

Stock Price Prediction- Use case of Sentiment Analysis

Submitted towards partial fulfilment for the award of certificate of
Post Graduate Program in Business Analytics and Business Intelligence

By

Group 3

Chukka Priyanka Reddy- BACAUG18035

Prabu Manoharan- BACAUG18031

Prashanth A- BACAUG18033

Sudhahar Veeramohan- BACAUG18046

Yuvaraj Ganesan- BACAUG18051

Under the guidance of Jatinder Bedi



Batch – (PGPBABI.G.Aug'18)

Year of Completion (August' 2019)

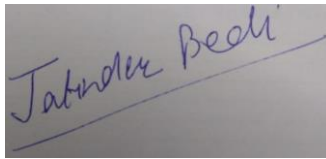
Acknowledgments

This is to certify that the project titled “Stock Price Prediction- Use Case of Sentiment Analysis”, submitted by the members of Group 3 for the partial fulfilment for the award of certificate PGP-BABI is a bonafide record of the work done by us during the period Aug’18 to Aug’19 is not submitted for any other Degree, Diploma, Title or Recognition earlier. The data collected for this project is also from authenticate source that is available in the market not for further use apart from this project.

CERTIFICATE

This is to certify that the participants Chukka Priyanka Reddy, Prabu Manoharan, Prashanth A, Sudhahar Veeramohan and Yuvaraj Ganesan who are the students of Great Lakes Institute of Management, has successfully completed their project on “Stock Price Prediction- Use case of Sentiment Analysis”

This project is the record of authentic work carried out by them during the academic year 2018- 2019.



Mentor's Name and Sign

Jatinder Bedi

Program Director

Name: Mr. P.K.Vishwanath

Date: 28/07/2019

Place: Gurugram

Table of Contents

S.No	Description	Page No.
1.	Domain and Context	6
1.1	Domain	6
1.2	Industrial Growth	6
1.3	Context	6
1.4	Objective	6
2.	Data	7
2.1	Data Source	7
2.2	Data Description	7
2.2.1	Stock Price Data	7
2.2.2	Twitter Data	8
3.	Data Analysis	10
3.1	Stock Price Analysis	10
3.2	Sentiment Analysis	11
3.2.1	Data Cleaning	11
3.2.2	Sentiment Score	11
3.3	Exploratory Data Analysis	12
3.4	Summary of Data Analysis	15
4.	Performance Evaluation metrics	16
4.1	Evaluation Metrics	16
5.	Machine Learning Models	17
5.1	Model Selection	17
5.2	Logistic Regression	17
5.3	K-Fold Cross Validation	18
5.4	Random Forest	19
5.5	XGBoost Model	21
5.6	Support Vector Machine	22
5.7	K-Nearest Neighbour	24
5.8	ARIMA	25
5.9	Summary of ML models	27
6.	Final Classification Model	28
7.	Summary	29
7.1	Conclusions	29
7.2	Challenges	29

7.3	Limitations	29
8.	Appendix	29
9.	References	29

1. Domain and Context

1.1 Domain

Predicting stock price movement with consideration of various factors that affect them, can give any trader/ investor an edge while making decisions. As every second passes during live market hours, the prices of the securities can change dramatically and move with an upward or downward trend. Machine Learning and Deep Learning models can provide a helping hand to predict the prices of stocks. Sentiments also plays an important role in the market movements and these models can help us incorporate text data to provide us with an accurate prediction of the stock prices.

1.2 Industrial Growth

Trading of stocks dates' back to early 19th century in India, where a group of brokers used to gather and an open outcry of stocks would be done for exchange. By 1957, Indian Government recognised this business and was named as Bombay Stock Exchange.

Likewise, the rest of world have stock exchanges and total amount that gets transacted is around \$80 trillion, where US contributes to around 43% of the total business.

Historically the Bombay Stock Exchange had an open outcry floor trading exchange, which then switched to an electronic trading system developed by CMC Ltd. in 1995. It took the exchange only 50 days to make this transition. This automated, screen-based trading platform called BSE On-Line Trading (BOLT) had a capacity of 8 million orders per day. The BSE has also introduced a centralized exchange-based internet trading system, BSEWEBx.co.in to enable investors anywhere in the world to trade on the BSE platform.

Apart from the Stock exchanges for equities, India also has a dedicated Commodities exchange and derivatives exchanges for trading.

1.3 Context

As the stock exchanges have turned to electronic trading, they are also slowly adopting automated trading in the stock exchanges. To supplement automated trading, a statistical model that can predict the stock price movement on daily basis would be of great help to the traders. Machine Learning and Deep learning with text analytics is gaining importance in the Finance industry to predict prices.

1.4 Objective

The objective of this project is to predict whether the stock shall be bearish or bullish using stock price historical data along with text sentiment data from Twitter. We shall also compare the model performance with Buy & Hold Trading Strategy.

2. Data

2.1. Data Source

This project is fully self-funded by the participants themselves. The data that is required for the project has been taken from the Bloomberg Terminal, Twitter feeds and NSE website.

2.2. Data Description

Our initial activity for the project started out with collecting price feeds and twitter sentiment feeds for the Infosys Stock.

2.2.1. Stock Price Data

The initial data of stock price was downloaded from NSE website for the period 1st Jan 2014 to 30th Apr 2019.

Pre-processed data of stock price is as below

Date	Symbol	Series	Prev Close	Open	High	Low	Last	Close	VWAP	Volume	Turnover	Trades	Deliverable Volume	%Deliverble
01-01-2014	INFY	EQ	3,485.65	3,492.00	3,499.00	3,462.05	3,465.40	3,468.00	3,473.69	184836	6.42E+13	11,876.00	51,658.00	0.28
02-01-2014	INFY	EQ	3,468.00	3,463.25	3,508.00	3,463.25	3,484.20	3,480.55	3,483.33	550593	1.92E+14	28,599.00	3,42,682.00	0.62
03-01-2014	INFY	EQ	3,480.55	3,475.00	3,575.00	3,454.25	3,561.20	3,565.15	3,539.91	1255793	4.45E+14	74,223.00	8,65,191.00	0.69
06-01-2014	INFY	EQ	3,565.15	3,575.10	3,580.50	3,468.60	3,510.90	3,517.90	3,520.33	1387291	4.88E+14	79,022.00	8,98,450.00	0.65
07-01-2014	INFY	EQ	3,517.90	3,519.05	3,525.90	3,446.70	3,460.00	3,457.15	3,478.47	964211	3.35E+14	66,664.00	7,12,188.00	0.74
08-01-2014	INFY	EQ	3,457.15	3,461.00	3,475.00	3,421.00	3,432.40	3,428.10	3,437.05	1123799	3.86E+14	94,416.00	7,88,138.00	0.70
09-01-2014	INFY	EQ	3,428.10	3,440.00	3,482.00	3,417.85	3,448.30	3,450.80	3,453.73	1573851	5.44E+14	1,22,460.00	8,88,041.00	0.56
10-01-2014	INFY	EQ	3,450.80	3,490.00	3,575.00	3,448.00	3,556.00	3,551.25	3,534.32	3782272	1.34E+15	1,90,570.00	17,39,108.00	0.46
13-01-2014	INFY	EQ	3,551.25	3,582.00	3,675.00	3,582.00	3,664.00	3,665.00	3,660.54	2139989	7.83E+14	1,32,647.00	14,48,502.00	0.68

Below are the data definition of important variables we will be using.

