# Machine Learning Engineer Nanodegree

## Capstone Project Proposal: Forecasting Stock Price Movement Direction

### 1. Introduction

Predicting the trend of future stock price has always been an attractive topic in many fields including trading, finance, statistics and computer science. The primary objective is to determine the market timing to buy, hold, or sell a certain stock. The task is challenging because there are many uncertainties involved and many factors that influence the stock price. These include macroeconomic factors such as political events, firms' policies, general economic conditions, investors' expectations, institutional investors' choices, movement of other stock market, and psychology of investors etc.[1]. Therefore, stock market prices are susceptible to quick changes and are generally regarded as dynamic, non-parametric, chaotic and noisy in nature [2].

Financial markets around the world are becoming more integrated as the results of globalization. When subprime mortgage crisis in U.S. spiraled out of control, it caused the meltdown of the US economy and subsequently other countries. U.S. stock market suffered significant drop as shock over the United Kingdom voters' move to exit the European Union in 2016 [3]. These are the evidences that no financial market is completely isolated today. Political instability, economic factors, war and terrorism events in oversea could influence the performance of domestic markets. It is envisioned that the prices of global stock markets are correlated to each other.

### 2. Problem Statement

The objective of this project is to predict whether the close price of S&P 500 today will be higher or lower than yesterday. It is formed as machine learning classification problem with output be '1' if S&P 500 close is predicted to be higher than yesterday, '0' otherwise.

The main idea is to use major world stock indices as input features for the machine learning model. Stock market from around the globe has different closing time with U.S. market. For instance, Hong Kong's Hang Seng Index closes 12 hours before S&P index. Its stock price data is available before the beginning of the U.S. market trading time. This project would like to explore whether the correlation of oversea stock indices and U.S. market could be utilized as significant predictor for future movement of S&P 500.

## 3. Datasets and Inputs

For this project, the following stock indices will be obtained from Yahoo Finance via Python module 'pandas_datareader'[4]:

| Index | Description |
|---|---|
| S&P 500 | Stock market index based on the market capitalizations of 500 large companies having common stock listed on the New York Stock Exchange (NYSE) or NASDAQ. |
| Dow Jones Industrial Average | Price-weighted average of 30 significant stocks traded on the NYSE and the NASDAQ. |
| NASDAQ Composite | Market capitalization-weighted index of approximately 3,000 common equities listed on the Nasdaq stock exchange. |
| London FTSE | Share index of the 100 companies listed on the London Stock Exchange with the highest market capitalization. |
| Frankfurt DAX | Stock market index consisting of the 30 major German companies trading on the Frankfurt Stock Exchange. |
| Tokyo Nikkei | Stock market index comprised of Japan's top 225 blue-chip companies traded on the Tokyo Stock Exchange. |
| Hong Kong Hang Seng | Market capitalization-weighted index of 40 of the largest companies that trade on the Hong Kong Exchange. |

The following figure shows the snapshot of stock index data retrieved from the web:

| Date | Open | High | Low | Close | Adj Close | Volume |
|---|---|---|---|---|---|---|
| **2014-10-23** | 25.100000 | 25.270000 | 25.100000 | 25.270000 | 25.270000 | 1300 |
| **2014-10-24** | 25.139999 | 25.190001 | 25.139999 | 25.190001 | 25.190001 | 15300 |
| **2014-10-27** | 25.028000 | 25.028000 | 25.028000 | 25.028000 | 25.028000 | 300 |
| **2014-10-28** | 25.440001 | 25.510000 | 25.440001 | 25.510000 | 25.510000 | 3800 |
| **2014-10-29** | 25.660000 | 25.670000 | 25.660000 | 25.670000 | 25.670000 | 13600 |

Each stock index has the following attributes:

a) Date: in days

b) Open: price of the stock at the opening of the trading

c) High: highest price of the stock during the trading day

d) Low: lowest price of the stock during the trading day

e) Close: price of the stock at the closing of the trading

f) Adj Close: price of the stock at the closing of the trading adjusted with dividends and stock splits.

g) Volume: amount of stocks traded during a given trading day

## 4. Solution Statement

Several supervised learning classification techniques will be used to address the stock price prediction problem, namely Decision Tree, Logistic Regression, Support Vector Machine and Multilayer Perceptron Neural Network. The input data to the classifier is the stock price of major world indices, while the output is the predicted trend of S&P 500. The output is '1' if the market is predicted to close higher than yesterday and '0' otherwise. The performance of these classification models will be compared to select the best model.

## 5. Benchmark Model

After the classification model has been trained, its predicted output will be used to generate buy/sell signals in a trading strategy that trades S&P 500. The performance of the trading strategy will be benchmarked against a low risk buy-and-hold strategy of S&P 500. The trading strategy based on the prediction outcomes could be consider success if its portfolio return is higher compared with buy-and-hold strategy.

## 6. Evaluation Metrics

Accuracy and F1 Score will be used to quantify the performance of the classification model. They are defined as follow:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Predictions\ Made}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

To compare the portfolio performance of buy-and-hold strategy and the trading strategy based on the prediction results, Sharpe Ratio and return rate will be used. These metrics are defined as follow:

$$Return\ = \frac{Ending\ Value\ of\ Portfolio}{Beginning\ Value\ of\ Portfolio} - 1$$

$$Sharpe\ Ratio = \frac{\bar{r}_p - r_f}{\sigma_p}$$

where:

$$\bar{r}_p = Mean\ return\ of\ portfolio$$
$$r_f = Risk\ free\ rate$$
$$\sigma_p = Standard\ deviation\ of\ portfolio$$

## 7. Project Design

The proposed solution will be implemented based on the following workflow:

a) Data retrieval
   - Historical stock price data will be retrieved from Yahoo Financial by using Python library.

b) Data Preprocessing
   - Filling missing values and normalization of data will be performed in this stage.

c) Feature Extraction/Dimension reduction
   - The correlation of various global stock prices and S&P 500 will be studied to select strong predictors for the model. Dimension reduction will be performed if necessary to reduce the input dimension of the model.

d) Model construction
   - After splitting the data into training set and testing set, the training dataset will be used to build the classification models.

e) Model Evaluation and Optimization
   - The performance of various classification models will be evaluated and compared by using testing dataset. The parameters of the models will be fine-tuned by using grid search method to improve the prediction accuracy.

f) Trading Strategy Evaluation
   - A trading strategy based on buy/sell signals generated by prediction output will be implemented. Its portfolio return and Sharpe ratio will be compared with buy-and-hold strategy.

**References**


[1] T. Z. Tan, C. Quek, and G. S. Ng, "BIOLOGICAL BRAIN-INSPIRED GENETIC COMPLEMENTARY LEARNING FOR STOCK MARKET AND BANK FAILURE PREDICTION1," *Computational intelligence*, vol. 23, pp. 236-261, 2007.

[2] Y. S. Abu-Mostafa and A. F. Atiya, "Introduction to financial forecasting," *Applied Intelligence,* vol. 6, pp. 205-213, 1996.

[3] *U.S. stocks hammered as Brexit shock rocks markets*. Available: https://www.usatoday.com/story/money/markets/2016/06/24/brexit-bombshell-torpedoes-global-markets/86323890/

[4] *pandas-datareader Documentation*. Available: https://pandas-datareader.readthedocs.io/en/latest/