# Machine Learning Engineer Nanodegree

## Financial Time-Series Prediction via Machine Learning

*Bryan Chik*

*15th August 2017*

##### I. Definition

#### Project Overview

Throughout the past 50 years, there has been numerous attempts and a rich discussion on financial markets mechanics and looking drivers of asset returns. However, one drawback of classical methods in predicting asset returns are most models are linear in nature. Therefore, most models have bad prediction power, as they cannot cater any non-linearity with the markets. Machine learning techniques could overcome this problem and qualifies as a good candidate to try solving the asset price prediction problem.

In this project, I would like to explore the possibility of applying machine learning techniques to predict Hong Kong Hang Seng index’s next day price movement, given closing prices of other 42 global exchange indices. A trading strategy will be designed based on the results of this research.

Initially I have obtained daily closing price of Hang Seng index and 42 other selected exchange indices globally from data provided by FactSet, starting from 1st January 2010 to 30th June 2017. A full list of indices can be found in the appendix.

The data is then trained to various model candidates and use for model performance evaluation and prediction.

#### Problem Statement

In this project, I am predicting Hang Seng Index’s next day movement, which has 3 states: Up, Neutral and Down. This problem is a classification problem.

“Up” label is defined as the daily price movement compared to yesterday’s close is greater than 0.5%. “Down” label is defined as daily price movement compared to yesterday’s close is less than -0.5%. While “Neutral” label is defined as not “Up” or “Down”.

4 classification algorithms will be utilized and see which one has the best predictability. Namely

* Logistic Regression
* AdaBoost
* Feed-Forward Neural Network
* LSTM

The performance of the above models are compared against a benchmark model, which in this case we are using Gaussian Naïve Bayes model, using 2 evaluation metrics for classification problems:

* F-beta Score (beta = 0.6)
* Precision

I anticipate there is at least 1 model would be better than the benchmark model, with higher precision and F-beta score. This in turn meaning the model does have higher prediction power, and can be used as an input to make trading decisions based on the prediction.

#### Metrics

2 specific metrics are chosen for evaluating the model performance:

* F-Beta Score (beta = 0.6)
* Precision

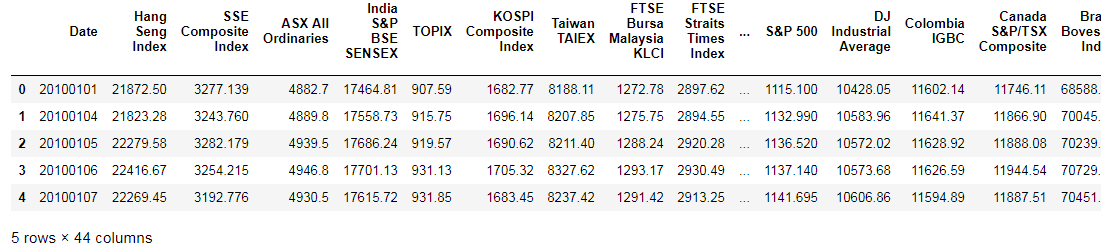
F-beta score is an overall measure that describes model accuracy, taking into account precision and recall. Note that I have placed special emphasize on precision with a beta value larger than 0.5. The reason behind is that in practical trading, it is easier to go long than short, so therefore, we want to predict better the “Up” labels.

For the same reason, precision is included in the evaluation metrics.

##### II. Analysis

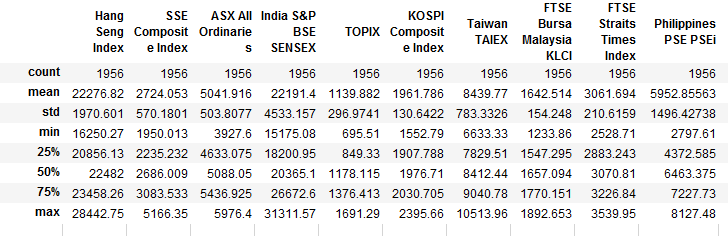
#### Data Exploration & Visualization

The data that I used to study is sourced from [FactSet Research System’s](http://www.factset.com/) investment management workstation terminal. I have downloaded daily closing index levels of 43 global equity exchange indices, including Hang Seng index, in the form of a csv file. The first column is the date index, starting from earliest to closest. A snippet of the data is shown below.



*Figure 1: First 5 rows of the raw index close data*

Next, I have generated some basic descriptive statistics of the data set. This is an attempt to get some knowledge around the distributional properties and shape of the data, and with the aim to determine any data pre-processing is required or not. An excerpt of the output is shown below.



*Figure 2: Descriptive Statistics of Asia Pacific indices*

There are some important observations within the table:

1. We have around 1,956 sample data points
2. Different indices have different levels. For example, mean Hang Seng level is 22,276.82, while for Japan TOPIX index the mean level is around 1,139.88
3. Different indices have different volatility profiles and range

From the above observation, particularly 2 and 3, imply that the data is not directly usable as they are at different levels or scale. As a result, the data has to be normalized