cash and Cash Equivalents	Total Non-Current Assets	Total Liabilities	Total Current Liabilities	Payables and Accrued Expenses, Current	Trade and Other Payables, Current	...	Total Operating Profit/Loss	Non-Operating Income/Expense, Total
0	2013	111057000.0	77936000.0	29976000.0	22006000.0	33121000.0	47980000.0	39057000.0	18908000.0	15314000.0	...	15000.0	-1421000.0
1	2014	183670000.0	113239000.0	40364000.0	32175000.0	70431000.0	68658000.0	48601000.0	36403000.0	31181000.0	...	5409000.0	-927000.0
2	2015	273475000.0	153928000.0	35960000.0	28074000.0	119547000.0	108077000.0	74080000.0	37537000.0	29040000.0	...	13666000.0	-2498000.0
3	2016	322318000.0	157552000.0	50268000.0	50224000.0	164766000.0	94934000.0	59973000.0	49977000.0	37956000.0	...	23184000.0	-2181000.0
4	2017	452984000.0	229697000.0	82972000.0	82936000.0	223287000.0	119708000.0	70708000.0	67673000.0	51997000.0	...	87162000.0	-2636000.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4589	2019	299424000.0	244920000.0	175848000.0	67940000.0	54504000.0	146431000.0	131853000.0	15286000.0	6230000.0	...	-75807000.0	-417000.0
4590	2020	402227000.0	266784000.0	171937000.0	54275000.0	135443000.0	237568000.0	165620000.0	19721000.0	6901000.0	...	-85665000.0	2712000.0
4591	2021	421507000.0	293740000.0	186594000.0	94110000.0	127767000.0	249594000.0	187997000.0	16799000.0	6626000.0	...	-73862000.0	2561000.0
4592	2022	441252000.0	328335000.0	215389000.0	113507000.0	112917000.0	270645000.0	219297000.0	21010000.0	9207000.0	...	-96176000.0	-1822000.0
4593	2023	668598000.0	518552000.0	386245000.0	203239000.0	150046000.0	571438000.0	311619000.0	26922000.0	6011000.0	...	-112455000.0	-74933000.0

4594 rows × 40 columns

Dataframes transformation Y

The previous procedure is for the training data (X). The data on which they will learn is the stock performance during the year 1/1/20xx - 31/12/20xx+1. Therefore, the prices that the stocks had at the beginning and at the end of the year (downloaded form yahoo finance: https://finance.yahoo.com) were extracted and the performance (labeled data - Y) was calculated as follows:

$$return = y\_endofyear - y\_startofyeary\_startofyear$$

In order to convert the algorithm to a classification problem, we re-labeled the return column into zeroes and ones, 0 for return less than 10% and 1 for more than 10%.

This data was taken from Yahoo finance and was in this format:

ROG.csv - Σημειωματάριο

Αρχείο Επεξεργασία Προβολή

```
Date,Open,High,Low,Close,Adj Close,Volume
1985-01-01,5.937500,7.125000,5.625000,7.000000,6.903252,285200
1985-02-01,7.000000,7.687500,6.937500,7.437500,7.334704,158000
1985-03-01,7.500000,7.562500,7.000000,7.093750,6.995705,210400
1985-04-01,7.093750,7.156250,5.906250,5.906250,5.824618,231600
1985-05-01,5.906250,7.187500,5.750000,7.093750,6.995705,551200
1985-06-01,7.093750,7.187500,6.218750,6.468750,6.379344,635600
1985-07-01,6.468750,6.562500,5.937500,5.937500,5.855436,798800
1985-08-01,5.937500,6.031250,5.125000,5.187500,5.115802,274400
1985-09-01,5.187500,5.250000,4.312500,4.968750,4.900075,907600
1985-10-01,4.937500,4.937500,4.437500,4.718750,4.653531,300000
1985-11-01,4.718750,5.437500,4.718750,4.968750,4.900075,492800
1985-12-01,4.968750,5.250000,4.625000,5.031250,4.961712,875200
1986-01-01,5.031250,5.687500,5.031250,5.031250,4.961712,451600
1986-02-01,5.031250,5.562500,5.031250,5.500000,5.423983,288800
1986-03-01,5.437500,6.125000,5.156250,5.750000,5.670527,591200
1986-04-01,5.750000,5.812500,4.937500,5.312500,5.239074,685600
1986-05-01,5.250000,5.687500,5.218750,5.437500,5.362347,315600
1986-06-01,5.437500,5.500000,5.281250,5.406250,5.331528,229600
1986-07-01,5.406250,5.468750,4.781250,4.843750,4.776803,270000
1986-08-01,4.843750,5.250000,4.593750,4.937500,4.869257,267200
1986-09-01,4.906250,5.437500,4.843750,4.843750,4.776803,366000
1986-10-01,5.000000,5.000000,4.406250,4.437500,4.376168,262000
1986-11-01,4.468750,4.562500,4.312500,4.531250,4.468623,452400
1986-12-01,4.468750,4.843750,4.468750,4.562500,4.499440,545600
1987-01-01,4.562500,5.312500,4.562500,5.156250,5.084984,733600
1987-02-01,5.156250,6.375000,5.125000,6.375000,6.286890,564000
1987-03-01,6.375000,6.562500,5.531250,6.218750,6.132799,496800
1987-04-01,6.187500,6.468750,5.875000,5.937500,5.855436,268000
1987-05-01,5.937500,6.156250,5.781250,6.000000,5.917072,251600
1987-06-01,6.031250,6.343750,5.812500,6.187500,6.101981,174800
1987-07-01,6.156250,6.375000,6.000000,6.218750,6.132799,288000
```

For each year we obtained 12 values (one for each first of the month) and we only needed the first and last of each year (1/1/yyyy and 31/12/yyyy). So we kept only the Date and Open columns and only the rows for the month of January (or if some were listed on the exchange e.g. November then we kept that date). We considered the price a stock has on 1/1/xxxx to be the first for that year and also the last for the previous year. Finally, we created one more column for the year and another that had the stock's performance using the above formula. The dataframe ended up like this:

Final dataset

The final dataset was essentially a merge of X's with Y's based on the year. Those rows that did not have a return were necessarily left out of the dataset. In addition to the existing features, new ones were created by combining the existing ones based on spearman correlation (details in the repo - path: **src/Dataset_Exploration/Feature_Engineering.ipynb**). We removed some features that were not very important financially and/or did not have a high correlation with the return. We also removed years and tickers (stock names) to avoid bias and ended up with the dataset we used.

	Open	Date	Year	return
0	9.840	2019-11-01	2019	0.006098
1	9.900	2020-01-01	2020	0.344141
2	13.307	2021-01-01	2021	-0.092207
3	12.080	2022-01-01	2022	-0.508278
4	5.940	2023-01-01	2023	0.341751
5	7.970	2024-01-01	2024	NaN

Target

The objective of the model is as follows: To make returns and stocks classified through financial data so that each data point that comes in can be categorized into the following two classes:

- **Below 10% return**
- **Above 10% return**

The reason 10% was chosen is that the stock market, as a whole over the long term and on average, returns ~10% so we are asking through our classification to give us the stocks with the highest probability of returning above 10% (i.e. beating the market over the long term).

We selected about 500 companies from the technology sector with a market capitalization over 50 million $ and downloaded X data (three statements) from morningstar.
For Y, respectively for the above ~500 companies, we downloaded them from yahoo finance. We used the opening stock price on 1/1/20xx as the price the stock ended up having in the previous year 12/31/20xx-1 to calculate the return. Our dataset does not pose any significant class imbalance.

Dataset Analysis

Given the broad nature of this dataset, we could perform several types of analysis, such as:

Descriptive Statistics: Provide a summary of the main statistical measures for each column (mean, median, standard deviation, etc.).

Correlation Analysis: Determine the relationships between different financial metrics (e.g., between total assets and total equity, cash flows from operating activities, and net income).

Segment Analysis: If the dataset includes identifiers that allow for segmentation (such as industry sectors or company names), we could analyze financial metrics by these segments.

DESCRIPTIVE ANALYSIS

The descriptive statistics for the dataset provide a comprehensive overview of various financial metrics across all entries. Here are some key observations:

Total Assets: The mean value is approximately 3.54 trillion, with a standard deviation indicating significant variation among companies.

Total Liabilities and Equity: Both show wide ranges, reflecting the diversity in size and financial structure of the companies included.

Financial Ratios: Ratios like the Price to Earnings Ratio, Price to Cash Flow Ratio, and others vary widely, suggesting a mix of industries and financial health among the companies.

Due to the extensive number of metrics and their technical nature, here are a few specific highlights:

Price to Earnings Ratio (P/E): The mean P/E ratio is around 1.37 with a standard deviation, indicating varied investor expectations and company earnings performance.

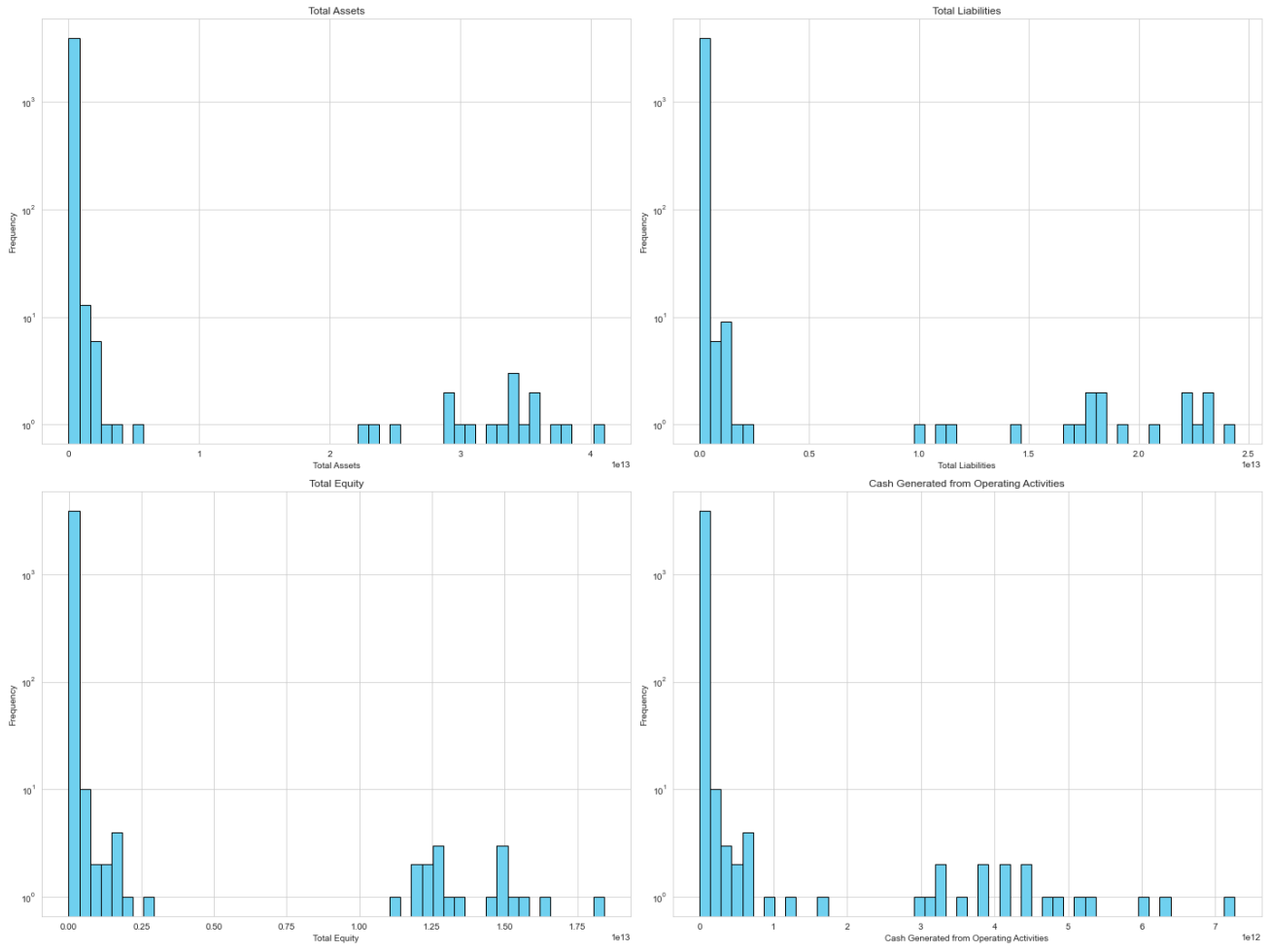
Total Revenue to Total Assets Ratio: This ratio has a mean value close to 0.92, suggesting, on average, companies are generating revenue close to the value of their assets.

Total Assets to Total Equity Ratio: With a mean of approximately 2.53, this indicates, on average, companies have more than double their equity in assets, which could suggest varying levels of leverage.

These statistics offer a snapshot into the financial health and operational performance of the companies in the dataset. For more detailed analysis or specific insights, further exploration into individual metrics or segments of data is required.

	count	mean	std	min	25%	50%	75%	max
Total Assets	3944.0	1.749797e+11	2.199246e+12	1.129660e+05	3.489132e+08	1.234200e+09	5.516662e+09	4.098981e+13
Total Current Assets	3944.0	5.937442e+10	7.169121e+11	8.987000e+04	1.957612e+08	6.149625e+08	2.322485e+09	1.318707e+13

Histograms can provide a clearer view of the distribution of values, including the spread and central tendency, without being as affected by extreme values as boxplots can be.



