Final Project Report  **Machine Learning on Stock** **Market**chine Learning on Stock Market Data

short line

Instructor- Srikanth Krishnamurthy

TA- Pranjal Jain

Team 9 -Tapadyuti Maiti, Khushbu Parekh

August, 2017

[**Problem Statement**](#_4lqp25cx7kth) **2**

[**Summary**](#_59uob76loooz) **4**

[**Part 1: Data Wrangling**](#_u2lbr22vzvod) **5**

[Data Consideration :](#_qgfhbjwnczkq) 5

[Data Import from all the major Market Indicators around the World](#_m02uql9fcpg9) 5

[Programmatically download the data:](#_oyznv1nuwf13) 6

[Challenges Faced :](#_oy8l7b50c9sg) 7

[Data PreProcessing](#_wtfh9qd77b1q) 7

[**Exploratory Data Analysis**](#_d0ipve5h6ild) **10**

[DOW30](#_3ui9jutfqjjt) 10

[Univariate Analysis](#_7r5ub2rau4wd) 14

[Bivariate Analysis](#_8sg3jkldvr1c) 19

[Data Normalization](#_td2ykb3948jm) 19

[Top 5 Stocks](#_nxu6yg8cn3dg) 23

[Univariate Analysis](#_qd8dqk8dh4ok) 25

[Bivariate Analysis](#_x2reyazdo8vo) 28

[Data Normalization](#_ygttdqx4j1k0) 28

[**Feature Engineering**](#_ltweqva18mi9) **30**

[Recursive Feature Elimination](#_4dkx6klp8wax) 32

[**Prediction Modelling**](#_ihcd0chp2zi0) **34**

[DOW 30](#_dlxwom8t3upg) 34

[Prediction with Logistic Regression](#_nv0l54tyz2xh) 34

[Prediction with Random Forest](#_iloykggdmqf4) 39

[Prediction with Neural Network](#_kx28ktwr8vwv) 47

[Top 5 Stocks](#_3la3h0u9tg4q) 56

[**Choosing Best Model :**](#_utbrtevdz7zq) **61**

[**Twitter Sentiment Analysis**](#_q6faoq52pm7c) **62**

[**Classification**](#_s8oo2i19ns5p) **63**

[**Test Cases**](#_dxbydl9ts8qf) **63**

[**Web Application:**](#_97w7ri51zyyi) **64**

[Mailing Functionality :](#_4dpphmk1vo6s) 65

[**Technology Stack**](#_8tmeox38jsws) **66**

[**Contribution**](#_l8pfvehcgljx) **68**

[**Appendix :**](#_wh6f7bnhftyx) **69**

[Added Feature:](#_qleqps2ao7e1) 69

# 

# 

# 

# Problem Statement

Stocks Prices are totally random and unpredictable and it is highly volatile in nature. We would to predict and classify tomorrow's exchange probable closing price and its indicator of going high and low. As we know successful prediction of a stock’s future Price could yield significant profit.

Dataset:

Yahoo Finance : Stocks Insight Data https://finance.yahoo.com/lookup

Quandl : WIKI - https://www.quandl.com/data/WIKI-Wiki-EOD-Stock-Prices

Twitter Data - Tweets regarding each Company https://twitter.com/

Google Trends - Trending Words https://trends.google.com/trends/

# 

# Summary

We have come upon a successful model which can predict on selected Stocks and DOW 30 Stock Market Indicator closing price with MAPE 1.3 % by training Feed Forward Neural Network.

We have also made Classification Model which can classify tomorrow's HIGH and LOW for S&P 500.

Following report will convey all steps that we have followed to make this project and experimentation done over the time to achieve it.

# 

# Data Wrangling

Data Download

## Data Consideration :

We focusing on the DOW 30 - Dow Jones Industrial Average (DJIA) Index as the Market Indicator of USA . As it is the general case that whenever something extraordinary thing happens, a lot of factors get affected because of it. Through this project, we are trying to prove the same, whether the daily news has any impact on the stock market or not.

Nowadays, no financial market is isolated. Economic data, political perturbation and any other oversea affairs could cause dramatic fluctuation in domestic markets. A “bad day” on the Australian or Japanese exchange is going to heavily affect Wall Street opening and trend. In the light of the previous considerations the following predictors have been selected:

## **Data Import from all the major Market Indicators around the World**

NASDAQ Composite (^IXIC Yahoo Finance) - USA

Dow Jones Industrial Average (^DJI Yahoo Finance) -USA

Frankfurt DAX (^GDAXI Yahoo Finance) - Germany

London FTSE-100 (^FTSE Yahoo Finance) - UK

Paris CAC 40 (^FCHI Yahoo Finance) - France

Tokyo Nikkei-225 (^N225 Yahoo Finance) - Japan

Hong Kong Hang Seng (^HSI Yahoo Finance) - China (H/K)

Australia ASX-200 (^AXJO Yahoo Finance) - Australia

We have considered all major market indicator around the globe which may affect the DOW 30 market indicator.

We are analysing the top performer from the following sectors , so as to analyse how they co relate with each , if so and how all companies from the all sectors are performing.

HEALTH - Jhonson & Jhonson

FINANCE - Bank of America

TECHNOLOGY - Apple

ENERGY - Exxon Mobil Corporation

TELECOMMUNICATION - AT&T

## Programmatically download the data:

We are using pandas\_datareader package of Python to download the data from yahoo finance .As the data is open we do not need any username and password or authentication process involved.

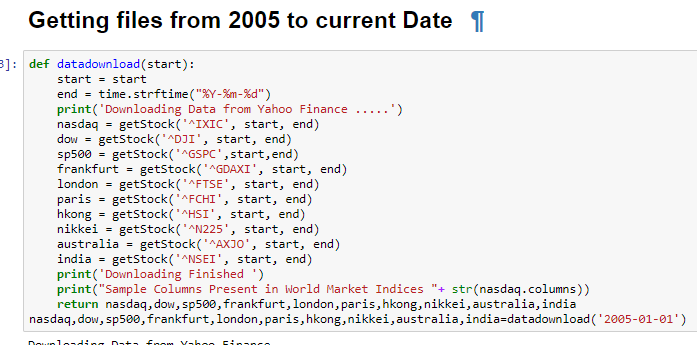
Few Insights of the method of the request that we are using :

Start date - 2005/1/1

End Date - Current Date

We have handled error also if there is connection issue from the Yahoo , then it will take up the local repository data.





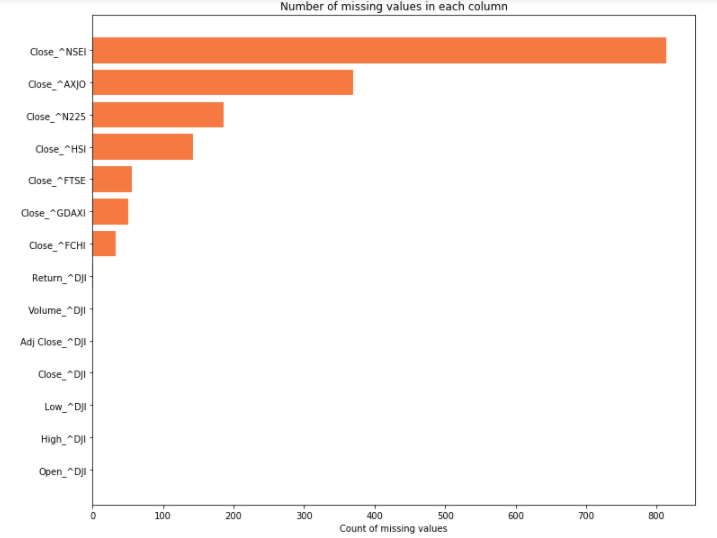
### Challenges Faced :

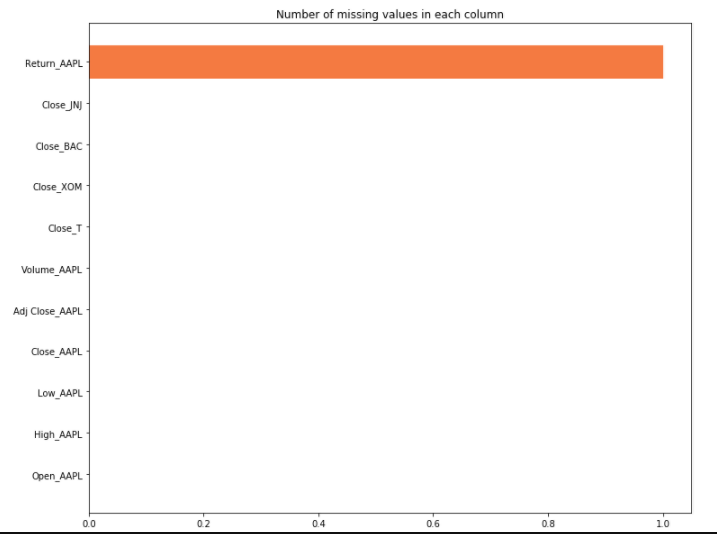
We have noticed few times the Yahoo link would not respond.

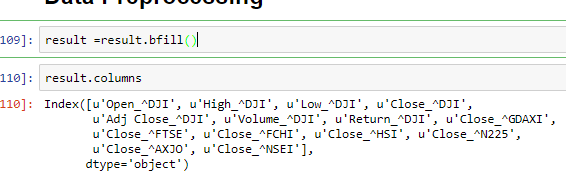
Solution: a.Wait for 30 min to get reactivate the link or work with last repository that you have saved.

## Data PreProcessing

We have preprocessed the data in the data frame itself and removed the trade date where no trade has happened and filling the missing values with the before day trade.







# Exploratory Data Analysis

## DOW30

WE are first visualising the data how the close price has been changed throughout the time.

We have plotted the DOW 30 , France , and Germany Market

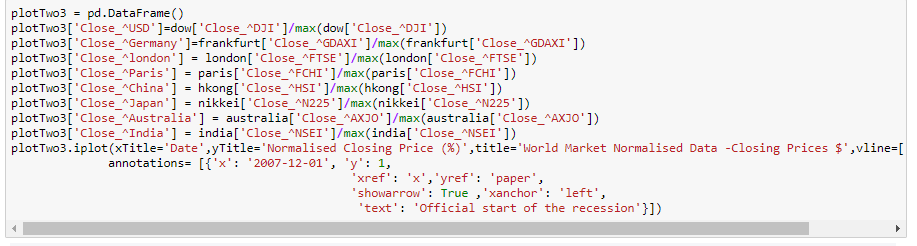


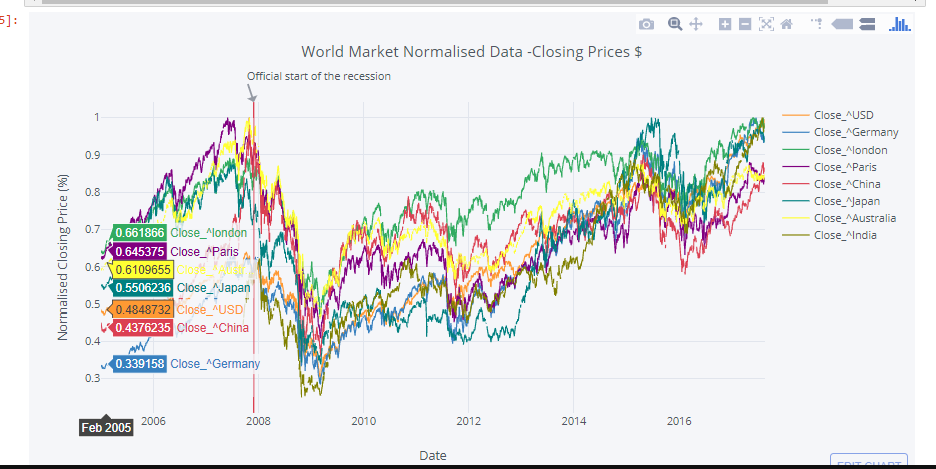
We observe that there is steep drop during the recession time frame , but DOW 30 have over the time have captured market again and better compared to France and Germany .



We clearly observe China’s Market have been hit severely during the Recession time frame compared with other markets.

WE would try to Normalize the data and then vizualise it.

