**Data Modelling and Presentation**

**By – Piyush Bhatt**

**Table of Contents**

Abstract ............................................................................................................................................ 3

Introduction ..................................................................................................................................... 3

Methodology ................................................................................................................................... 3

Setting the Research goal ................................................................................................................ 4

Retrieving data ................................................................................................................................. 4

Data preparation .............................................................................................................................. 4

Data Exploration .............................................................................................................................. 4

Data Modelling ................................................................................................................................ 8

Results ............................................................................................................................................. 9

KNN............................................................................................................................................…. 9

Confusion Matrix.............................................................................................................................. 9

Classification Report........................................................................................................................ 9

Decision Tree..................................................................................................................................... 10

Confusion Matrix.............................................................................................................................. 10

Classification Report......................................................................................................................... 10

Discussion ....................................................................................................................................... 10

Parameter Selection ........................................................................................................................ 11

KNN................................................................................................................................................. 11

Decision Tree..................................................................................................................................... 11

Conclusion ....................................................................................................................................... 11

References ........................................................................................................................................ 12

**Abstract:**

Prediction of stock index movement will always be a challenging task. In this project we are trying to predict next day’s stock index movement with some useful features. For this different feature sets are being used. Predicting whether the stock index closing price will be lower or higher than the previous day.

The direction of the next closing price is being predicted by comparing the performance of the 2 supervised machine learning techniques namely, KNN (nearest neighbours) and Decision tree. The study focuses on the cross validation of the 2 techniques.

To predict the target variable which is LABEL in this dataset several constraints were placed on it. And finally, all the prediction with there accuracy of train and test data with all the visualisations were produced and analysed to explore the results.

**Introduction:**

Forecasting of stock market has been a vital topic of discussion in the market because of its monetary gain. Stock market is a place where companies trade their shares and a huge amount of capital is invested with a great risk. Efficient market hypothesis states that the market is efficient and cannot be predicted and stock market forecasting disapproves it by stating that financial markets are predictable.

In this project we have predicted whether the stock index closing price will be lower or higher than the previous day. By using 13 variables including the target variable. Selecting the NASDAQ stock exchange, it is an American market exchange, second largest in the world by market capitalization. In this report visualizations among the variables is carried out to determine the target variable and also checked the predictor which impact the target variable most in the prediction.

Relationship among the attributes and the hypothesis will be explained ahead in the report.

**Methodology:**

Dataset has been selected from the website **Kaggle** website, an online community of data scientist and machine learners.

Data Source: <https://www.kaggle.com/jmq19950824/stock-index-prediction-both-labels-and-features>

Dataset attributes:

* Label (Target Variable): 0(Down) and 1(UP), whether close price will be up or down.
* Open: amount at which the stock exchange will open.
* High: Highest Stock Price.
* Low: Lowest Cost Price.
* Close: amount at which the stock will close.
* Volume: volume of transaction that day.
* Interest Rate: the rate that is earned in terms of constant-[purchasing-power](https://www.nasdaq.com/investing/glossary/p/purchasing-power) dollars.
* Exchange Rate: value of one currency for the purpose of conversion to another.
* VIX: Volatility Index, measure stock market expectation of volatility.
* Gold: Price of gold.
* Oil: Price of oil per barrel.
* TEDSpread: difference between interest rate on interbank loans and on short term U.S government debt (T-bills).
* EFFR: Effective federal Funds Rate

**Data Science Process**

1. Setting the research goal
2. Retrieving the data
3. Data Preparation
4. Data Exploration
5. Data Modelling
6. Presentation and automation

**Setting the Research Goal**

The goal of this project is to develop a model through the predictor attributes and predict the label whether the stock index closing price will be lower or higher than the previous day.

Technique like data visualisation has been used to represent the relationships between attributes and their impact on the target variable. Model will be trained to identify if stock index closing price will be lower or higher.

**Retrieving Data**

Dataset was loaded into **python,** using jupyter notebook to write and run the python codes. All the libraries were read using the read.csv command.

