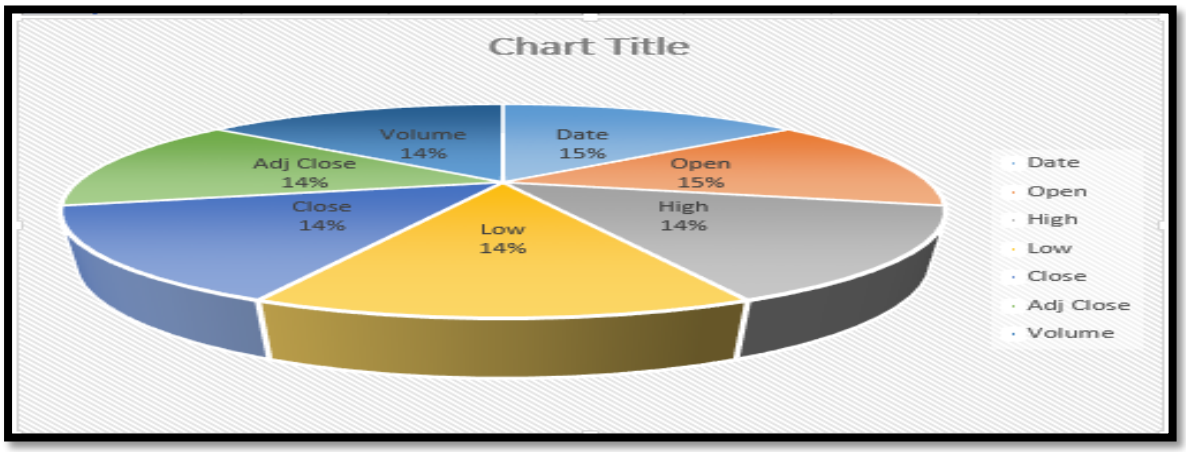
Analytics Report of the Stock Price Prediction

1. **Overview of Dataset:-** The Dataset is comprised of the various feature and the task is to predict the future price of the the stocks .The data is of the form 2335*7 both for Google and AMD .The various attributes are described below:-

   1. Date :-It consist of the various dates that are recorded for the stock measure
   2. Open:-It consist of the opening of the stock for the days .It had positive and negative impact
   3. Close:- It consist of the closing of the stock for the days .It had positive and negative impact
   4. High:- It describe the Peak achieved by stock for the days .It had positive and negative impact.
   5. Low:- It describe the Depth achieved by stock for the days .It had positive and negative impact.
   6. Volume:-It describe the number of shareholder of the day and share sold.

	Date	Open	High	Low	Close	Adj Close	Volume
0	2009-05-22	198.528534	199.524521	196.196198	196.946945	196.946945	3433700