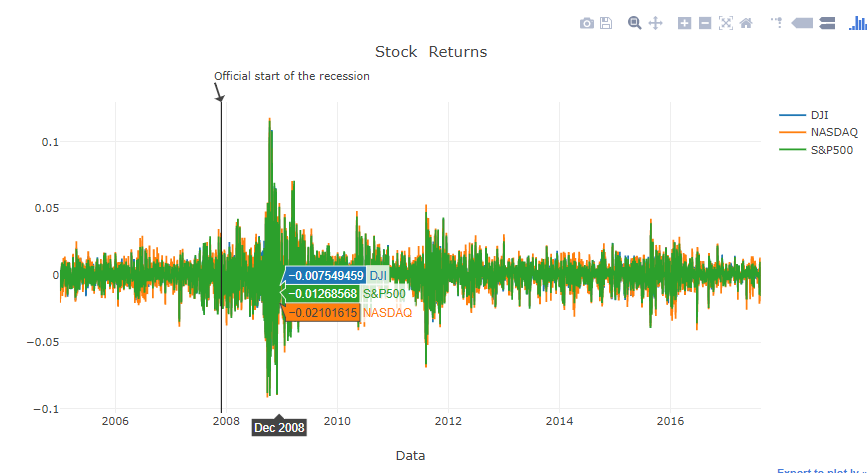




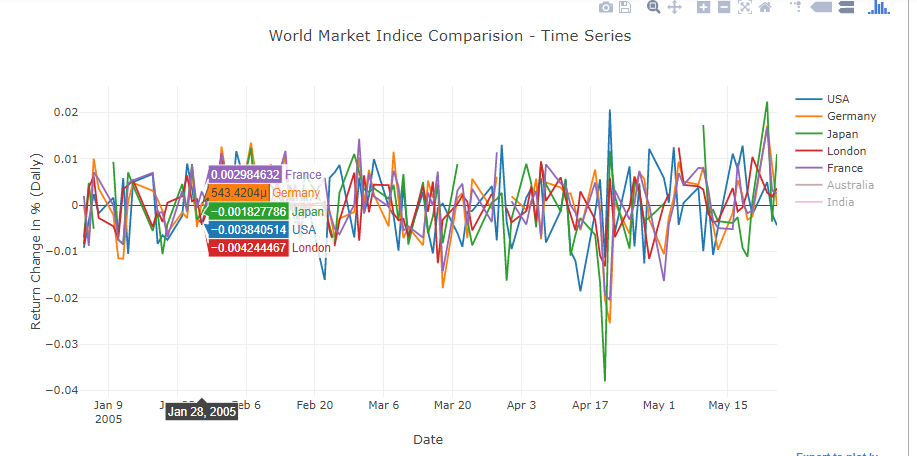
As we can see London Market has performed relatively well after the recession, whereas DOW 30 and others have been following the same trends over the time

WE would plot now the returns of DOW 30 , NASDAQ, S&P 500 .

Daily returns = ClosePrice (t)- ClosePrice(t-1) / ClosePrice(t-1) \* 100



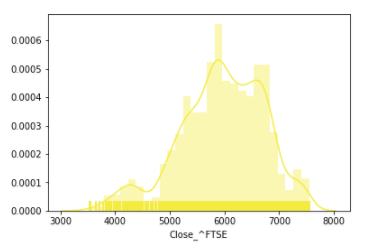
We can see that all three USA market indicator vary in same way.So we can say that these three market indicators are highly co related with each other.

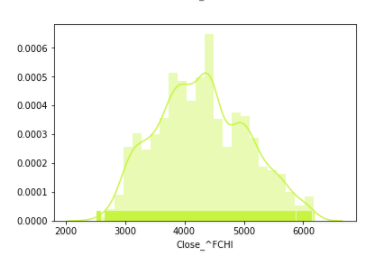


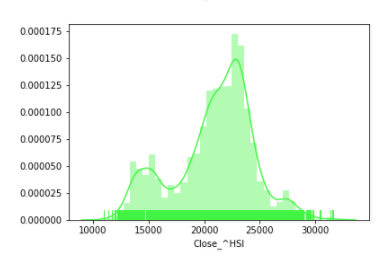
All though it is not clear how each market is being co related with this graph, but we can see few spikes which are relatively related with each other and oppositely also.

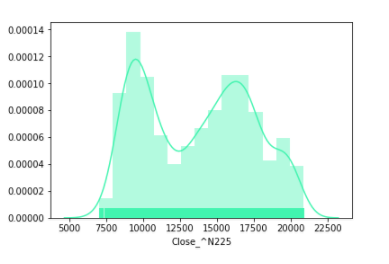
### Univariate Analysis

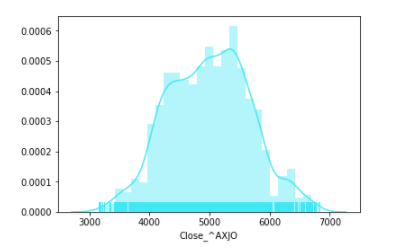
### 

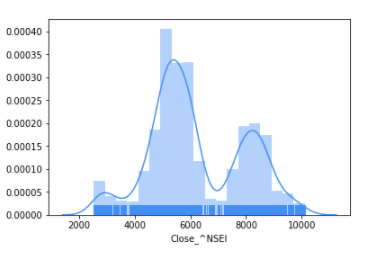


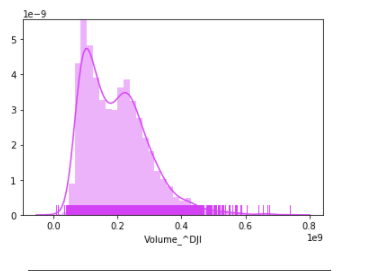


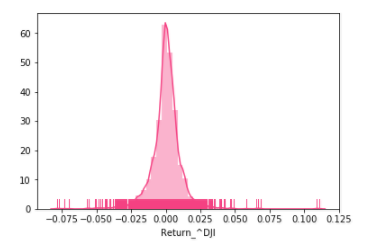




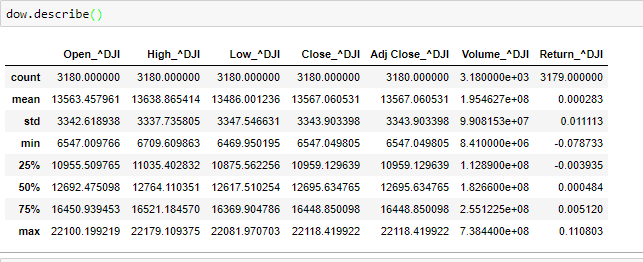




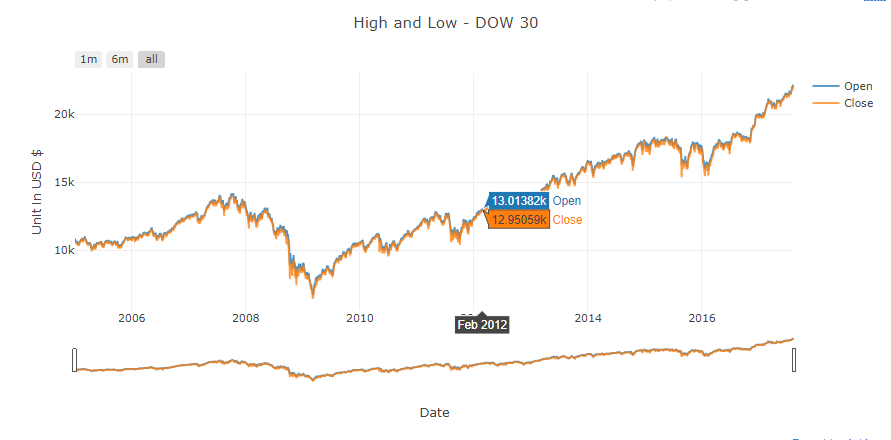




We will focus on the DOW 30 data how its variable has been changing throughout the time.



Open Price and Close Price :

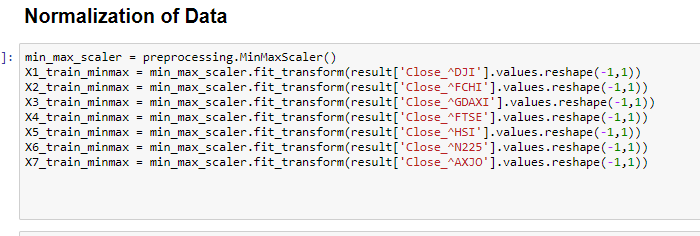


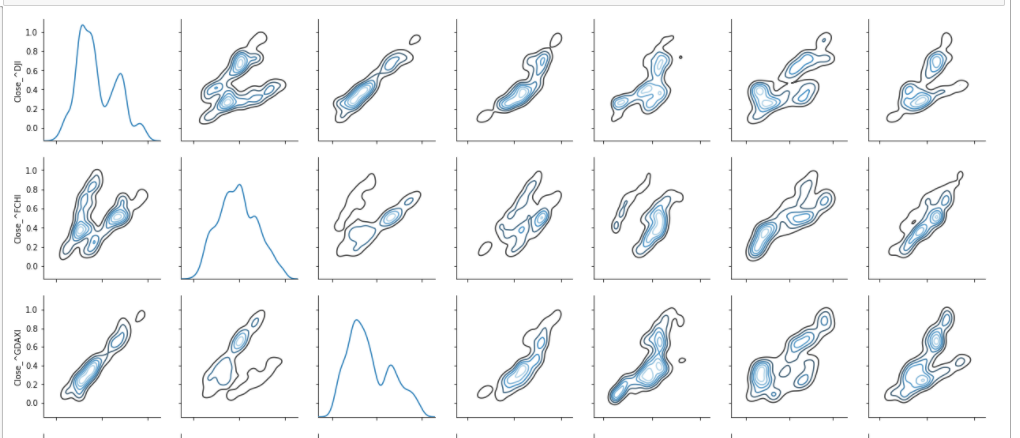
### Bivariate Analysis

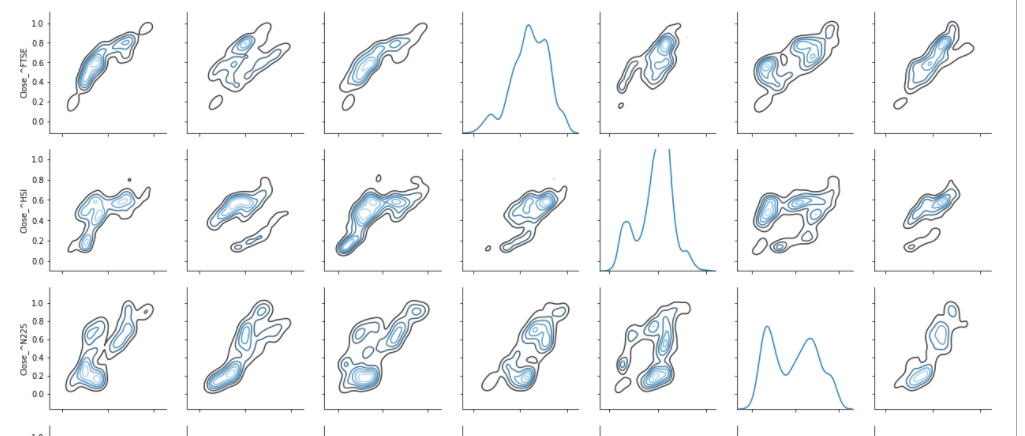
#### Data Normalization

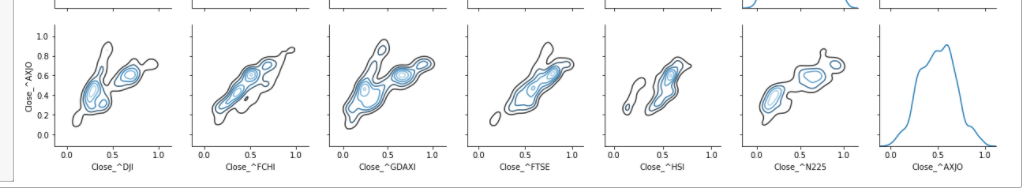
Before we do the bivariate analysis , we need to normalize the data to compare with each other

We are using min Max Scaler from the Scikit Preprocessing library.









I tmay intimidating to solve the puzzle which one is co related with one anaother , but if we closely look the below market are failry co related :

