# Capstone Project Proposal

**Predicting Stock Price Movement Using Machine Learning**

## Domain Background

Financial institutions, such as investment companies, trading firms and hedge funds, build various financial models to analyze market information in order to make better investment and trading decisions. The increased accessibility of computational power and various kinds of market data have enabled advanced technics, such as machine learning, to be employed in problem-solving in the finance domain. Besides, being an engineer in a trading firm, I am keen in applying the machine learning knowledge gained in this course to examine and solve a problem in the finance domain.

## Problem Statement

Analyzing and predicting trading prices of financial instruments have been a challenging but intriguing task for many analysts and researchers. A model that effectively predicts the direction of future stock price movement will aid investors and traders in making more profitable investments and being better at managing the risks of their portfolios.

## Solution Statement

The availability of historical market data and machine learning frameworks, such as TensorFlow, Keras, Scikit-learn and PyTorch, make machine learning a plausible and increasingly popular approach to predicting stocks price movements. In this project, I aim to build a stock price movement classifier that predicts the direction of stock price movement, namely up or down, of selected stocks N-days ahead based on historical data. In view of the time series financial data used, I will explore several models suited for time series prediction including Recurrent Neural Networks (RNN), such as Gated Recurrent Unit (GRU) and Long Short-Term Memory neural network (LSTM), under the Keras framework . I will also evaluate the performance of a Random Forest classifier using scikit-learn. After a few iterations and model tuning, the best-performing model will be implemented in the stock price movement classifier.

**Datasets and Inputs**

The stock price movement classifier takes daily historical trading data over a certain date range for selected stocks as inputs to the training of a machine learning model. The classifier outputs whether the stock price will move up or down N-days ahead from a selected date . The historical stock prices will be retrieved from Yahoo! Finance via the yahoofinancials or web API. The data consists of trading date (Date), opening price (Open), highest and lowest price traded in the day (High/Low), closing price adjusted for stock splits (Close), closing price adjusted for splits and dividends (Adjusted Close), and finally the trading volume (Volume). The target for prediction is the direction of movement of the adjusted close price, which is used in the end of day PnL (Profit and Loss) calculation. However, the returns, instead of prices, will be used as model inputs in order to satisfy stationarity and generate more accurate results. Additional features will also be generated for exploratory data analysis and model training.

**Benchmark Model**

Naïve method, which is often used for evaluating the accuracy of time-series forecasting, is chosen as the benchmark model. In the absence of seasonality, the naïve method is based on a random walk and each prediction is simply equal to the last observed value, in this case the direction of the previous day's stock price movement. Although naïve method itself might be overly simplistic, it serves as a useful benchmark for other forecasting methods and machine learning models in financial time series prediction. A machine learning model needs to beat the naïve method in order to be considered practical.

**Evaluation Metrics**

This project deals with a Classification problem as the target for prediction is the direction of stock price movement (up or down). Accuracy, which is calculated as the total correct predictions (true positives + true negatives) divided by the total number of predictions, will be used as an evaluation metric. To account for any potential imbalanced data, such as more up movements or more down movements, Recall and Precision scores may also be calculated and examined.

**Project Design**

The project workflow is planned as follows:

1. Initial research and selection of stocks to be used for machine learning
2. Model training and evaluation via Jupyter Notebook:
    - Retrieve historical stock prices from Yahoo! Finance
    - Exploratory data analysis and feature engineering
        - Generate stats, plots, correlations, etc and examine the data
        - Create features such as returns and other technical indicators such as Moving Average (MA) and Relative Strength Index (RSI)
        - Create prediction targets based on the sign of daily return
    - Data-preprocessing for machine learning
        - Train/Test data preparation in consideration of the dimensionality requirements of different models: Random Forest requires 2D inputs while RNN requires 3D inputs
    - Model training, tuning and evaluation
        - Several models such as Random Forest, Gated Recurrent Unit (GRU), Long Short-Term Memory neural network (LSTM) will be examined and evaluated against the benchmark model
        - Machine learning techniques such as hyperparameter tuning may be explored
3. Write a python script to implement the stock price movement classifier using the best-performing model. The script takes user inputs via command line arguments including:
    - stock names and a date range for model training
    - a stock name and a date for stock price movement prediction
4. Write a project report detailing all the above steps including the design, analysis, implementation of the project, as well as results, conclusions and any potential enhancements to the model training or project implementation.

In conclusion, predicting stock price movement using machine learning is an interesting problem to solve in the finance industry. I look forward to applying the machine learning skills gained in this course to complete this project successfully.