1	2009-05-26	196.171173	202.702698	195.195190	202.382385	202.382385	6202700
2	2009-05-27	203.023026	206.136139	202.607605	202.982986	202.982986	6062500
3	2009-05-28	204.544540	206.016022	202.507507	205.405411	205.405411	5332200
4	2009-05-29	206.261261	208.823822	205.555557	208.823822	208.823822	5291100

2. **Data Description:-** Below is the described effectiveness and the data description :-



3.Data Exploration :- Below is the complete description of the datatype along with checking the NAN and missing values:-

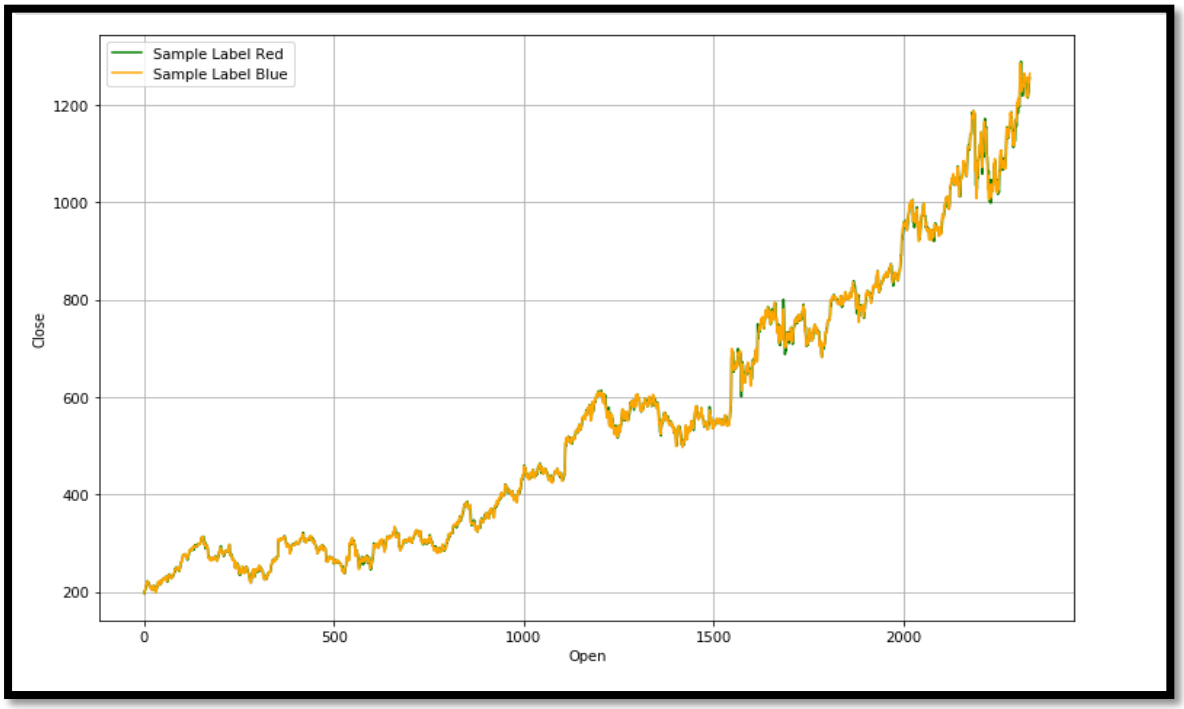
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2335 entries, 0 to 2334
Data columns (total 7 columns):
Date            2335 non-null object
Open            2335 non-null float64
High            2335 non-null float64
Low             2335 non-null float64
Close           2335 non-null float64
Adj Close       2335 non-null float64
Volume          2335 non-null int64
dtypes: float64(5), int64(1), object(1)
memory usage: 127.8+ KB
```

As there is no null entry so data is complete and no missing value or any NAN value.