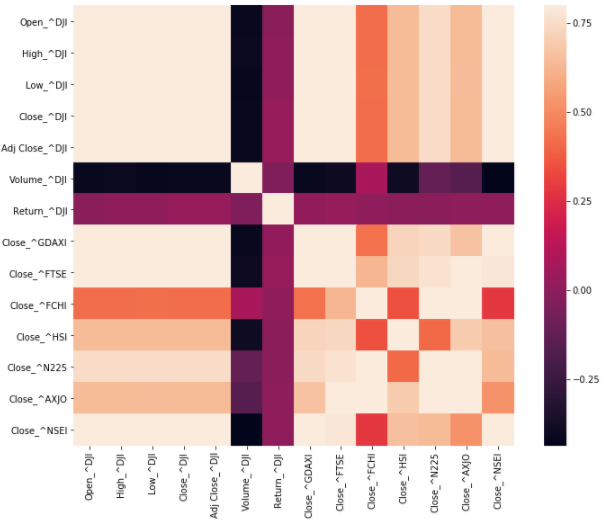
USA

France

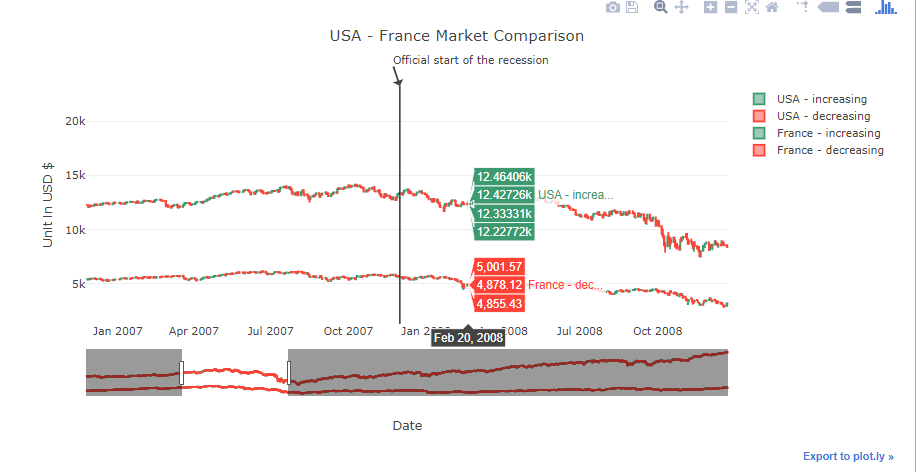
Australia

China

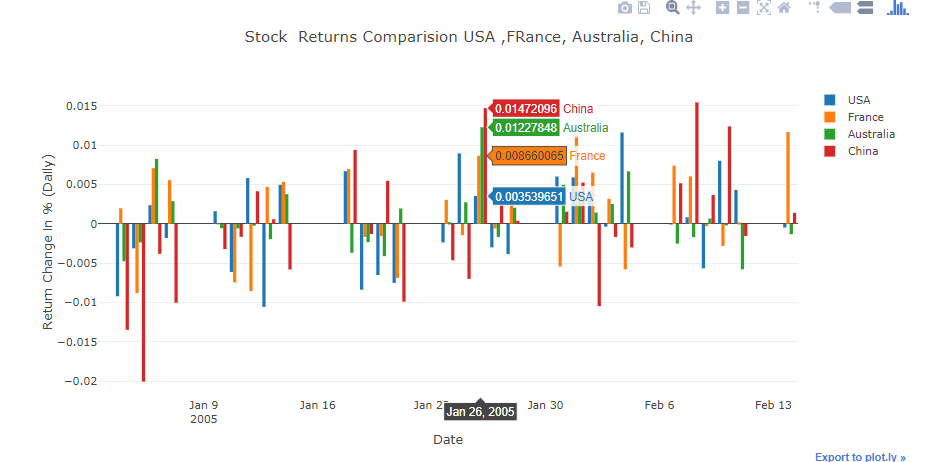
Plotting Co relation Matrix :



We have plotted most co related markets USD and France Green Shows - Increasing from previous Day and Red Shows Decreasing from Previous Day .



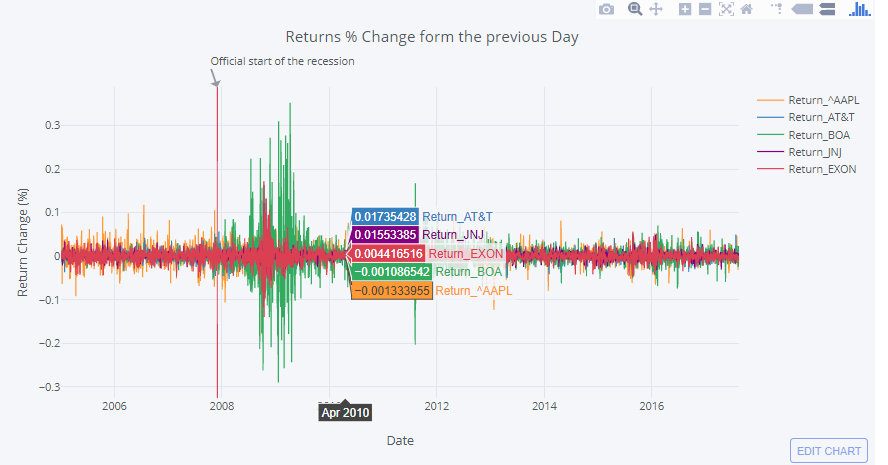
Another Graph to show how top co related Market are varying with DOW 30



## Top 5 Stocks

We will first visulise the close Price fo TOP 5 Stocks around the different Markets in USA .



WE will plot the returns:   


AS we can see there is similar change in the return among all the Stock Market companies during the Recession Time , we can relatively make a choice of trading during any time.

Auto corelation :

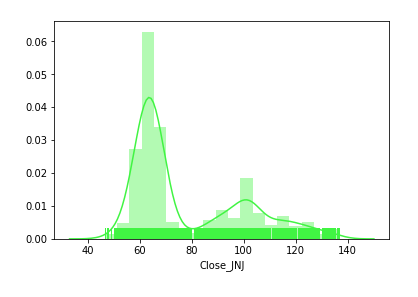


So in a lag of 1000 days , there is a strong co relation with the present index. For all the companies.

### Univariate Analysis

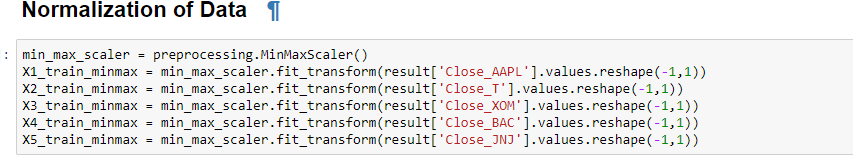
### 

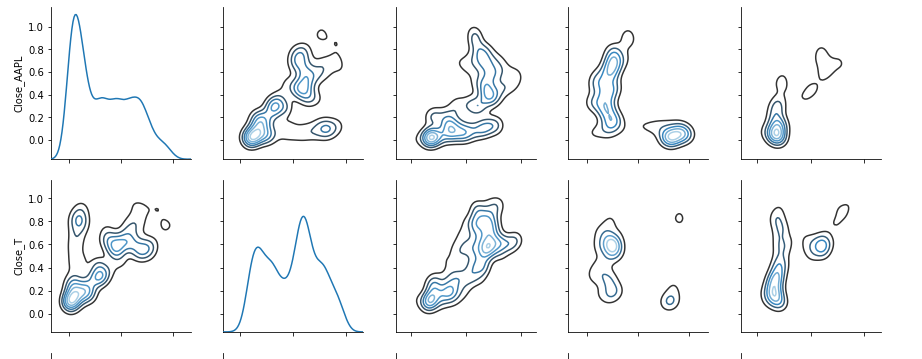


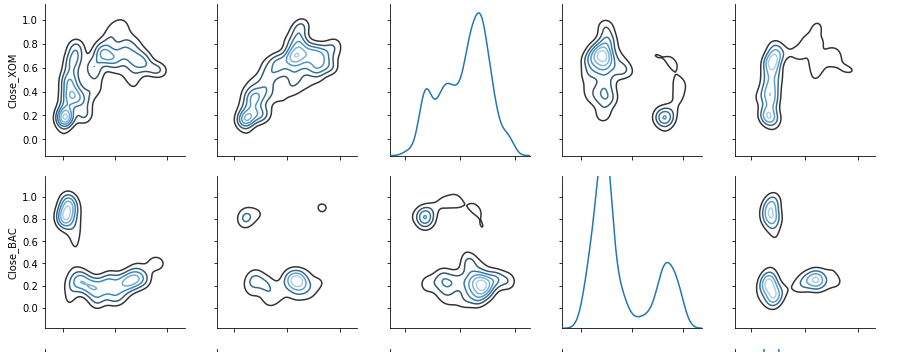


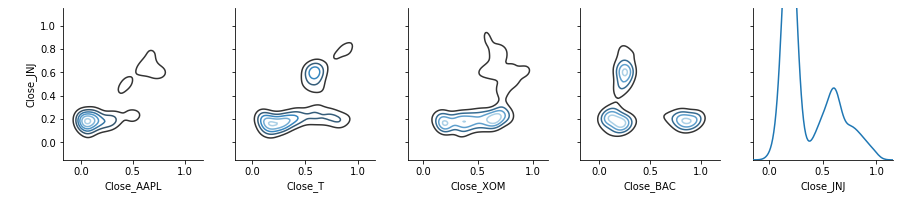
### Bivariate Analysis

#### Data Normalization











# Feature Engineering

We are adding various parameters by using TA - lib, as we tried modelling with the below parameters :

Open , Volume,Daily Returns

To predict Close Price . we were getting very less accuracy and high MAPE score.So we take into account of the following parameters :

SMA : Simple Moving Average

MOM:Momentum

RSI - Relative Strength Index

STOCHF: Stochastic Oscillator

WMA: Weighted Moving Average

T3:Triple Exponential Moving Average (T3)

CORREL :Pearson's Correlation Coefficient (r)

NATR :Normalized Average True Range

ADOSC :Chaikin A/D Oscillator

MACD: Moving Average Convergence/Divergence

WILLR : William % R - Stochastic %K. The values range from zero to 100

CLOSE PRICE : Stock Market Close Price

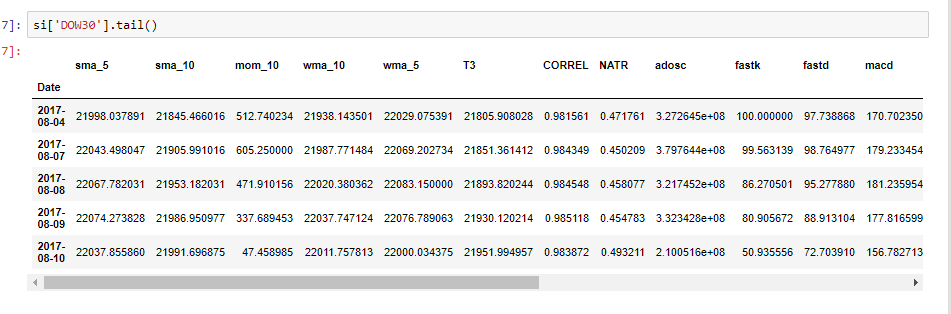
We are adding these function with the help of **TA-Lib**(Technical Analysis Library) is an open-source software library of technical analysis indicators. The library provides about 125 function for pattern recognition.

We are also labelling the percentage Change > 0 as 1 else 0.





Dataset after adding features :



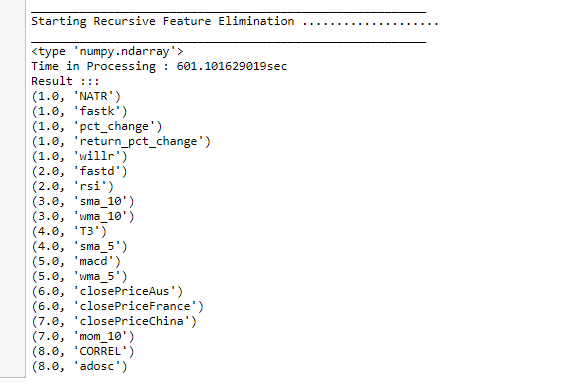
As we are not sure how these parameters added would help in determining predicting closing price , we will find out which features are suited the best.

### Recursive Feature Elimination

We are using scikit feature library’s RFE function with Logistic Regression as the estimator to give us the rank of out Features to predict Closing Price.



And we got the result :



We also tried for Recursive feature elimination with cross-validation , but it was taking more than 10 Hours , on our testing machine in Google Cloud.

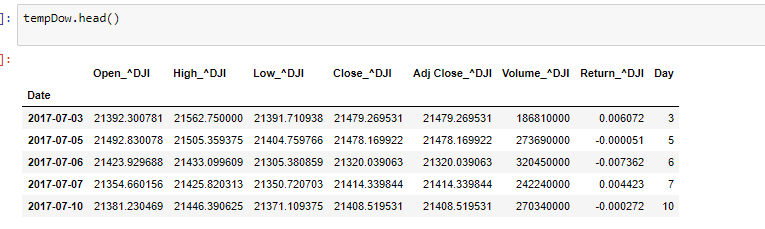
# 

# **Prediction Modelling**

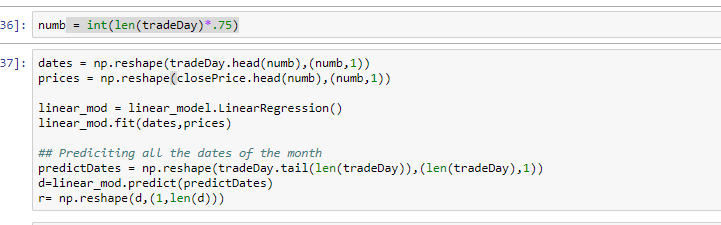
## DOW 30

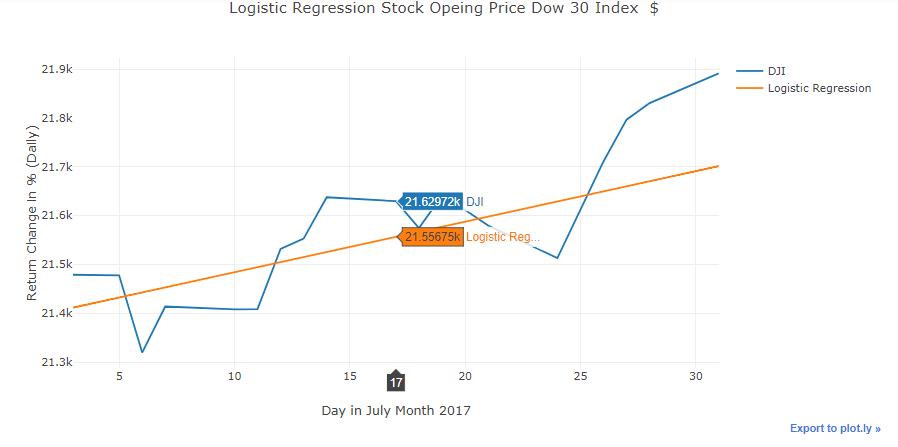
### Prediction with Logistic Regression

Sample Input given for linear regression

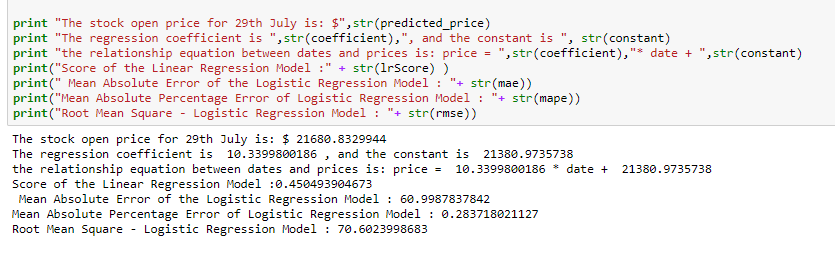


Data Splitting 75 - 25 of the Data :