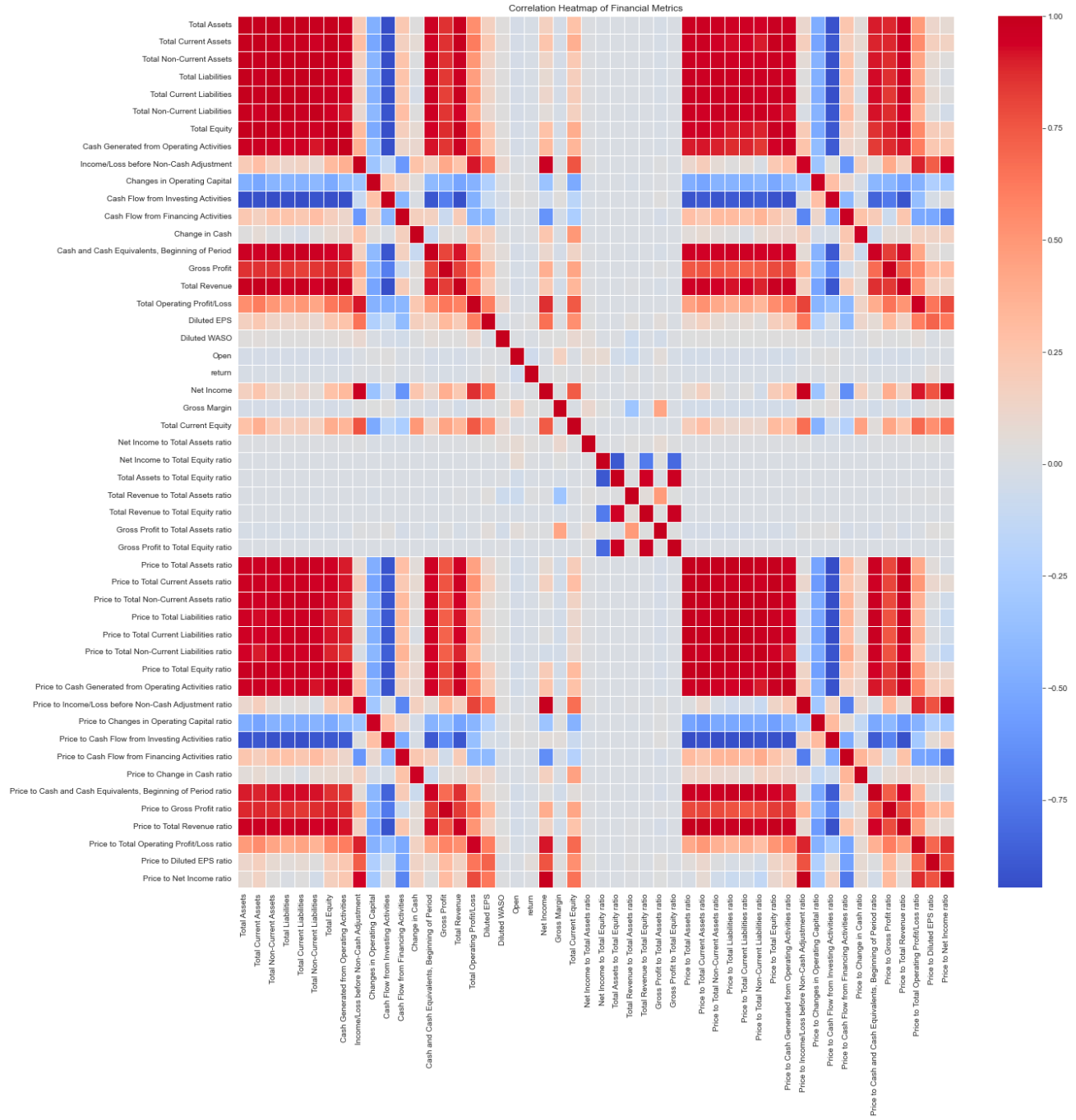
Each histogram shows the frequency distribution of values for these metrics on a logarithmic scale for the y-axis. This approach helps in visualizing the distribution's spread, central tendency, and the presence of any outliers or skewness in the data.

The histograms reveal how the data points are distributed across different ranges for each metric, providing insights into the financial structure and operations of the entities represented in the dataset.

CORRELATION ANALYSIS

To provide a different perspective on the dataset, we can create a correlation heatmap. A correlation heatmap visualizes the correlation coefficients between all pairs of numerical variables in the dataset, offering insights into how these financial metrics relate to one another. High positive or negative values indicate strong relationships, while values close to zero suggest little to no linear correlation.

We create a correlation heatmap for the dataset.



In this visualization:

Colors closer to red indicate a positive correlation, meaning that as one metric increases, the other tends to increase as well.

Colors closer to blue represent a negative correlation, indicating that as one metric increases, the other tends to decrease.

Colors near the middle of the color spectrum (lighter shades) suggest little to no correlation between the metrics.

This heatmap can be instrumental in identifying patterns and relationships that may not be immediately apparent. For instance, strong correlations (either positive or negative) between

certain metrics could point to underlying financial principles or industry-specific practices that influence these relationships.

Correlation analysis can help us understand the relationships between different financial metrics in the dataset. By examining the correlation coefficients, we can identify which pairs of variables are positively correlated, are negatively correlated, or show little to no linear relationship.

For this analysis, we'll focus on a selection of key financial metrics to keep the analysis manageable and insightful. We consider the following metrics for correlation analysis:

Total Assets

Total Liabilities

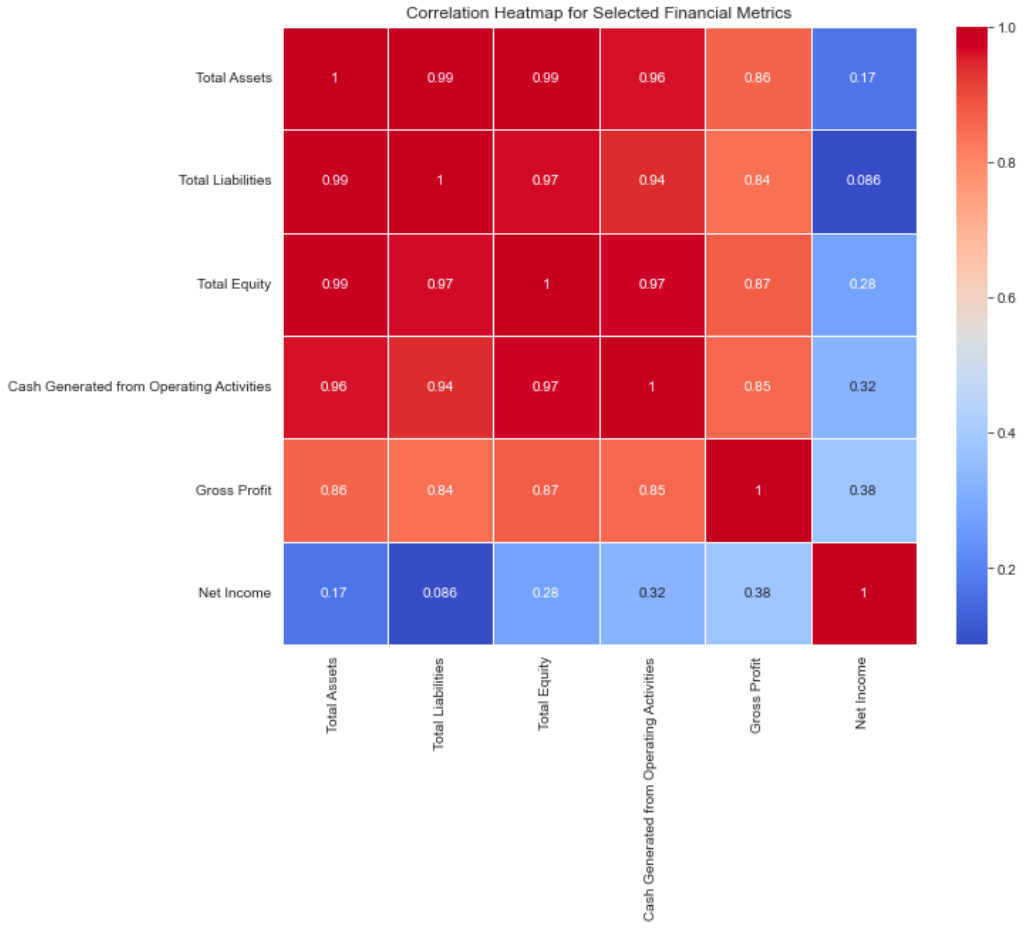
Total Equity

Cash Generated from Operating Activities

Gross Profit

Net Income

We'll calculate the correlation coefficients for these metrics and visualize them using a heatmap for easier interpretation. This will help us identify potential relationships that might be of interest for further analysis or decision-making.



The correlation heatmap shows how each pair of metrics is related. Here are some key takeaways:

Total Assets and Total Liabilities: There's a strong positive correlation, indicating that as the total assets of a company increase, its total liabilities tend to increase as well. This relationship is expected, as larger companies often have both higher assets and liabilities.

Total Assets and Total Equity: Also positively correlated, suggesting that companies with more assets generally have more equity. This is consistent with the accounting equation where Assets = Liabilities + Equity.

Cash Generated from Operating Activities and Net Income: This shows a positive correlation, indicating that companies with higher net income typically generate more cash from their operating activities, which is a good sign of operational efficiency.

Gross Profit and Net Income: There's a positive correlation between these two metrics, suggesting that higher gross profits often lead to higher net income. This relationship is intuitive, as gross profit is a major component of net income before subtracting expenses and taxes.

These correlations can provide valuable insights into financial management and operational effectiveness.

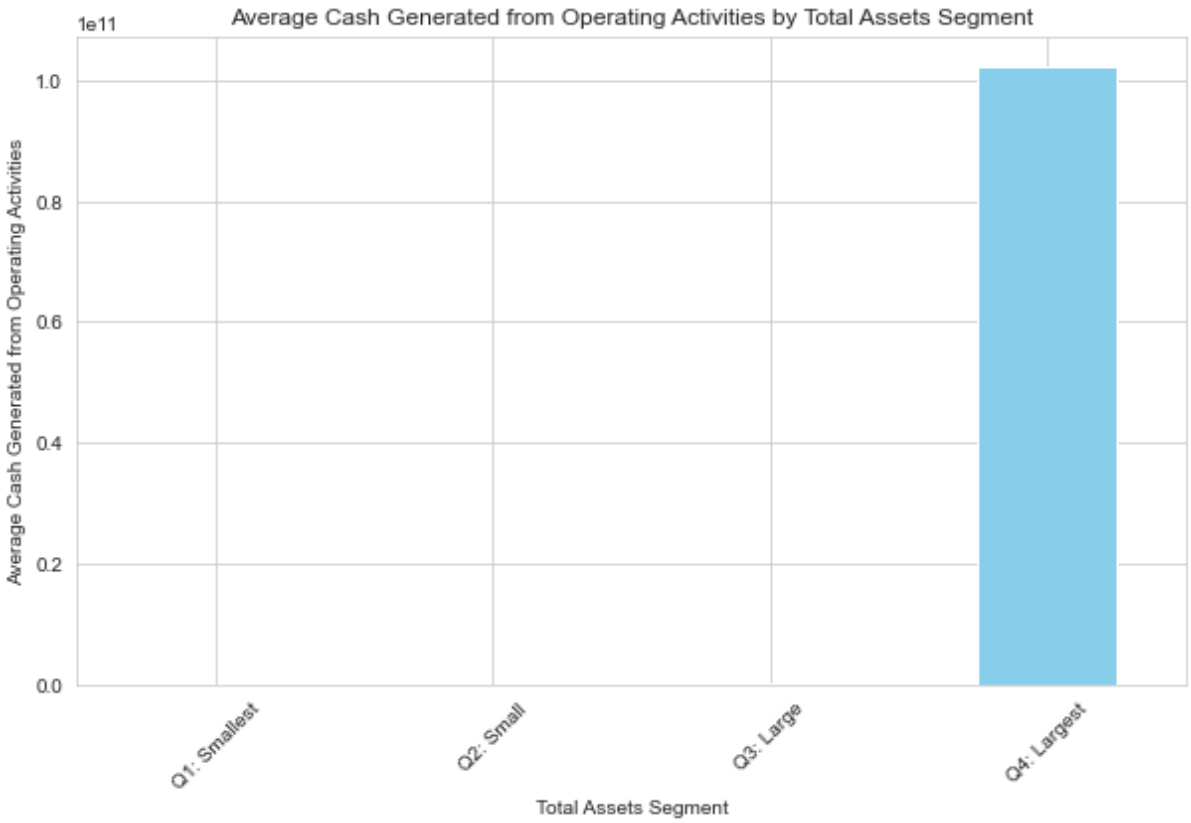
SEGMENT ANALYSIS

For a segment analysis, we typically need a categorical variable that divides the dataset into different groups or segments. This type of analysis allows us to compare financial metrics across different segments to identify trends, outliers, or significant differences that could inform strategic decisions or further analysis.

One approach will be to create segments based on a quantitative metric, categorizing companies into groups based on their size or performance.

We proceed with a simple segmentation based on a financial metric. For example, we will segment companies into groups based on their Total Assets, Net Income and then analyze differences in Cash Generated from Operating Activities, across these segments.