4.Closure look at High and Low:- The High and Low both are significant feature that particularly describe that what is extreme positive value and extreme negative value for a day. The below is the description of the high vs low:-

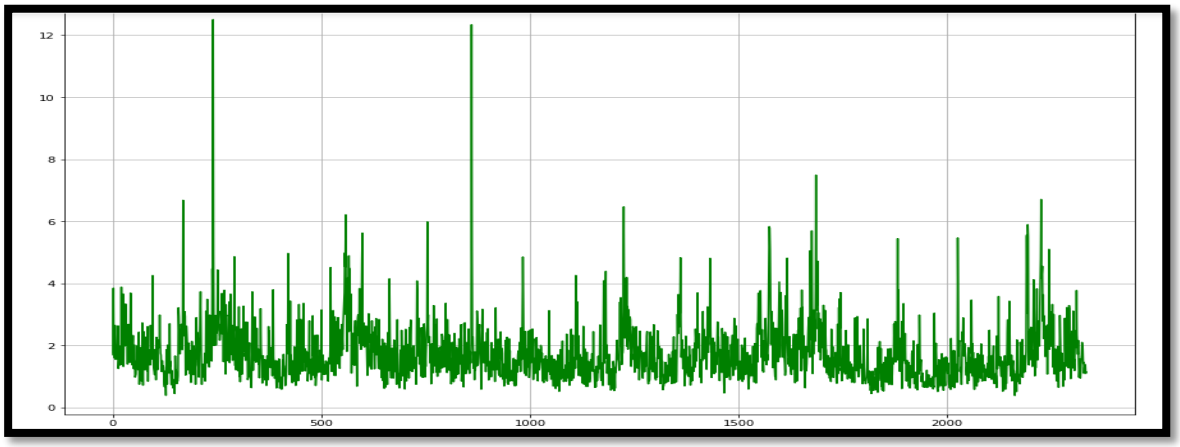
=) X axis describe the High and is described by blue line over graph
=) Y axis describe the Low and is described by the orange over the graph

5. Closure look at Open and Close:-The both features Open and Close are very significant as they are defing the time interval for the stock and also because it describe on a day with which price stock is opened and by the end of day what is the stock during stock closing. The below is decribed relationship between the open and close:-



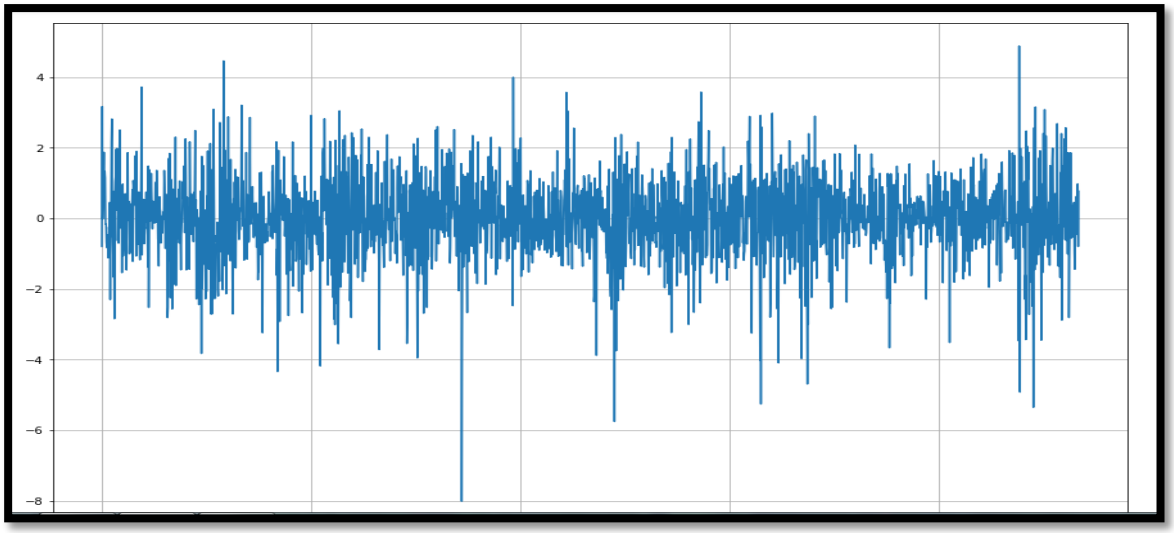
=) X axis describe the opening of the stock and is shown by the green line
=) Y axsis describe the closing the stock and is shown by the orange line
For many values the opening is less than closing and vice-versa though.

