**Machine Learning Engineer Nanodegree**

**Capstone Proposal**

Partha Deka  
February 13th, 2017

**Proposal**

**Domain Background**

I have chosen the Investment and Trading Capstone project provided as one of the options. I wanted to use my Machine Learning knowledge to solve a real problem and where I can learn the domain too. I have always been interested in investment and trading as an amateur and now I want to use my machine learning knowledge and see if it could help in predicting stock prices in the short-term future. I extensively took help of the free courses – “Machine learning for Trading” and “Time Series Forecasting” for this Capstone Project.

**Problem Statement**

Build machine learning models that learn from historical stock price attributes and predict the stock price on a future date (any day after the last training date), I am only predicting the Adjusted Closing pricing.

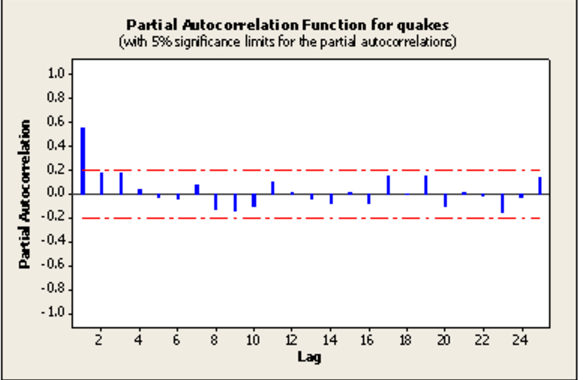
Building few models such as Linear Regression, KNN Regression which learn from historical stock data, attributes such as: Open, High, Low, Close, Volume, Adjusted Closing and “Adjusted Closing price” n days ahead in future. After the model is built on the historical data, it would predict the Adjusted Closing price for a date (only one date) which is n days ahead in future from the last date of training. Please note splitting the datasets has to be done in the ascending order of dates manually, and not using the SKLEARN test train split module which randomly split the data. In the stock price prediction world, we must not train a model on future datasets and try to predict the stock prices for past dates (Source - “Machine learning for Trading” Udacity course)

I am also using Non-seasonal ARIMA (model for Time series prediction) based model. This model is based on three terms: AR –Autoregressive - P, I – differencing - d, MA – Moving Average - q.  [Parameters](https://en.wikipedia.org/wiki/Parameter) p, d, and q are non-negative integers, p is the order (number of time lags) of the [autoregressive model](https://en.wikipedia.org/wiki/Autoregressive_model), d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the [moving-average model](https://en.wikipedia.org/wiki/Moving-average_model). Three items should be considered to determine a first guess at an ARIMA model: a time series plot of the data, the ACF, and the PACF.

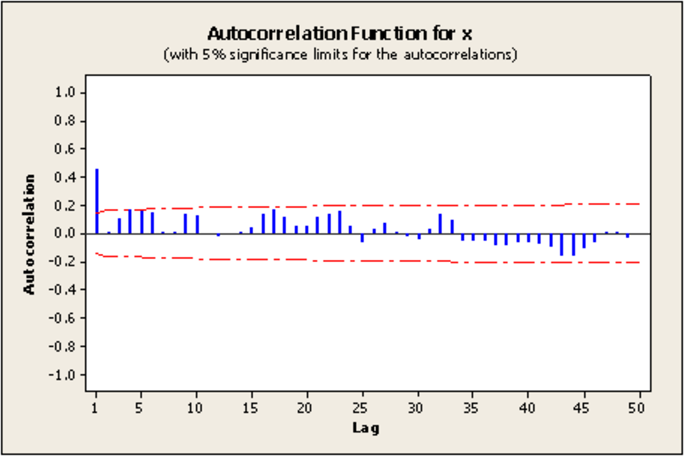
AR model: **Identification of an AR model is often best done with the PACF.** the theoretical PACF “shuts off” past the order of the model.  The phrase “shuts off” means that in theory the partial autocorrelations are equal to 0 beyond that point

MA model: For an MA model, the theoretical PACF does not shut off, but instead tapers toward 0 in some manner.  A clearer pattern for an MA model is in the ACF.  The ACF will have non-zero autocorrelations only at lags involved in the model.

Sample PACF for an AR model below where the PACF shuts off past the 1st order so, AR =1 in this case:



Sample ACF for a MA model below where MA =1



**Datasets and Inputs**

I am downloading all historical Stock data from [Yahoo Finance](https://finance.yahoo.com/)

In this section, the dataset(s) and/or input(s) being considered for the project should be thoroughly described, such as how they relate to the problem and why they should be used. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included with relevant references and citations as necessary It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.

**Solution Statement**

*(approx. 1 paragraph)*

In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).

**Benchmark Model**

*(approximately 1-2 paragraphs)*

In this section, provide the details for a benchmark model or result that relates to the domain, problem statement, and intended solution. Ideally, the benchmark model or result contextualizes existing methods or known information in the domain and problem given, which could then be objectively compared to the solution. Describe how the benchmark model or result is measurable (can be measured by some metric and clearly observed) with thorough detail.

**Evaluation Metrics**

*(approx. 1-2 paragraphs)*

In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).

**Project Design**

*(approx. 1 page)*

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

**Before submitting your proposal, ask yourself. . .**

* Does the proposal you have written follow a well-organized structure similar to that of the project template?
* Is each section (particularly **Solution Statement** and **Project Design**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
* Would the intended audience of your project be able to understand your proposal?
* Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?
* Are all the resources used for this project correctly cited and referenced?