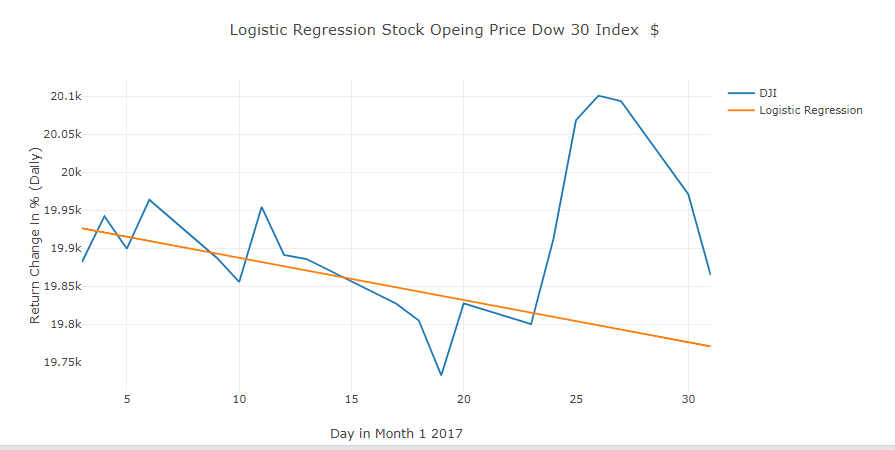


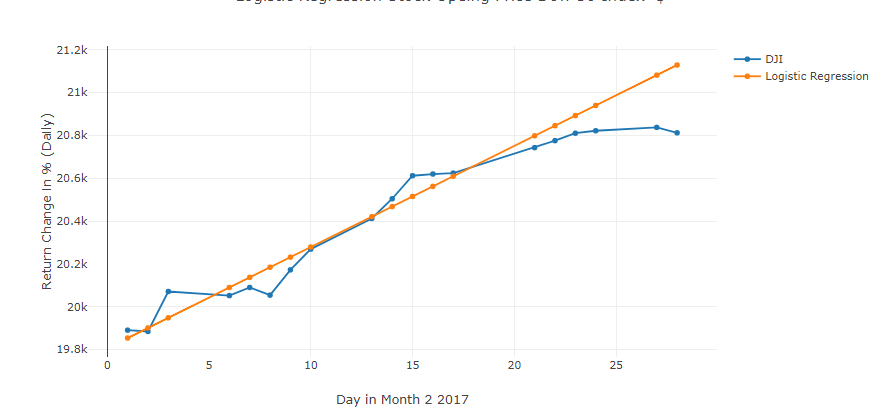
WE tried to predict for next 20 days by shifting the time period. Of th input and tallying with the current indices.

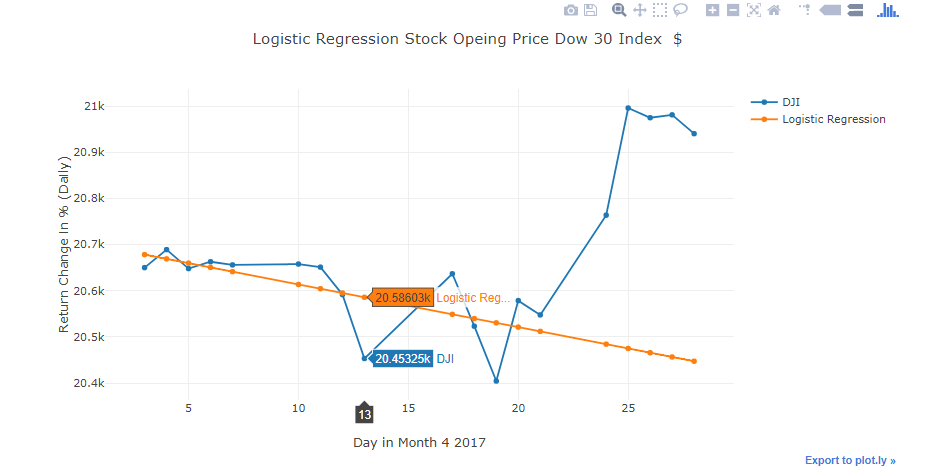


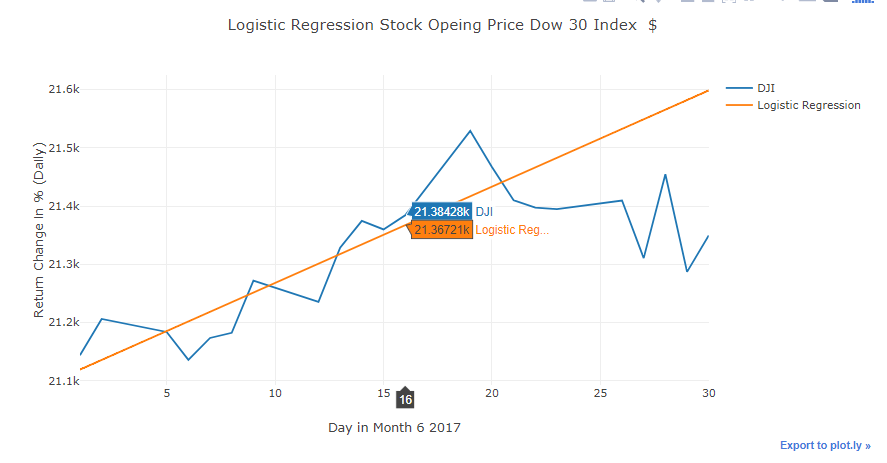
We are testin our model on monthyl basis as input is from 1- 30 and output is Closing price.

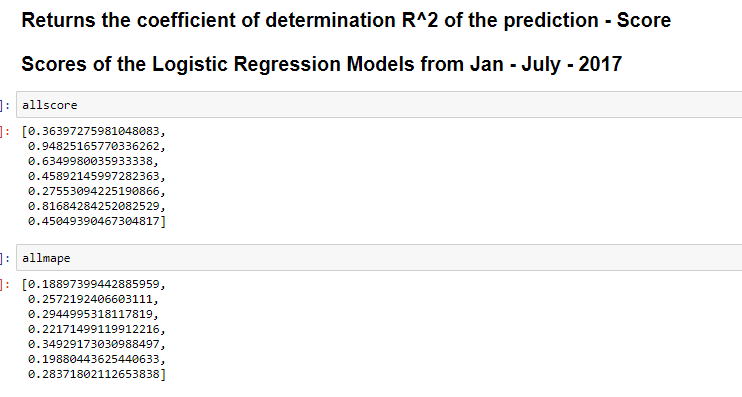
Below are screenshot for the all months in 2017:









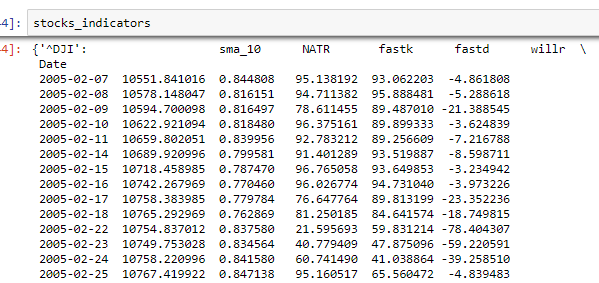


**Nutshell :**

|  |  |
| --- | --- |
|  | **Linear Regression** |
| MAPE | 60.88 |
| MAE | 28.37 |
| RMSE | 70.602 |
| Score | 0.45 |

### Prediction with Random Forest

In Random Forest we are giving input as added features.

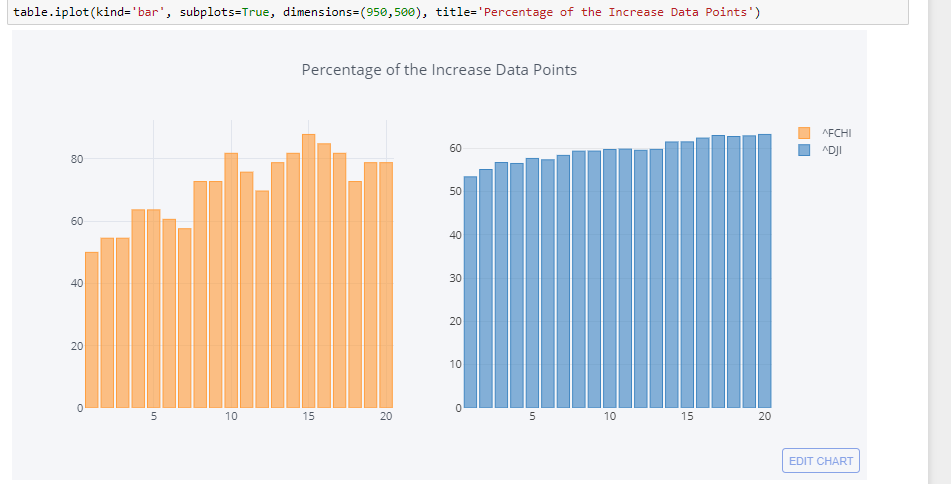


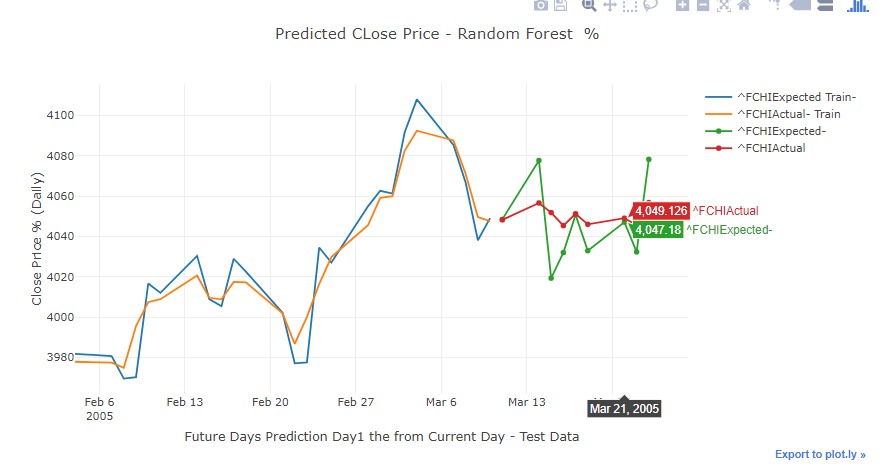
But before proceeding to the data Modelling , we will check check if our DAta is balanced or not .This will be done on the basis of how many number of Percentage Change Label of 0 and 1 varies in our data.