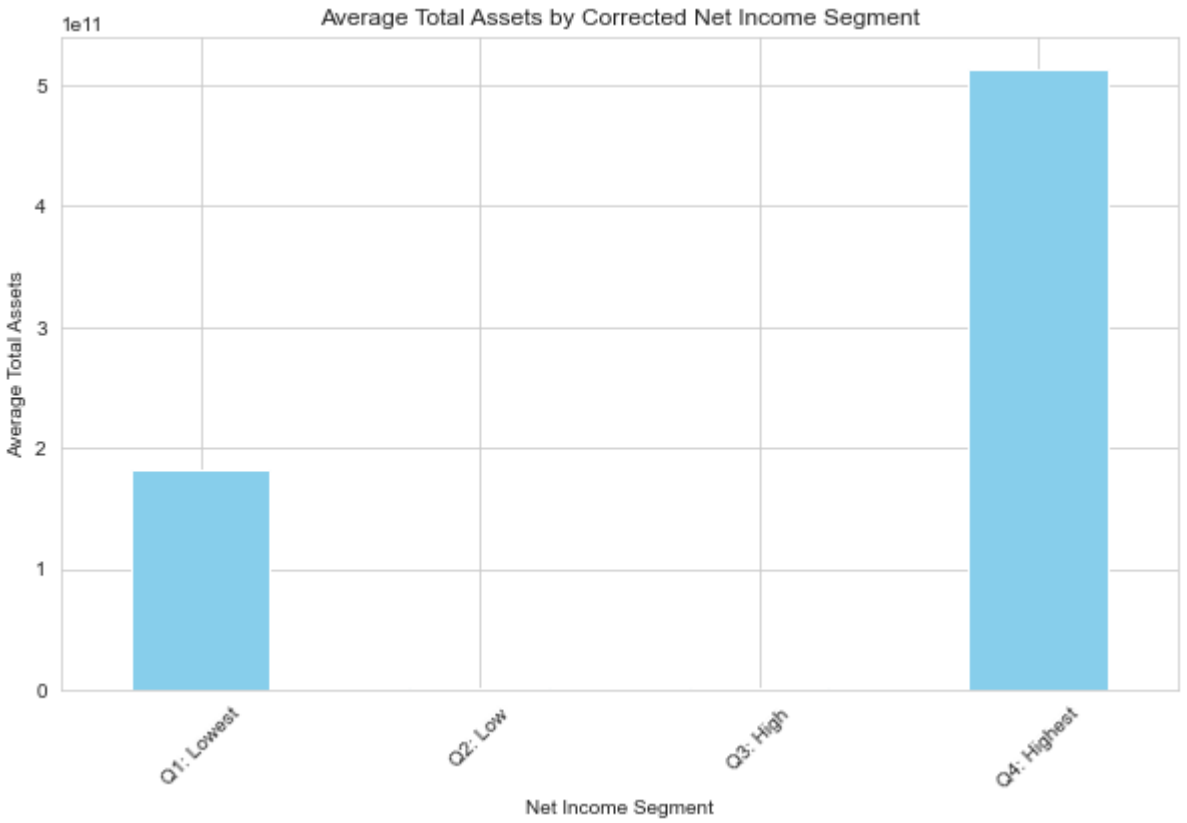
Total Assets



The bar chart visualizes the average Cash Generated from Operating Activities for companies segmented by their Total Assets into quartiles: Q1 (Smallest), Q2 (Small), Q3 (Large), and Q4 (Largest). The analysis reveals significant differences across segments:
Q1: Smallest companies have an average of approximately 1.33 million in Cash Generated from Operating Activities.
Q2: Small companies show a substantial increase, with an average of around 53.21 million.
Q3: Large companies further increase to an average of approximately 257.65 million.
Q4: Largest companies have a dramatically higher average of about 102.16 billion.
This analysis suggests a strong positive relationship between the size of a company (as measured by Total Assets) and its ability to generate cash from operating activities. The largest companies (Q4) generate substantially more operating cash flow than smaller companies, highlighting the impact of scale on operational efficiency and financial performance.
Such insights indicate that larger companies might have better cash flow stability and operational efficiency, which are important factors for investment and strategic decisions.

Net Income

We adjust the binning strategy to ensure it accommodates all values in the Net Income column appropriately, starting from the minimum value in that column rather than a fixed -1.



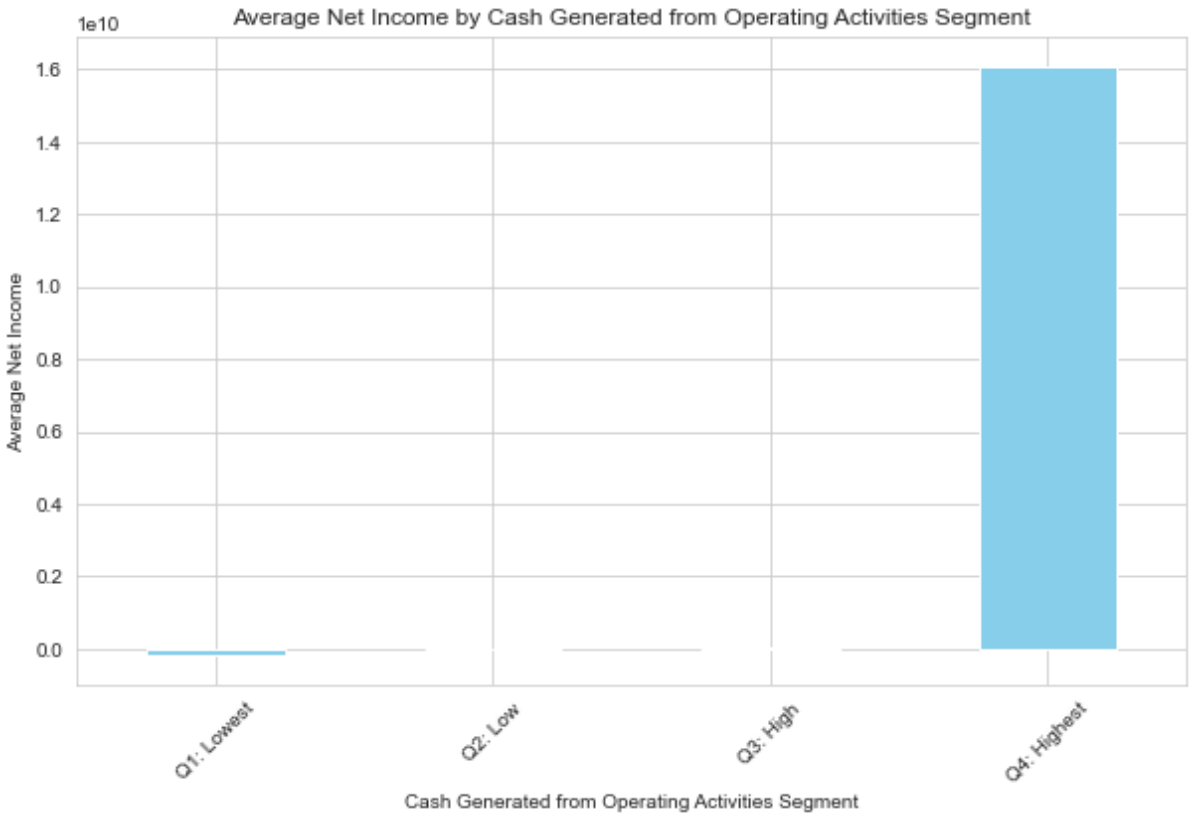
The corrected bar chart visualizes the average Total Assets for companies segmented by their Net Income into quartiles: Q1 (Lowest), Q2 (Low), Q3 (High), and Q4 (Highest). This segmentation reveals interesting insights:

Q1: Lowest segment shows an exceptionally high average Total Assets value, approximately 182.14 billion. This could indicate that some companies with large assets might have experienced losses or low net income during the period, potentially due to significant investments, depreciation, or other factors affecting net income negatively.

Q2: Low and Q3: High segments have average Total Assets of approximately 1.62 billion and 2.20 billion, respectively. These figures suggest a moderate level of assets relative to their net income categories.

Q4: Highest segment demonstrates a significantly higher average of Total Assets, about 513.95 billion, indicating that companies with the highest net income also tend to have large asset bases. This could reflect successful operations, efficient asset utilization, or a combination of high-revenue-generating activities and effective cost management.

Cash Generated from Operating Activities



The bar chart illustrates the average Net Income for companies segmented by their Cash Generated from Operating Activities into quartiles: Q1 (Lowest), Q2 (Low), Q3 (High), and Q4 (Highest). The analysis uncovers distinct patterns across these segments:
Q1: Lowest: Companies in this segment have an average Net Income of approximately -189.26 million, indicating that companies generating the least cash from operating activities are, on average, experiencing losses.
Q2: Low: This segment shows a smaller average loss in Net Income, around -11.58 million, suggesting a slight improvement in profitability compared to the lowest quartile.
Q3: High: Companies in this segment have an average Net Income of approximately 74.09 million, indicating that higher cash generation from operating activities correlates with positive net income.
Q4: Highest: The highest quartile demonstrates a significantly larger average Net Income of about 16.08 billion, highlighting a strong positive relationship between the ability to generate cash from operating activities and profitability.
This segmentation analysis underscores the crucial role of cash generation from operating activities in a company's financial health. Notably, companies in the highest segment for cash generation are also those with the highest net income, reinforcing the importance of efficient operations and cash flow management for profitability.

FURTHER ANALYSIS

Based on what we've already explored, here are a few additional analyses that could yield valuable information:
1. Profitability Ratios Analysis: Delve into ratios like Return on Assets (ROA), Return on Equity (ROE), and Profit Margin to evaluate how efficiently companies are generating profit relative to their assets, equity, and revenues.

2. Liquidity Ratios Analysis: Examine metrics such as the Current Ratio and Quick Ratio to assess the companies' short-term financial health and their ability to cover short-term obligations.
3. Leverage Ratios Analysis: Analyze ratios like Debt to Equity and Interest Coverage to understand the companies' debt levels and their ability to meet financial obligations, which is crucial for assessing financial risk.
4. Operational Efficiency Analysis: Look into metrics like Inventory Turnover and Receivables Turnover to evaluate how efficiently companies manage their assets to generate sales.
5. Impact of Operational Activities on Financial Performance: Further explore the relationship between cash flows from operating activities and other financial metrics to understand how operational efficiencies translate into overall financial performance.
6. Outlier Detection and Analysis: Identify companies that are outliers in terms of financial performance or ratios. Understanding why certain companies are outliers could uncover insights into exceptional strategies, market conditions, or operational practices.
7. Benchmarking Analysis: Compare key metrics against industry averages or standards to identify areas of strength and opportunities for improvement.

1.PROFITABILITY RATIOS

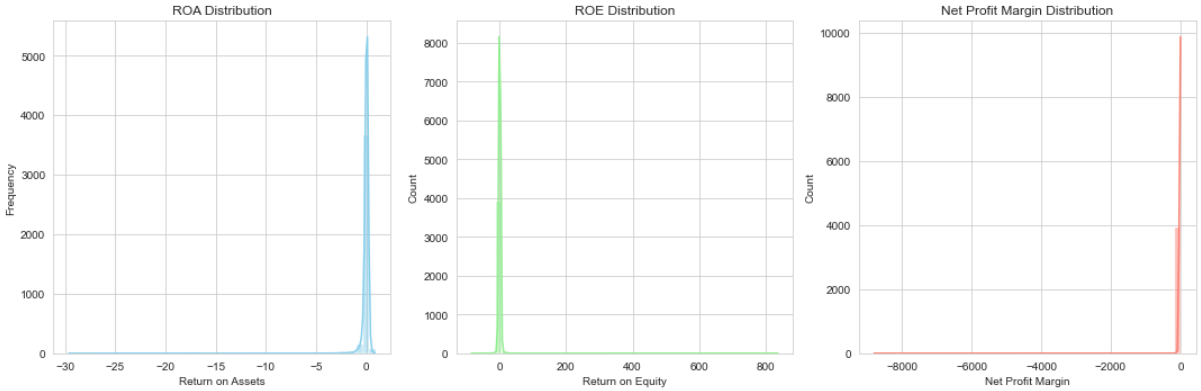
Profitability ratios are key indicators of a company's financial performance, focusing on its ability to generate earnings relative to its revenue, assets, equity, or other financial metrics. Common profitability ratios include:

Return on Assets (ROA): Measures how efficiently a company uses its assets to generate profit.
ROA = Total Assets / Net Income

Return on Equity (ROE): Indicates how effectively a company uses shareholders' equity to generate profits.
ROE = Total Equity / Net Income

Net Profit Margin: Shows the percentage of revenue that remains as profit after all expenses are paid.
Net Profit Margin = Total Revenue / Net Income

We calculate these profitability ratios for the dataset and then analyze the results.



ROA	ROE	Net Profit Margin	
count	3944.000000	3944.000000	3944.000000
mean	-0.042087	0.258469	NaN
std	0.609774	14.124193	NaN
min	-29.633810	-85.408469	-inf

ROA	ROE	Net Profit Margin	
25%	-0.060681	-0.106870	-0.096148
50%	0.022189	0.048224	0.027821
75%	0.072659	0.149543	0.109373
max	0.858180	837.618358	inf

The analysis of profitability ratios—Return on Assets (ROA), Return on Equity (ROE), and Net Profit Margin—provides insights into the financial performance of companies within the dataset:

ROA Distribution: The Return on Assets varies significantly among companies, indicating differences in how efficiently assets are used to generate profit. The distribution suggests a mix of performance, with the majority of companies having ROA values close to 0, indicating that most companies generate modest returns on their assets.

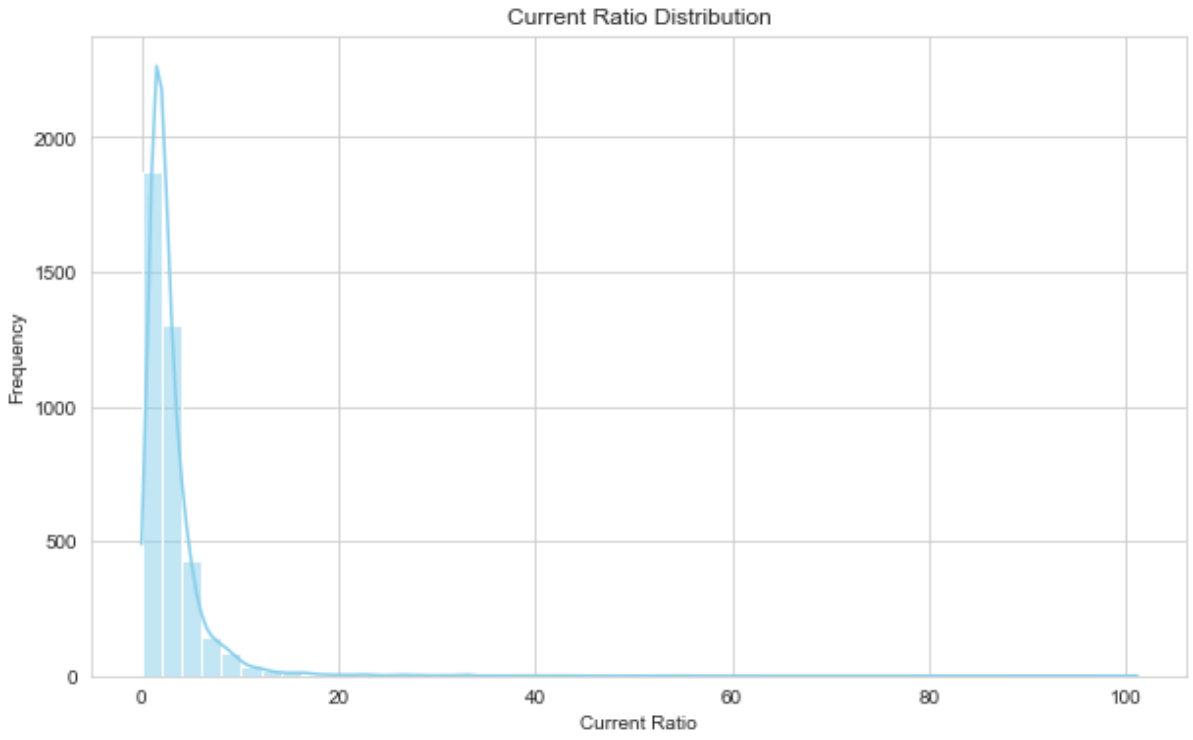
ROE Distribution: The Return on Equity shows a wide range, including extreme positive values, which signifies that equity efficiency in generating profits varies widely among companies. Some companies demonstrate very high ROE, potentially indicating high profitability relative to shareholders' equity or low equity levels.

