# Machine Learning Nanodegree Capstone Project

# Stock Value Prediction

*Nabin Acharya*

## Domain Background:

Using Financial Models from the past to make future investment profitable returns is a common trend. Investment firms, Hedge funds and individuals use this model all the time to determine when and how to move investments so profits can be maximized and losses can be minimized.

Thousands and thousands of companies are traded by stock in stock markets around the world with hedge funds, investment firms , mutual funds and individuals.

There are many different factors that affect the price of the stock and it is very difficult to predict this with a good level of accuracy.

One of the most important decisions in finance is to buy and sell a stock at a certain price. Predicting the price of a stock based on historical data and other industrial factors is a very important challenge.

## Personal Motivation:

Most retirement funds, banks, even high-schools today rely a lot on the stock markets for profits. One of the prominent trends in the financial world is to take profits as soon as possible to meet trends influenced by various industrial trends.

Financial domain changes lives tremendously, and machine learning can help that understanding a lot given so much historical data is available to solve some complex problems to help grow personal wealth.

Research papers:
https://www0.gsb.columbia.edu/faculty/ptetlock/papers/Kelley_Tetlock_Aug16_Retail_Short_Selling_and_Stock_Prices_with_Appendix.pdf
https://arxiv.org/pdf/1603.00751.pdf
http://www.vatsals.com/essays/machinelearningtechniquesforstockprediction.pdf
https://www.researchgate.net/publication/259240183_A_Machine_Learning_Model_for_Stock_Market_Prediction

## Problem Statement:

Given the how a stock has traded over a period of time,  and the information about short trading and the trends in the industry that the stock belongs to, predict the price of the stock in the future.

## Datasets and inputs:

Financial Historical data for this project will be obtained from Yahoo Finance using the Pandas DataReader. Dataset for the begin_date and the end_date for the stock training data will be fetched from Yahoo finance and saved as a csv spreadsheet file.
Data collected on each stock ticker will be

Date of Trading
Volume of Trading
Price Opened
Price closed
High
Low

For each trading day.

Information on short sales will be obtained from:
http://www.batstrading.com/market_data/shortsales
Stocks used in this project will be GOOG, AAPL, AMZN and NFLX
Dates used for the Training begin_date will be 1/1/2017 and end_date will be 5/12/2017
Prediction date will be 6/12/2017

Data collected will be used as is. We will add features like Short-Ratio that we will calculate from the data obtained from batstrading.com , and Moving average that we will calculate.

The begin_date and end_date will be split into sets of each month ( 5 in this case )
Older data is used as a training data and newer data will be used as a testing data.

## Solution Statement :

In this project the user has choices for the following:

Begin_date: Date to start looking at the stock trading data set
End_date : Last date to pick up the stock traded data set
Prediction_date : Date to predict the stock value
Industrial Ticker List: List of other stock ticker symbols related to this stock

We will use short-ratio based on the number of stocks shorted vs number of stocks traded. We will also calculate the moving average and calculate the industrial trends based on industrial ticker list provided.

Prediction_date will always be after the end_date .

Data will be first obtained from Yahoo Finance and processed for basic trading data for each day of trading for :

Date of Trading
Volume of Trading
Price Opened
Price closed
High
Low

Data will then be obtained from batstrading.com for Volume of trade and Volume of shorts to determine the shorts-ratio for the stock

Moving average feature will be calculated.

2 techniques will be used:

Technique 1:

1. Use SVM for each month of data to create a prediction data for the next month.
2. Use K-Fold in the prediction data with the stock data from next month to create a good model to run the svm predict from.

Technique 2:

1. Use Neural net from Keras.
2. Use each month's data as training and next month as testing data
3. Use the neural net as prediction

Technique will be a choice to the user.

At the time of writing this proposal we are not sure which technique will produce a better result in prediction, so both of the techniques are listed here.

## Benchmark model:

- Better the score returned by SVM out-of-the-box benchmark
- Make prediction so it is within 10% of the actual value

Compare with the stock prediction with the online site:

- http://www.stock-forecasting.com

## Evaluation Metrics:

We will use the MAE ( Mean Absolut Error ) as described in the Wikipedia as:
https://en.wikipedia.org/wiki/Mean_absolute_error

Mean Absolute Error = (1/N) * Sigma( i from 0 to N ) ( prediction_value - real_value )

The MAE measures the mean of the difference between the prediction value and the real value.
For this type of a problem , this appears to be a good performance metric.

Root Mean squared Error will also be used

```
from sklearn.metrics import mean_squared_error
RMSE = mean_squared_error(y, y_pred)**0.5
```

This is because RMSE gives relatively high weight to a large error value. This makes it so RMSE is useful when larger error deltas are not desirable on predicted values.

## Project Design:

Data Gathering

This is by using Yahoo Finance with Pandas DataReader
Short data is obtained from batstrading.com

Initial Data Processing

Data Exploration is done by looking at the stock data obtained
stock-Ratio is calculated
Moving average is calculated
Feature selection is done in Date, Closing price, Opening price, Volume of trading

Train/Test Data K-Fold Cross Validation

Support Vector Machines

We will use the SVC type available in scikit-learn

Neural Network

User Query ⇒ Neural Network ⇒ Result

Or

User Query ⇒ SVM ⇒ Result