**STOCK PRICE PREDICTION**

**Team Members:**

Aditya Pawar ([pawar.ad@husky.neu.edu](mailto:pawar.ad@husky.neu.edu)) Dhairya Jaiswal ([jaiswal.d@husky.neu.edu](mailto:jaiswal.d@husky.neu.edu)) Ranga Chari ([vinjamuri.r@husky.neu.edu](mailto:vinjamuri.r@husky.neu.edu))

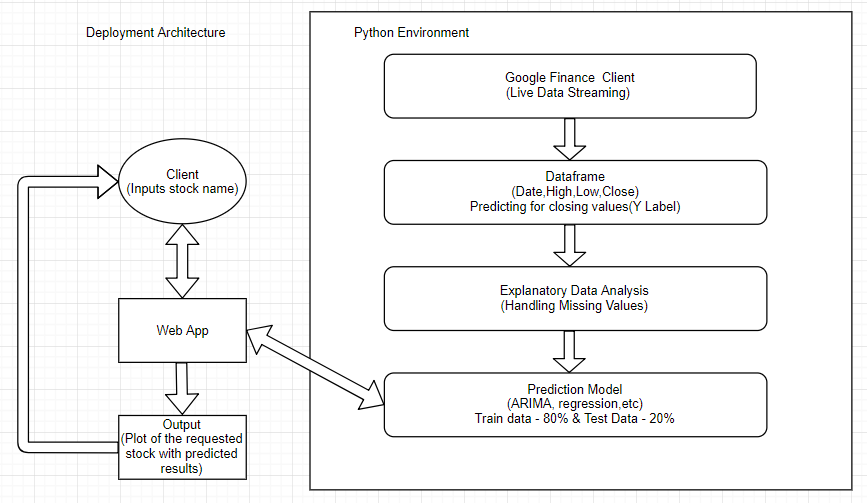
**Introduction:**

Predicting the stock price trend by interpreting the seemly chaotic market data has always been an attractive topic to both investors and researchers. Among those popular methods that have been employed, Machine Learning techniques are very popular due to the capacity of identifying stock trend from massive amounts of data that capture the underlying stock price dynamics.

**OBJECTIVE:**

The main objective of the dissertation is to examine the important and potential factors/predictors that could impact the stock market and develop a system for the investors and researchers in stock prediction.

**Work Flow:**

****

**Dataset:**

### Here are the highlights and word definitions of the column names so as any reader can take as reference to understand how to read stocks:

### 1. Open: The price at the beginning of the trading day

### 2. High: The highest price the stock reached during the day

### 3. Low: The lowest price the stock reached during the day

### 4. Close: The final closing price of the stock for the day which becomes the opening price for the next day.

### 5. Volume: Volume is the number of shares that have been traded that is sold or brought throughout the day. It is important to understand Volume: A stock with low volume is thinly traded as it lacks liquidity, i.e., it would be difficult to buy or sell the stocks of that company easily. Similarly, if there is a stock that has a volume higher or more than the average trade-off dealing in this case can be volatile as the price fluctuation would be drastic and quick.

### 6. Adjust Closing Price: An adjusted closing price is a stock’s closing price on a given day of trading that has been amended to include any distribution and corporate actions that occur before the next day opening. The adjusted closing price is a useful tool when examining historical returns because it gives analyst an accurate representation of the firm’s equity value beyond the simple market price.It includes the corporate actions such as stock splits, dividends/distribution, and right offerings.

### 

### Explanatory Data Analysis and Handling Missing Values:

So, we are using Google finance module/package to update our dataset with new stock data on a daily basis but our original dataset had a time column attributing 4’O Clock closing time and another adjusted closing volume column which have been dropped because the new data being appended by the google finance package does not have these columns.

**Models Implemented:**

We are using the basic time series approach to predict stock trends, a popular and widely used statistical method for time series forecasting is the ARIMA model.

**ARIMA MODEL:**

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data.

An [ARIMA model](https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average) is a class of statistical models for analyzing and forecasting time series data.

It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

* **AR**: *Autoregression*. A model that uses the dependent relationship between an observation and some number of lagged observations.
* **I**: *Integrated*. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
* **MA**: *Moving Average*. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each of these components are explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

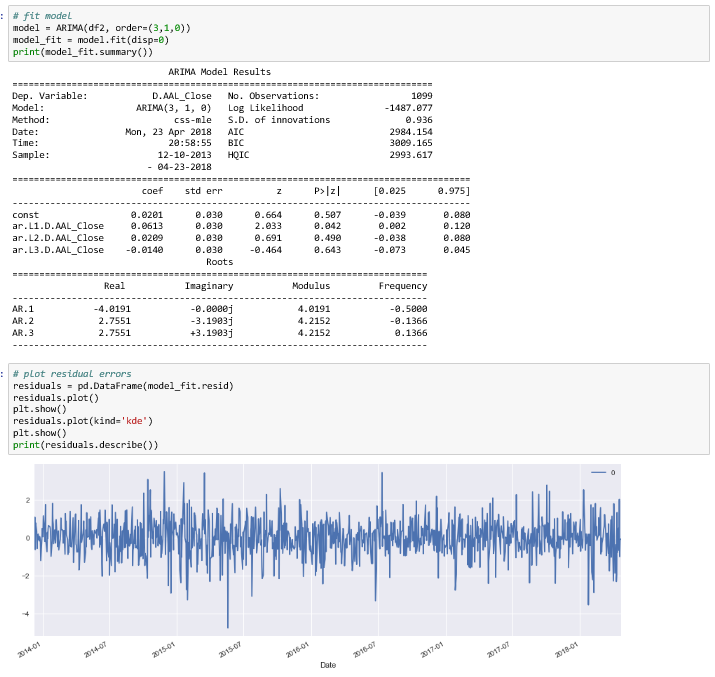
The parameters of the ARIMA model are defined as follows:

* **p**: The number of lag observations included in the model, also called the lag order.
* **d**: The number of times that the raw observations are differenced, also called the degree of differencing.
* **q**: The size of the moving average window, also called the order of moving average.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model.

A value of 0 can be used for a parameter, which indicates to not use that element of the model. This way, the ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model.

Adopting an ARIMA model for a time series assumes that the underlying process that generated the observations is an ARIMA process. This may seem obvious but helps to motivate the need to confirm the assumptions of the model in the raw observations and in the residual errors of forecasts from the model.



**LINEAR REGRESSION MODEL:**

Linear regression is a statistical approach for modelling relationship between a dependent variable with a given set of independent variables. Linear regression models are a good starting point for regression tasks. Such models are popular because they can be fit very quickly and are very interpretable.



**RANDOM FOREST MODEL:**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

****

**Results:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL | Test MSE | TEST R2 | TRAIN RMSE | TRAIN R2 |
| ARIMA | .913 | .915 |  |  |
| Linear Regression | .1518 | .9966 | .1336 | .997 |
| Random Forest | .2525 | .9942 | .042 | .999 |

**References:**