7. Closure look at the High-low % :- The change in percentage in the High to low plays as a significant feature in finding out the significant change in stock price. Below is the description of the high low percentage change:-



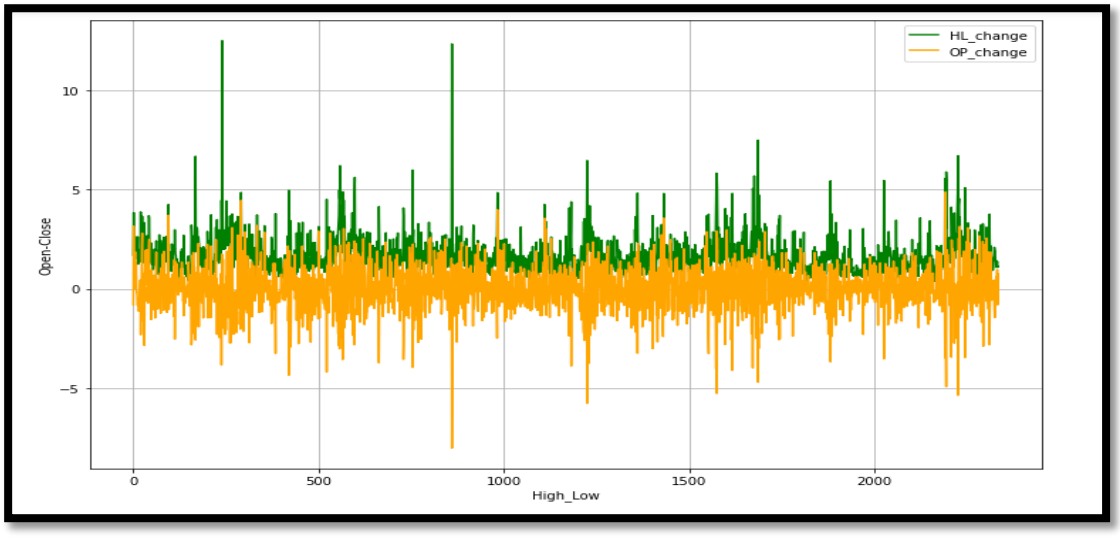
=)The maximum high to low change is in range 12-13%
=)The minimum change in the high to low is 1.2%

8. **Closure look at the Open-Close % :-**The open to close percentage is going to describe out the profit and loss for each count .Below is the complete description of the open to close percentage change:-



=) The maximum profit gain in the single day is :-5.1%
=) The maximum loss monitored in single count is :-8%

9. **Closure look at both % changes:-**Both changes describe the profit and the loss and complete deviation from each count to another count . The merged data of both the percentage change describe the best visual representation of the stock analysis. Below is visual description of the the data:-



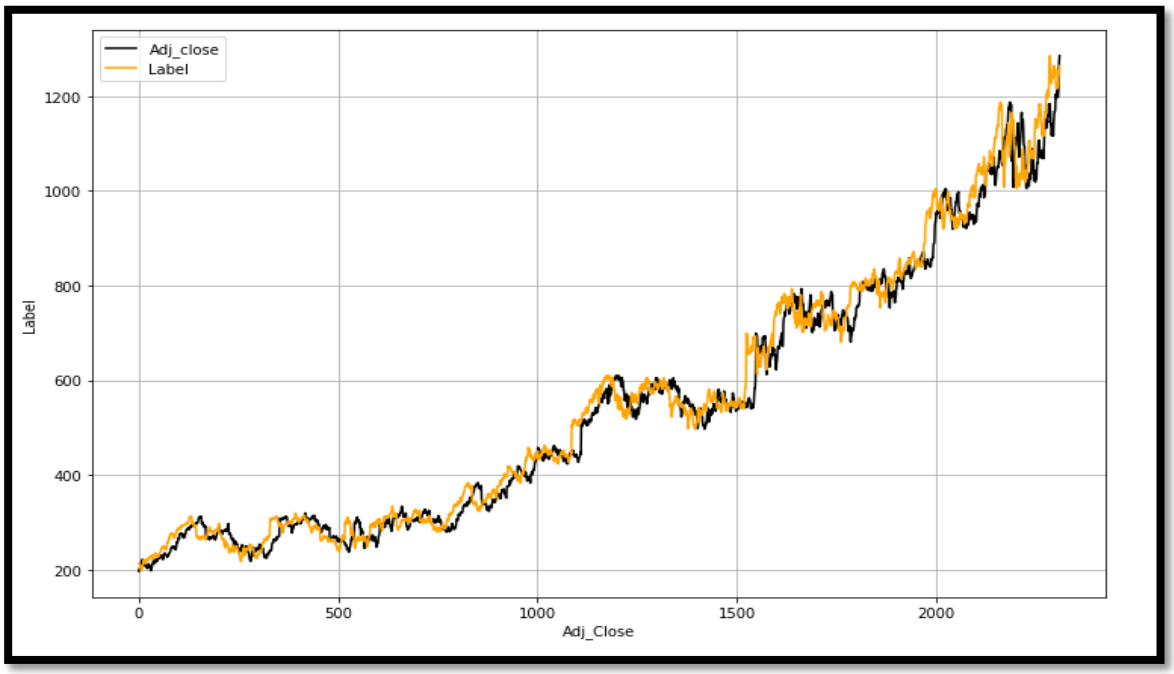
