**Domain Background:**

Stock market and share trading has always been my interest. During my undergraduate studies, I was always curious to know how stock prices vary over time and what is the reason behind the variation. I was very keen to know what makes people trust a company’s stock and what is the factor that big stock traders trust while investing big money in stocks. Warren Buffet has been my inspiration always and I have been observing his investment and business actions put up in news.

I tried to relate lot of factors to predict some common stocks from the NASDAQ and Bombay stock exchange in India but succeeded only for a few stocks and not very often. Despite being an electrical engineer, I have always been working on the software and working on ideas that would be great for stock prediction. After I got to know about machine learning and how it works, I believe I have found my solution to the stock prediction problem. Working on a machine learning algorithm to predict stock and improving the accuracy to get as close as possible is my goal for this project.

**Problem Statements:**

Predicting stock prices of firms participating in New York stock exchange is the problem statement of this proposal. New York stock exchange is one of the largest stock markets in the world with a market capital value of 19.3 trillion [1]. Predicting the stock prices of companies that trade in New York stock exchange would help realize the profits and losses in investment and would assist a stock trader or investment firms to choose and trade wisely. Machine learning is a mathematical solution to this problem of predicting stock prices. The idea is to solve the stock prediction problem using a mathematical approach of machine learning.

**Datasets and Inputs:**Large amount of stock data is required to develop an accurate model that predicts stock prices precisely. Quandl is one tool we will be using for this project to extract historical stock data of New York stock exchange. The project would be developed in python using quandl library which provides 12917 datasets to work on [2]. The stock data available on quandl has open price, close price, high and low price are the features available. Extra features would be created as per requirement. Labels would be next day open price, same day close price as per the requirement. Validation set can be extracted from the real-time stock trading that occurs every day. The daily opening and closing prices can be compared to the predicted prices by the model.

**Solution Statement:**

To address the issue mentioned above in the problem statement, a model would be developed using one of the classifiers like gradient descent, Adam or Naive Bayes to predict the closing price, next day opening price, highest price of stock of a firm of choice for a day or by weekly given the previous stock data and its features.

**Benchmark model:**

The developed model can be evaluated against existing models. Some papers published on machine learning for stock prediction claim accuracy of up to 79% and these models would be compared to evaluate the model and improve. Ideally the goal is to develop a model that predicts the closing prices with an accuracy of 70-75%. The model would be specific to the New York stock market and hence would be optimized to train and test faster than a generic model.

**Evaluation Metrics:**

The model can be evaluated against the daily stock closing prices, a validation set can be prepared with real time data to check the accuracy of the prediction.

**Project Design:**

The first step of this project is collection of stock data using quandl library for python. All the required data related to the New York Stock exchange for stock prediction is extracted using the code XYNS [3] and data is split into train, test and validation datasets.

The second stage is to selectively use important features prioritizing them accordingly in order and using five to six different models like gradient descent or Naïve Bayes to compare the scores, behavior and time taken to train, predict. Then the best model is chosen and a check for overfitting is done by feeding real time stock data, pre-defined validation data set.

If the bench marks are not reached, the model is overfitting and hence an algorithm would be developed to prevent overfitting or generalize the model. The model can be compared to the ground truth using a graph plotted in real time when the predicted value is outputted.

**References:**

[1]. https://en.wikipedia.org/wiki/New\_York\_Stock\_Exchange

[2]. https://www.quandl.com/data/XNYS-New-York-Stock-Exchange-Prices

[3]. https://www.quandl.com/data/XNYS-New-York-Stock-Exchange-Prices