Date: Trading date

Open: Opening price of the stock

High – Highest price the stock traded during the trading hours of the day

Low – Lowest price the stock traded during the trading hours of the day

Close – Closing price of the stock at the end of trading hours

Volume – Total number of shares traded for the day

2.2.2 Twitter Data

We collected tweets for Infosys for past 5 years from 1st Jan 2014 to 30th April 2019. The main key word that we used was “Infosys”, but we also included details from retweets, mentions, id, hashtags and many more. The keywords used for filtering are devised with extensive care and tweets are extracted in such a way that they represent the exact emotions of public about the company over a period using a python library called Get-OldTweets which is a wrapper for the Twitter Search API. Each tweet contains information about user, no. of retweets, tweet text of 140 characters and the time when it was made. The sample of the raw file that we downloaded is as below.

id	permalink	username	to
1.00197E+18	https://twitter.com/yosim15107/status/1001973935603437568	yosim15107	
1.00197E+18	https://twitter.com/Nauth_Financial/status/1001973340725182464	Nauth_Financial	
1.00197E+18	https://twitter.com/carleenhaylett/status/1001966975692701697	carleenhaylett	HimanshuDPA
1.00196E+18	https://twitter.com/UdacityForBiz/status/1001962738011246594	UdacityForBiz	
1.00196E+18	https://twitter.com/acenik10/status/1001960781624369152	acenik10	
1.00195E+18	https://twitter.com/DisrupTVShow/status/1001954598490378240	DisrupTVShow	
1.00195E+18	https://twitter.com/AubreyUT/status/1001950632557924357	AubreyUT	
1.00195E+18	https://twitter.com/sumit_mit/status/1001945487665258496	sumit_mit	
1.00193E+18	https://twitter.com/subrakd7/status/1001932923782578178	subrakd7	
1.00193E+18	https://twitter.com/AutomationAnywh/status/1001928108457582594	AutomationAnywh	
1.00192E+18	https://twitter.com/VegettoEX/status/1001923537710538752	VegettoEX	
1.00192E+18	https://twitter.com/immgfairness/status/1001921059925495810	immgfairness	
1.00192E+18	https://twitter.com/immgfairness/status/1001920549432578050	immgfairness	
1.00192E+18	https://twitter.com/ishan_gupta/status/1001918464376926211	ishan_gupta	UdacityINDIA
1.00192E+18	https://twitter.com/Ashishy96289929/status/1001917723302137856	Ashishy96289929	ashokk_cs
1.00191E+18	https://twitter.com/NarayanaParsa/status/1001914720272576512	NarayanaParsa	KTRTRS
1.00191E+18	https://twitter.com/SportBizAwards/status/1001913228941721601	SportBizAwards	
1.00191E+18	https://twitter.com/Nauth_Financial/status/1001912655978868738	Nauth_Financial	

text	date
@blrcitytraffic @BlrCityPoliceat around 5:15 am the bus TN bus was very rash and nearly dashed our Infosys office cab...	2018-05-30 23:49:36+00:00
Infosys Finacle forms blockchain-based trade network in India https://nauthfinancial.com/2018/05/infosys-finacle-forms-bl	2018-05-30 23:47:14+00:00
Very much looking forward to learning more @HimanshuDPA and to meeting the @Infosys team. See you in Vegas!!! https://	2018-05-30 23:21:57+00:00
Infosys bets on self-driving carts to shape its future http://ow.ly/S3lI30kgac1 via @BloombergQuint @UdacityForBiz @Udaci	2018-05-30 23:05:06+00:00
Many #Indian #IT top #executives I dealt with only understood the #enterprise #software & related business models, It's like	2018-05-30 22:57:20+00:00
Check out Friday's show for interviews with @ImAmyO @cloudera, @imravikumars of @Infosys, & @jonep @diginomica ht	2018-05-30 22:32:46+00:00
Great list of upcoming guests on @DisrupTVShow: @ImAmyO @cloudera, @imravikumars @Infosys, @jonep @diginomica,	2018-05-30 22:17:00+00:00
#marketbasket #infosys #sapphirenow https://lnkd.in/egap9hj	2018-05-30 21:56:34+00:00
What the Hell you are speaking @kanhaiyakumar Sirji.... Azim Premji is associated with Infosys..?!!!! From when did Azim Pr	2018-05-30 21:06:38+00:00
Thank you again to our amazing #ImagineNewYork18 sponsors: platinum partners @Accenture @Deloitte @EYnews @Genp	2018-05-30 20:47:30+00:00
Dooly Bravo Land (Nintendo Famicom, Daou Infosys, 1992). Son Goku murders a cute dinosaur. pic.twitter.com/PQKvJr1fCY	2018-05-30 20:29:20+00:00
You are the managing director of Infosys so why cant you hire Americans ?	2018-05-30 20:19:30+00:00
Uou are running a scam H1B program, you dont want Indians to become citizens or #H4EAD because you cannot exploit ther	2018-05-30 20:17:28+00:00
Incredible to work with awesome leaders like @imravikumars & @SudipSingh1 at @Infosys and build the talent for autonom	2018-05-30 20:09:11+00:00
See how we do it @Infosys #PegaWorld #Infosys #Digitize_Core #amplifyExperience https://twitter.com/ashokk_cs/status/1	2018-05-30 20:06:14+00:00
Infosys... Ki velle mundhu.. Akkada ap students vuntaaru ekkuvu.. Telagaanaa students okkadu vundadu..	2018-05-30 19:54:18+00:00
Shortlist for Best Business Serving Sport (over 60 Employees): Portview Fit-Out, Reading Room, Kukri Sports, http://ADI.TV , C	2018-05-30 19:48:23+00:00
Infosys Finacle forms blockchain-based trade network in India https://nauthfinancial.com/2018/05/infosys-finacle-forms-bl	2018-05-30 19:46:06+00:00

retweets	favorites	mentions
0	0	@blrcitytraffic @BlrCityPolice
0	0	
3	2	@HimanshuDPA @Infosys
1	2	@BloombergQuint @UdacityForBiz @UdacityINDIA
0	0	
0	1	@ImAmyO @cloudera @imravikumars @Infosys @Jonerp @diginomica @rwan0 @ValaAfshar
1	3	@DisrupTVShow @ImAmyO @cloudera @imravikumars @Infosys @Jonerp @diginomica @luke_beatty @Brandfolder @gdh
0	1	
0	0	@kanhaiyakumar
8	14	@Accenture @Deloitte @EYnews @Genpact @Infosys @ISG_News @PwC @TCS
8	20	
1	0	
0	0	
1	7	@imravikumars @SudipSingh1 @Infosys
0	0	@Infosys
0	0	
1	2	
0	0	

hashtags	geo
#digitaltransformation #selfdrivingcar	
#Indian #IT #executives #enterprise #software #Infy	
#DisrupTV	
#DisrupTV	
#marketbasket #infosys #sapphirenowhttps	
#ImagineNewYork18	
#H4EAD	
#PegaWorld #Infosys #Digitize_Core #amplifyExperiencehttps	
#SBA2018	

After the data was collected, the process was to analyse the data and understand which of the data columns that we have collected is actually needed for our final analysis. The process is explained in detailed in the upcoming sections.

3. Data Analysis

3.1 Stock Price Analysis

After analysing the excel copy of stock prices, we understood that the prices that was derived from the NSE website is not sufficient for the analysis. For further analysis in the project, the original data was changed to the below derived variables. The values under Price Lag1 to Price Lag5 was derived from the Close price of the stock.

The thought process on how we calculated the Price Lag value in percentage terms—In case of Lag1, we considered the difference between the previous day price and current day price. In case Lag2, the difference is calculated between 2 days' prices

Date	Close	HighClose	OpenClose	Price_Chng	Price_Lag1	Price_Lag2	Price_Lag3	Price_Lag4	Price_Lag5	Volume_Chng	Target
13-01-2014	458.10	0.00	0.02	0.03	0.03	0.01	-0.01	-0.02	-0.01	-0.77	1
14-01-2014	460.80	0.00	0.01	0.01	0.03	0.03	0.01	-0.01	-0.02	-1.31	1
15-01-2014	464.00	0.00	0.00	0.01	0.01	0.03	0.03	0.01	-0.01	0.57	1
16-01-2014	465.60	0.00	0.01	0.00	0.01	0.01	0.03	0.03	0.01	0.03	1
17-01-2014	466.20	0.01	0.00	0.00	0.00	0.01	0.01	0.03	0.03	0.31	1
20-01-2014	468.70	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.03	-1.63	1
21-01-2014	469.80	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.19	1
22-01-2014	470.70	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	-0.79	1
23-01-2014	474.10	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.14	0
24-01-2014	469.80	0.01	0.00	-0.01	0.01	0.00	0.00	0.01	0.00	-0.15	0

New variables created using NSE stock data:

HighClose - Difference between High and Close price of the day $[(High - Close)/Close]$

OpenClose - Difference between Open and Close price of the day $[(Close - Open)/Open]$

PriceLag1 – Previous day stock price change in percentage term

PriceLag2 – 2 days before stock price change in percentage term

PriceLag3 – 3 days before stock price change in percentage term

PriceLag4 – 4 days before stock price change in percentage term

PriceLag5 – 5 days before stock price change in percentage term

Volume_Chng – Percentage of volume change compared to previous day $[(Curr\ Day\ Vol - Prv\ Day\ Volume)/Prv\ Day\ Volume]$