Net Profit Margin Distribution: The Net Profit Margin also varies, with some companies achieving high margins. However, the presence of negative and infinite values (as indicated by -inf and inf in the summary statistics) suggests that some companies incur losses or have very low revenue bases, affecting their profit margins.
The summary statistics reveal mean values for ROA and ROE, but the Net Profit Margin mean is not displayed due to the presence of infinite values, which occur when companies have zero or near-zero revenue, leading to undefined profit margins.
This analysis highlights the diversity in financial health and efficiency among the companies. Profitability ratios are crucial for assessing a company's ability to generate earnings.

2. LIQUIDITY RATIOS ANALYSIS

Liquidity ratios are key financial metrics that assess a company's ability to meet its short-term obligations. They are crucial for evaluating the financial health of a company, especially its solvency and risk of default in the near term. The most common liquidity ratios include:
Current Ratio: Measures a company's ability to pay off its short-term liabilities with its short-term assets.
Current Ratio = Total Current Assets / Total Current Liabilities

Quick Ratio (Acid-Test Ratio): Similar to the current ratio but excludes inventory from assets, providing a stricter sense of the company's short-term liquidity.
Quick Ratio = (Total Current Assets − Inventory) / Total Current Liabilities
Since the dataset does not explicitly include an Inventory column, we'll focus on the Current Ratio for our liquidity analysis. We calculate the Current Ratio for the companies in the dataset and analyze the results.



	count
3944.000000	
mean	3.048305
std	3.619227
min	0.021640
25%	1.415706
50%	2.123764
75%	3.469413
max	101.129076

The Current Ratio distribution and its summary statistics provide insights into the short-term liquidity of companies within the dataset:

Mean: The average Current Ratio across all companies is approximately 3.05, suggesting that, on average, companies have about three times more current assets than current liabilities. This indicates a generally healthy liquidity position across the dataset.

Standard Deviation: With a standard deviation of approximately 3.62, there's considerable variability in the Current Ratio among companies, pointing to differences in liquidity management and operational contexts.

Minimum and Maximum: The Current Ratio ranges from a low of about 0.02 to a high of over 101.13, highlighting extreme variations in liquidity positions. Some companies may be operating with very tight liquidity (close to the minimum), while others have a significant buffer of current assets over liabilities.

Quartiles: The 25th percentile (Q1) is approximately 1.42, the median (Q2) is around 2.12, and the 75th percentile (Q3) is about 3.47. This indicates that 50% of the companies have a Current Ratio between 1.42 and 3.47, considered a healthy range for managing short-term obligations. The histogram shows the distribution of the Current Ratio, with a right-skewed pattern indicating that while most companies maintain a Current Ratio that suggests adequate liquidity, a significant

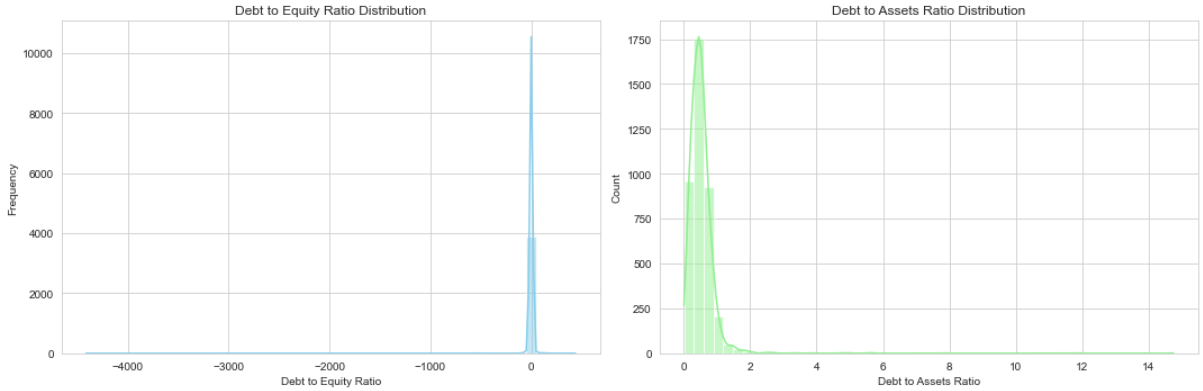
number of companies have higher ratios, which could imply excessive liquidity or inefficient use of assets.
This analysis underscores the importance of liquidity management in ensuring a company can meet its short-term obligations.

3. LEVERAGES RATIOS ANALYSIS

Leverage ratios are important financial metrics that assess a company's debt levels relative to its equity or assets, offering insights into its financial structure and risk profile. Common leverage ratios include:

Debt to Equity Ratio: Measures the company's financial leverage by comparing its total liabilities to its shareholders' equity.
Debt to Equity Ratio = Total Liabilities / Total Equity

Debt to Assets Ratio: Indicates the proportion of a company's assets that are financed by debt.
Debt to Assets Ratio = Total Liabilities / Total Assets
These ratios help us understand the extent to which a company is using debt to fund its operations and growth, and the potential risk this debt poses.
We calculate both the Debt to Equity Ratio and the Debt to Assets Ratio for the companies in the dataset and analyze the results.



	Debt to Equity Ratio	Debt to Assets Ratio
count	3944.000000	3944.000000
mean	-0.039255	0.526163
std	73.085890	0.465057
min	-4413.210031	0.009994
25%	0.371994	0.309454
50%	0.796862	0.475045
75%	1.534838	0.648899
max	439.236220	14.739935

The leverage ratios—Debt to Equity Ratio and Debt to Assets Ratio—provide insights into the companies' use of debt in their financial structures:

Debt to Equity Ratio:
The distribution shows a wide range, with the mean near -0.04, affected by extreme values. The negative and extremely high values suggest some companies have negative equity (more liabilities than assets) or very low equity relative to their debt, which is unusual and might require further investigation.
The standard deviation is quite high, indicating significant variability among companies in how much debt they use relative to their equity.

Debt to Assets Ratio:
The mean value of approximately 0.53 suggests that, on average, companies finance over half of their assets through debt, indicating a moderate level of leverage.
The distribution is more centralized than the Debt to Equity Ratio, with a standard deviation suggesting variability but within a more typical range for corporate finance.
The histograms illustrate the distributions of these ratios, showing how companies within this dataset vary in their reliance on debt financing. While most companies have Debt to Equity and Debt to Assets Ratios within a reasonable range, the presence of outliers with extremely high or negative values indicates diverse financial strategies and situations.
These leverage ratios are crucial for assessing the financial risk and capital structure of companies. High leverage ratios can indicate a high risk of financial distress, especially in adverse market conditions, but they can also signal aggressive growth strategies financed through debt.

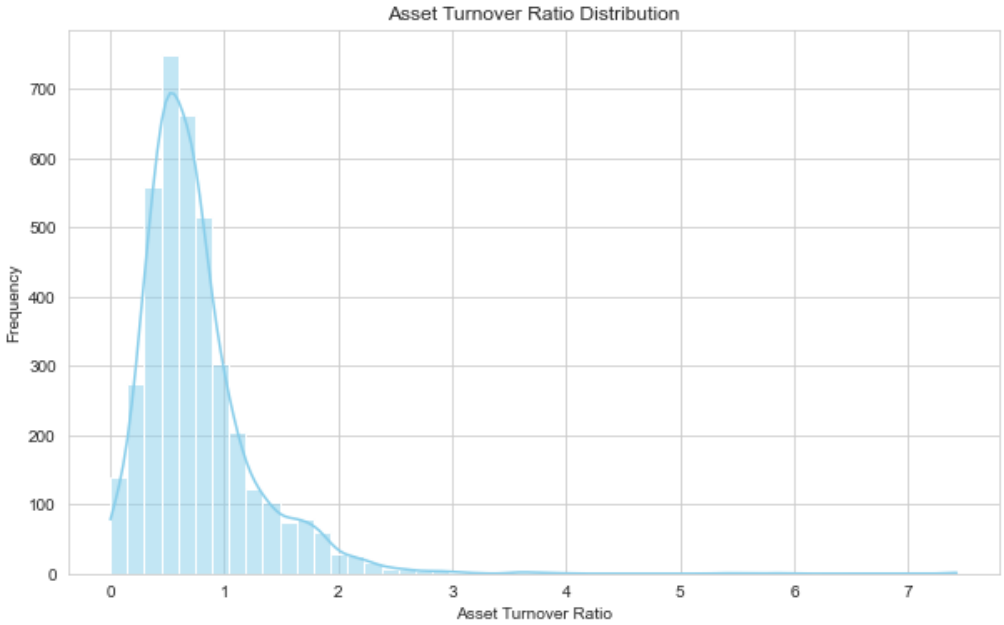
4. OPERATIONAL EFFICIENCY ANALYSIS

Operational efficiency analysis involves examining how well a company uses its resources to generate income and manage its operations. Key metrics in this analysis often include:

Inventory Turnover Ratio: Measures how many times a company's inventory is sold and replaced over a period. It's a key indicator of inventory management efficiency and sales performance.

Asset Turnover Ratio: Indicates how efficiently a company uses its assets to generate sales.
Asset Turnover Ratio = Total Revenue / Total Assets

Receivables Turnover Ratio: Measures how efficiently a company collects its receivables or the efficiency of its credit policy.
Given the data available, we will focus on the Asset Turnover Ratio, which we can calculate with the provided information. This ratio will give us insights into how effectively the companies are using their assets to generate revenue.



count	3944.000000
mean	0.751285
std	0.499529
min	0.000000
25%	0.448425
50%	0.650677

75% 0.918541
max 7.427476

The Asset Turnover Ratio analysis provides insights into how effectively companies within the dataset are using their assets to generate revenue:

Mean: The average Asset Turnover Ratio is approximately 0.75, indicating that, on average, companies generate $0.75 in revenue for every dollar of assets they own. This suggests a moderate level of operational efficiency across the dataset.

Standard Deviation: With a standard deviation of about 0.50, there's considerable variability in the Asset Turnover Ratio among companies, indicating diverse efficiency levels in using assets to generate sales.

Minimum and Maximum: The ratio ranges from 0 (for companies with no revenue or extremely high assets relative to their revenue) to about 7.43, showing a significant disparity in operational efficiency. A maximum value of 7.43 is exceptionally high, suggesting that some companies are extremely efficient in generating revenue with relatively few assets.

Quartiles:
The 25th percentile (Q1) is approximately 0.45, indicating that 25% of companies generate $0.45 or less in revenue for every dollar of assets.
The median (Q2) is about 0.65, showing that half of the companies generate $0.65 or less in revenue for every dollar of assets.
The 75th percentile (Q3) is around 0.92, suggesting that 75% of companies generate $0.92 or less in revenue for every dollar of assets.
The histogram illustrates the distribution of the Asset Turnover Ratio, which is somewhat right-skewed, indicating that while most companies operate with moderate efficiency in asset utilization, a few outliers achieve exceptionally high turnover ratios.
This analysis highlights the importance of asset management in operational efficiency. Companies with higher asset turnover ratios are likely utilizing their assets more effectively to generate sales, indicating streamlined operations and potentially higher profitability. In contrast, lower ratios may suggest underutilization of assets or inefficiencies in converting assets into revenue.

