Group members:

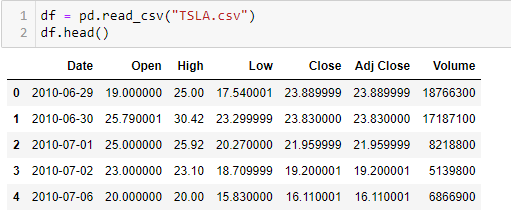
Muhammad abdullah(p19-0036)  ABDULLAH TAHIR(P19-0067)

Stock prediction using machine learning

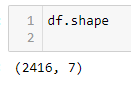
**Importing Dataset**

The dataset we will use here to perform the analysis and build a predictive model is Tesla Stock Price data. We will use OHLC(‘Open’, ‘High’, ‘Low’, ‘Close’) data from 1st January 2010 to 31st December 2020 which is for 11 years for the Tesla stocks.

You can download the CSV file from: https://www.kaggle.com/datasets/timoboz/tesla-stock-data-from-2010-to-2020 (you will have to replace “-” in date with “/”).



From the first five rows, we can see that data for some of the dates is missing the reason for that is on weekends and holidays Stock Market remains closed hence no trading happens on these days.

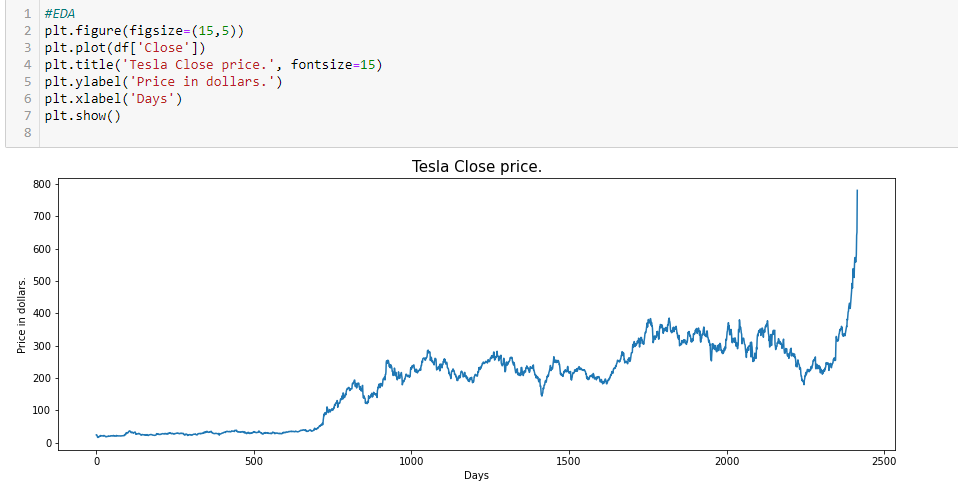


From this, we learned that there are 2416 rows of data available and that we have 7 different features or columns for each row.

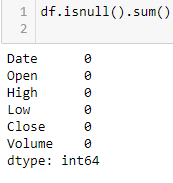
**Exploratory Data Analysis**

EDA is an approach to analyzing the data using visual techniques. It is used to discover trends, and patterns, or to check assumptions with the help of statistical summaries and graphical representations.

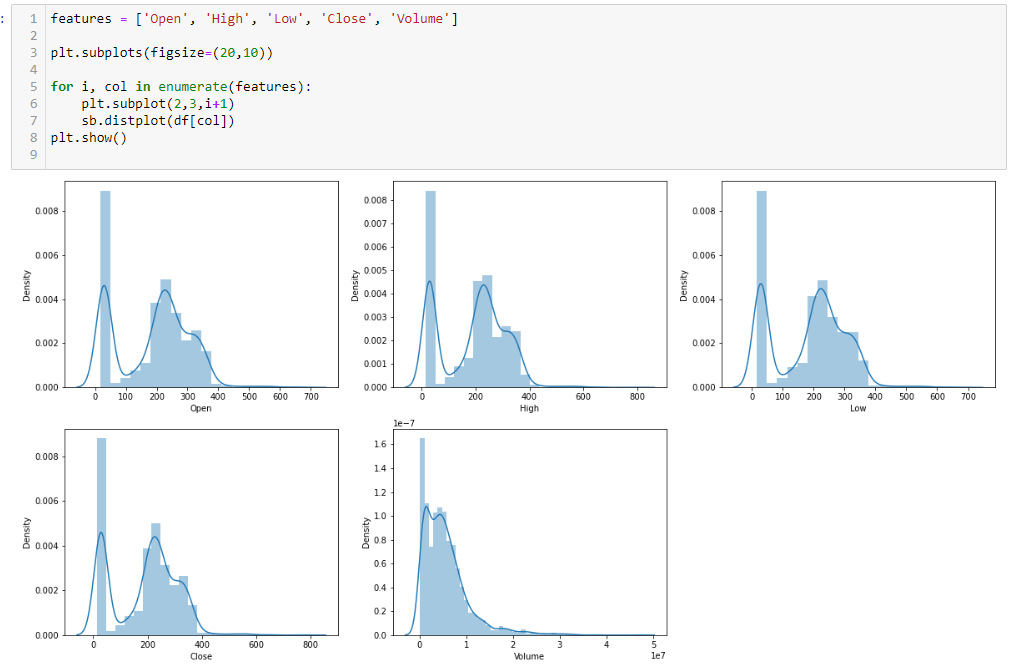
While performing the EDA of the Tesla Stock Price data we will analyze how prices of the stock have moved over the period and how the end of the quarters affects the prices of the stock.



