NASDAQ FInancials

The dataset for this project was collected from kaggle and originate from Nasdaq Financials. The dataset contains New York Stock Exchange historical metrics extracted from annual SEC 10k filings(2012-2016) should be enough to derive most of popular fundamental indicators.

In this project focus is on clustering and applying unsupervised learning algorithms to find the best candidate algorithm that accurately predicts whether a company has net profit or net loss. To do that, I shall transform Net Income column into a binary representation of whether or not a company made profit, where 0 represents loss and 1 represents profit.

Net Income is used for indicating a company's profit after all of its expenses have been deducted from revenues. This number appears on a company's income statement and is also an indicator of a company's profitability.

Data attributes:

- Ticker Symbol
- Period Ending
- Accounts Payable
- Accounts Receivable
- Addl income/expense items
- After Tax ROE
- Capital Expenditures
- Capital Surplus
- Cash Ratio
- Cash and Cash Equivalents
- Changes in Inventories
- Common Stocks
- Cost of Revenue
- Current Ratio
- Deferred Asset Charges
- Deferred Liability Charges
- Depreciation
- Earnings Before Interest and Tax
- Earnings Before Tax
- Effect of Exchange Rate
- Equity Earnings/Loss Unconsolidated Subsidiary
- Fixed Assets
- Goodwill
- Gross Margin
- Gross Profit
- Income Tax
- Intangible Assets
- Interest Expense
- Inventory
- Investments

- Liabilities
- Long-Term Debt
- Long-Term Investments
- Minority Interest
- Misc. Stocks Net Borrowings
- Net Cash Flow
- Net Cash Flow-Operating
- Net Cash Flows-Financing
- Net Cash Flows-Investing
- Net Income Net Income Adjustments
- Net Income Applicable to Common Shareholders
- Net Income-Cont. Operations
- Net Receivables
- Non-Recurring Items
- Operating Income
- Operating Margin Other Assets
- Other Current Assets
- Other Current Liabilities Other Equity
- Other Financing Activities
- Other Investing Activities
- Other Liabilities
- Other Operating Activities
- Other Operating Items
- Pre-Tax Margin Pre-Tax ROE
- Profit Margin Quick Ratio
- Research and Development
- Retained Earnings
- Sale and Purchase of Stock
- Sales General and Admin.
- Short-Term Debt / Current Portion of Long-Term Debt
- Short-Term Investments
- Total Assets Total Current Assets
- Total Current Liabilities
- Total Equity Total Liabilities
- Total Liabilities & Equity
- Total Revenue Treasury Stock
- For Year
- Earnings Per Share
- Estimated Shares Outstanding

Data exploration & Feature Transformation:

- Checking for null values and dropping those.
- Dropped columns which had no relevant information such as Unnamed:0,, Ticker Symbol and Period Ending
- Made sure all the columns are continuous which is what is needed for K-means clustering.

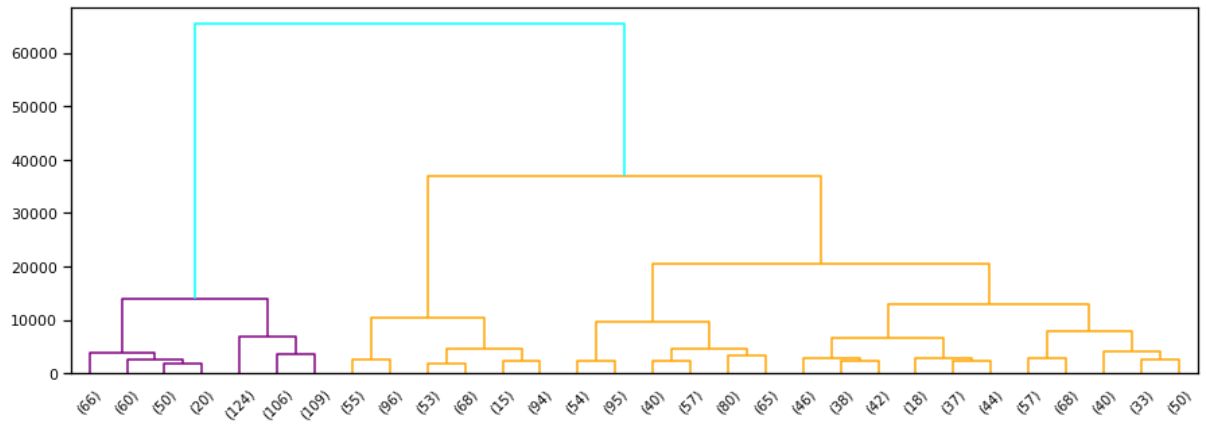
- Transformed Net Income (target) to binary values.
- Ensured that data is scaled and normally distributed.

Training Model:

- Fitting the model with 2 clusters initially
- Fitting 2 agglomerative clustering model 2 clusters (ward-link and complete-link clustering)
- Comparing the results
- Visualising the dendrograms produced by the agglomerative clustering.

Results:

Net Income	agglom_complete	agglom_ward	kmeans	number
0	**0**	**0**	**0**	8
			1	5
		1	**1**	89
1	**0**	**0**	**0**	287
			1	28
		1	**1**	1356
	1	**0**	**0**	8



Comparing the results, it can be seen that I am able to predict profit better than loss which is what I expected, given that I have more data for companies with profit (1:1679 vs 0:102). The best algorithm for predicting loss is the Complete-Link agglomerative clustering model

and for predicting profit K-Means clustering seems to be the best candidate although Ward-link agglomerative clustering achieved nearly the same result.

Going Forth:

There are various parameters that can be tried out, and also other algorithms like DBSCAN can be tried out if it comes out as fast processing and also having less computational complexities.