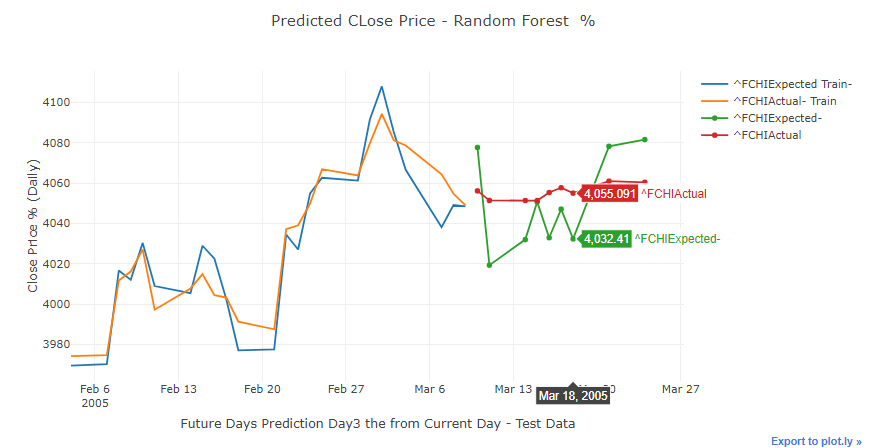
Data Balancing :

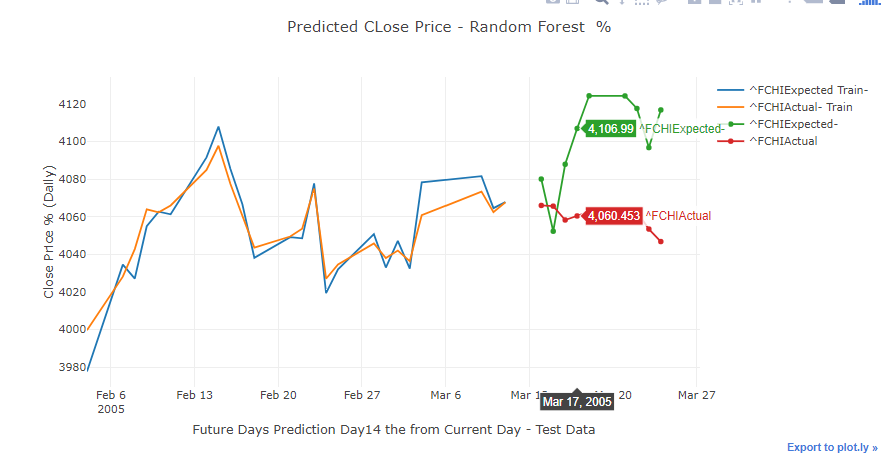
Following Code Explains how we have checked if the data is balanced or not

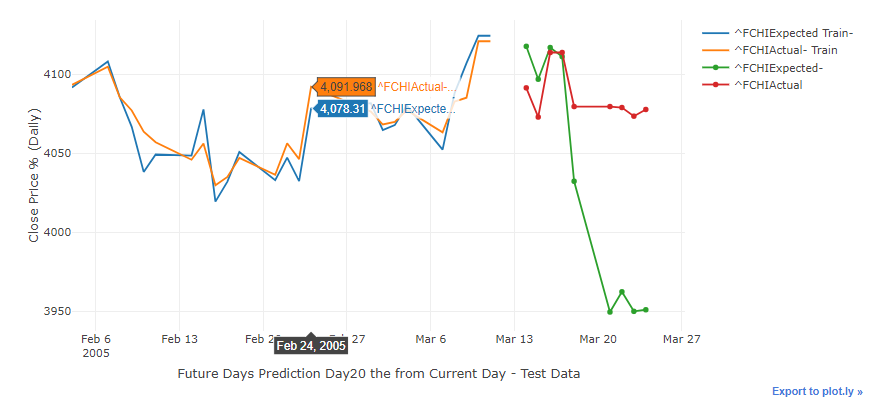


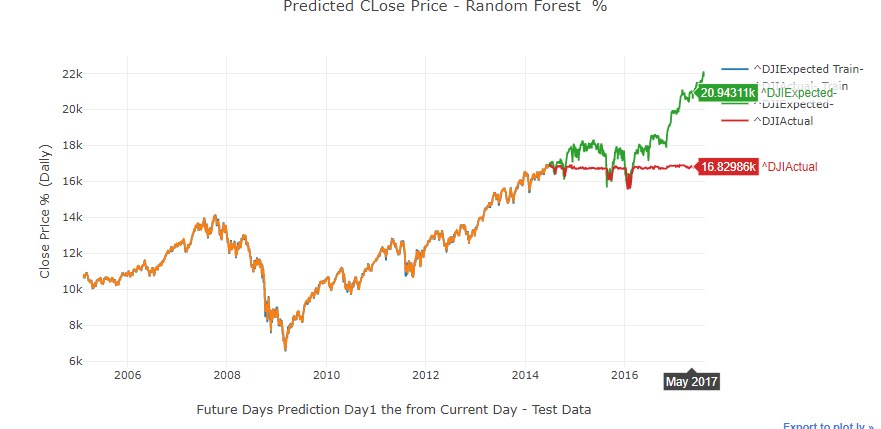


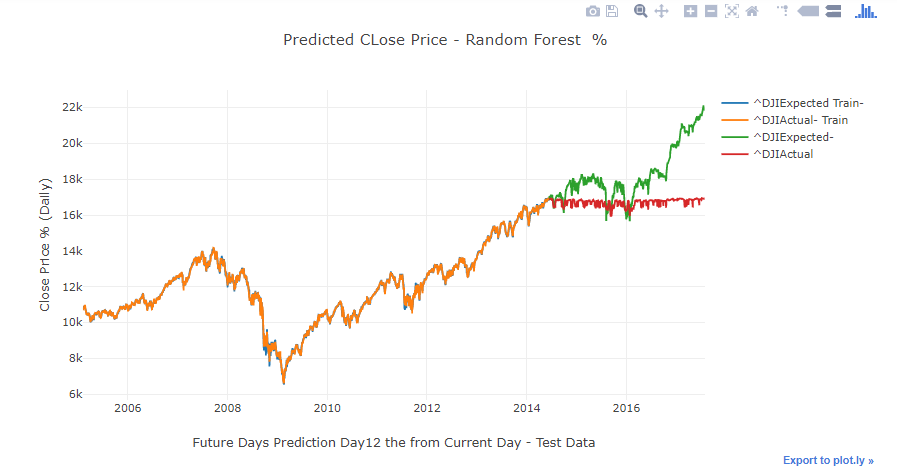


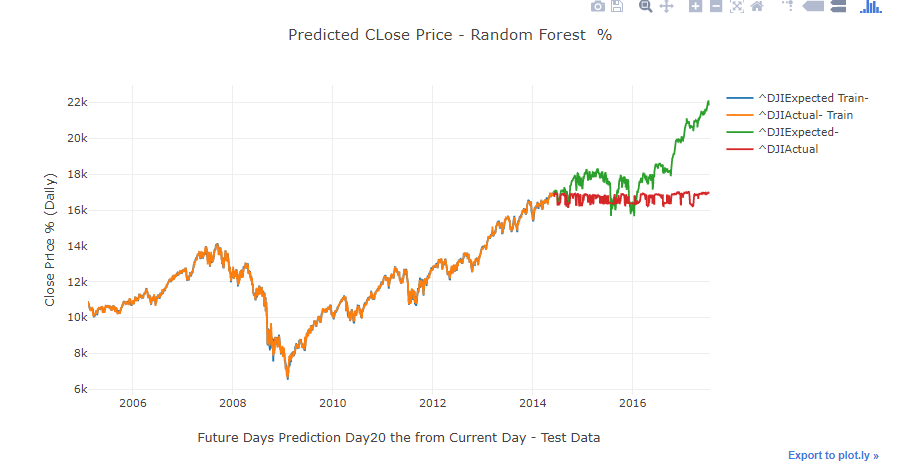


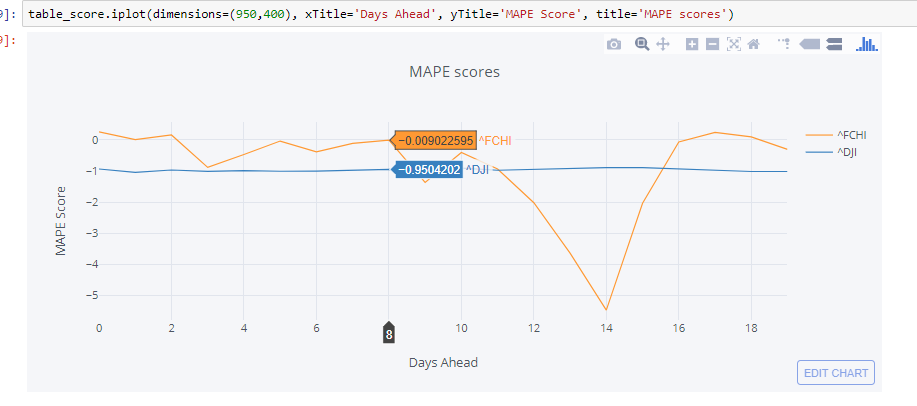


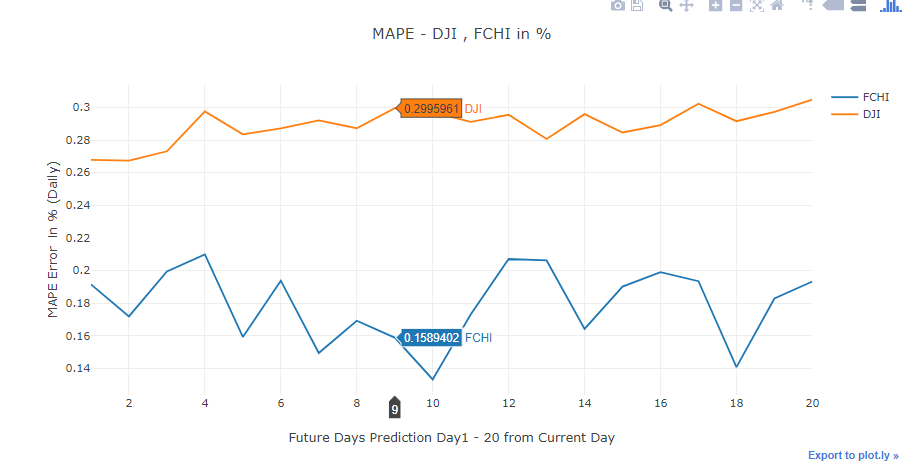


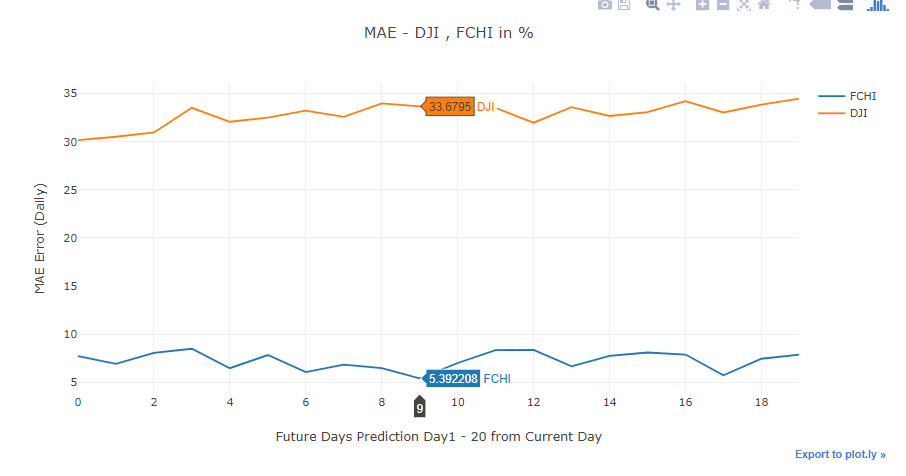


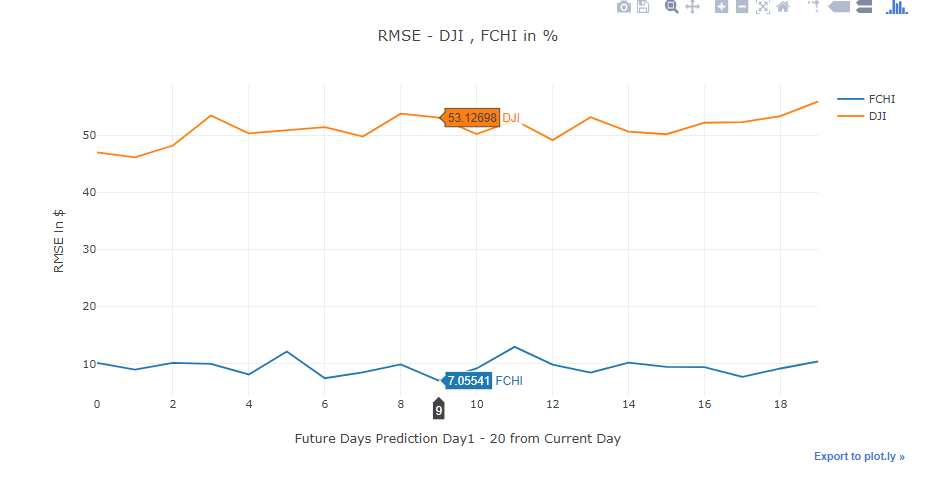












Nutshell:

|  |  |
| --- | --- |
|  | Random Forest |
| MAPE | 26.74 |
| MAE | 30.96 |
| RMSE | 48.51 |

### Prediction with Neural Network

First of all we Normalised our data to make from Four Column Data to Three Column data so that we can group it in K Means Clustering.

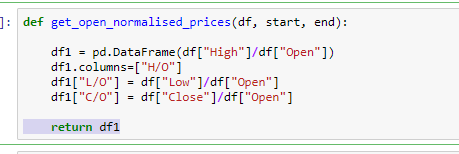
We have taken High, Low, and Close Columns normalized with respect to Open Column.