Target – Target variable. 1 – if next day close price is higher than today' close price. 0 – if next day close price is lower than today' close price

3.2 Sentiment Analysis

For further analysis of twitter text data to understand the sentiments, we found out that only “text” and “Date” was necessary. Using the R code we considered only the relevant columns for our analysis.

	text	date
	RBS deal loss: Infosys says won't cut 3,000 jobs, but will reallocate them in other projects: RBS announced I... http://bc.vc/apT5Wt8	17-08-2016
	Big News Network - Infosys to sack 3,000 techies as Scottish bank cancels deal http://bit.ly/2bBUS4u	17-08-2016
	RBS deal loss: Infosys says won't cut 3,000 jobs, but will reallocate them in other projects: RBS announced I... http://bc.vc/apT5Wt8	17-08-2016
	3,000 jobs to get affected post RBS project scrap: #Infosys - http://ecoti.in/0t6l1Y	17-08-2016
	If it is not a fact nothing is true in this world INDIAN JINGOISM HAS NO MATCH	17-08-2016

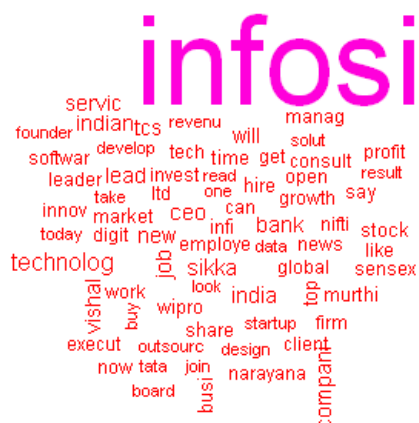
3.2.1 Data cleaning:

Twitter data are unstructured in nature, it is very import we clean the data before performing any analysis on it. We removed special characters, numbers, website links and punctuations from the twitter text to make is clean for further analysis.

	text	date
RBS deal loss Infosys says wont cutjobs but will reallocate them in other projects RBS announced I		17-08-2016
Big News NetworkInfosys to sacktechies as Scottish bank cancels deal		17-08-2016
RBS deal loss Infosys says wont cutjobs but will reallocate them in other projects RBS announced I		17-08-2016
jobs to get affected post RBS project scrap Infosys		17-08-2016
If it is not a fact nothing is true in this world INDIAN JINGOISM HAS NO MATCH		17-08-2016
BTIndia Sensex Nifty close lower Infosys erases lossespictwittercommIlgLcm		17-08-2016

3.2.2 Sentiment Score:

After the data was cleaned, we used “wordcloud” function to understand which words from the text had more significance.



Once understanding the importance of wrodccloud, we started the process to calculate the sentiment score using `get_sentiment` function from “syuzhet” package for each tweet. We also calculated the number of positive, negative and neutral tweets in each day. Finally, each day is tagged as positive, negative or neutral based on which category of the tweets are mostly posted on that day.

Then using one-hot encoding method, we have come up with 2 columns “Neg” and “Pos”.

	Date	Negative	Neutral	Positive	Neg	Pos
1	10-Jun-14	187	118	197	0	1
	09-Jun-14	90	60	81	1	0
	08-Jun-14	66	52	56	1	0
	07-Jun-14	87	358	171	0	0
	06-Jun-14	372	376	578	0	1

Negative	Neutral	Positive
87381	370511	312674

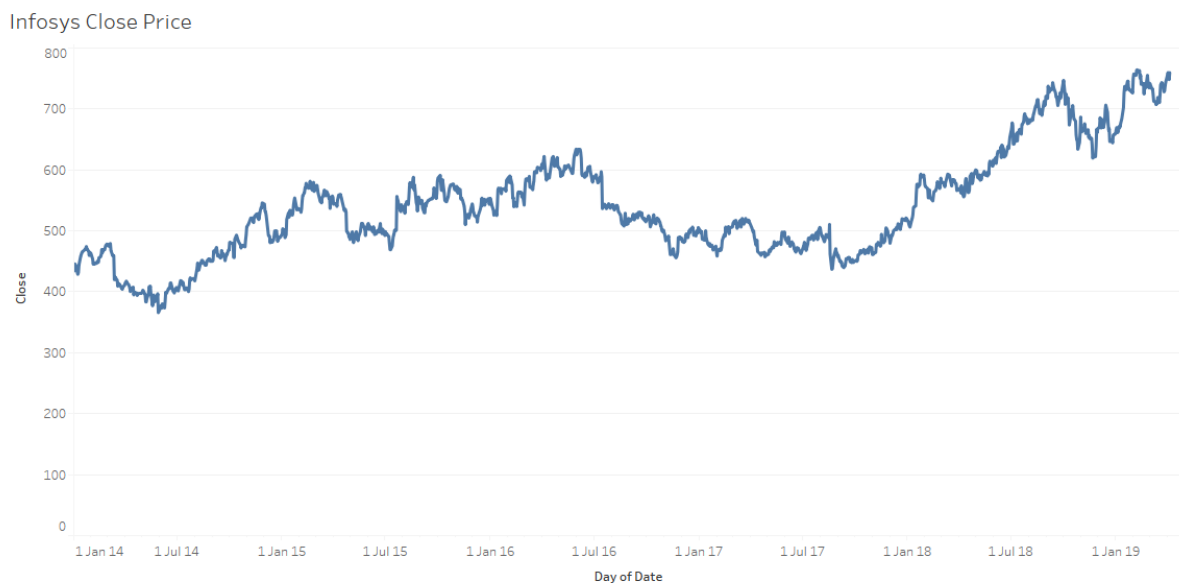
This sentiment variables “Pos” – Positive and “Neg” – Negative are used in machine learning models to predict stock price movement.

3.3 Exploratory Data Analysis

After the stock price data is uploaded, we need to understand the variance or pattern in the data for further processing. Any pattern recognition and deduction can be better understood through EDA process.

Infosys Close Price (1st Jan 2014 – 30th Apr 2019)

Figure 1:

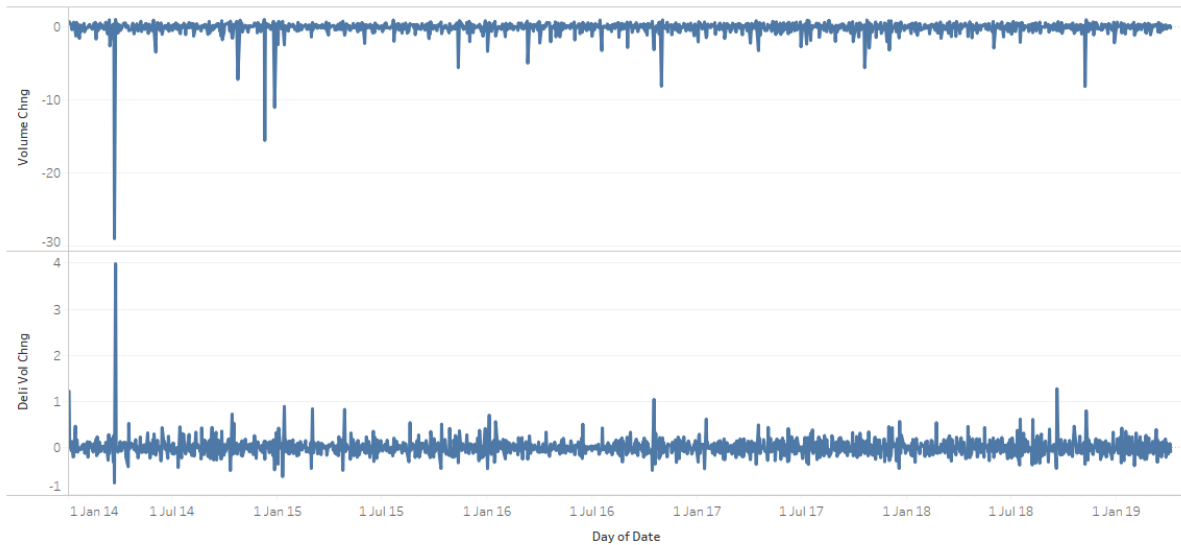


The above stock price chart show us the price movement over a period of 5 years. A careful observation of chart throws light on when the price moves upward, which can be related to quarterly results of the company.

