# Machine Learning Engineer Nanodegree

## Capstone Proposal

Wei Feng
February 4th, 2018

## Proposal

*This proposal is motivated and developed based on the suggested problem areas provided by Udacity Capstone Project Description – Investment and Trading*

### Domain Background

Stock price forecast is important for individuals, investment firms and hedge funds to make decisions in stock exchanges and portfolio management. Accurate stock price forecast is the key for investors to maximize their return.

There are basically two steams of analyses in stock price prediction: fundamental analysis and technical analysis (Xu, 2012). Fundamental analysis focuses on the entire stock market as a whole or determines a company's stock price based on underlying factors that affects the company's actual value; technical analysis tries to predict a stock price only based on its past trend (usually time serious analysis techniques).

This project focuses on technical analysis of stock price because it is free of individual company's information collection, screening and processing and therefore it is more transferable to other stocks. The study focuses on short-term 7-day stock price forecast, which is very useful for short-term investors to find the stock that has the highest forecast increase/decrease in the next 7-day so that the investor can decide which stock to buy or sell. This project does not intend to provide optimization recommendations for investors on portfolio management.

### Problem Statement

The goal of this project is to build a stock price forecaster that takes stock name and historical date range as input from users and output forecast closing stock price for the next 7 days.

## Datasets and Inputs

The data used in this project comes from Yahoo! Finance, the data include the following basic information for each stock: Date, Open (price), High (price), Low (price), Close (price adjusted for splits), Adj Close (price adjusted for both splits and dividends) and Volume. Therefore, Adj Close will be used as the label. These data will be used in the stock price prediction model training as well as cross-validation and testing. In this project, five stocks (e.g. FB, AAPL, AMZN, GOOG, MSFT) and the benchmark S NASDAQ will be chosen and the past two years' data will be used.

These historical stock closing price data provide the basic information to develop features such as momentum, volatility, Bollinger Bands, Sharp Ratio, etc. which may be used in the models.

## Solution Statement

Since stock price is a continuous variable, the learning algorithms that will be used and compared in this project include: linear regression, k-nearest mean, random forest, or an ensemble of these algorithms.

The data will be split into three sets: training, validation and test. Training set will be used to develop model, validation set will be used to fine-tune the model to reduce overfitting, and the test data will be used to evaluate the performance between benchmark model and the proposed model.

## Benchmark Model

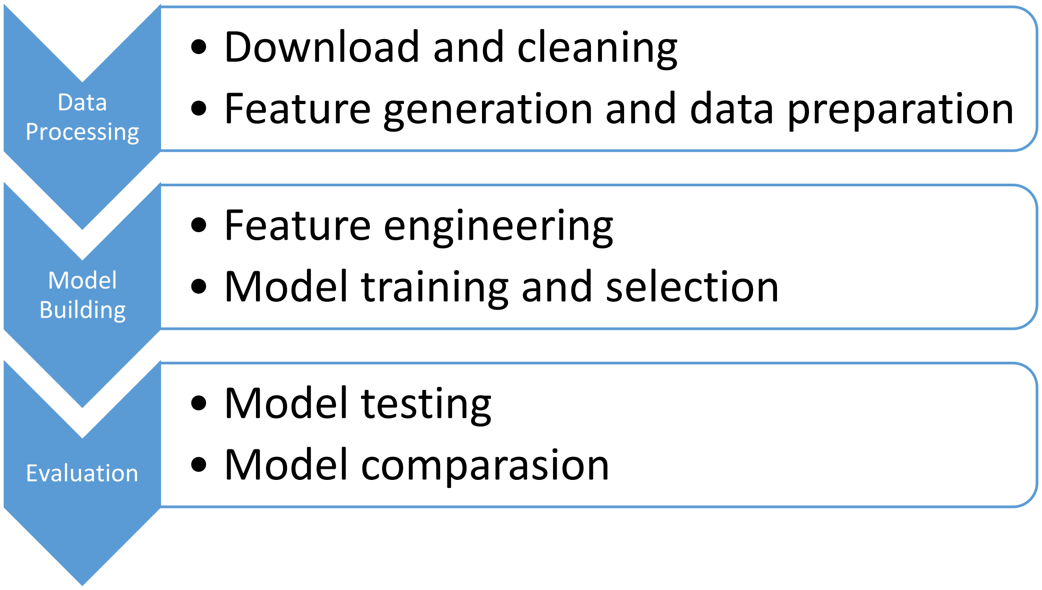
There are numerous models for stock price forecast, in this project, two benchmark models will be used: a simple moving average (SMA) model as the naive benchmark model; and the default model used by QuantDesk (developed by Lucena Research) as another benchmark model. The default QuantDesk model is a linear regression model with more than 500 available indicators, for simplicity, the benchmark linear regression model in this project uses SMA, SMA Momentum, Sharpe Ratio, Volatility, as well as SP500/NASDAQ SMA Change and SP500/NASDAQ Volatility as the predictors.

## Evaluation Metrics

The performance of the two benchmark models and the proposed models can be evaluated by Root Mean Squared Error (RMSE) between the forecast closing prices and actual adjusted closing prices. (Machine Learning for Trading, Tucker Balch, Lesson 03-03)

$$RMSE = \sqrt{\frac{\sum\left(Y_{actual} - Y_{forecast}\right)^{2}}{N}}$$

## Project Design



- Download and cleaning
- Feature generation and data preparation
- Feature engineering
- Model training and selection
- Model testing
- Model comparasion

### Data Processing

- Download historical data for five stocks (FB, AAPL, AMZN, GOOG, MSFT) and the benchmark (NASDAQ) over the past two years.
- Remove rows with empty adjusted closing prices

### Model building

- For each stock and the benchmark, compute relevant features such as SMA, Momentum SMA, SMA change, Volatility, Sharp Ratio, deviation from Bollinger Bands, etc.
- Construct a new dataset for each stock, in which each row contains the Date and adjusted closing price and all of the computed features for the stock and the benchmark (NASDAQ). Date is not a feature here but will be used to split data chronologically.
- Split data into training, validation and test sets. Due to the characteristics of the stock data, these three data sets will be split chronologically in a rolling basis so that no future information is used in prediction. Validation data are always later than training data (at least one day) and testing data are always later than validation data (at least one day). For example, if feature uses data 5/1 – 5/31,

label date is 6/7 in training data, validation data can use feature based on data 5/2 – 6/1 with label date as 6/8, similarly for testing data.

- Automate the above steps by writing a function that takes stock name as the input and output the three data sets for the corresponding stock.
- Two algorithms will be used: random forest and XGBoost.
- For each algorithm, select the best features and parameters based on cross-validation results. For example, for random forest, find the best number of maximum features, number of trees and minimum leaf sizes, for XGBoost, find the best learning rate, maximum depth, etc.

**Evaluation**

- Test the performance of the models based on testing data and select the one with the best performance over all five stocks
- Compare the performance with the two benchmark models and analyze the performance difference.

## Reference

Xu Y., "Stock Price Forecasting Using Information from Yahoo Finance and Google Trend", UC Berkeley, 2012.