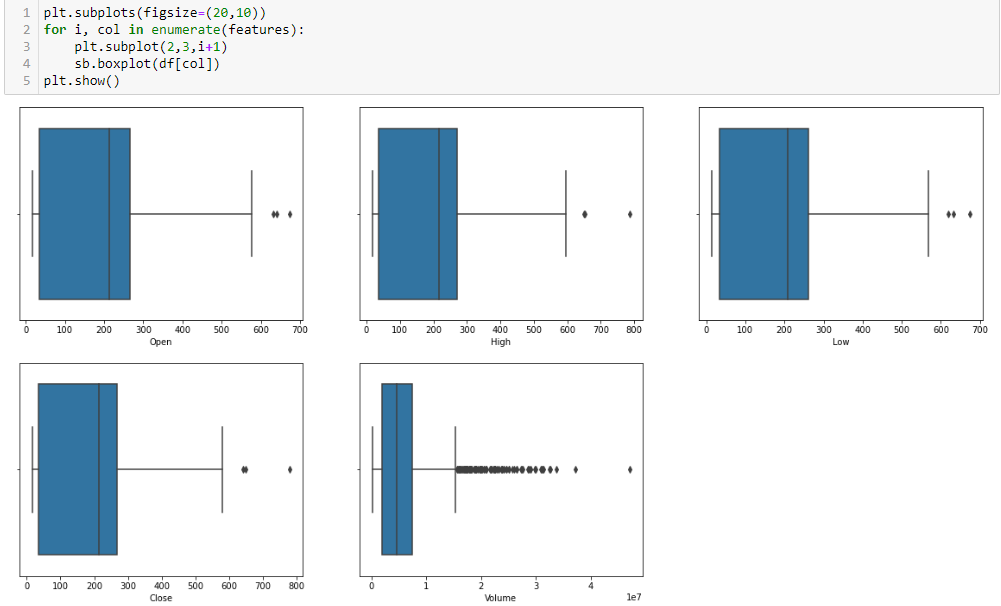
All the rows of columns ‘Close’ and ‘Adj Close’ have the same data. So, having redundant data in the dataset will not help, so we’ll drop this column before further analysis.



This implies that there are no null values in the data set provided.



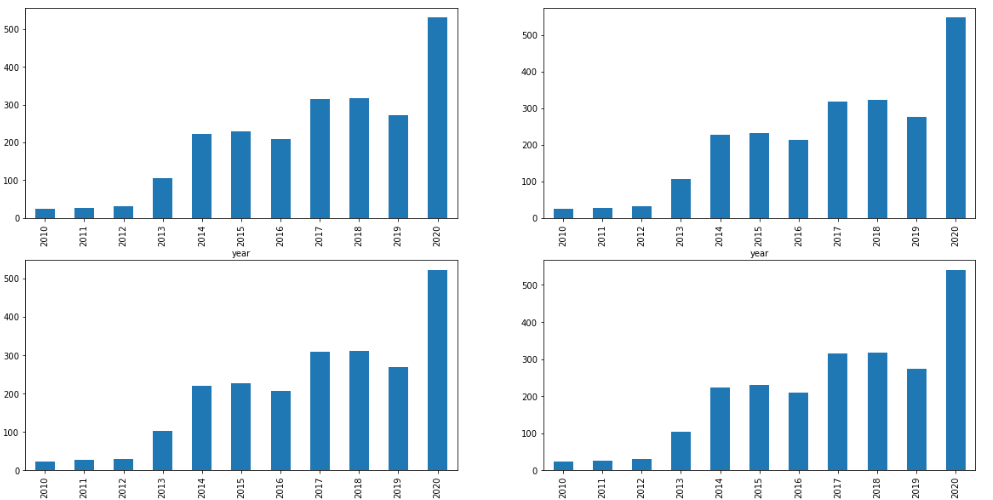
In the distribution plot of OHLC data, we can see two peaks which means the data has varied significantly in two regions. And the Volume data is left-skewed.