Daily Volume Change

Figure 2:

Daily Volume Change

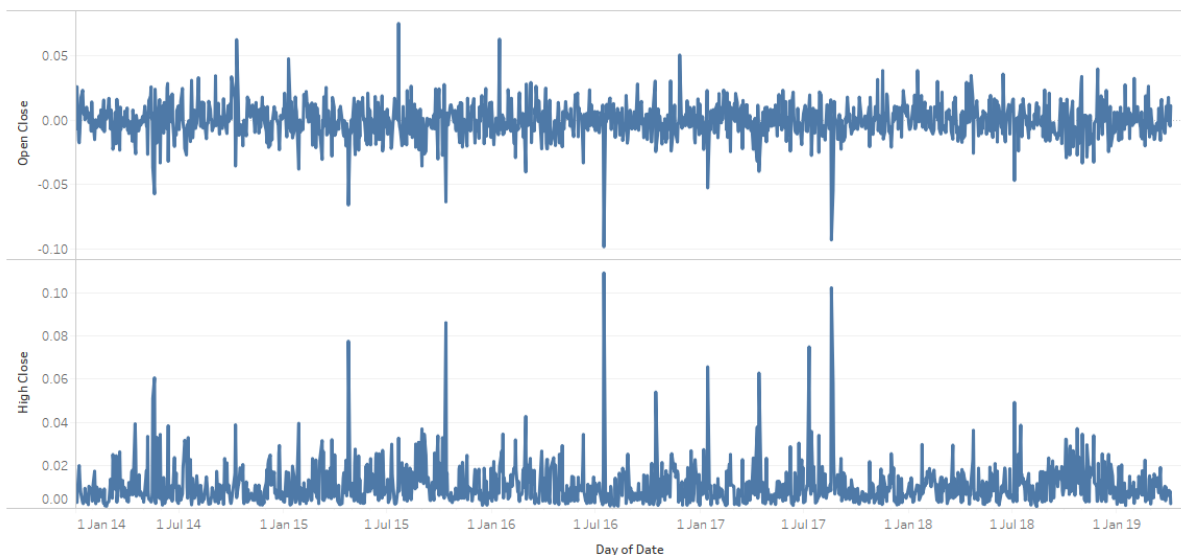


The above chart shows us the daily volume change in the market that is traded. As a particular day during the month Mar-Apr'14, we see strong increase in the daily volume change, which can be attributed to a corporate action event that would have taken place.

High-Close & Open-Close

Figure 3:

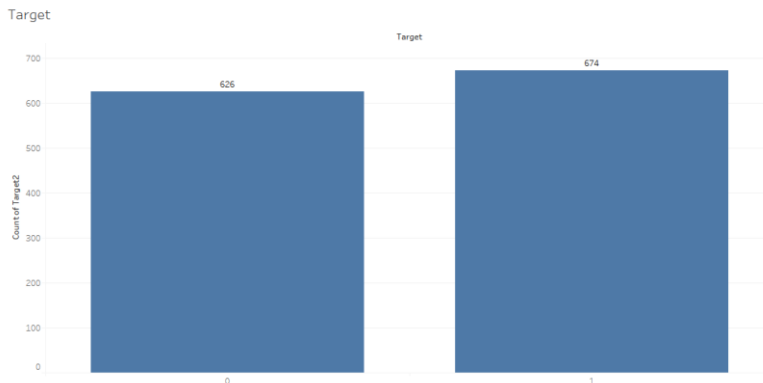
High-Close & Open-Close



This chart shows us the pattern of prices that was provided under the column of High-Close and Open-Close

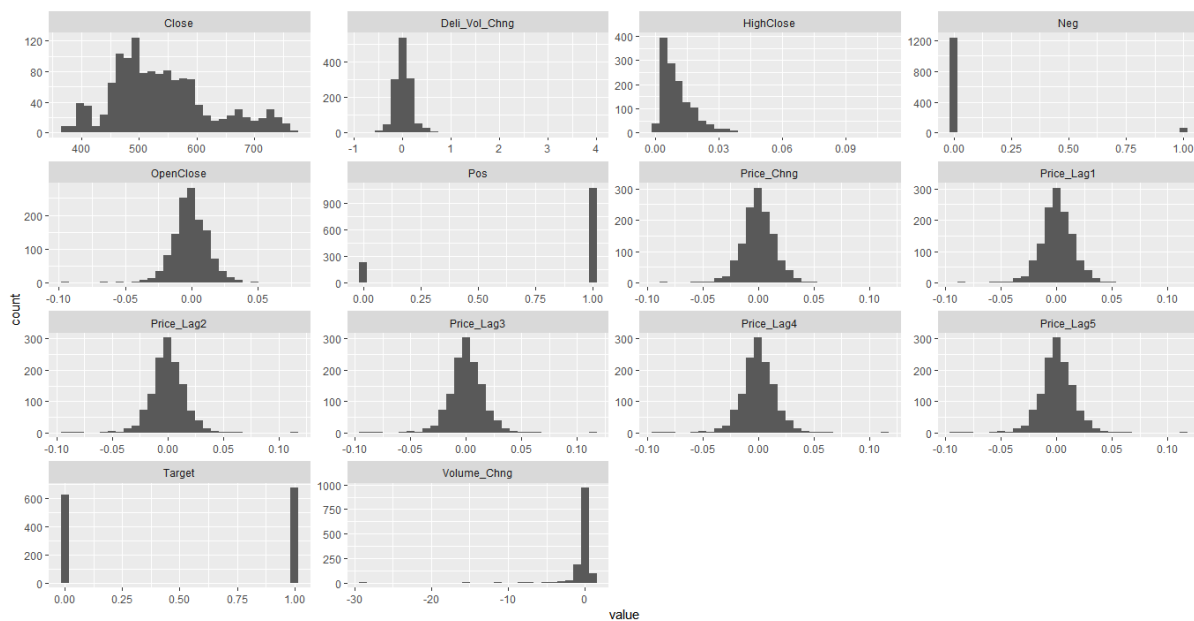
Target Variable

Figure 4:



Histogram of the variables

Figure 5:

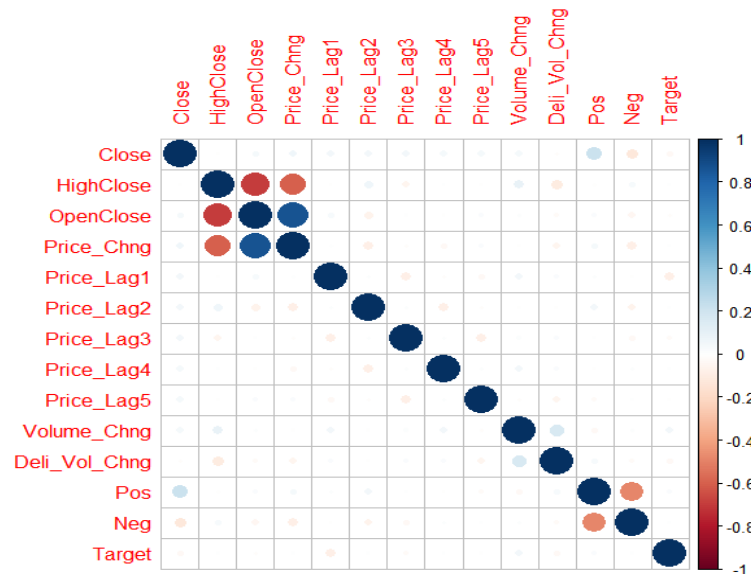


The above series of histograms were made to understand if the data that we had considered is normally distributed or is there any tuning that needs to be done.

As we can clearly see that the derived variables have a uniform distribution, which is further considered for our analysis.

Correlation Plot

Figure 6:



Observations from EDA:

- Daily volume change and Delivery Volume changes are mostly within range. Though we could see few abnormal changes they are not alarming.
- Target variable has good balanced 2 classes.
- From histogram, it is clear all the numeric variables except High-Close are normally distributed. And, much of outlier issues.
- The correlation plot shows that the variables High Close and Open Close, High Close and Price Change and Open Close and Price Change are highly correlated. We need to aware of this and will decide whether to removing highly correlated variables while building machine learning models.