5. IMPACT OF OPERATIONAL ACTIVITIES ON FINANCIAL PERFORMANCE

Analyzing the impact of operational activities on financial performance involves examining how cash flows from operating activities correlate with other key financial metrics. This can reveal insights into the efficiency of a company's core business operations and their effect on overall financial health. Key metrics to consider in this analysis include:

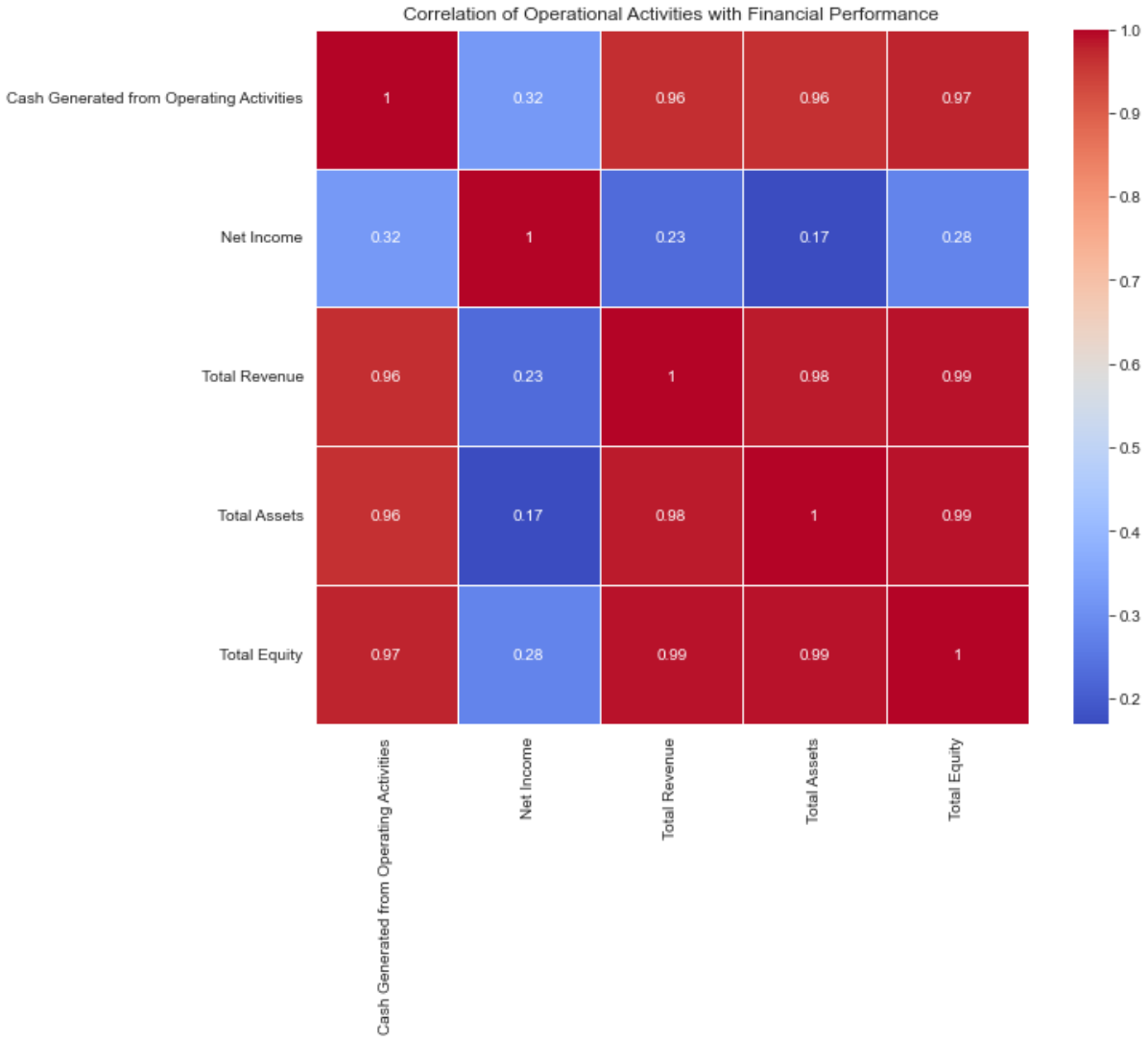
Cash Generated from Operating Activities: A direct measure of the cash inflows and outflows from a company's core business operations.

Net Income: The bottom line of a company's income statement, representing the total profit after all expenses have been deducted from revenues.

Total Revenue: The total income generated from the company's activities before any expenses are subtracted.

Total Assets and Total Equity: Indicators of the company's financial scale and the value provided to shareholders.
By examining the relationship between cash generated from operating activities and these metrics, we can assess the effectiveness of a company's operational management in contributing to its financial performance.

We perform a correlation analysis to explore how cash generated from operating activities is related to Net Income, Total Revenue, Total Assets, and Total Equity. This will help us understand the impact of operational activities on financial performance.



The correlation heatmap illustrates the relationships between cash generated from operating activities and several key financial performance metrics:

Cash Generated from Operating Activities and Net Income: There's a positive correlation, indicating that companies with higher cash flows from operating activities tend to have higher net income. This suggests that efficient operational management positively impacts profitability.

Cash Generated from Operating Activities and Total Revenue: The positive correlation here shows that higher operational cash flows are associated with higher total revenues, reinforcing the idea that effective operations are crucial for generating sales and income.

Cash Generated from Operating Activities and Total Assets: The correlation is positive, albeit likely weaker than with net income and revenue. This relationship indicates that larger asset bases can support or require higher operational cash flows, reflecting the scale of operations.

Cash Generated from Operating Activities and Total Equity: Similar to total assets, the positive correlation with total equity suggests that companies with higher equity, which can indicate more

resources available for operations or investment, tend to generate more cash from their operating activities.

These correlations highlight the fundamental role of operational activities in driving financial performance. Efficient operational management not only contributes directly to profitability (as shown by the relationship with net income) but also supports broader financial growth and stability, as evidenced by its impact on revenue, assets, and equity.
This analysis underscores the importance of optimizing operational activities to enhance overall financial health, suggesting that companies focusing on improving their operational efficiency can see significant benefits in their financial performance.

6.OUTLIER DETECTION & ANALYSIS

Outlier detection can reveal companies with financial metrics significantly different from the majority, potentially indicating unique operational efficiencies, strategic decisions, or financial risks. For this analysis, we'll focus on identifying outliers within key financial metrics that we've discussed: Total Assets, Net Income, and Cash Generated from Operating Activities. These metrics can provide insights into the financial health, profitability, and operational efficiency of companies.
We'll use the Interquartile Range (IQR) method for outlier detection, which identifies outliers as observations significantly below the first quartile (Q1) or above the third quartile (Q3) by a certain factor of the IQR. The IQR is the difference between the 25th and 75th percentiles of the data.
We proceed with identifying outliers in these metrics and briefly analyze any significant outliers detected.
{'Total Assets': 586,
 'Net Income': 814,
 'Cash Generated from Operating Activities': 654}
The outlier detection analysis revealed the following number of outliers across key financial metrics:

Total Assets: 586 outliers were identified, suggesting that these companies have asset levels significantly higher or lower than the majority. High outliers might represent very large companies or those with extensive capital investments, whereas low outliers could be smaller companies or those with minimal assets.

Net Income: 814 outliers were detected, indicating companies with exceptionally high or low profitability. High net income outliers may represent highly profitable companies, possibly due to successful operations, unique competitive advantages, or favorable market conditions. Low outliers could indicate companies experiencing losses, possibly due to operational challenges, high costs, or adverse market conditions.

Cash Generated from Operating Activities: 654 outliers were found, highlighting companies with significantly high or low cash flows from operations. High outliers might indicate companies with efficient operations and strong cash generation capabilities, while low outliers could suggest companies facing operational inefficiencies, low sales, or high operating costs.

The scatter plots above illustrate the distribution of data points and outliers for three key financial metrics: Total Assets, Net Income, and Cash Generated from Operating Activities. In each plot:
The **blue** points represent regular data points within the expected range.
The **red** points highlight the outliers, which are significantly above or below the majority of data points.

Observations:

Total Assets: The outliers are scattered across a wide range of values, indicating companies with exceptionally high or low total assets compared to the rest. High outliers might represent large corporations with substantial assets, while low outliers could be smaller companies.

Net Income: Similar to Total Assets, the outliers in Net Income show a mix of extremely profitable companies and those incurring significant losses. These outliers may represent companies with exceptional events or operational efficiencies impacting their profitability.

Cash Generated from Operating Activities: Outliers here represent companies with unusually high or low cash flows from their core business operations. High outliers could indicate strong operational efficiency or sectors with high cash generation capabilities, while low outliers might reflect operational challenges or industries with tighter cash flows.
These visualizations provide a clear picture of how certain companies stand out from the rest in terms of their financial metrics, underscoring the importance of further investigation into the reasons behind these outlier values. Understanding the context behind these outliers can offer valuable insights into operational effectiveness, industry dynamics, and financial strategies.

7.BENCHMARKING ANALYSIS

Benchmarking analysis involves comparing a company's performance metrics against industry standards or competitors to identify areas of strength and opportunities for improvement. For a meaningful benchmarking analysis, we typically need:

Benchmark Metrics: Identifying key metrics for comparison, such as profitability ratios, operational efficiency ratios, liquidity ratios, and leverage ratios, among others.
Given the limitations of our dataset, specifically the lack of explicit industry or sector information, a direct benchmarking analysis against industry standards may not be feasible. However, we can still perform a form of internal benchmarking by comparing companies within our dataset against the overall averages or quartile distributions for selected metrics. This can help identify top performers or outliers in various financial metrics.

For this analysis, We focus on comparing companies' ROA (Return on Assets) and Current Ratio against the dataset's overall averages. These metrics were chosen because they provide insights into operational efficiency and liquidity, respectively, which are important across all industries. We'll calculate the overall average for these metrics and then identify companies that significantly exceed or fall below these averages.
{'Average ROA': -0.042087070635861486,
'Average Current Ratio': 3.0483052264297417,
'Top Performers ROA (>75th percentile)': 986,
'Low Performers ROA (<25th percentile)': 986,
'Top Performers Current Ratio (>75th percentile)': 986,
'Low Performers Current Ratio (<25th percentile)': 986}
The benchmarking analysis against the dataset's averages for ROA (Return on Assets) and Current Ratio yields the following insights:

Average ROA: The overall average ROA is approximately -0.042, suggesting that on average, companies in the dataset are experiencing a slight loss relative to their assets. This average could be skewed by outliers or specific sectors within the dataset.

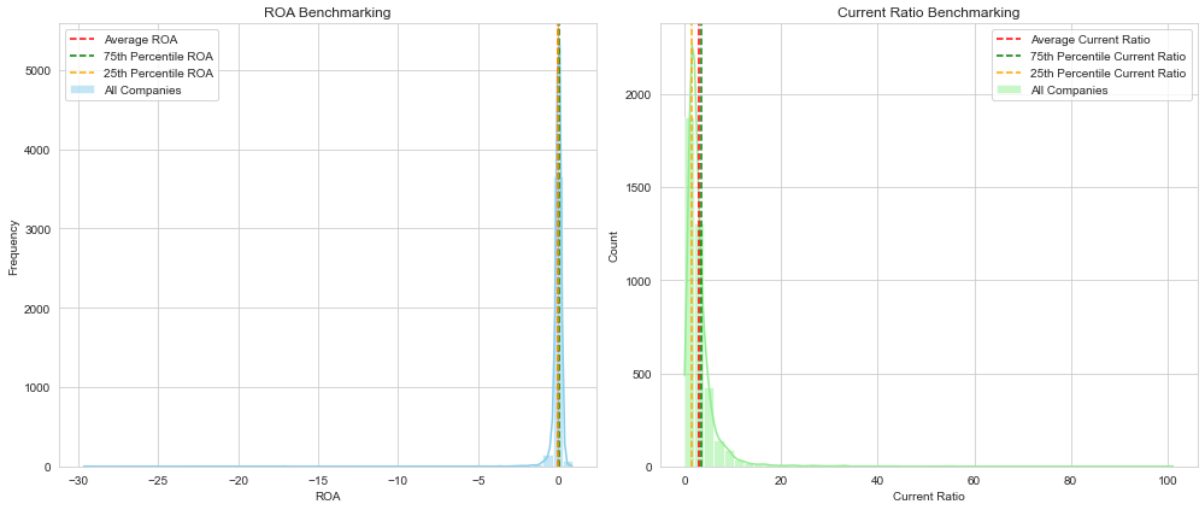
Average Current Ratio: The average Current Ratio is about 3.05, indicating that companies have, on average, three times more current assets than current liabilities. This suggests a generally healthy liquidity position across the dataset.

Top Performers in ROA (above the 75th percentile): 986 companies are identified as top performers based on their ROA, significantly exceeding the dataset's average and showcasing superior operational efficiency in using assets to generate profit.

Low Performers in ROA (below the 25th percentile): Another 986 companies are categorized as low performers in terms of ROA, falling below the 25th percentile. These companies might be facing operational challenges or inefficiencies in asset utilization.

Top Performers in Current Ratio (above the 75th percentile): Similarly, 986 companies exceed the 75th percentile for the Current Ratio, indicating strong liquidity and the ability to cover short-term obligations.

Low Performers in Current Ratio (below the 25th percentile): 986 companies fall below the 25th percentile in Current Ratio, potentially indicating liquidity challenges or higher reliance on short-term liabilities.
This form of internal benchmarking highlights companies that stand out in terms of operational efficiency and liquidity management within the dataset. Top performers in these metrics may exemplify best practices or competitive advantages in their operations and financial management, while low performers may need to investigate underlying issues or consider strategic changes to improve their standings.



The graphs illustrate the benchmarking analysis for ROA (Return on Assets) and Current Ratio, highlighting how companies compare to average and percentile thresholds:

ROA Benchmarking:

The distribution of ROA values among all companies is shown, with the overall average ROA marked by a red dashed line. The 75th and 25th percentiles are indicated by green and orange dashed lines, respectively, delineating top and low performers in operational efficiency.
This visualization helps identify where the bulk of companies lie relative to these benchmarks, with top performers exceeding the green line and low performers falling below the orange line.

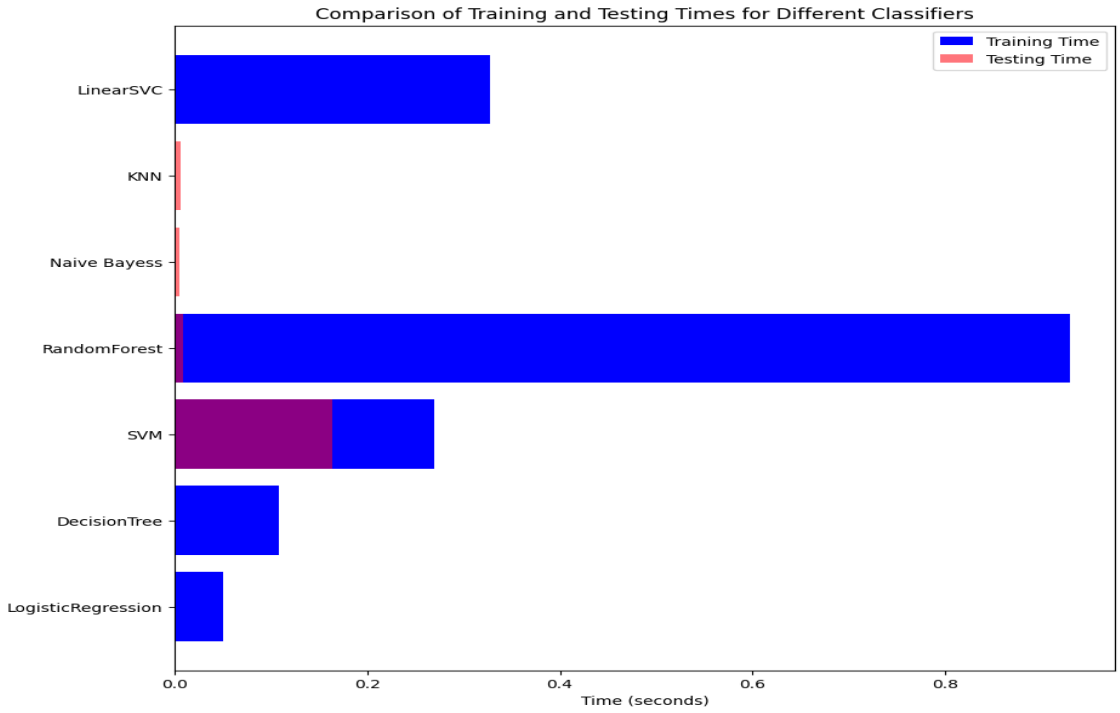
Current Ratio Benchmarking:
Similarly, the distribution of Current Ratio values is displayed. The average Current Ratio is also marked by a red dashed line, with the 75th and 25th percentile thresholds shown in green and orange, respectively.
The graph provides a clear view of liquidity management across the dataset, highlighting companies with strong liquidity positions (above the green line) and those that might face liquidity challenges (below the orange line).

