STOCK PREDICTION USING MACHINE LEARNING

**Aim:** The aim of the project is to determine the price of the given stock based on the previous data using machine learning algorithms.

**Basic terms:**

Before predicting the price of a given stock, we need to first know about some basic terms related to the stock market.

As we look on the sample data set that we have, we have the following terms.

**Price:** A share **price** is the price of a single share of number of saleable stocks of a company, derivative or other financial asset. The stock price is the highest amount, someone is willing to pay for the stock, or the lowest amount that it can be bought for.

**Open Price:** The **opening price** is the price at which a security first trades upon the opening of an exchange on a trading day. The price of the first trade for any listed stock is its daily opening price.

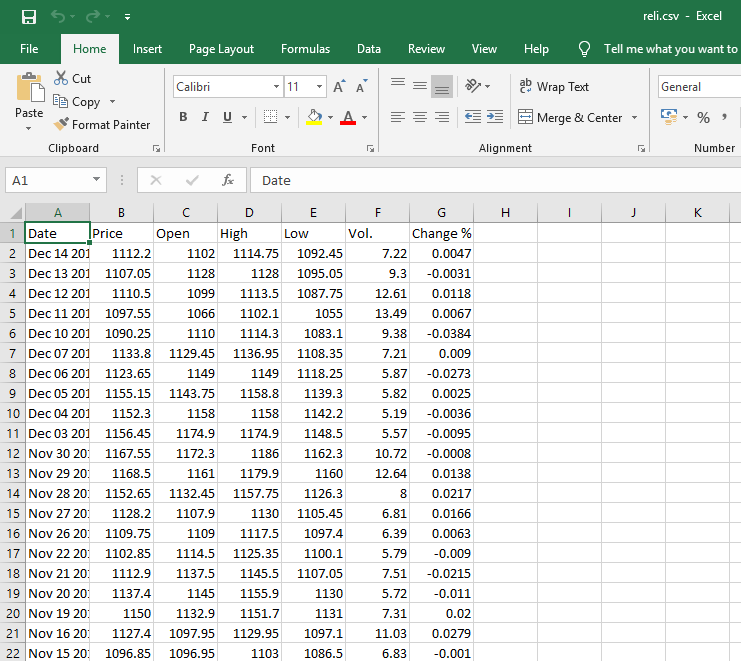
**High Price:** The **high price** is the highest price recorded in the specific period of time, like the highest recorded price of the stock in the given day.

**Low Price:** The **low price** is the lowest price recorded in the specific period of time, like the lowest recorded price of the stock in the given day.

**Volume: Volume** is commonly reported as the number of **shares** that changed hands during a given day.

**Change: Change** percent is the difference between today’s opening price and the closing price of the previous day.

Below is the screenshot of the data file with the columns described above.

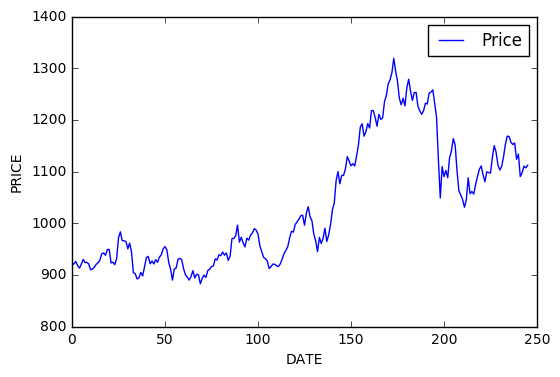


Screenshot of the data of the stock prices

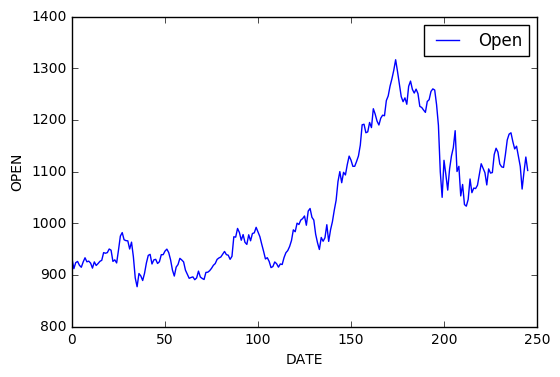
**Pre-Processing:**

Now as we have the required data, we need to pre-process it and clean it before proceeding further. The first step is to eliminate or convert the **categorical data** into numerical form. Here **‘Date’** is a categorical data and hence we need to ‘Label Encode’ the data or eliminate it. Since ‘Label Encoding’ and ‘One Hot Encoding’ takes a lot of columns and as we are limited to small dataset, it’s better to eliminate the column and instead add another column denoting the **“index”**. Each date is assigned with an index value with increasing order, i.e, 1st Jan’s index is **“0”** and 2nd Jan’s index is **“1”** and so on.

The next step is to plot the graphs for ‘dates’ vs different parameters is shown.



Date vs Closing Price



Date vs Opening Price

Since all the plots are closely similar, we consider date vs open plot.

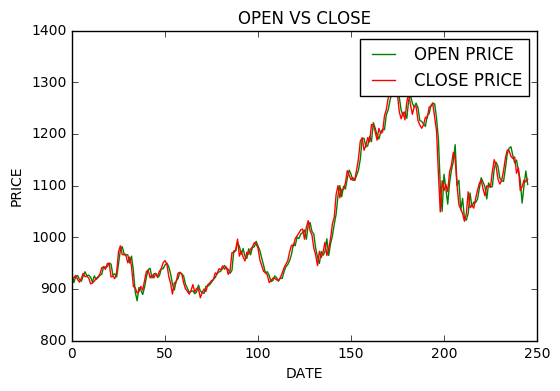
**Prediction:**

Now we have our dataset pre-processed. The next and very important thing is to divide our dataset into ‘features’ and ‘labels’. ‘features’ are the values we already know or pre-determined and ‘labels’ are the values we need or predict.

Since we already know ‘date’ and ‘Open’ price, they become our features and using these values, we will first predict ‘Price’ and ‘Change%’.

Afterwards we can use these predicted parameters to predict further secondary values such as ‘Volume’, ‘High’, ‘Low’ etc.

We have predicted the ‘Price’ parameter using ‘Open’ as there is a close relation between them. We can get to this conclusion by plotting a graph between ‘Price, Open vs Date’.



Open Price vs Closing Price

Our next step in pre-processing is to **‘Feature Scale’** the data. Since some features are big like ‘Price’ when compared to ‘Change%’, Feature scaling helps us to overcome biasing which happens due to larger value columns.

Now we can predict the required data. Though we can predict using many different algorithms, the algorithm used here to predict is ‘Multiple Linear Regression’.

We divide dataset in the ratio 1:5. The first part is for checking our results and the second part is for the training. Then we calculate the score of the algorithm to check it’s accuracy.