**Data Preparation**

Retrieving data in typically raw form and may not be directly usable-

* Checking whether the loaded data is equivalent to the data in the source CSV file (using Shape command).
* Data types were checked with the **dtypes** command
* Dealing with all the potential issues/errors such as – typos, extra whitespaces, missing values.
* Extra whitespaces were removed using the strip command.
* None impossible values were found.
* Missing or null values were replaced by the mean of the attribute.

**Data Exploration**

**Single Variable Plots:**

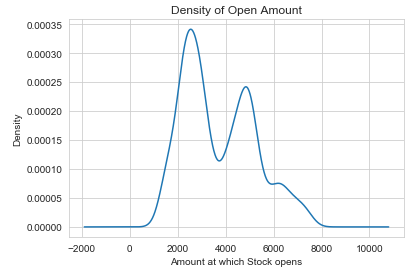
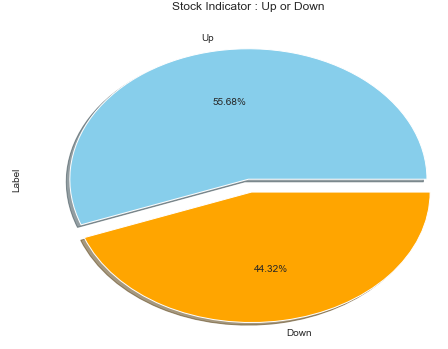


Fig-1: Stock Indicator (Label) Fig-2: Density plot of Open Amount

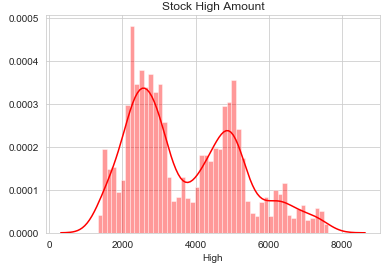
 

Fig-4: Histogram of Stock high Amount Fig-5: Timeseries graph of exchange rate

* Fig-1- shows pie chart distribution of Stock Indicator (Label:0 means down,1 means up)
* Fig-2- shows density plot of open value which is $2000.
* Fig-3- shows highest stock price went till $8000.
* Fig-4- Exchange rate started at a higher rate but gradually decreased with time.