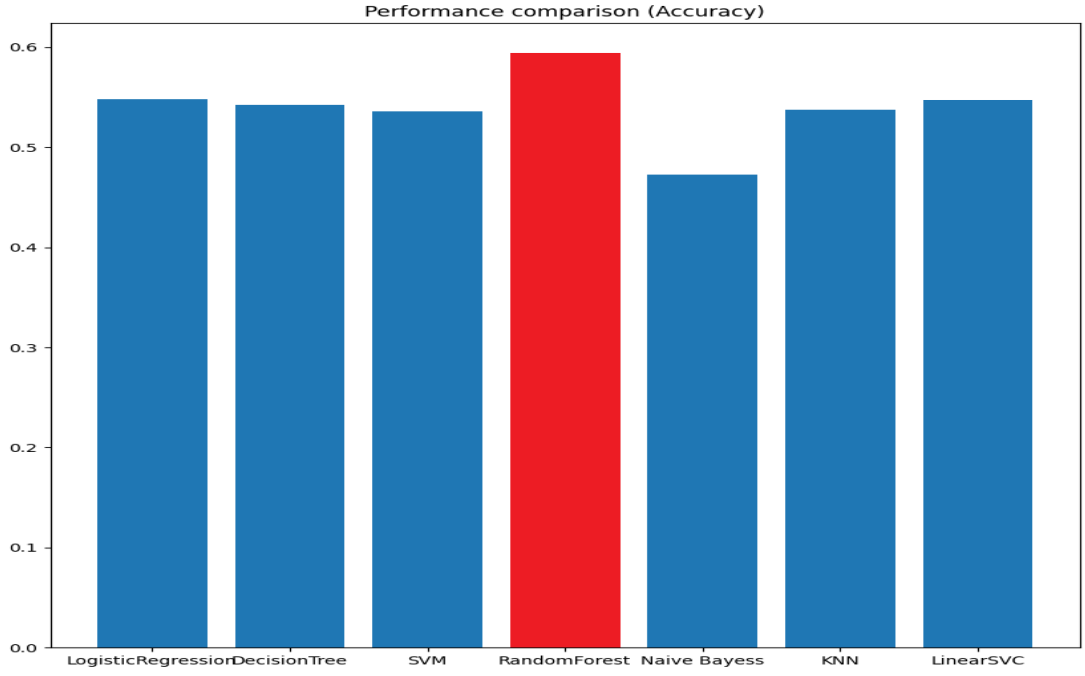
Feature and Method Exploration

We experimented with several ML methods and tested how features impact the overall result of the classification result. Our features were sequential and numerical. These types of features are handled well by SVM and RandomForest Methods which performed the best and ultimately were selected for the classification task. The data are also complex numerical relations, that are well explained by SVM's and more complex models.

We performed several benchmarks for different ML methods with the benchmark being classification accuracy and training/testing time. We experimented with different subsets of features and scaling (using *StandardScaler*) of the features to observe the impact they have on the models performance. As we will also explain later, the data are quite non-separable and even ensemble or more complex methods struggle to separate them. As a result, feature selection and scaling doesn't have a very

significant impact but we adopted them anyways as they reduced model complexity and slightly boosted the generalization capability of the models.





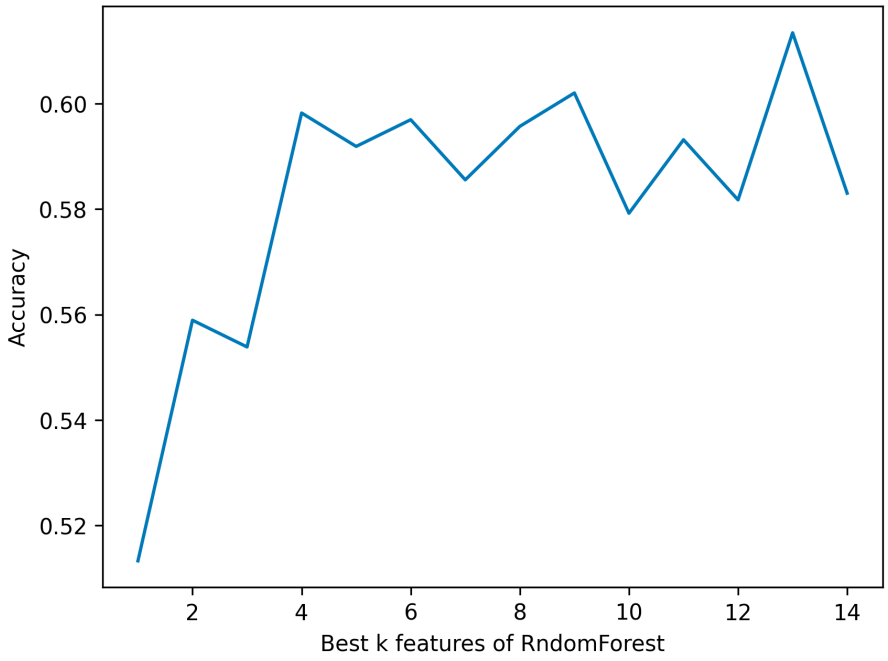
We selected to proceed with the Random Forest and SVM methods as they performed the best (we couldn't allow for any performance drop as the problem is already hard), even though they had large training times.

The rest of the plots can be found on the *Exploration.ipynb* notenook.

Fature Selection

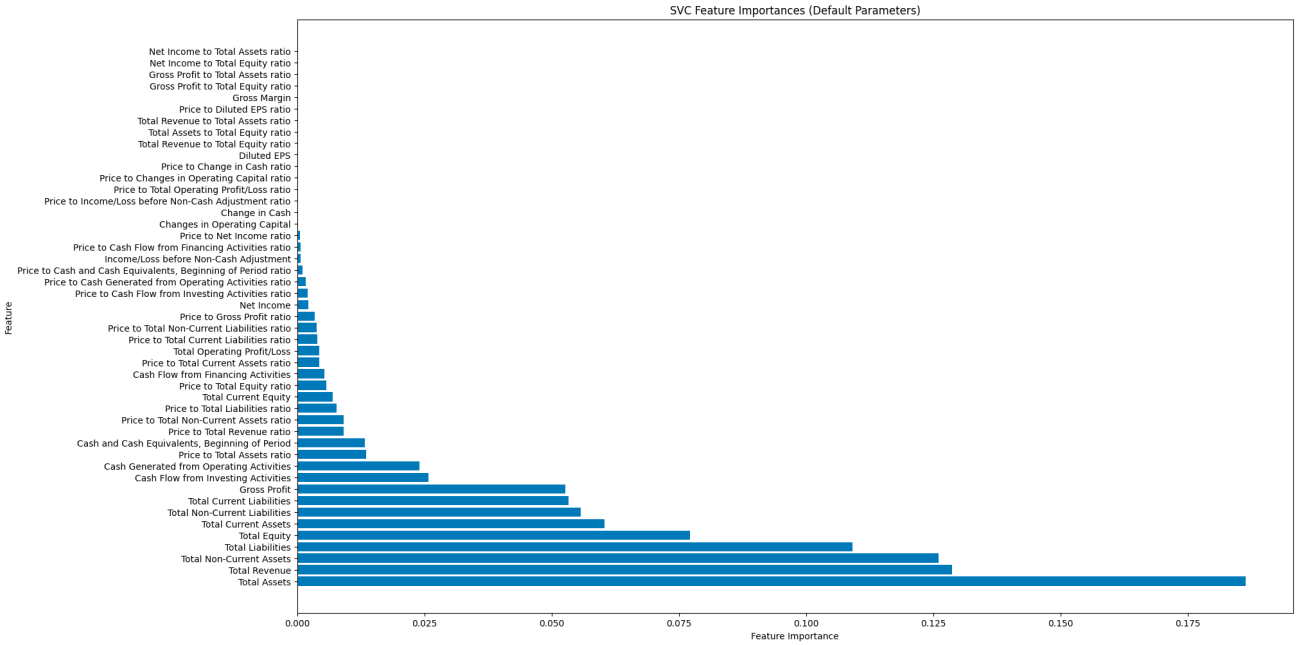
We performed feature selection using two methods: via a wrapper method that selected features based on a Random Forest model and via the *get_support* attribute of a fitted svm model. These two attempts were performed to get the sense of the features selected by different approaches. The final 5 best features where selected using a filter method *SelectKBest* based on the same Random Forest model. Interestingly, any subset of features yielded about the same performance, indicating the difficulty of the problem and the not so strong correlation of the features on the target class. The final dataset was formed by these 5 feature columns which was then scaled. Also removed from the dataset were features that had low variance using *VarianceThreshold*.

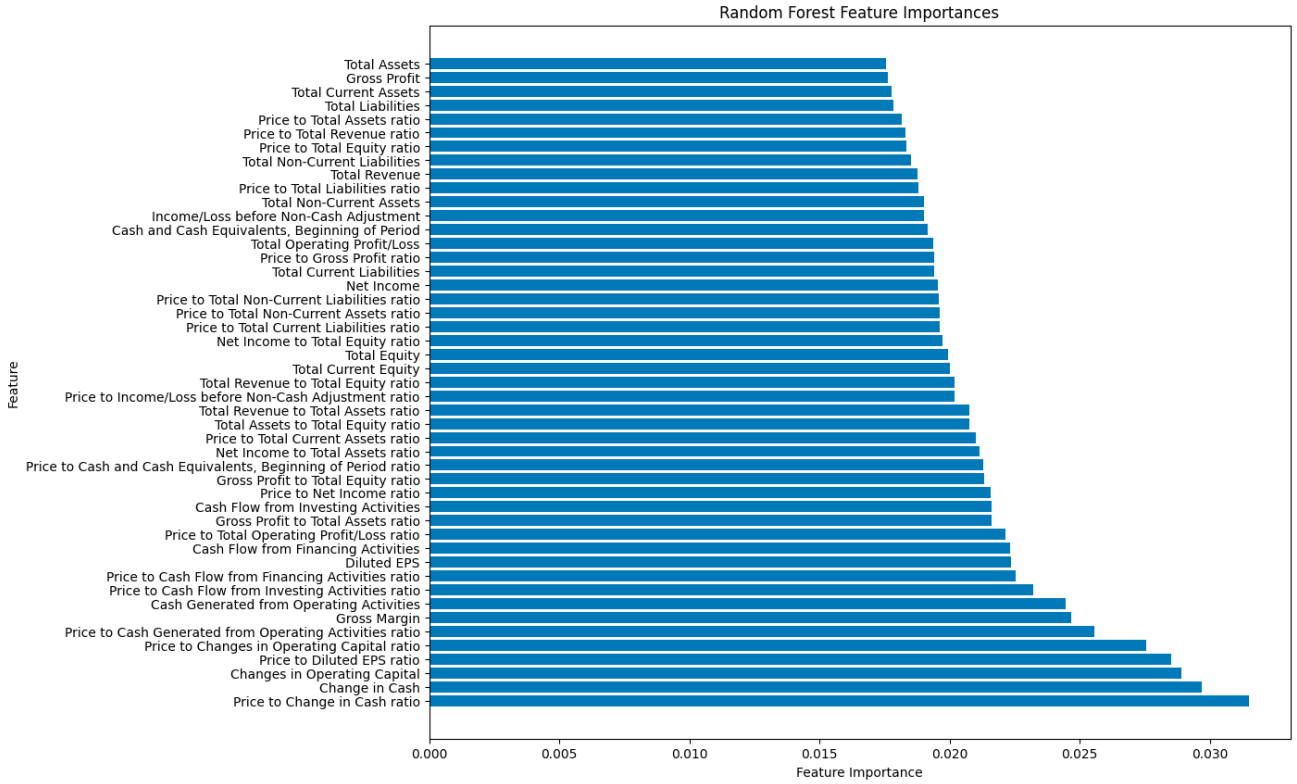
This is illustrated by the following plot that shows the performance based on the number of best features selected.