3.4 Summary of Data Analysis

After further classification of new variables from the stock price and cleaning od data from twitter feed we classified them as positive or negative tweets and provide values as 1 and 0. With this data we can try below machine learning methodology to predict stock price direction.

- a) Build baseline classification models like Logistic Regression, KNN, Random Forest, etc
- b) Check the model accuracy and validate the with resampling method(K-Fold)
- c) Build a time series model and convert the predictions into binary so that the results can be compared with other models
- d) Try ensembling method to build final model and check the accuracy

4. Performance Evaluation metrics

4.1 Evaluation Metrics:

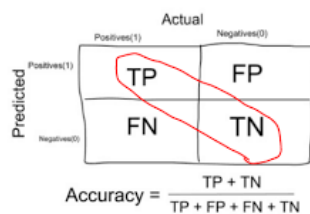
Given our approach to solving the problem, classification of variables and correctly identifying both Positive and Negative tweets is important for further analysis, we have chosen Accuracy and F1 Score as our model evaluation metrics. We will also use Confusion Matrix, AUC and ROC for broad level comparison of models.

Confusion Matrix:

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	Positive Predictive Value	$a/(a+b)$
	Negative	c	d	Negative Predictive Value	$d/(c+d)$
		Sensitivity	Specificity	Accuracy = $(a+d)/(a+b+c+d)$	
		$a/(a+c)$	$d/(b+d)$		

Accuracy:

It is the ratio of number of correct predictions to the total number of input samples.



F1 Score:

F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is.

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

5. Machine Learning Models

5.1. Model Selection

As we know that we are dealing with a classification problem, we can already choose some basic models (or classifiers). We can test these models with our data to get preliminary idea on how these models behave. The Baseline algorithms which can be tested initially are listed as follows:

1. Logistic Regression
2. KNN
3. Random Forest
4. XGBoost
5. Support Vector Machine (SVM)

Apart from the above classification models, we will also try ARIMA timeseries model and convert the results into binary.

Finally, we will be building an ensemble algorithm at the end to test their performance.

5.2. Logistic Regression

Logistic Regression is used when the dependent variable (target) is categorical. It models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail. In the logistic model, the log-odds (the logarithm of the odds) for the value labelled "1" is a linear combination of one or more independent variables; the independent variables can each be a binary variable or a continuous variable.

We started the model with all independent variables except sentiment variables. Through the continuous iteration of the model, we were able to narrow down to 3 variables which are significant in predicting the price direction.

Significance of the model was confirmed by log likelihood test.

None of the variables had high correlation this was confirmed by VIF.

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	148	142
1	298	312

Accuracy : 0.5111
 95% CI : (0.4779, 0.5442)
 No Information Rate : 0.5044
 P-Value [Acc > NIR] : 0.357

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.05106	0.06832	0.747	0.4548
Price_Lag2	0.52464	4.17392	0.126	0.9000
Volume_Chng	0.12128	0.06968	1.741	0.0818
Deli_Vol_Chng	-0.58302	0.35760	-1.630	0.1030

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Then we added sentiment variables to the model and tested it. As, expected the model accuracy has increased from 0.51 to 0.55.

Confusion Matrix and Statistics

```

                Reference
Prediction  0    1
           0 219 176
           1 227 278

```

```

                Accuracy : 0.5522
                95% CI : (0.5191, 0.585)
    No Information Rate : 0.5044
    P-Value [Acc > NIR] : 0.002284

```

Coefficients:

```

                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.10130    0.16181  -0.626  0.53129
Price_Lag1    -12.73562    4.40336  -2.892  0.00382 **
Volume_Chng     0.14344    0.07219   1.987  0.04694 *
Deli_Vol_Chng  -0.58918    0.35926  -1.640  0.10101
Pos              0.18413    0.17931   1.027  0.30449
Neg              0.28493    0.30853   0.923  0.35575
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Observations:

- “Price_Lag1” is highly significant
- Volume_Chng and Deli_Vol_Chng are significant at 10% error rate
- Twitter sentiment variables “Pos” and “Neg” are not statistically significant
- Accuracy has marginally decreased to 0.554 (highest of all models)
- There is no multicolliniarity issue
- P value of lrtest is less than 0.05. Hence, this model is significant.

We have developed reasonably stable model with 5 independent variables. Let’s perform cross validation and check how well our model performs.

5.3. K-Fold Cross Validation

Below is the result of 10-Fold cross validation of Logistic Regression model with identified 5 independent variables. Accuracy of the model is very close to our model 6 which indicates this model is reasonably stable for unknown values.

```

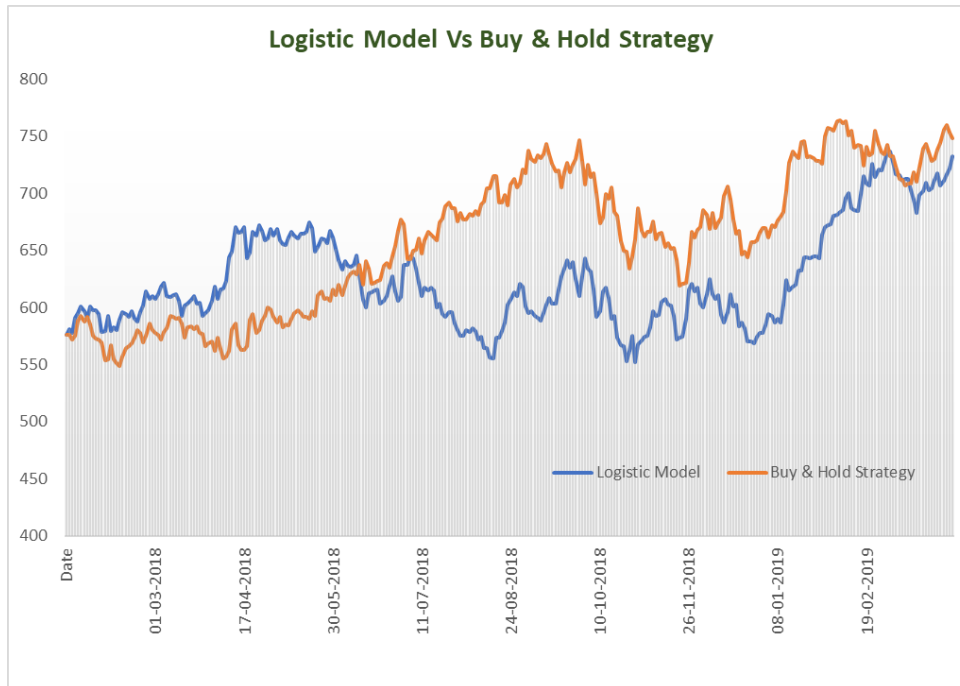
Generalized Linear Model

900 samples
 5 predictor
 2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 810, 810, 810, 810, 810, 810, ...
Resampling results:

Accuracy  Kappa
0.5410877 0.08069394

```



Observations:

The logistic model failed to achieve the buy & hold strategy performance.

5.4. Random Forest

Random forest, like its name implies, consists of large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest comes out with a class prediction and the class with the most votes becomes our model's prediction. The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners.

We used "mlr" package in R to develop a Random Forest and Xtreme Gradient Boosting models. We used only 5 independent variables that were identified to be significant in logistic regression model.

After building baseline random forest model, we tried random grid search with 5 Fold validation to find the optimal parameters for the model.

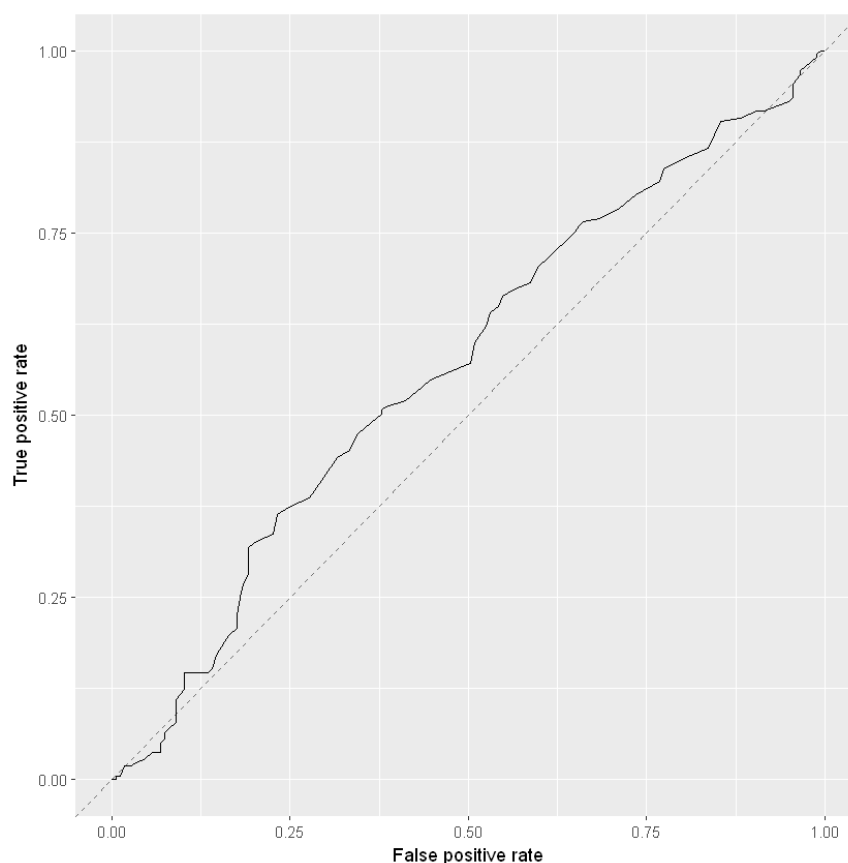
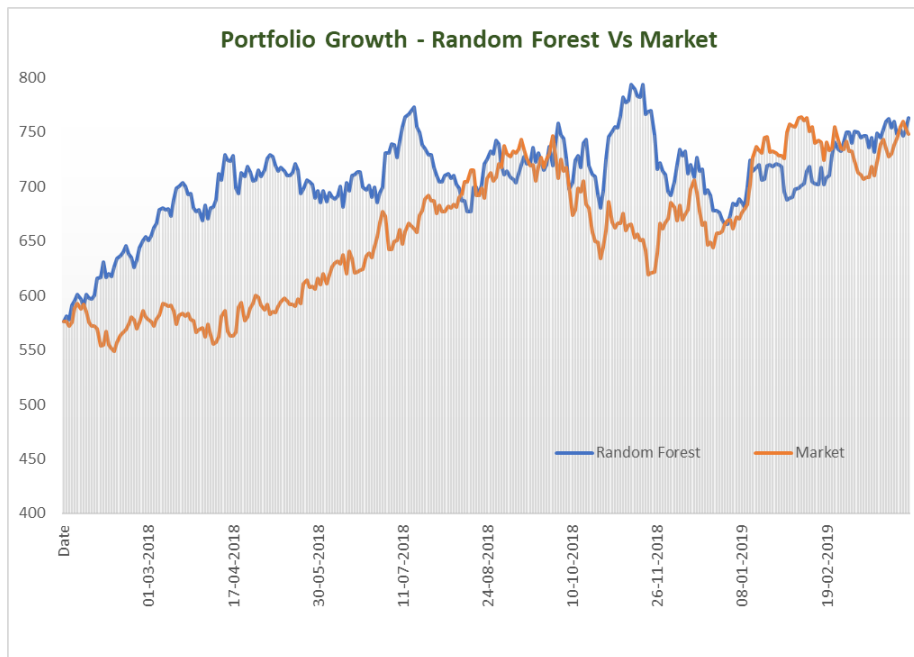
Confusion Matrix and Statistics

```

      Reference
Prediction  0   1
      0 102  75
      1 101 116

      Accuracy : 0.5533
      95% CI : (0.5027, 0.6031)
      No Information Rate : 0.5152
      P-Value [Acc > NIR] : 0.07179

```



Observations:

- Based on the 5 Fold cross validation, Number of trees – 250, No of Variables at each tree – 3 are the optimal values identified from model tuning
- Model accuracy on the train data is 0.62 and on test data is 0.55. It shows model is not stable.
- This model unstability is reflecting on the portfolio growth performance chart. Model failed to beat the buy&hold strategy performance.

5.5. XGBoost Model

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

It has an inbuild cross validation feature, we tried 3-Fold cross validation. And, did random grid search with 100 iterations.

Below are the optimal parameters identified from the grid.

```
acc.test.mean: 0.536666666666667
```

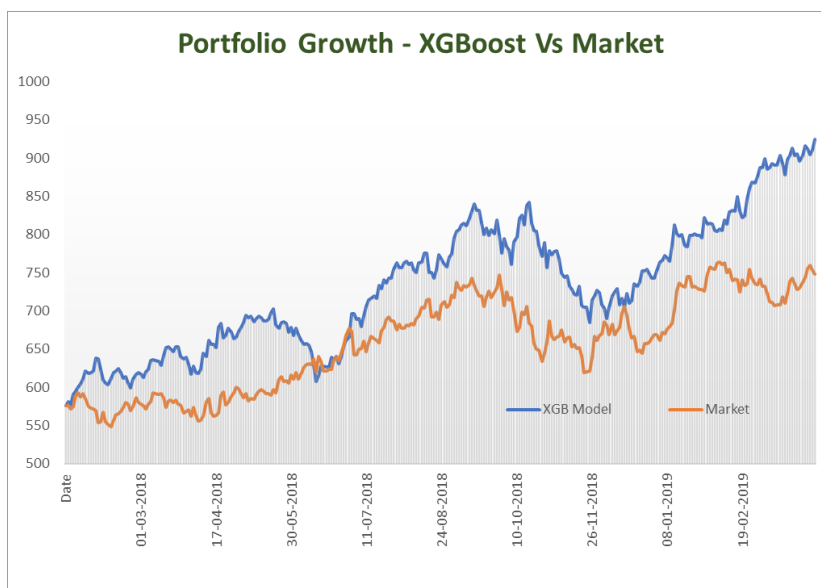
```
$nrounds
513
$max_depth
10
$lambda
0.655163732299116
$eta
0.225949774053413
$subsample
0.705872889584862
$min_child_weight
2.11109115555882
$colsample_bytree
0.700365474983119
```

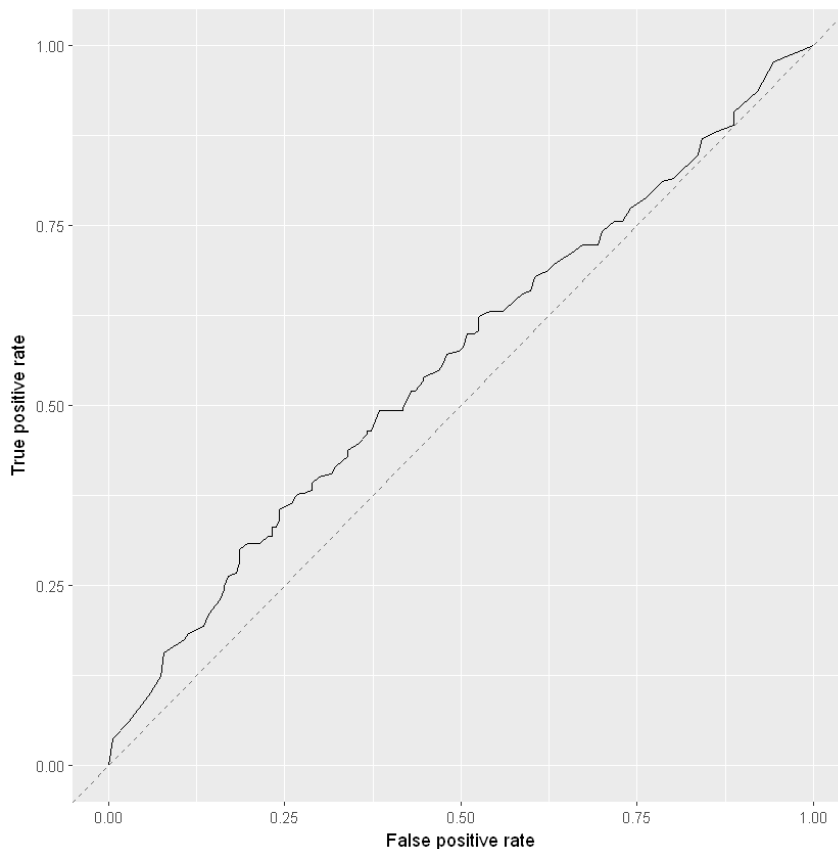
Confusion Matrix and Statistics

```

      Reference
Prediction  0   1
      0  102  75
      1  108 109

      Accuracy : 0.5355
      95% CI : (0.4849, 0.5856)
      No Information Rate : 0.533
      P-Value [Acc > NIR] : 0.48031
```





Observations:

- Model achieved 0.53 accuracy on both train and test data. It shows the stability of the model to unseen data.
- Though the model accuracy is not good compared to the logistic regression and Random Forest models, it managed to achieve better P&L because the model was able to correctly predict the big price moves on both directions.

5.6. Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data (*supervised learning*), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

There are many kernel available under SVM, for this model we tried both “linear” and “radial” kernels.

The model with “radial” kernel gave better accuracy, so did further tuning of the model with 10-Fold cross validation and came up with the below optimal parameters for the final SVM model.