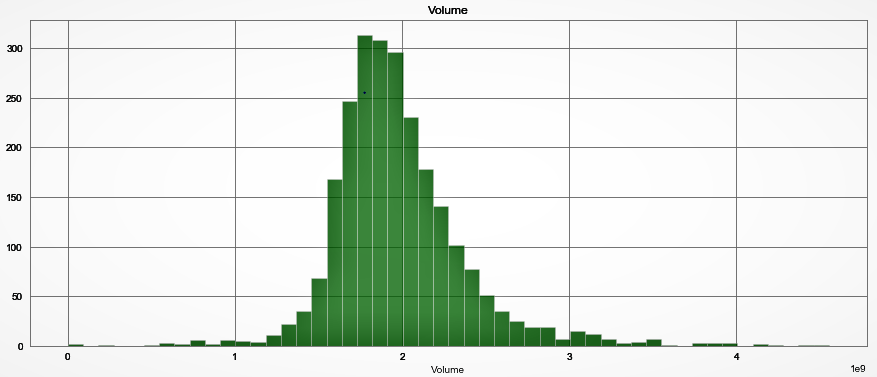
 

Fig-5: Box plot of Close Amount Fig-6: Histogram of volume attribute

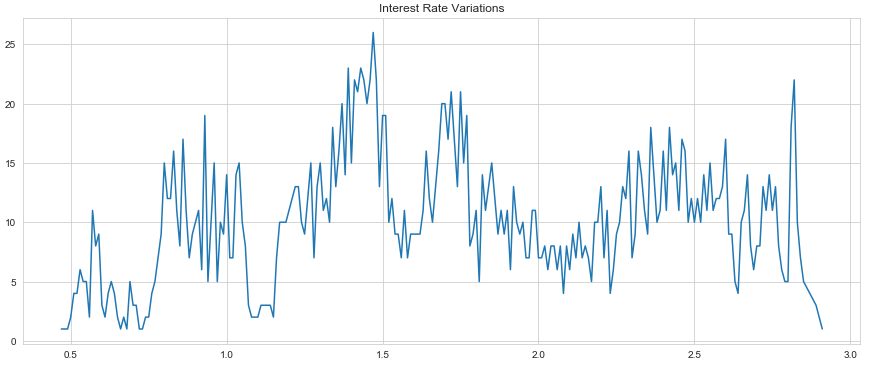


Fig- 7: Interest Rate Variations

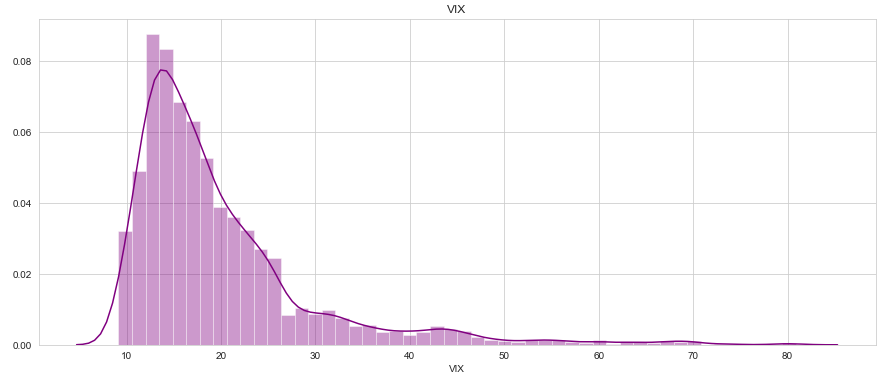


Fig-8: Distribution of VIX indicator



Fig-9: Distribution of Gold Price variation

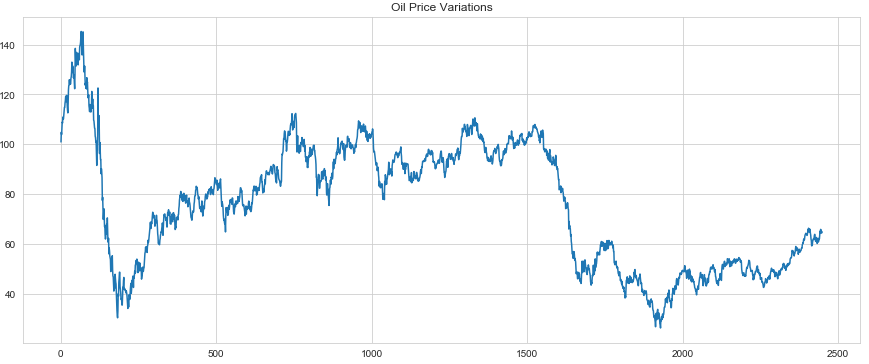
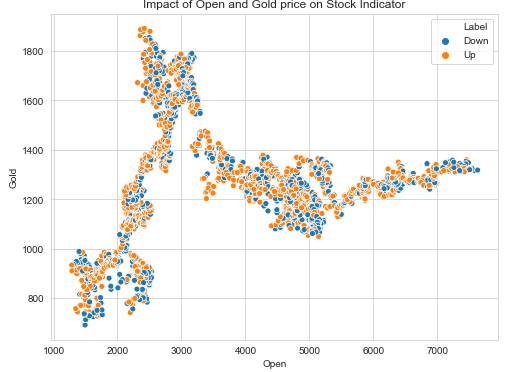
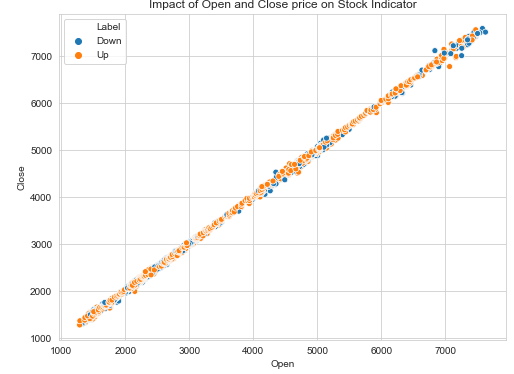
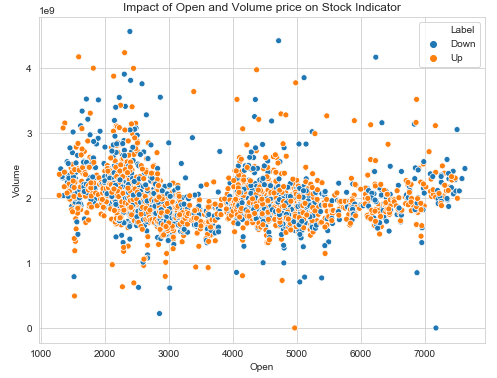
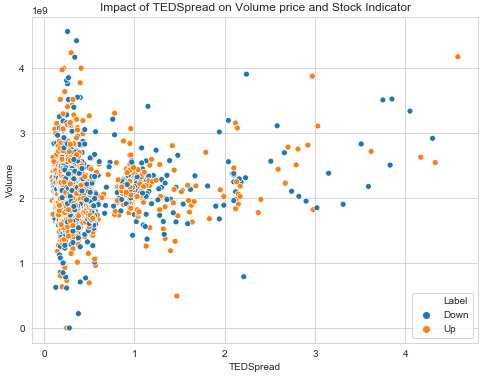