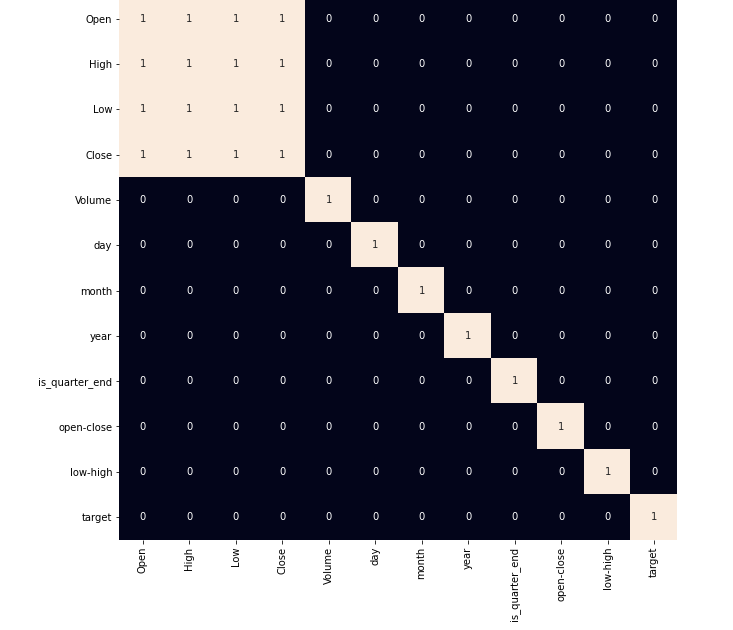
From the above boxplots, we can conclude that only volume data contains outliers in it but the data in the rest of the columns are free from any outliers.

**Feature Engineering**

Feature Engineering helps to derive some valuable features from the existing ones. These extra features sometimes help in increasing the performance of the model significantly and certainly help to gain deeper insights into the data.



From the above bar graph, we can conclude that the stock prices have doubled from the year 2013 to that 2014 and from 2019 to 2020.



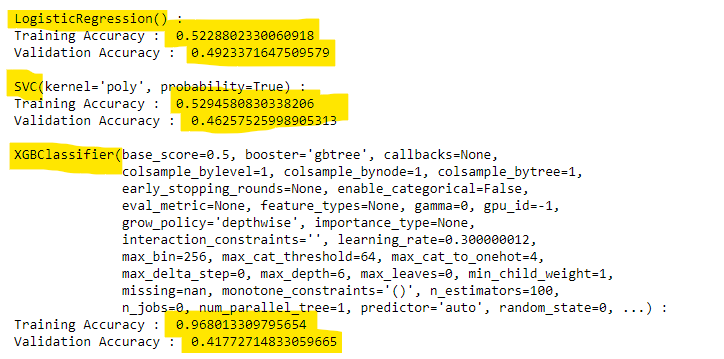
From the above heat map, we can say that there is a high correlation between OHLC which is pretty obvious, and the added features are not highly correlated with each other or previously provided features which means that we are good to go and build our model.

**Data Splitting and Normalization**

After selecting the features to train the model on we should normalize the data because normalized data leads to stable and fast training of the model. After that whole data has been split into two parts with a 90/10 ratio so, that we can evaluate the performance of our model on unseen data.

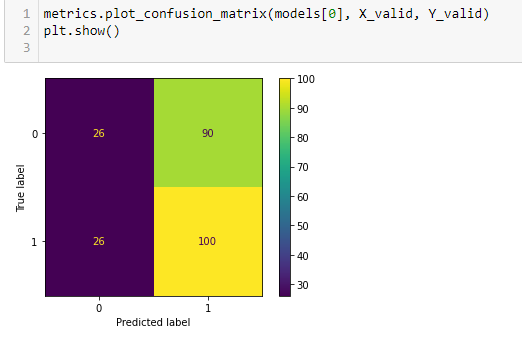
**Model Development and Evaluation**

Now is the time to train some state-of-the-art machine learning models(Logistic Regression, Support Vector Machine, XGBClassifier), and then based on their performance on the training and validation data we will choose which ML model is serving the purpose at hand better.

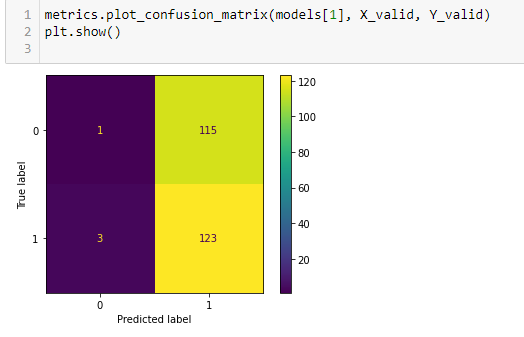
Among the three models, we have trained XGBClassifier has the highest performance but it is pruned to overfitting as the difference between the training and the validation accuracy is too high. But in the case of Logistic Regression, this is not the case. 

Now let’s plot a confusion matrix for the validation data.

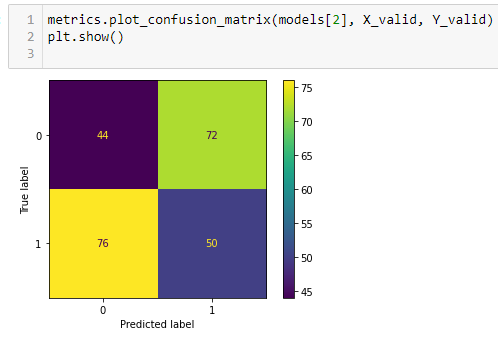
1. Logistic Regression



1. SVC



3) XGBClassifier

****

**Conclusion:**

We can observe that the accuracy achieved by the state-of-the-art ML model is no better than simply guessing with a probability of 50%. Possible reasons for this may be the lack of data or using a very simple model to perform such a complex task as Stock Market prediction.