10. Correlation :- The correlation will best find the features for selecting suitable attributes for modelling .The below is the correlation of the same:-

	Open	High	Low	Close	Adj Close	Volume	HL_change	OP_change
Open	1.000000	0.999879	0.999815	0.999687	0.999687	-0.573989	-0.067720	-0.001950
High	0.999879	1.000000	0.999805	0.999844	0.999844	-0.572042	-0.061926	0.007988
Low	0.999815	0.999805	1.000000	0.999874	0.999874	-0.577802	-0.079350	0.011078
Close	0.999687	0.999844	0.999874	1.000000	1.000000	-0.575422	-0.071887	0.020148
Adj Close	0.999687	0.999844	0.999874	1.000000	1.000000	-0.575422	-0.071887	0.020148
Volume	-0.573989	-0.572042	-0.577802	-0.575422	-0.575422	1.000000	0.442151	-0.089211
HL_change	-0.067720	-0.061926	-0.079350	-0.071887	-0.071887	0.442151	1.000000	-0.167198
OP_change	-0.001950	0.007988	0.011078	0.020148	0.020148	-0.089211	-0.167198	1.000000

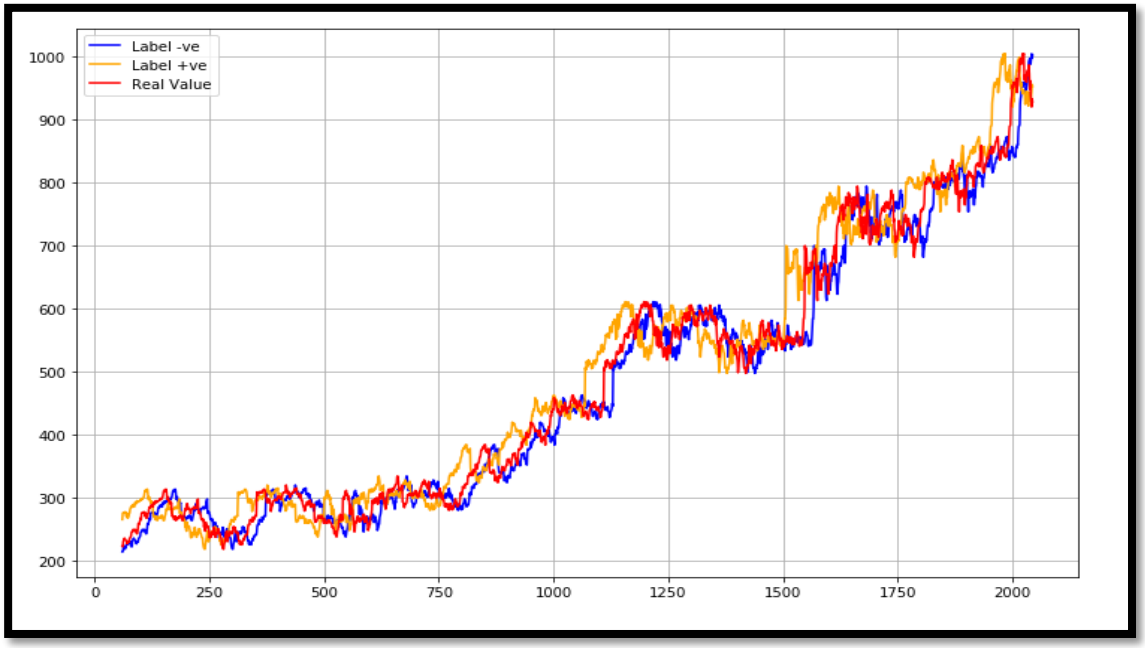
=) Open ,High ,Low, Close are such attribute whose correlation is nearly 1 so we cannot use these as it lead to overfitting of the model.