Fig-10: Distribution of Oil Price variation

* Fig-5: shows close amount distribution.
* Fig-6: shows distribution of the volume attribute.
* Fig-7: interest rate variation over the period of time
* Fig-8: volatility distribution of the stock
* Fig-9: gold price variation over the period of time
* Fig-10: oil price variation over the period of time

**Relationship between two attributes**

 Fig-11: Open and Gold price vs Stock Indicator Fig-12: Open and Close price vs stock indicator

 Fig-13: Open and Volume vs Stock Indicator Fig-14: TEDSpread vs Volume and Stock Indicator

* Fig-11: impact of Open and Gold price over stock Indicator is expected to be down.
* Fig-12: impact of Open and Close price over stock indicator is a linear towards label :down.
* Fig-13: impact of Open and Volume over stock indicator is more expected to be label :low.
* Fig-14: impact of TedSpread over stock indicator and Volume expected to be Up.

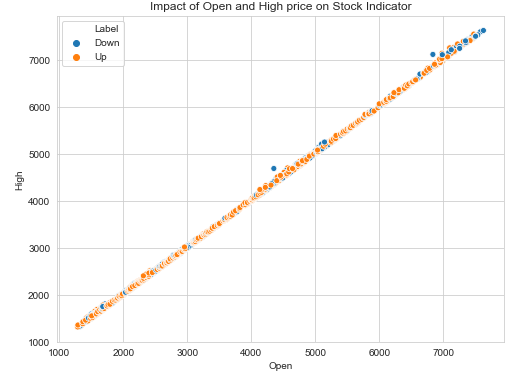
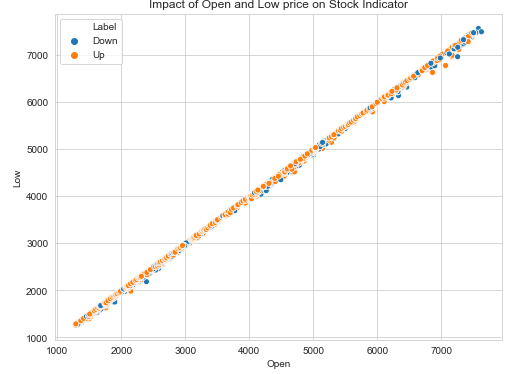
 

Fig-15: Open and High price vs Stock Indicator Fig-16: Open and Low-price vs Stock Indicator

