**Background:**

There are not many fields as lucrative and dynamic as the stock market. While predicting stock values flawlessly can’t yet be done even by a maestro, an increase in computational power and storage capacity have enabled us to try our hand at algorithmic predictions. In this paper, we aim to utilize the available financial data to classify if the stock values of a certain company will rise or diminish at the end of the following year. Such a prediction will give invaluable insights and will enable leaders to optimistically utilize the findings to propel their firm in the correct direction. Similarly, investors can use such a model to make informed decisions while buying stock.

**Introduction:**

We will be utilizing financial data of various companies from 2014 to 2018. The Dataset contains 200+ attributes, ranging from Sectors the companies are involved with, revenue and department wise financials. We aim to use multiple Machine Learning models to classify and predict if the company will have an overall increase in stock values in the upcoming year, or if it will decrease in value.

**Use cases:**

**Civilians:** users can use such a model to help complement their intuitions and knowledge to predict the stock values and make an informed decision while purchasing stocks. *Please note that even if the model will perform exceptionally well within our current dataset, the predictions should always be used as a piece of additional information.*

**Leaders:** Company leaders can use such a model to know which attributes are the ones most contributing to the company’s financial evaluation and which not so much. Hence resources can be pushed to pain points, and they can expect a growth in stocks accordingly.

**Dataset:**

The Dataset is available in the following files containing company data from different years.

* 2014\_Financial\_Data.csv
* 2015\_Financial\_Data.csv
* 2016\_Financial\_Data.csv
* 2017\_Financial\_Data.csv
* 2018\_Financial\_Data.csv

Overall, when combined the dataset has around 20K rows and 225 columns. Some columns contain nan cells (missing values). Additionally, the dataset contains outliers that are recognized in our data cleaning phase.

**Experiment Setup:**

The team have been collaborating on GitHub and working in Jupiter Notebook. We are using Python are our language. Below are a few major python libraries we are using for Data Analysis and Machine Learning.

Numpy, pandas, seaborn, matplotlib, sklearn, statsmodels

**Method:**

We start with Data Cleaning of the input data so that the dataset contains records that can be worked on by ML algorithms without erratic behavior.

**Data Cleaning:**

1. We start with loading all the input data streams (5 CSVs) into panda Data Frames.
2. All data frames (DFs) are then combined to make a common input Dataset.
3. Few columns are renamed, and unnecessary columns are then removed.
4. Null(*nan*) values cleaning: Columns having more than 20% null values are removed from the dataset. Similarly, rows with any null value is removed from the dataset. Final shape after operation: (12804, 171)
5. We observed that a few columns have more than 50% values 0. We hence deleted all such attributes having more that 50% values as 0 hence contributing to sparce Dataset.
6. One Hot Encoding is applied on Sector columns as it was the only column with a string value.
7. If the column contains just one value, its deleted.
8. Duplicate rows are checked for and deleted.

Next, we do **Exploratory Data Analysis (EDA)** on the data to find out patters and relations between the attributes. We aim to get more insights and understand the data further using the analysis results.

**EDA:**

1. We check if the output class (Attribute to be predicted) is balanced or not. As the output class contains 56% profitable and 44% lossy companies, we conclude that the data is balanced.
2. We then find correlation between different attributes and arrange/plot them according to positive and negative correlation.
3. We then plot Box plots to visually depict data variance.

We then proceed to outlier treatment. And employ multiple methods for outlier detection and below are the observations:

**Outlier Detection and Treatment:**

1. Standard deviation method: Number of identified outliers are 3948.
2. IQR Method : Just 12 records were identified as non-outliers.
3. 99-1 percentile method: non-outlier observations: 10864.

**Results:**

**Conclusion:**