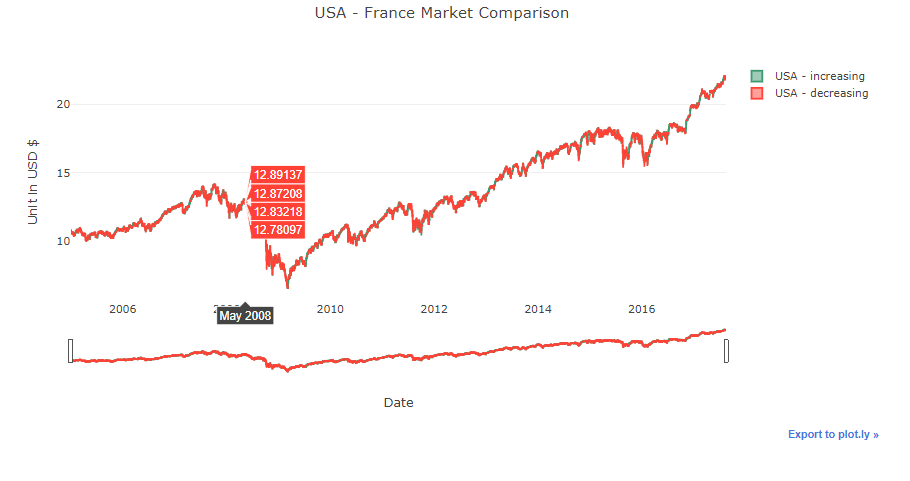
After this step WE choose what would best number K to choose from the Bend Graph.   
As a result we get K = 5 as the best ,

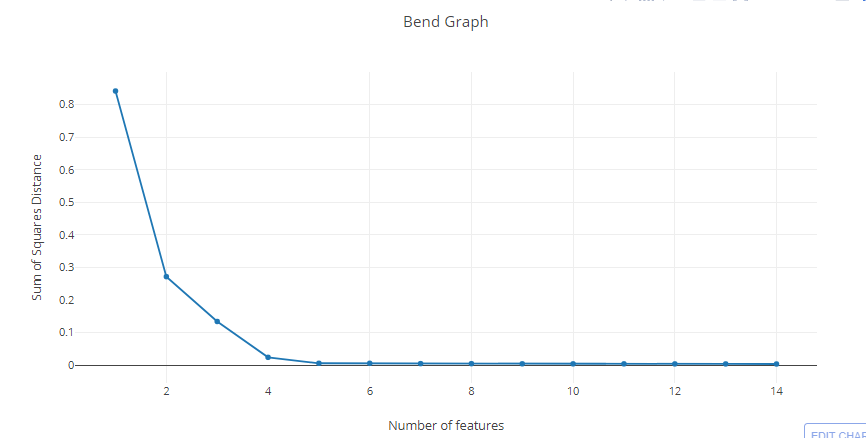
After this step we have clustered our data into 5 groups.

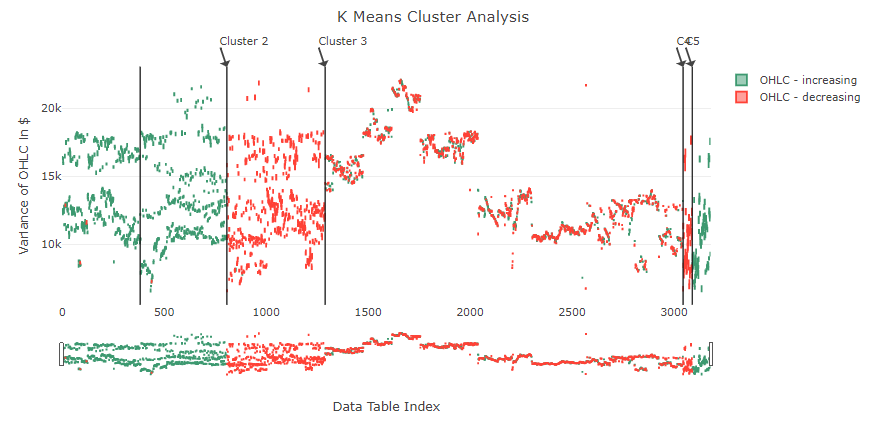
These Groups are of similar Market Fluctuation OHLC Would in a same variance level among each other. AS a next we will predict the current available OHLC to fall in which Cluster by K means Library (Scikit ) .   
We need to find out what could be the next cluster follows with maximum probability.   
To find out that we need to calculate the probability of next cluster to follow , we need to count all the pair and find out which has the highest count among others.

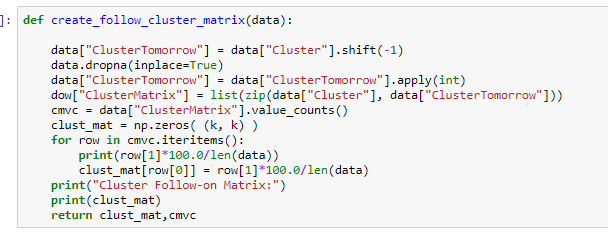
The next step would be modelling the Neural Network, where we have Set up 5 - 11 -11 -1 as the network with 200 epochs and activation function as ReLu. (Rectinear Unit)

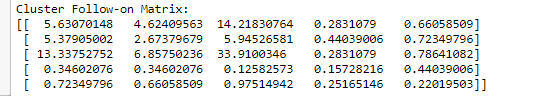




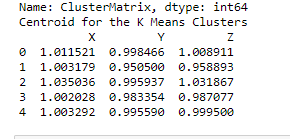


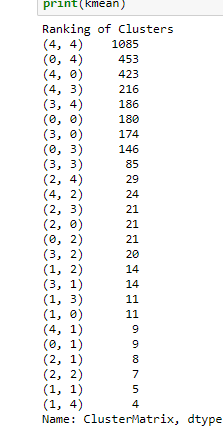


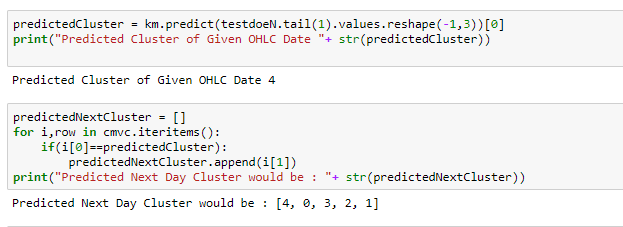


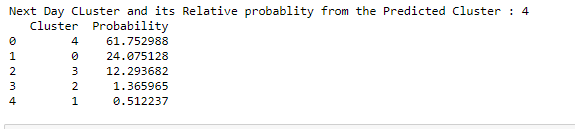


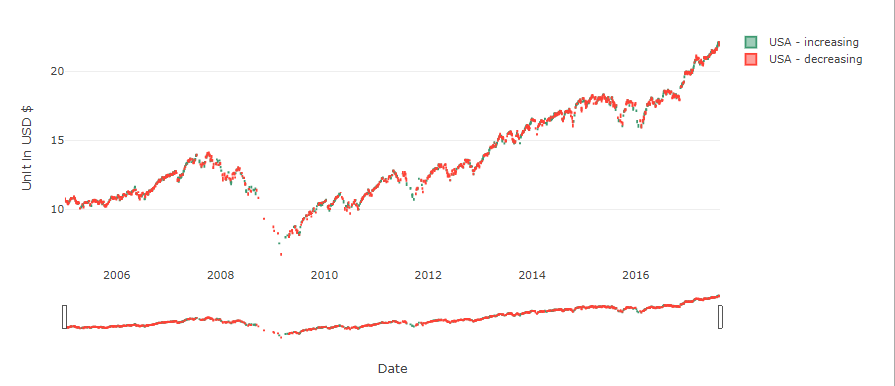


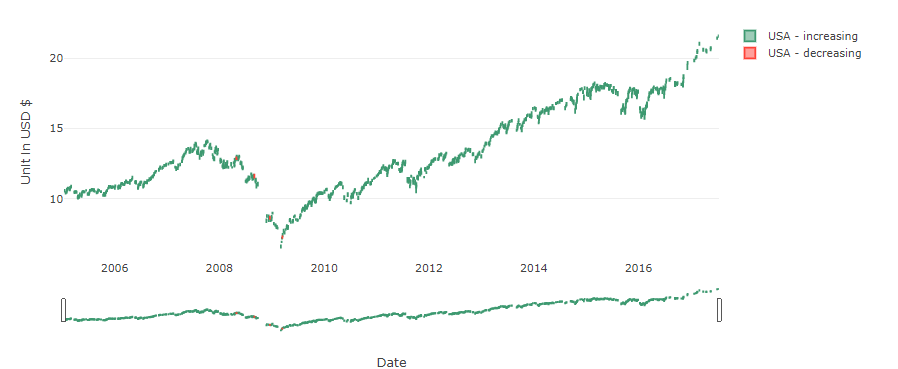


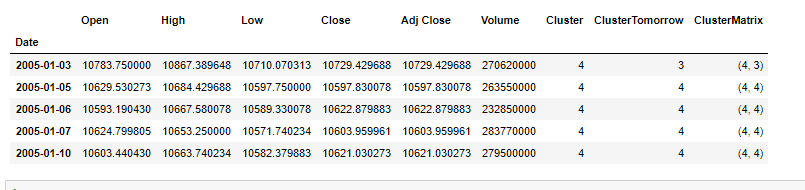


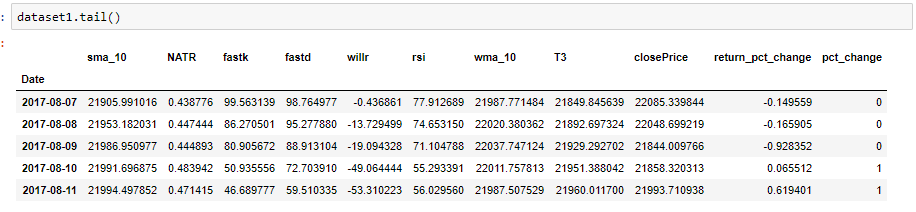


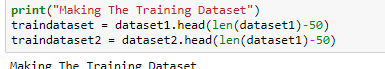


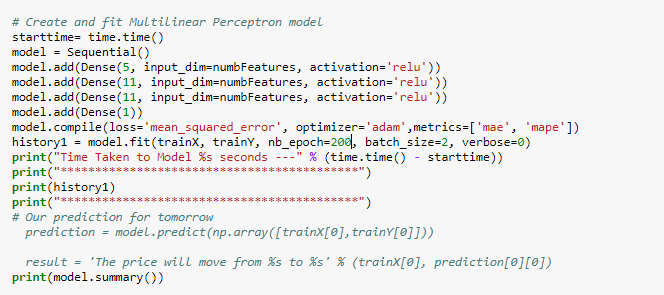


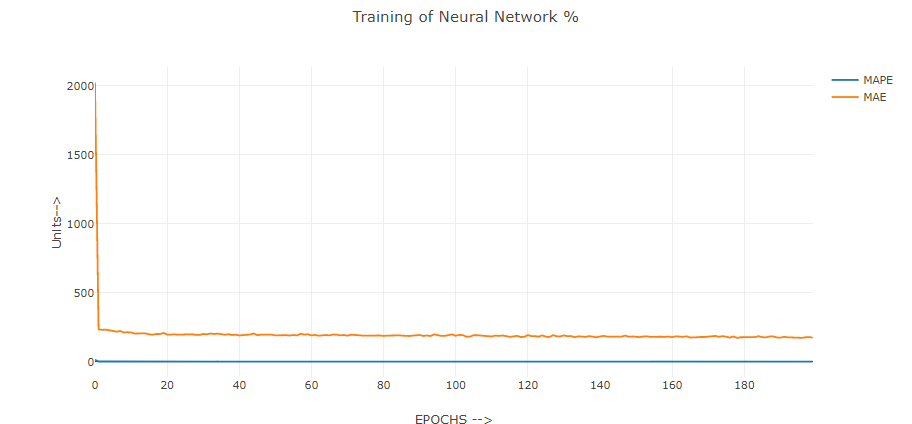




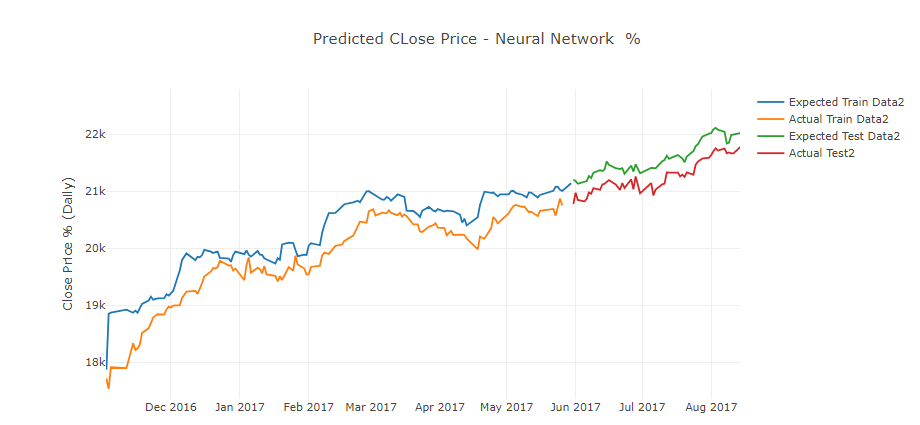














|  |  |
| --- | --- |
|  | Neural Network |
| Loss | 69678.38 |
| MAPE | 1.32 |
| MAE | 174.36 |