```
Parameter tuning of 'svm':
- sampling method: 10-fold cross validation
- best parameters:
  cost gamma
  0.5 0.05
- best performance: 0.5188889
```

SVM model performed well compared to the other classification models. Below are the final hyper parameters after model tuning. Model achieved **accuracy of 0.54** on test data.

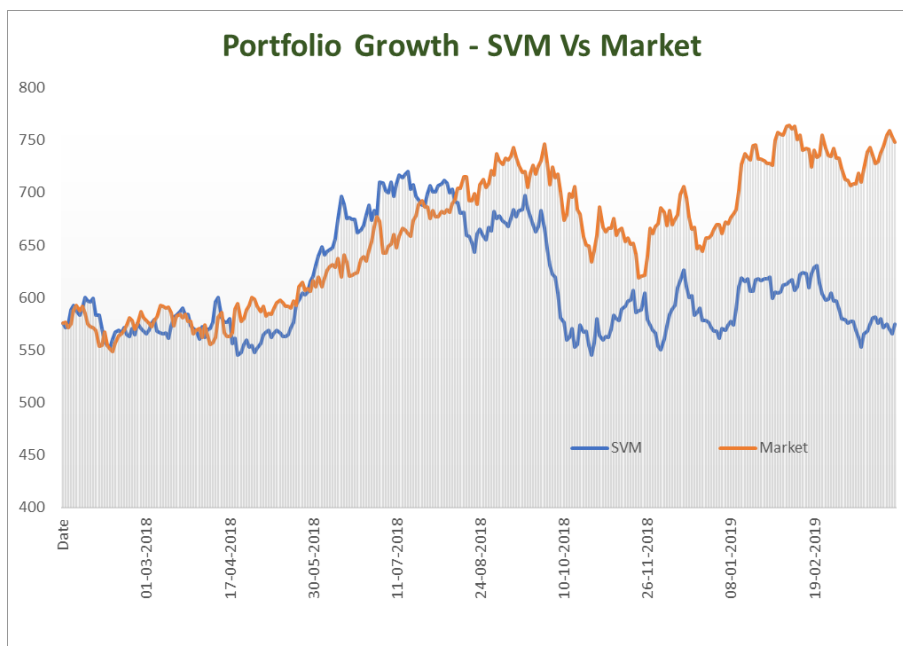
Confusion Matrix and Statistics

```

Reference
Prediction  0  1
0      74 103
1      77 140

Accuracy : 0.5431
95% CI : (0.4925, 0.5931)
No Information Rate : 0.6168
P-Value [Acc > NIR] : 0.99877

```



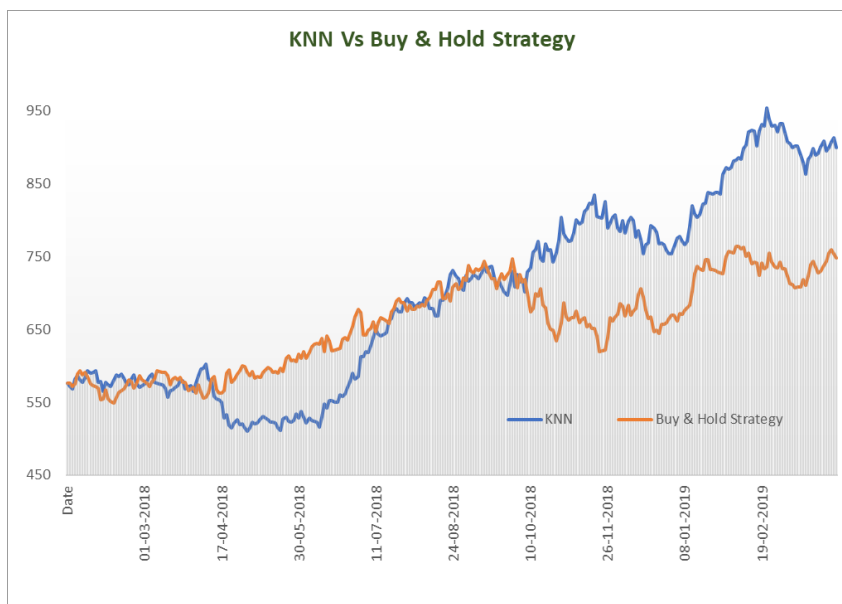
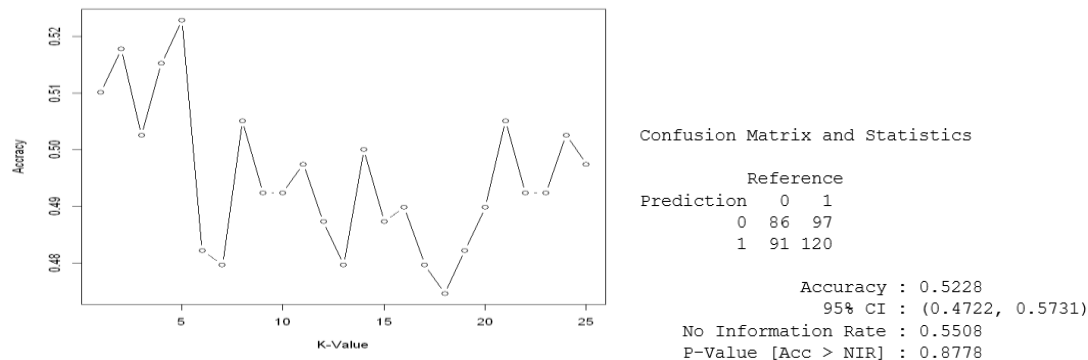
Observations:

Though the model achieved good classification accuracy, it did not perform well on Portfolio growth.

5.7. K-Nearest Neighbour

K Nearest Neighbour (KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution. In other words, the model structure determined from the dataset.

In KNN, K is the number of nearest neighbours. The number of neighbours is the core deciding factor. We tried KNN model with various K values. Below is the graph of model accuracy for various k value.



Observations:

- K value of 5 gave better accuracy
- Model performance beat the buy & hold strategy performance by good margin

5.8. ARIMA

ARIMA is an acronym that stands for Auto Regressive Integrated Moving Average. It is a generalization of the simpler Auto Regressive Moving Average and adds the notion of integration.

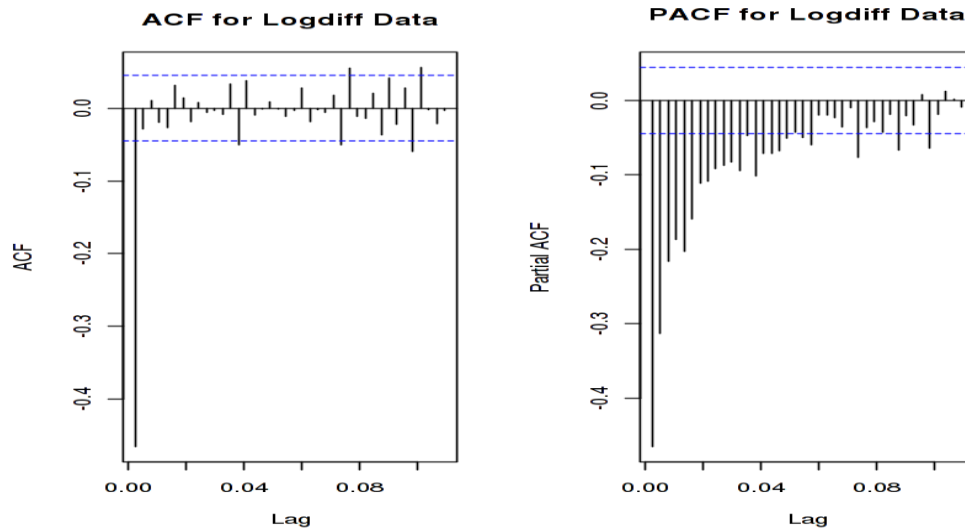
This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

- **AR:** *Auto regression.* 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.



From time series decomposing chart, we can see it has only Trend component and there is no clear Seasonal component in the series.



Results of Dickey-Fuller Test:

Test Statistic	-27.333927
p-value	0.000000
# Lags Used	1.000000

The Dickey-Fuller test for stationarity confirms the time series is stationary.

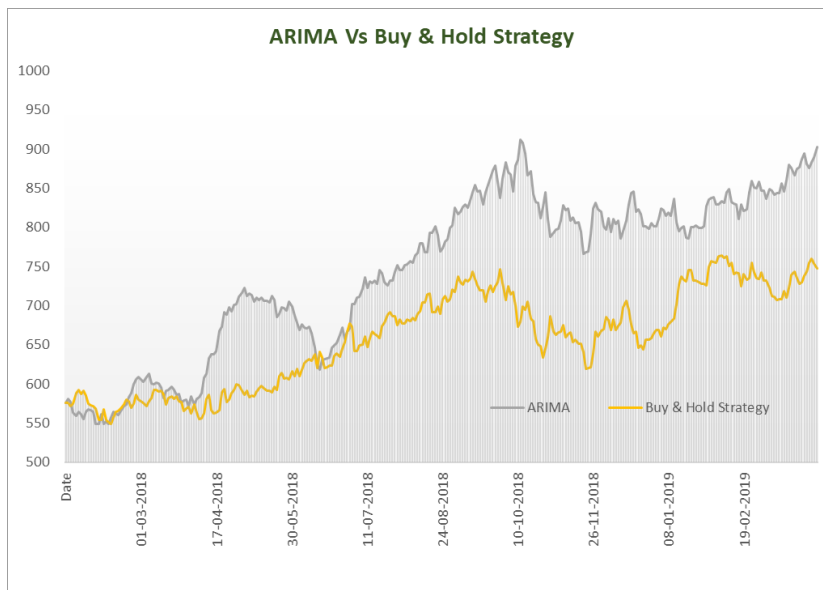
We used p,d,q (12,1,0) for the ARIMA model. With a slight change to the typical ARIMA model, we predicted stock price for next one day at a time. And, converted the predicted value to binary using the direction of prediction. If tomorrow's predicted value is higher than today's closing price, then the direction is 1 else 0.

	Actual	Predictions	Actual	Predictions	ActualTrend	PredictionTrend
0	576.350	574.074368	576.350	574.074368	Down	Down
1	571.725	574.909326	571.725	574.909326	Up	Down
2	575.000	570.710152	575.000	570.710152	Up	Down
3	588.550	574.773665	588.550	574.773665	Up	Down
4	592.900	588.418793	592.900	588.418793	Down	Down

Using this binary value, we calculated the Accuracy and other evaluation metrics.