=) HL_change , Op_change and volume are the most reliable attributes for modelling.

11. Shifting the values by 10 :- We usually shifted the values by providing the new column as Label that get value by seeing the last 20% value of adj_close. Below is the visual representation:-



12. **Finding upper and lower bound:-**Next step for stock prediction is the finding the upper and lower bound of the Adj_close feature so as to defin the range for prediction. In order to do so I made 2 new column label and label1 that had a value by shifting 20% value positively and negatively both to get a limit range:-

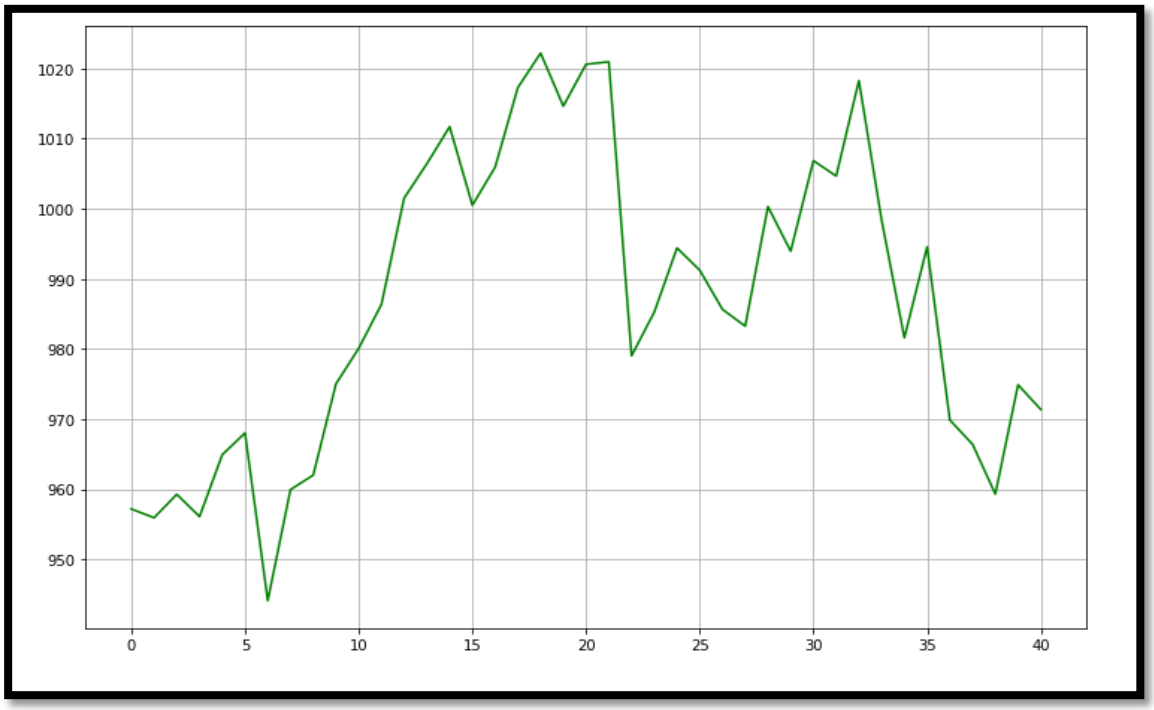


=) Red color=Real Value
=) Orange Color=Upper bound
=) Blue color=Lower Bound

13. **Modelling:-** I have used two model SVM and Linear Regression . The SVM is quite failure because during modelling it is loosing identity of most the features .Regression is the best option as it goes by comparing each relationship and best work for numerical data. The various technique used for the Modelling are:-

1. Linear Regression
2. SVM
3. Decision Tree
4. Random Forest
5. K-Nearest Neighbours

Below is the visual representation of the prediction using regression as accuracy of the 96% over the cross validation over 70:30 ratio. Below is the Visual representation of the predicted value are:-



BY:-Abhishek Joshi
To:-Verzeo.in
Email:-cu.16bcs3171@gmail.com
College:-Chandigarh University