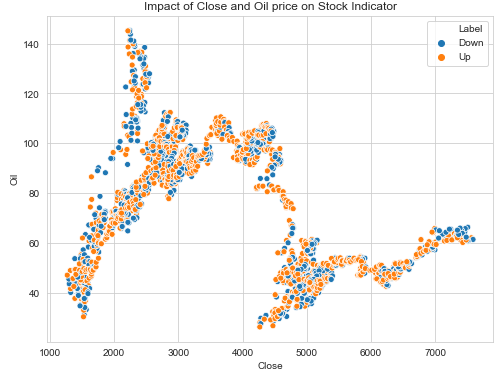
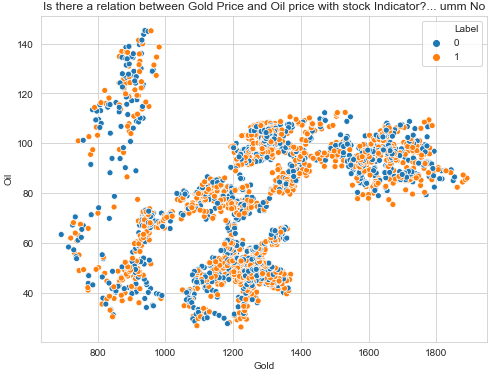
`  

Fig-17: Close and Oil price on Stock Indicator Fig-18: Gold Price and Oil price with stock Indicator

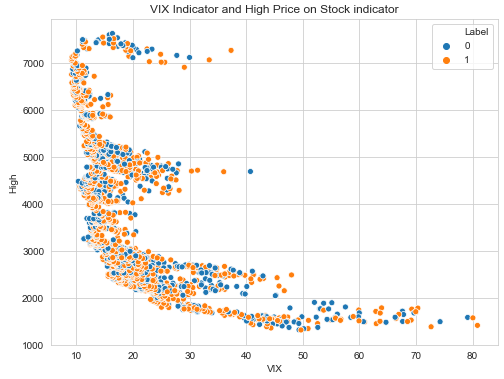
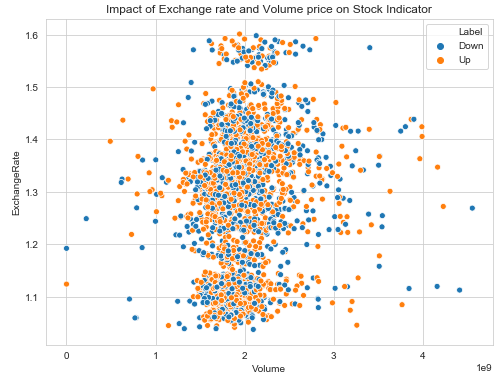
 

Fig-19: Vix and High Price vs Stock Indicator Fig-20: Exchange rate and Volume price vs Stock Indicator

* Fig-15: impact of Open and high price over stock indicator is linear-Down.
* Fig-16: impact of Open and low price over stock indicator is linear-UP
* Fig-17: impact of Close and Oil over stock indicator is expected to be down price for label.
* Fig-18: impact of Gold and Oil price over stock indicator cannot be related.
* Fig-19: impact of VIX and High price over stock indicator is expected to be high.
* Fig-20: impact of Exchange Rate and Volume price over stock indicator to be equal or similar.

**Data Modelling**

Data Modelling is one of the important phase of Data Science process. In the following step the best model is fit to the data among all the candidate models to get most accurate results.

The modelling phase consists of four steps:

* Feature engineering and model selection.
* Training the model.
* Model validation and selection.
* Applying the trained model to unseen data.

KNN (Nearest neighbour), and Decision Tree are the most commonly used classification Models, to generate the accurate prediction results. High accuracy model is not an easy task so its important that we use all parameters to get the good model.

Splitting the dataset into two parts:

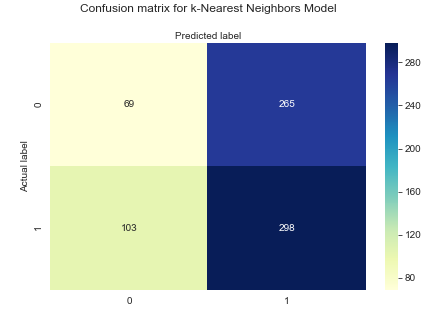
* Training dataset
* Test dataset

**Results**

**KNN**

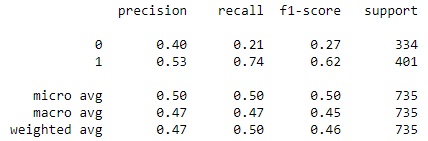
Applying knn modelling:

**Confusion matrix:**



* Total Predictions were 735.
* Out of total predictions classifier predicted (69 times Down and 298 times Up).