```
1 accuracy_score(test_df['ActualTrend'], test_df['PredictionTrend'])
0.5652173913043478
```

	precision	recall	f1-score	support
Down	0.52	0.49	0.50	134
Up	0.60	0.63	0.62	165
avg / total	0.56	0.57	0.56	299



Observations:

- Original series was not stationary, so did one order differencing to make it stationary
- (12, 1, 0) are the p,d,q parameters used in the model
- Classification accuracy of the model is 0.565
- Looking at the portfolio growth graph, it is clear the model is very stable
- Model achieved significantly higher portfolio growth compared to buy & hold strategy

5.9. Summary of ML models

We tried five classification models and one time series model. From the above results we could see that few models performed better than baseline (Buy & Hold strategy) and few failed to meet the baseline performance. Below is the summary of model accuracies on train and test dataset.

Model	Train			Test		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Logit	0.55	0.49	0.61	0.54	0.45	0.62
Random Forest	0.63	0.63	0.61	0.55	0.50	0.60
XGB	0.54			0.54	0.48	0.59
SVM	0.57	0.58	0.56	0.54	0.49	0.57
KNN				0.52	0.48	0.55
ARIMA				0.57	0.57	0.58

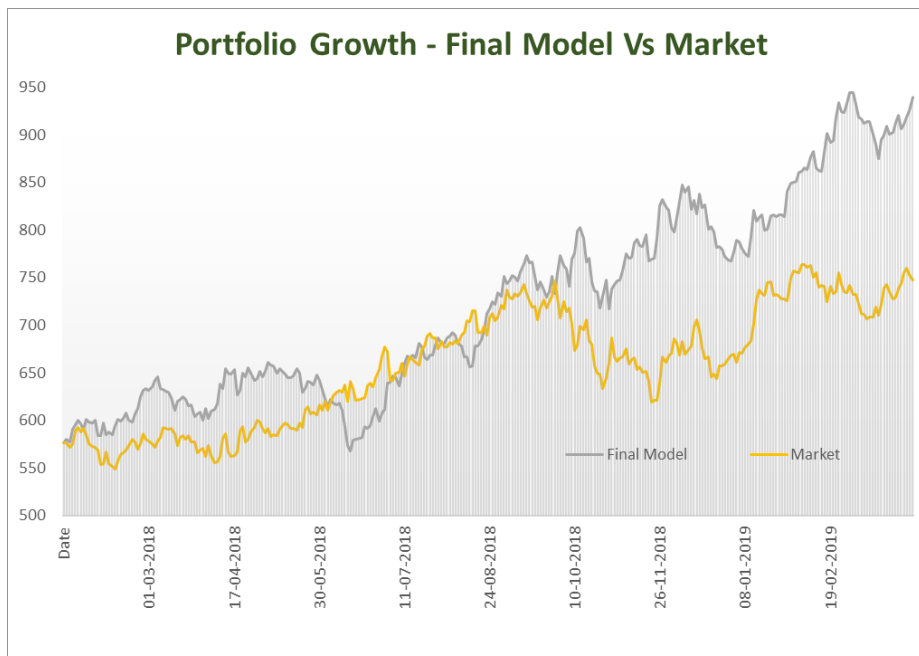
ARIMA and XGB models performed well and were stable throughout the test period. If we had to choose one model, we would choose ARIMA as it performed well both in terms of classification accuracy and portfolio growth performance.

Next we will try building a final model using voting method to make our model more robust.

6. Final Classification Model

Individual models performed with varied classification accuracy and portfolio growth performance. Now, we will try to build a robust model with better accuracy and portfolio growth performance.

We will use voting method to build a final classification model. For this we will use the predicted values of Logistic Regression, Random Forest, XGBoost, KNN and ARIMA models. We have not included SVM model in the voting process because we need to have odd number of models and SVM performed very poorly as individual model.



As expected, the final model performed very well compared to the other individual models. And it is also very stable and robust throughout the testing period. Model achieved classification accuracy of **0.57**.

7. Summary

7.1 Conclusions

- Sentiment Analysis proved helpful in predicting the stock price movement, when we added the sentiment variables to the model, we were able to improve the classification accuracy of Logistic Regression model by almost ~4%.
- Among the tried classification models, KNN performed well in portfolio growth performance.
- Random Forest and SVM models achieved good classification accuracy however failed to impress in portfolio growth performance.
- As an individual model, ARIMA time series model performed well both in terms of classification accuracy and portfolio growth performance.
- The final model built using voting method outperformed all other individual models both in classification accuracy and portfolio growth performance.

7.2 Challenges

- Collecting twitter data was difficult, the data size was huge, and it was a time-consuming process to get the data from twitter.
- Due to team members' comfortability on various ML tools we had to use multiple tools like R, Python, Excel, Tableau to complete the entire project. Hence, consolidation of the project was little challenging than expected.

7.3 Limitations

- Initially we thought of including other social media (News, blogs, ...) sentiment also to the model, however due to the time constraint and unavailability of free data we dropped this option
- We only considered historical price data and twitter sentiment data for building the model, other financial and economical features that affect stock price are not included due to either they are not freely available or very time-consuming

8. Appendix

All codes that was done for this project is either through R or Python. The EDA that I projected is done using Tableau. All the codes files are being shared separately as ".ipynb" files and not included in the report.

9. References

NSE Website- <https://www.nseindia.com/>

Twitter Feeds- <https://twitter.com/>

Bloomberg Terminal