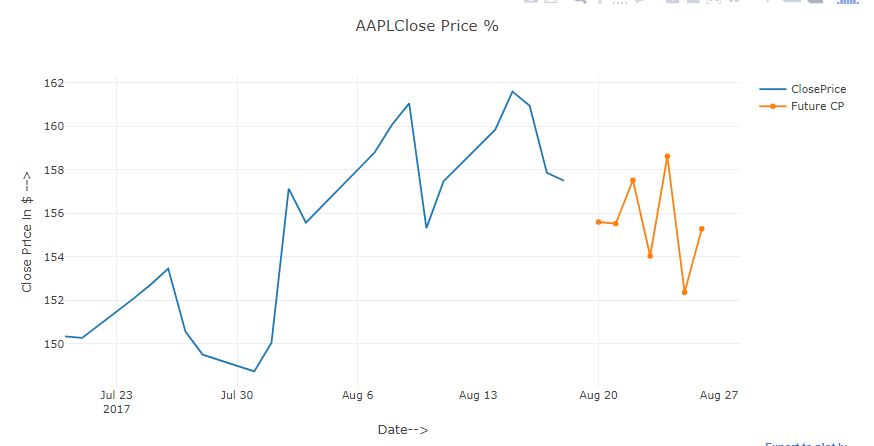
### Top 5 Stocks

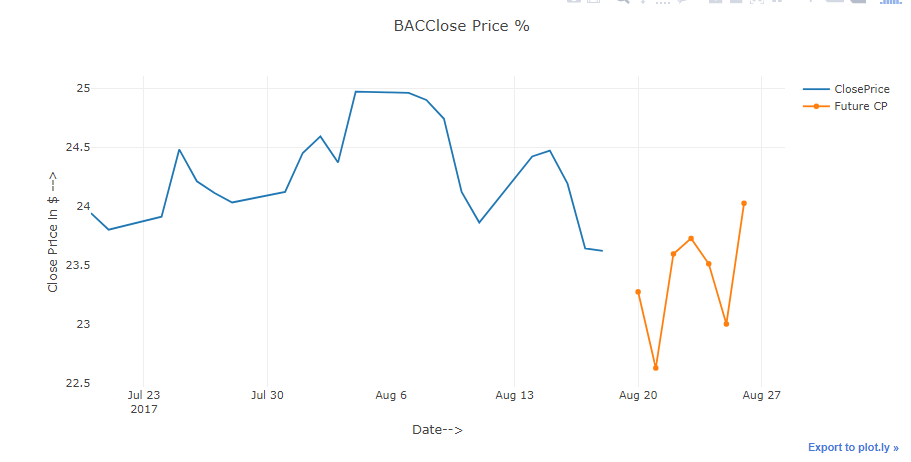
Same procedure have been followed as like Neural Network,

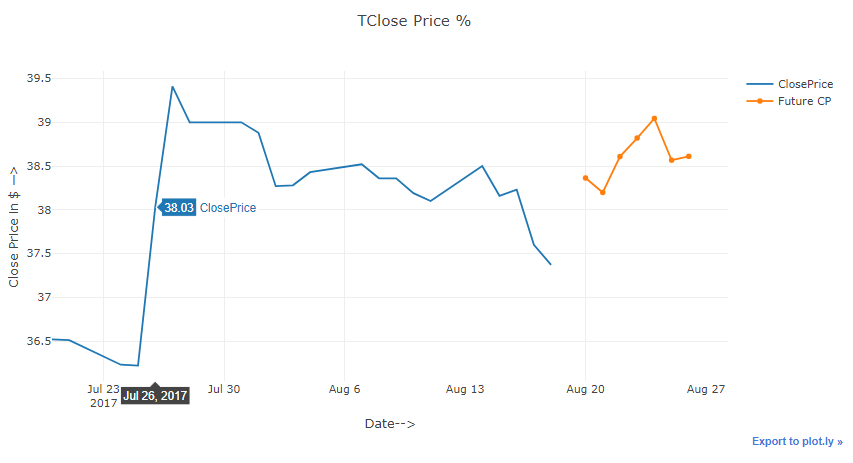
WE go t the below result for Day 1 Prediction :

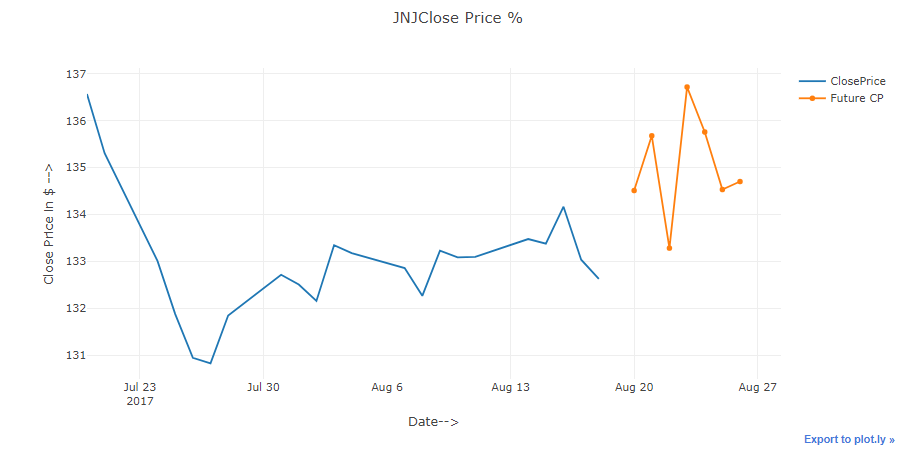
|  |  |  |  |
| --- | --- | --- | --- |
|  | loss | MAE | MAPE |
| AAPL | 4.588755445 | 1.508158267 | 3.050526 |
| XOM | 3.157263338 | 1.292062547 | 1.636211 |
| T | 0.549901262 | 0.530865484 | 1.624365 |
| JNJ | 2.001679601 | 1.013141569 | 1.301066 |
| BAC | 0.615068567 | 0.508888402 | 2.467182 |

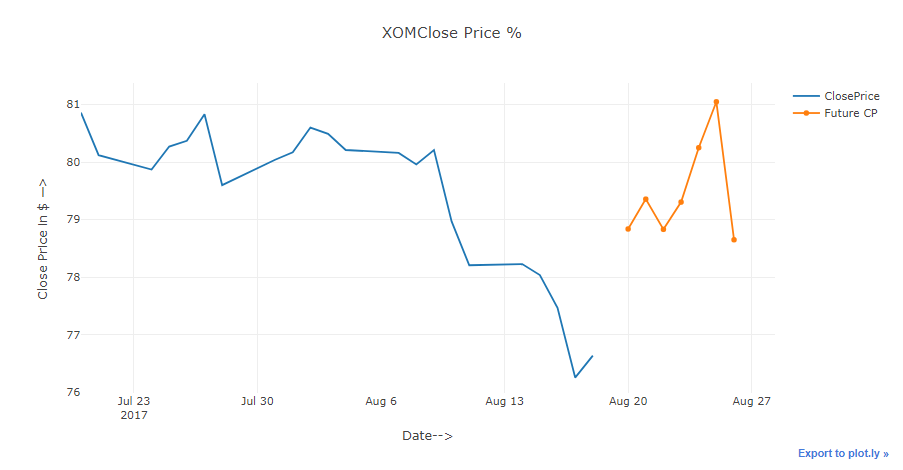
Below Graphs we got current forecast for next 7 days :





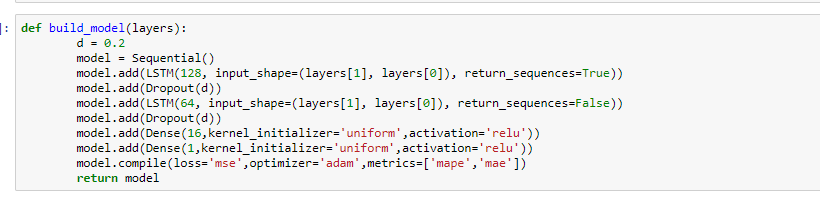






**LSTM**

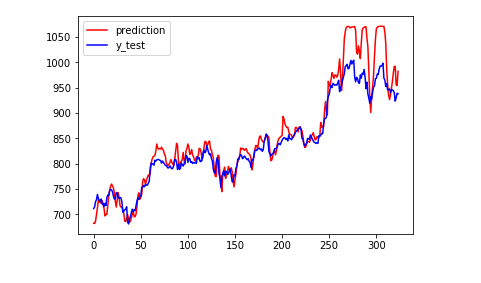
We have made try on the LSTM Model to find out how it is performing in the prediction,but we were running into overfitting , or sometimes not all fitting . We have used 128 - 64 - 16 - 1 layers of Neural Network , where first are of LSTM Layer , which can remember the sequence



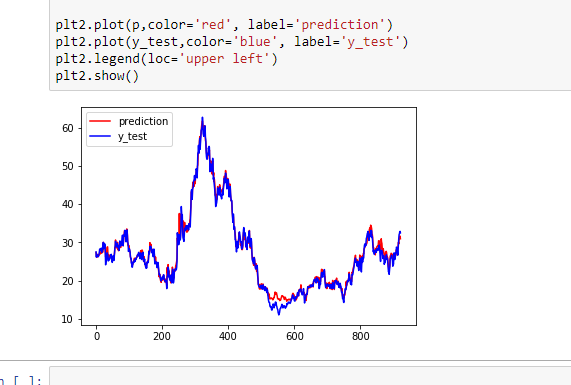
May be we need to train or make more sophisticated network to get a reliable model.

Below were our try result in graphs :

1. Google Stock



2. AAPL Stock

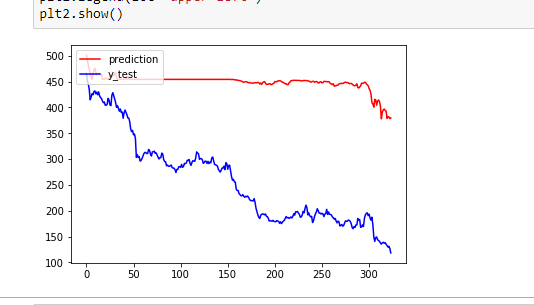


Scores for the APPLE

**MAPE MAE**

**6.73 6.94891**

3.**DOW 30 - ^DJI**



# Choosing Best Model :

As we have made the a rigorous modelling on the neural network trying to predict till 20 days ,Neural Network is proving to be the best among all other algorithm. Although we believe we can make the model more accurate by accounting many other stochastic variable in the market.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Loss** | **MAE** | **MAPE** |
| **AAPL** | **4.588755445** | **1.508158267** | **3.050526** |
| **XOM** | **3.157263338** | **1.292062547** | **1.636211** |
| **T** | **0.549901262** | **0.530865484** | **1.624365** |
| **JNJ** | **2.001679601** | **1.013141569** | **1.301066** |
| **BAC** | **0.615068567** | **0.508888402** | **2.467182** |
| **DJI** | **69678.38** | **1.32** | **174.36** |

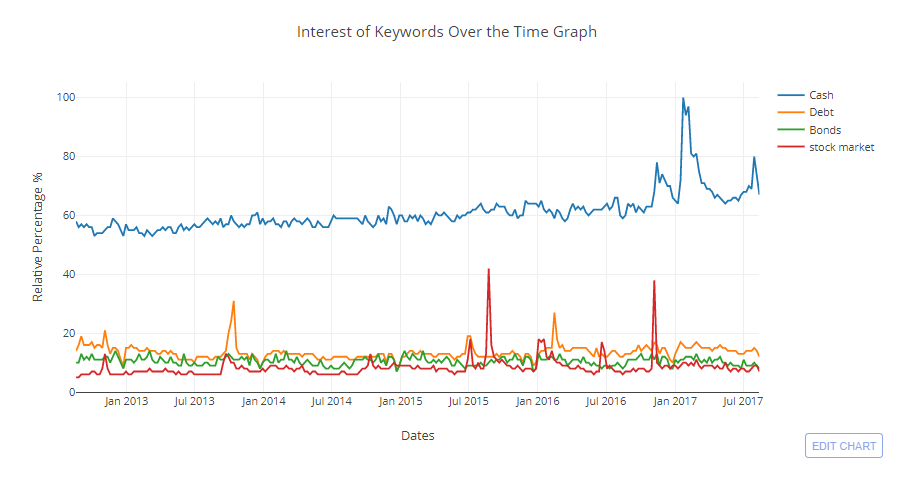
# Twitter Sentiment Analysis

We have essentially ignored the Market Volatility factor in our modelling process , to help our customer we are providing the Twitter Data related with the selected Stocks , which they can easily get an idea how the people are reacting about the stocks. This can be a really good analysis on how the people are positive about the company.

**Google Trends :**

We also did analysis how the below words vary in the google trends , it summarises the search result relatively with each other . Following words we have plotted over the time :

Cash Debt Bonds Stock Market

Unemployment Hedge Revenue Housing 

# 

AS we can see there is few correlation of search activity like Revenue and Unemployment, they have similar rate fashion of spike.This shows us People are highly active regarding this period where Market may fluctuate heavily.

