**Stock Predictive Model – On Equity Fundamentals**

# GOAL – predict if a stock is likely to go up given the fundamentals data

It has been a long-time discussion which people are always interested in: Can we use a company’s equity fundamentals data to predict stock price, or to know if the stock price is likely to go to go up next year? I would like to build a machine learning model to answer this question.

# DATASET – NYSE S&P 500 companies historical prices with fundamental data

The dataset that I use to build the model is NYSE Fundamentals from:

There are several challenges in this dataset:

* Limited data - The Fundamentals dataset only has annual data which means for a stock, there are only 5 records.
* Unaligned data - The Prices data is provided for daily; but the Fundamentals are annual.
* Incomparable data - By nature, all the fundamentals data are company specific and not scaled the same. The general conception is that the scale of the company does not correlate with the stock price as the stock price is per share basis.
* Many Attributes - The dataset Fundamentals has 78 fields

# SOLUTIONS

In order to tackle these issues, I needed to perform some data transformation. This includes:

* Convert the actual amount into year-over-year growth rate to solve the company scale issue and make the data comparable across companies
* Convert the daily price data into yearly average to match the fundamentals data format

# MODELING – criteria selection

I use a machine learning algorithm called **Random Forest** to build the predictive model. It is a supervised learning algorithm which provides a good estimate of feature importance. The model is trained to take the most important fundamentals features such as Operating Margin to make the prediction.

I discovered that the data are quite centralized and therefore the model does not perform well if we simply classify the “price will go up” using positive growth (greater than zero). This results high false positive rate. In order to solve this problem I use the growth threshold of **20%** so we can more clearly distinguish the cut off line and thus reduce the false positive rate significantly.

# PERFORMANCE EVALUATION

The overall accuracy of the model is around 73% with false positive rate of 1.5%. This is rather a conservative model in that I wanted to reduce the false positive rate as much as possible. The reason being, we probably are not that concerned if it predicts a stock to be “is not going up” and it turns out to be “is going up”; however, when the model predicts something to be “price is going up” we do very much want it to be accurate.