Feature Importance

We plotted the feature importances using the feature importance attributes of the classifiers

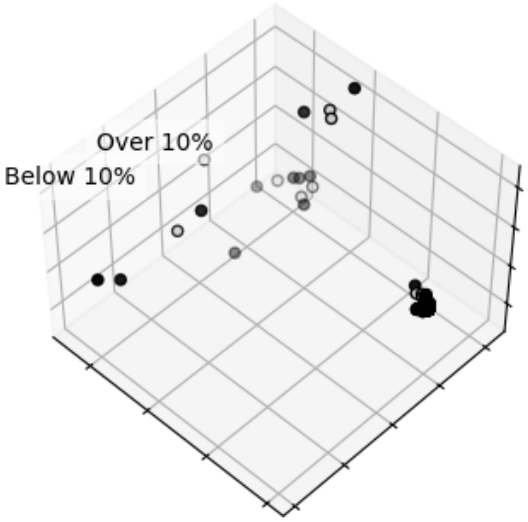




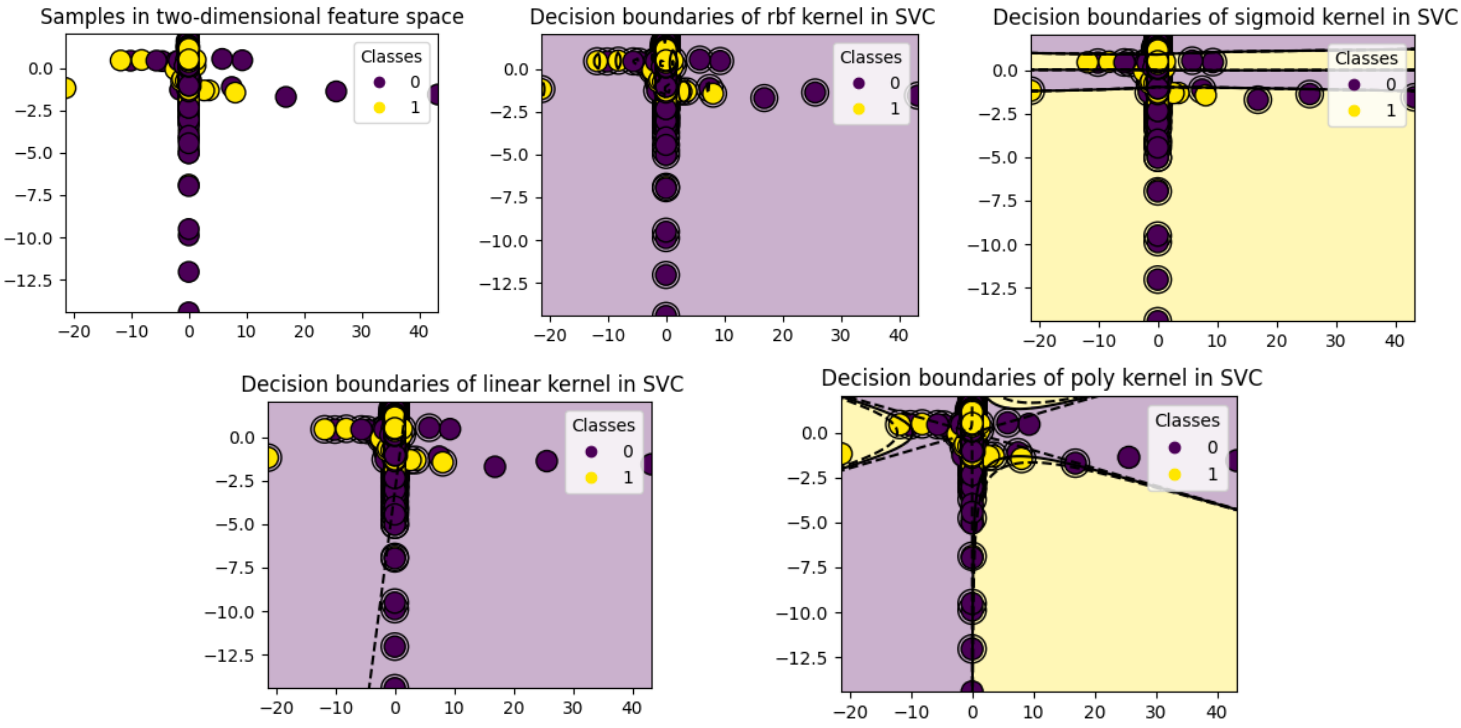
From these plots we can infer several remarks. Firstly we can see that the features do not have strong impact on the models performance and secondly and most importantly, features that were not considered important on the investment domain (ex *Price to Change in Cash ratio*) played a more significant role than other more general and important features (ex *Total Assets*).

Dataset Space

Using pca decomposition we were able to 3d visualize the 3 most important features of the dataset and we once again observe the inseparability of the data as most of the datapoints are clustered to one area. Nevertheless we tried to tune the models to yield as better of a result as possible.



The difficulty of the problem is excellently displayed by the decision boundaries of the SVM models for different kernels. As we can observe, the data are very difficult to be separated and classified even with complex models.



Modeling

As mentioned earlier, we proceeded with the SVM and the Random Forest methods.
We used the 5 best features of the data (scaled) and splitted the dataset into 80% training and 20% testing data using *train_test_split*. We did not use validation data as most of the validation was performed with *kfold cross validation* of the training data.

Model Training

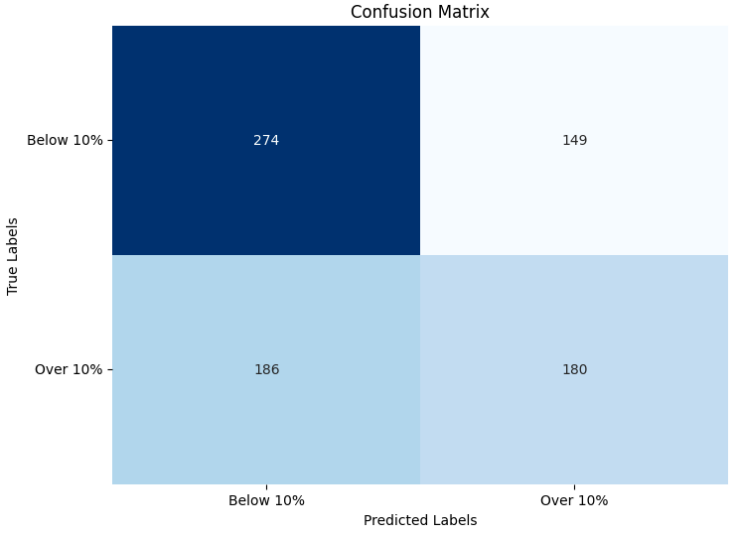
We trained several models and kept only the fine tuned Random Forest model which is loaded on the demo code.

Model Evaluation

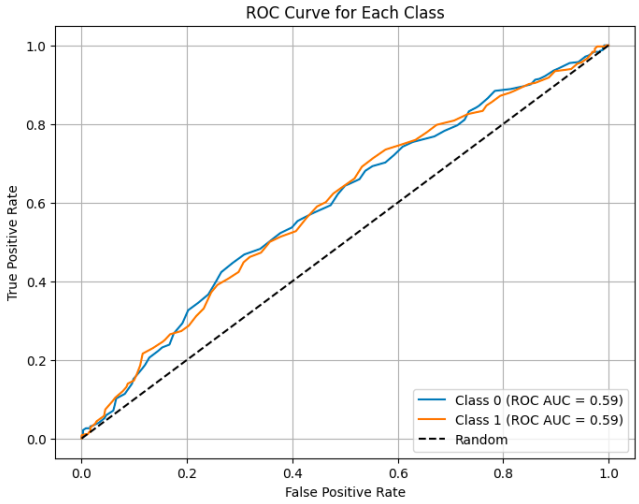
We computed the following evaluation metrics for the models we tested: accuracy, precision, recall and f-score. We also plotted several curves (roc-curve, precision-recall curve, confusion matrix and learning curve) and calculated auc scores for each model.

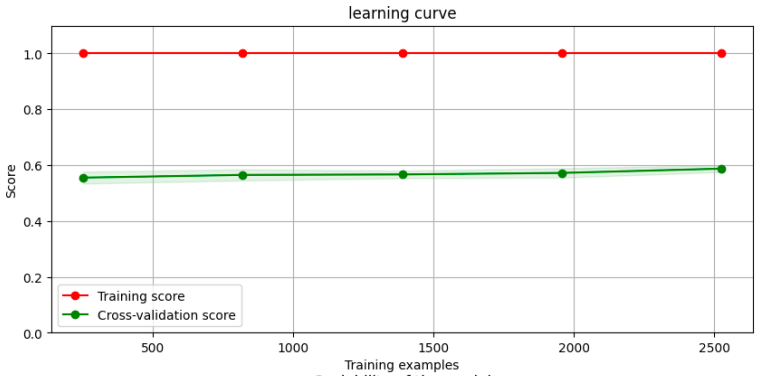
These were the performance metrics of a vanilla Random Forest model:

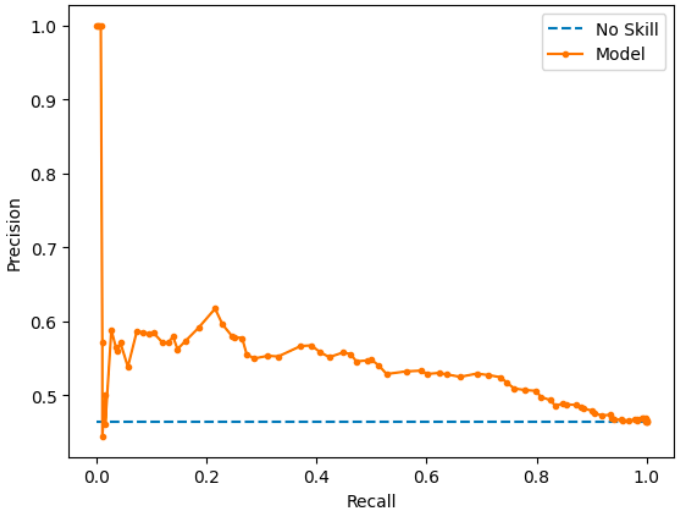
```
Cross validatation Accuracy:  0.58764017140443
              precision    recall  f1-score   support

   Below 10%       0.60      0.65      0.62       423
    Over 10%       0.55      0.49      0.52       366

    accuracy                           0.58       789
   macro avg       0.57      0.57      0.57       789
weighted avg       0.57      0.58      0.57       789

--------------------------------

Model Evaluation:
accurancy:  0.5754119138149556
precision:  0.5713823179595612
recall:  0.5754119138149556
f-measure:  0.5730056021668762
--------------------------------
```

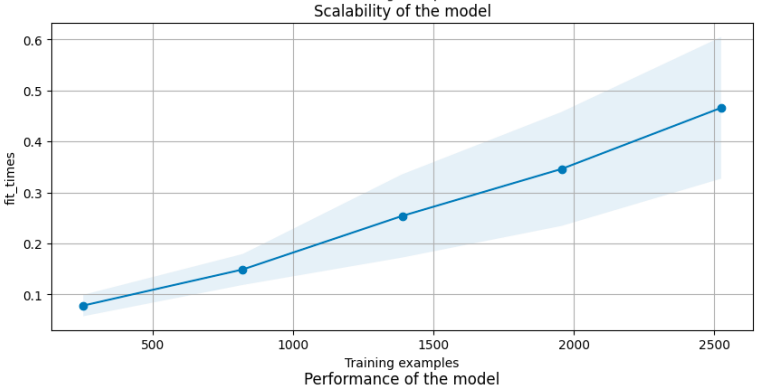


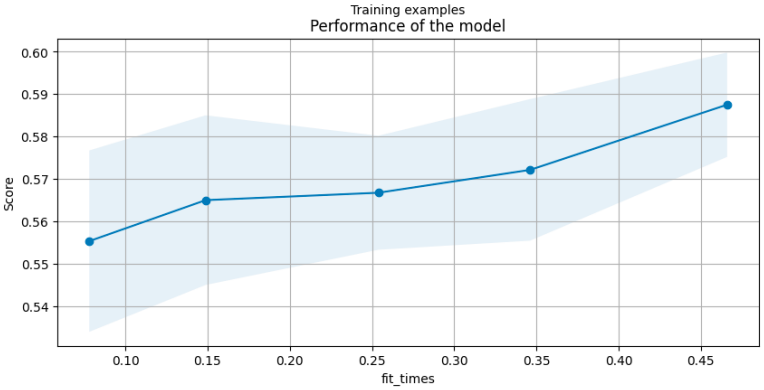
With an average ROC AUC score of 0.59, the model demonstrates slightly better than random performance in distinguishing between the positive and negative classes. An F1-score of 0.58 indicates a balance between precision and recall, suggesting that the model achieves reasonable performance in terms of both minimizing false positives and false negatives. However, it's essential to consider the class-specific metrics: a recall of 0.65 for one class suggests that the model effectively captures the majority of instances belonging to this class, while a lower recall of 0.49 for the other class indicates that the model struggles to identify all instances of this class, potentially leading to more false negatives. Overall, while the model shows moderate performance.