# 

# 

# Test Cases

Users can login in our website and can have the below test scenarios :

1. Get the Top 5 Stocks ‘s Predic tion for the Next 7 days with MAPE 1.3 %

Apple

Exxon Mobil

Jhonson & Jhonson

Bank of America

At&T

2. Get the DOW 30 Market Indicators Prediction for next 7 days with MAPE 1.3%

^DJI

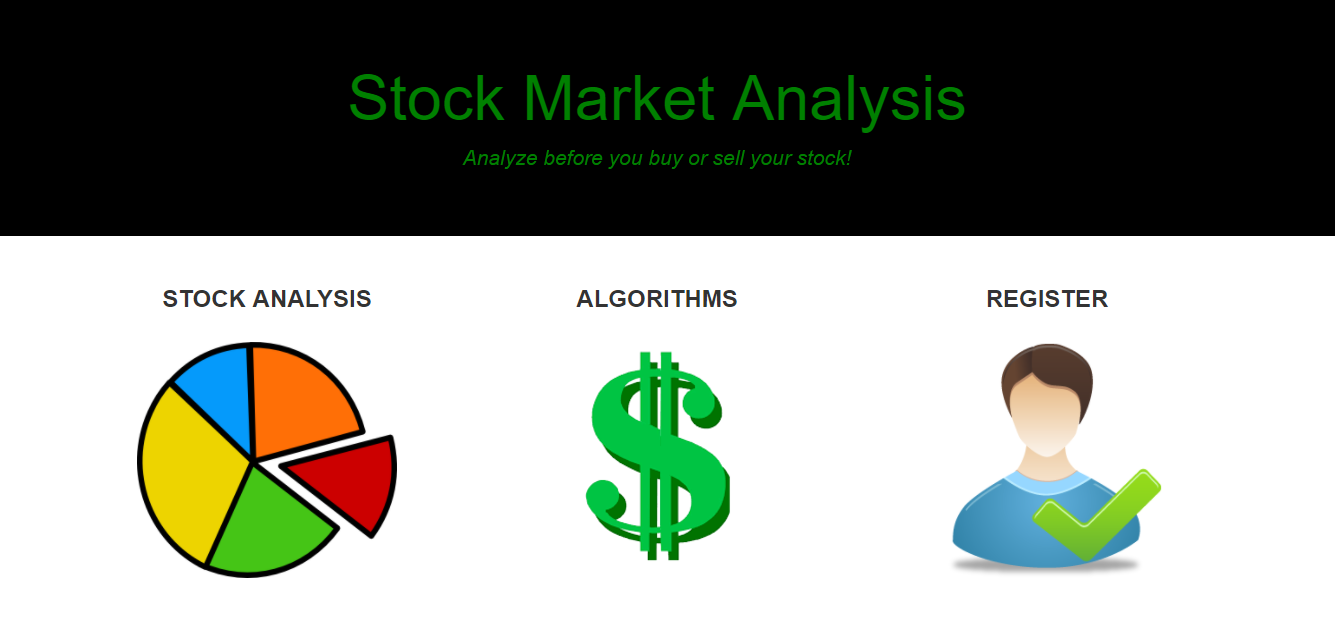
3. Get Comparison with the other World Market Indicator

4. Get Twitter Data Sentiment Analysis ,about what people are saying about all Top5 selected Stocks , which will help customer to judge market volatility and trend.

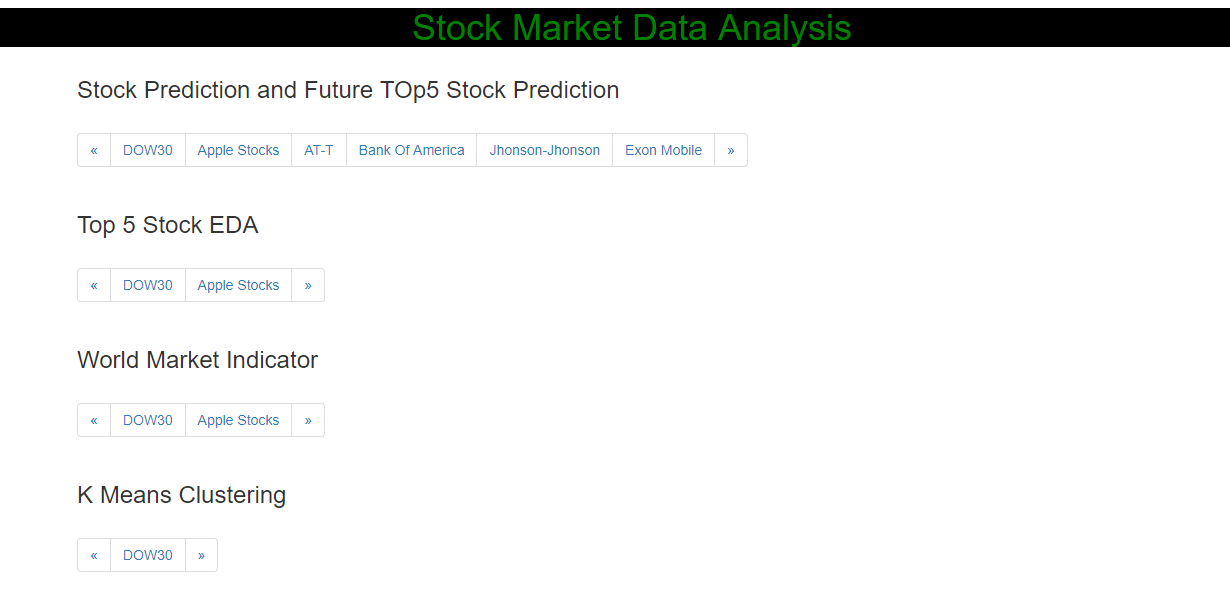
5. Users can Register themselves to get updated Forecast every day for the next Seven days.

# User Interface Application

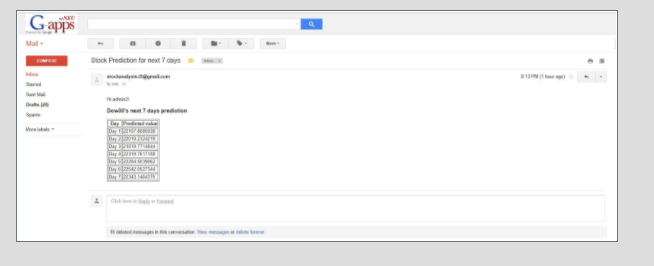
Following are the Screenshots of WEb Application Hosted on Amazon EC2 instance.



Stock Analysis Option to User :



### Mailing Functionality :

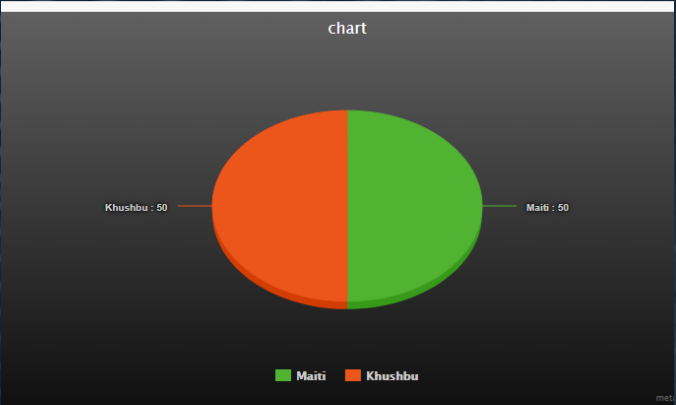


# Technology Stack

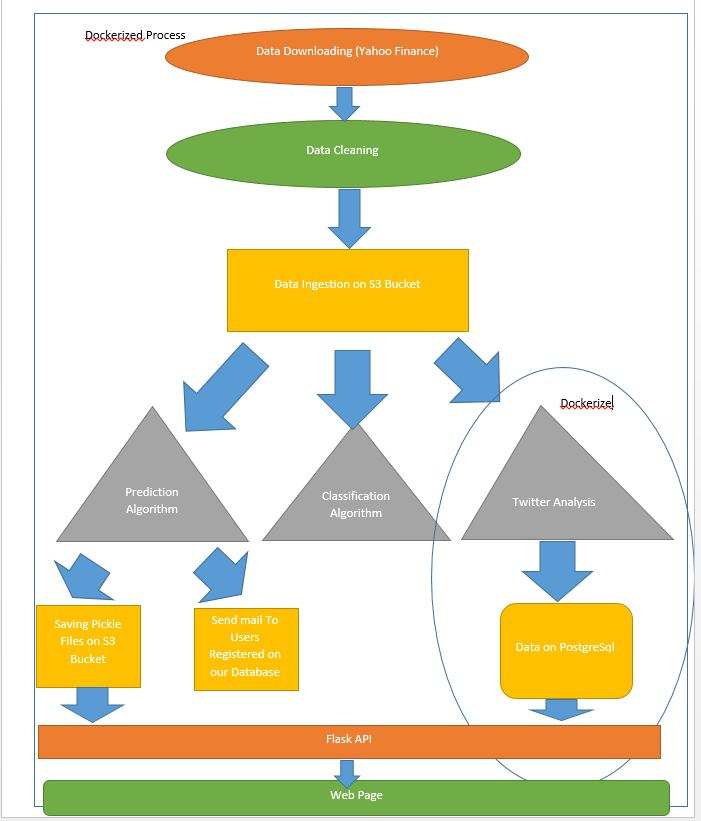
1. Amazon S3 Bucket
2. Amazon EC2 Instance
3. AmazonRDS-PostgreSql
4. Docker
5. Flask
6. Python
7. Jinja
8. Pickle File
9. Google Mailing Services
10. Google Cloud
11. Make

We are saving our models ias json and H5 py format . The h5py package is a Pythonic interface to the [HDF5](http://hdfgroup.org/) binary data format.

# Contribution

****

# Architecture Diagram:



# Future Scope:

We can more interactively engage our customers by mobile application and giving the on -live analysis , where our models have to be more efficient and accurate , which brings us to make a rigorous testing on the models , we can improve our features by delving into more sophisticated financial security rules like Monte Carlo and many other stochastic models. WE can expand our Stock Market indicator Prediction in future.Improving the model of LSTM can heavily impact on our models , and on top on that we can test few other interesting Machine learning Algorithm like Genetic Algorithm to account the randomness in the Stock Market.

# Appendix :

## Added Feature:

**SMA :** Simple Moving Average

**MOM:**Momentum

The Momentum is a measurement of the acceleration and deceleration of prices. It indicates if prices are increasing at an increasing rate or decreasing at a decreasing rate. The Momentum function can be applied to the price, or to any other data series.

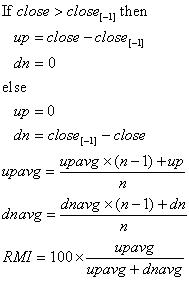
Formula:



**RSI** - Relative Strength Index

The Relative Strength Index (RSI) calculates a ratio of the recent upward price movements to the absolute price movement. The RSI ranges from 0 to 100. The RSI is interpreted as an overbought/oversold indicator when the value is over 70/below 30. You can also look for divergence with price. If the price is making new highs/lows, and the RSI is not, it indicates a reversal.

Formula:



**STOCHF:** Stochastic Oscillator

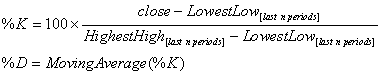
The Stochastic Oscillator measures where the close is in relation to the recent trading range. The values range from zero to 100. %D values over 75 indicate an overbought condition; values under 25 indicate an oversold condition. When the Fast %D crosses above the Slow %D, it is a buy signal; when it crosses below, it is a sell signal. The Raw %K is generally considered too erratic to use for crossover signals.

The Stochastic Indicator was developed by George C. Lane.

Terminology:

|  |  |
| --- | --- |
| Fast Stochastic | Refers to both %K and %D where %K is un-smoothed |
| Slow Stochastic | Refers to both %K and %D where %K is smoothed |
| Raw %K | Un-smoothed %K |
| Fast %K | Un-smoothed %K |
| Slow %K | Smoothed %K |
| Fast %D | Moving average of an un-smoothed %K |
| Slow %D | Moving average of a smoothed %K, in effect: a double smoothed %K |
| %D | Always refers to a smoothed %K (whether or not the %K itself is smoothed) |

Formula:



**WMA:**

**T3:**Triple Exponential Moving Average (T3)

**CORREL**

**NATR** Normalized Average True Range

**ADOSC** Chaikin A/D Oscillator

**MACD :Moving Average Convergence/Divergence**

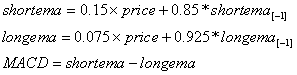
The Moving Average Convergence Divergence (MACD) is the difference between two Exponential Moving Averages. The Signal line is an Exponential Moving Average of the MACD.

The MACD signals trend changes and indicates the start of new trend direction. High values indicate overbought conditions, low values indicate oversold conditions. Divergence with the price indicates an end to the current trend, especially if the MACD is at extreme high or low values. When the MACD line crosses above the signal line a buy signal is generated. When the MACD crosses below the signal line a sell signal is generated. To confirm the signal, the MACD should be above zero for a buy, and below zero for a sell.

The time periods for the MACD are often given as 26 and 12. However the function actually uses exponential constants of 0.075 and 0.15, which are closer to 25.6667 and 12.3333 periods. To create a similar indicator with time periods other than those built into the MACD, use the Price Oscillator function.

The MACD was developed by Gerald Appel.

Formula:



**STOCHF**

WILLR The Williams %R is similar to an unsmoothed Stochastic %K. The values range from zero to 100, and are charted on an inverted scale, that is, with zero at the top and 100 at the bottom. Values below 20 indicate an overbought condition and a sell signal is generated when it crosses the 20 line. Values over 80 indicate an oversold condition and a buy signal is generated when it crosses the 80 line.

The %R indicator was developed by Larry Williams.

Formula:



CLOSEPRICE

RETURN PCT CHANGE

Reference:

https://www.nature.com/articles/srep01684/figures/3