**Classification Report:**

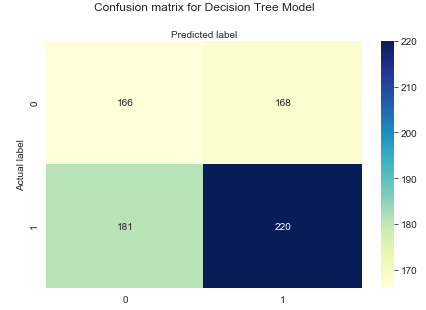


From the above results it can be said that correctly predicted instances are 0.47 and the fraction of relevant instances that are fully predicted are 0.50.

**Decision Tree**

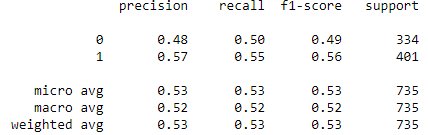
Applying decision tree modelling:

**Confusion Matrix**:



* Total Predictions were 735.
* Out of total predictions classifier predicted (166 times Down and 220 times Up).

**Classification Report:**



From the above results it can be said that correctly predicted instances are 0.53 and the fraction of relevant instances that are fully predicted are 0.53.

**Discussion**

KNN and Decision tree were used for data modelling to predict whether the stock market will open at a lower or a higher value than the previous day. In order to have best classifiers, it is necessary to find parameters that fit our model better.

In order to find better classifiers, further will look for parameters that better fitted our model.

**Parameter Selection**

Parameter selection details of each classifier are given below:

**KNN**

Used parameter sets to give good accuracy.

* N\_neighbours: 29
* Weights: default-uniform
* Metric: default- ‘minkowski’
* P: 2

Through cross validation best optimal values were found. If the results are observed then there are more chances of closing of stock exchange higher than the previous day. So uniform weights are used to eliminate unnecessary influence on the neighbours as uniform means all points in the neighbour are equally weighted. Default metric is the **minkowski**, and with p=2 is equivalent to the

standard Euclidean metric. 10-Fold cross validation is performed in which the data was slit ten times and every time one part was used as test data and other part was used as training data.

**Decision Tree**

Following et of parameter used:

* Minimum samples for a node split
* Minimum depth of tree
* Maximum features to consider split
* Maximum number of terminal nodes

One of the key challenges is overfitting to overcome that auto was selected. For minimum sample split we selected a range of 2 to 15 and for minimum sample leaf we selected 1 to 11 which the model selected 11 and for split 1. For choosing higher value of split is so that our model can learn and explore. Most optimal values of the parameters were found using the cross validation.

**Conclusion**

From the above observations of all the steps and good understanding of all the data science concepts a proper model was developed to predict whether the stock market will open at higher or a lower price than the previous day.

All the steps of the data science process were smoothly implemented for the good accurate results and data exploration with removal all the potential issues/errors in the data. Accurate parameters necessary for accurate modelling of the data. Custom parameter models gave better results then the default parameter models.

After training two different models to get the better and accurate results for the better fit to our dataset and it was found that **Decision Tree** to be the best fitted model to the dataset and has better weighted averages.

**References**

1. Week-6: Classification (3), Practical Data Science, RMIT
2. Dataset Source: <https://www.kaggle.com/jmq19950824/stock-index-prediction-both-labels-and-features>
3. Extracting the best features for predicting stock prices using machine learning:

<http://worldcomp-proceedings.com/proc/p2012/ICA4753.pdf>

1. Automatic prediction of stock price direction: <http://www.cs.rug.nl/~george/wp-content/uploads/2016/08/BaldacchinoCharlot.pdf>
2. Stock Prices Prediction Using Machine Learning: <https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/>
3. Using Machine Learning Classifiers to Predict Stock Exchange Index: <http://www.ijmlc.org/vol7/614-A101.pdf>