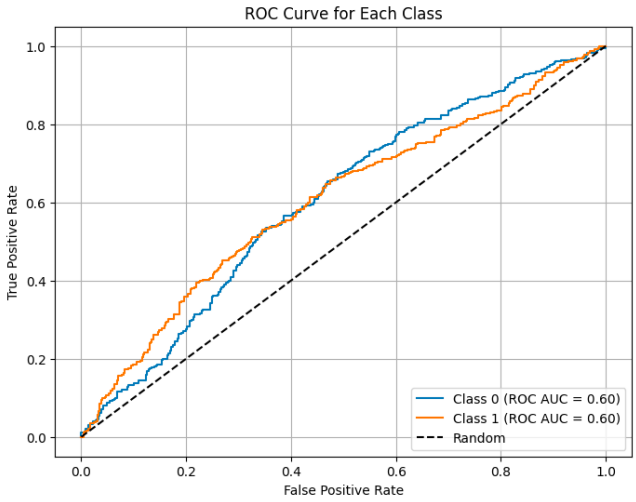


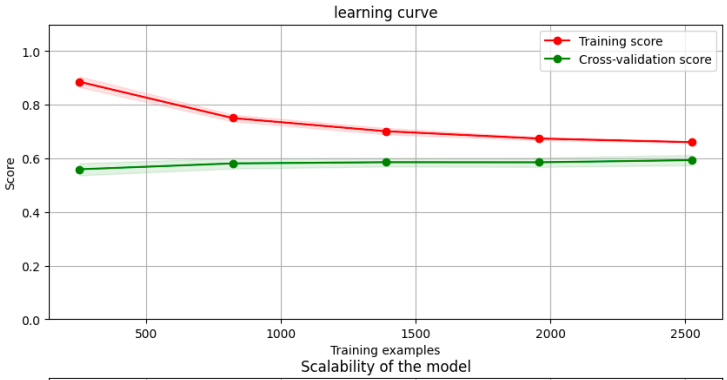


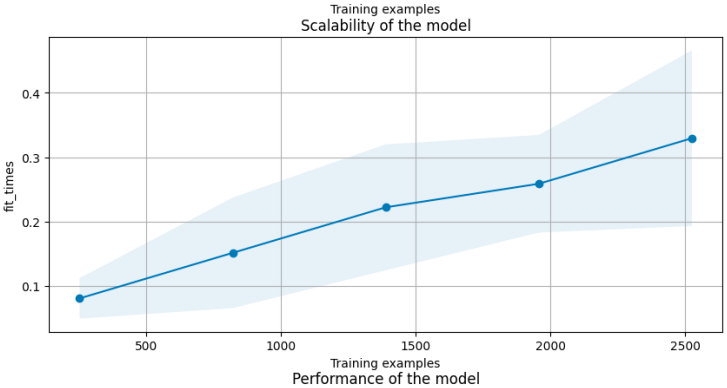
The learning curve of the model helps as visualize the factors affecting the models performance during training and also suggests that the model is **not overfitted.**

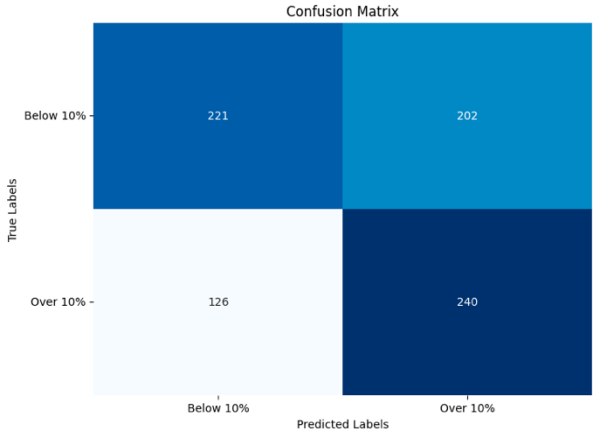
The fine tuned model performed similarly:

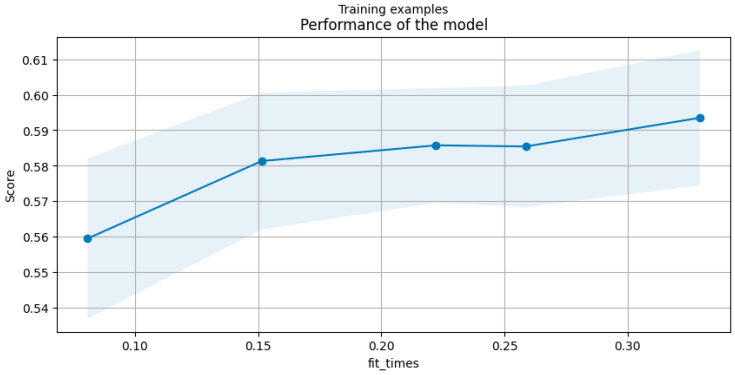
The performance boost is noticeable but not significant. Also the recall for each class is slightly inverted, possibly die to randomness.

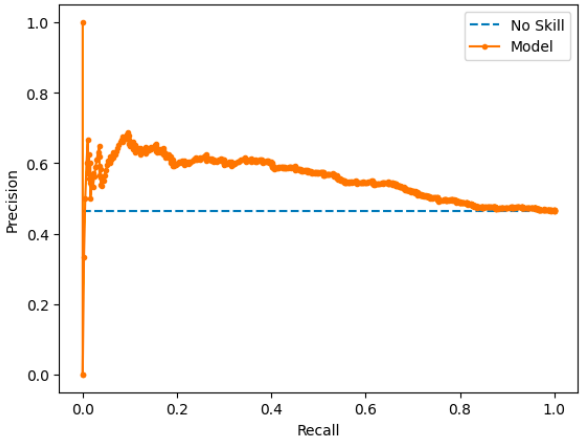












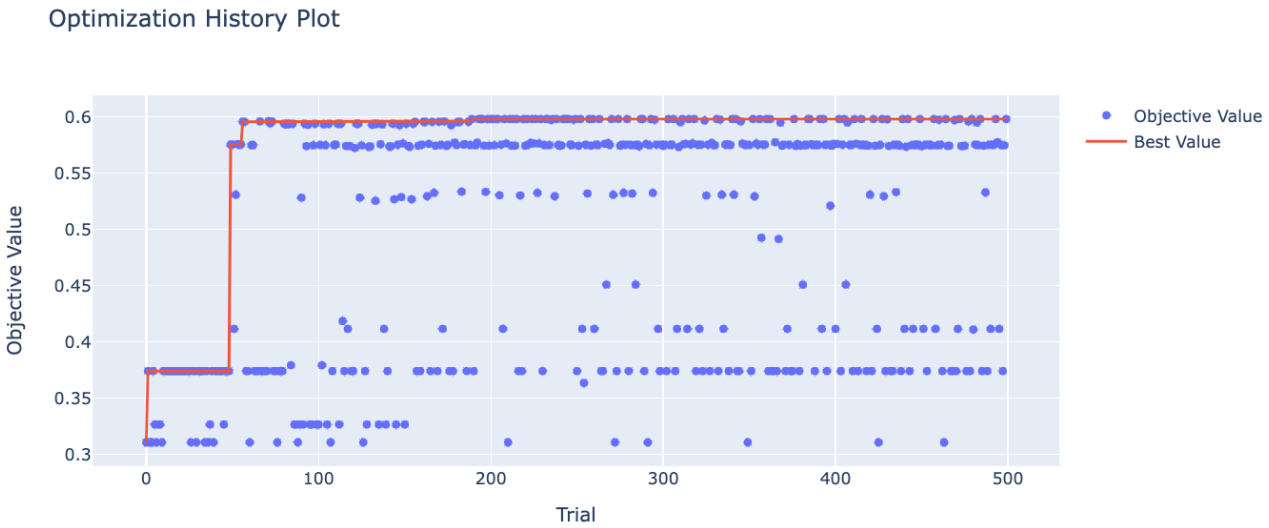

```
Cross validatation Accuracy:  0.5993674935119803
              precision    recall  f1-score   support

   Below 10%       0.64      0.52      0.57       423
    Over 10%       0.54      0.66      0.59       366

    accuracy                           0.58       789
   macro avg       0.59      0.59      0.58       789
weighted avg       0.59      0.58      0.58       789

------------------------------

Model Evaluation:
accurancy:  0.5842839036755386
precision:  0.5899370167042653
recall:  0.5842839036755386
f-measure:  0.5833190489065201
------------------------------
```

Our model achieves an accuracy of 58% for its total predictions, a precision of 54% for our above 10% returns and a precision of 64% for below 10% returns. Which in turn means that classifying a stock as an underperformer is easier for the model thus resulting in avoiding bad investing decisions. Also 66% recall for our first class implies that it also captures 2/3 of the actual stocks that overperform, meaning that we only miss 1/3 of overperformers when picking stocks to invest in.

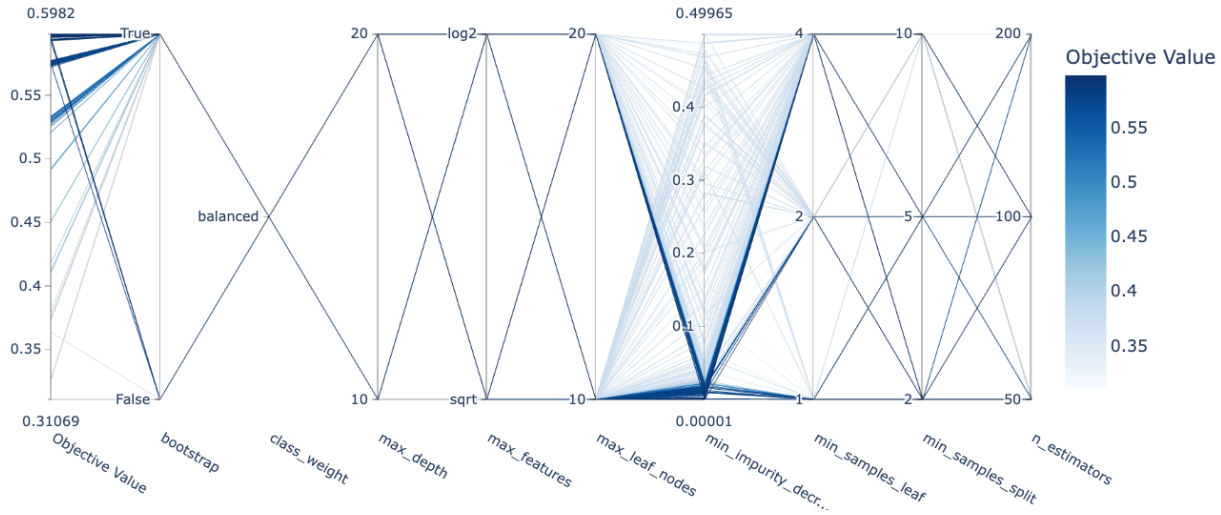
The metrics of the alternative tuned SVM model are available at the *Model.ipynb* notebook.

Model Tuning

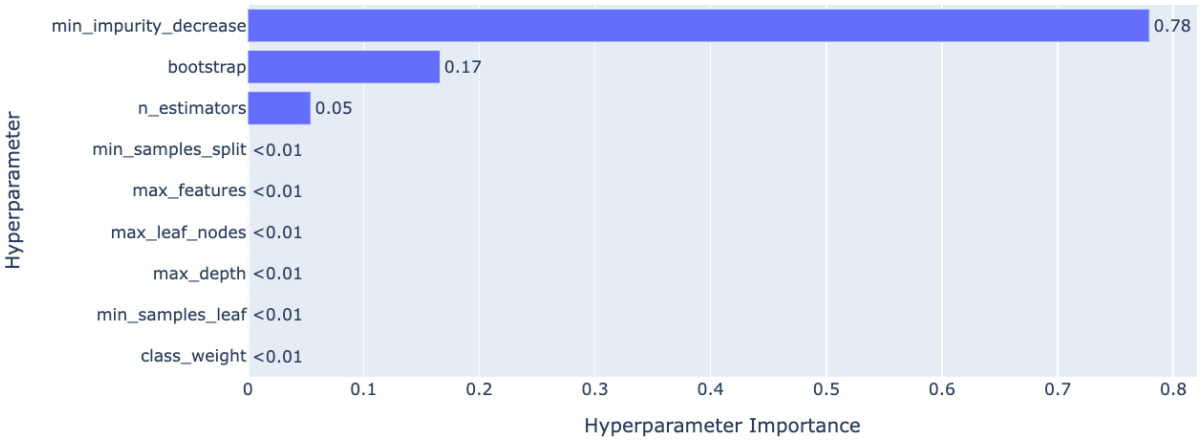
Two different fine tuning methods were used. Firstly we tuned an svm model using *GridSearchCV* and secondly we fine tuned a Random Forest Model using the *optuna* library. Different combinations of parameters were tested and the model's performance was recorded each time. More details about the tuning process can be found in the *Model.ipynb* notebook. Below are some plots generated by optuna showcasing the tuning process:

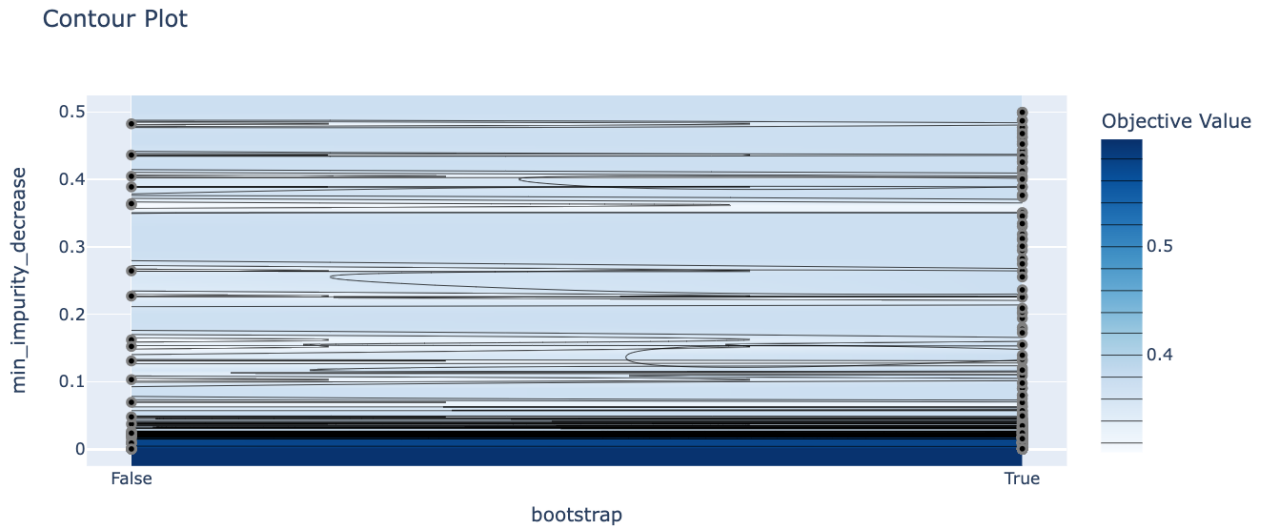


Parallel Coordinate Plot



Hyperparameter Importances





The importance of hyperparameters such as *min_impurity_decrease*, *bootstrap*, and *n_estimators* in a random forest model suggests their significant influence on the model's predictive performance. Specifically, a high importance score for *min_impurity_decrease* indicates the emphasis on controlling impurity decrease during tree-building, possibly implying a preference for simpler decision boundaries. The *bootstrap* parameter's importance suggests the importance of bootstrapping samples during training, indicating its role in enhancing model robustness and reducing overfitting. Lastly, the importance of *n_estimators* highlights the impact of the number of trees in the ensemble, implying the significance of model complexity and the trade-off between bias and variance.

Conclusions

A first look at the problem of stock investing using machine learning and financial data of companies results into promising results that could be somewhat useful in the investment world, thus leaving room for further improvement, mainly on the data used. We could, in the future, examine additional market domains and more features while we give emphasis on the feature curation and extraction to better explain the variance of a stock's return.