Report –

Embarking on a stock price prediction project aligns with my keen interest in navigating the challenge of deciphering market trends and honing predictive skills that could shape future investment strategies. And, Tesla, known for its innovations and market volatility, offers a captivating subject for such a project. I have conducted an in-depth analysis of Tesla's stock price data, spanning a decade from 2010 to 2020, encompassing OHLC, i.e., open, high, low, and close prices and volume of stock. These essential price points were collectively utilized to assess price dynamics, trends, volatility, and trading patterns within the stock and financial instruments. The data contains daily records of total 2416 data points. I have performed EDA to analyze how prices of the stock have moved over the period of time and how the end of the quarters affects the stock prices. The plot of Tesla closing price shows an upward trend with highest peak reaching in the year 2020. The distribution plot of OHLC data depicts two peaks which means the data has varied significantly in two regions. And the Volume of stock price is highly positively skewed signifying an uneven distribution of trading volumes, with a pronounced tail extending toward higher volume values. This pattern suggests that while most trading sessions exhibit relatively modest trading activity, a few sessions experience notably elevated levels of trading volume. The boxplot illustrating the volume of stock price data reveals a profusion of outliers, indicating a plethora of instances where trading activity significantly deviates from the norm. These outliers represent remarkable surges or plunges in trading volume that extend beyond the usual range. For predictive modeling, I have done feature engineering by creating two more columns, one by taking the difference between Low and High and other between Open and Close in similar fashion."Open minus Close" reflects whether a stock's closing price is higher or lower than its opening price, indicating potential shifts in market sentiment during the trading session. On the other hand, "Low minus High" measures the intraday volatility by quantifying the difference between the lowest and highest prices. A new column named as target has been created. The target value is an indicator variable that indicates whether in the next day Close stock price will rise or fall, if target is 1 then it is profitable to invest on buying stock. From pie chart of the target variable it is seen that around 48.9% cases the closing stock price has gone higher in the next day. Also, the data shows that prices and the volume of trades are higher in the months which are quarter end, i.e., March, June, September and December as compared to that of the non-quarter end months. The project aims to develop a classification model to predict whether tomorrow's stock price will rise or fall based on features such as open-close, low-high prices, and the is\_quarter\_end indicator. Prior to building the model, the correlation among these predictor variables is assessed. As no significant correlation is found in the heatmap, we are good to go.After train test split, I have applied 3 classification models -- Logistic regression, Support vector machine and Extreme Gradient Boosting among which I choose the Logistic regression model to be the best one as it shows the highest validation accuracy.

In depth analysis –

Data description :

* The dataset I will use here to perform the analysis and build a predictive model is Tesla Stock Price data. I will use OHLC(‘Open’, ‘High’, ‘Low’, ‘Close’) data from 1st January 2010 to 31st December 2020 which is for 10 years for the Tesla stocks.
* From the first five rows, I can see that data for some of the dates is missing the reason for that is on Iekends and holidays Stock Market remains closed hence no trading happens on these days.
* If I observe carefully I can see that the data in the ‘Close’ column and that available in the ‘Adj Close’ column. Hence, I drop the Adj column.
* There are total 6 columns and 2416 rows in the data.
* Data summary :

Open High Low Close Volume

count 2416.000000 2416.000000 2416.000000 2416.000000 2.416000e+03

mean 186.271147 189.578224 182.916639 186.403651 5.572722e+06

std 118.740163 120.892329 116.857591 119.136020 4.987809e+06

min 16.139999 16.629999 14.980000 15.800000 1.185000e+05

25% 34.342498 34.897501 33.587501 34.400002 1.899275e+06

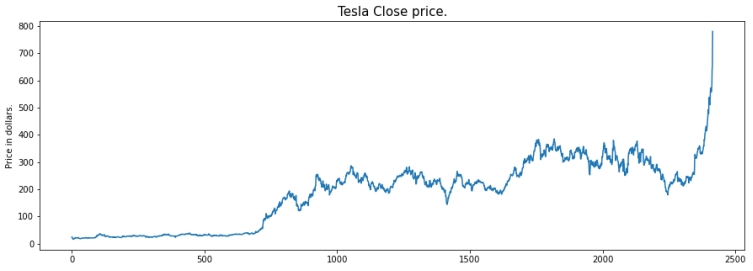
50% 213.035004 216.745002 208.870002 212.960007 4.578400e+06

75% 266.450012 270.927513 262.102501 266.774994 7.361150e+06

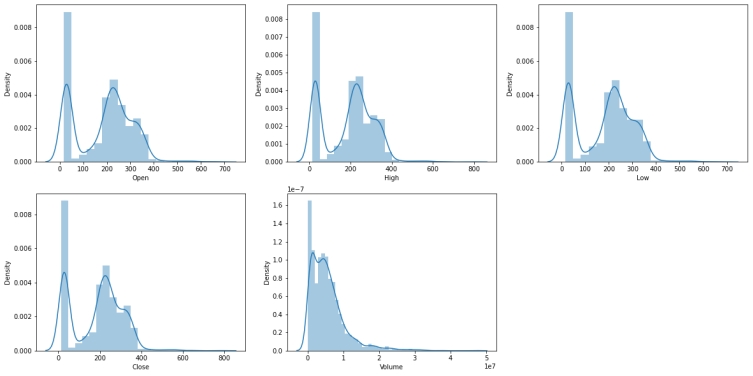
max 673.690002 786.140015 673.520020 780.000000 4.706500e+07

EDA :

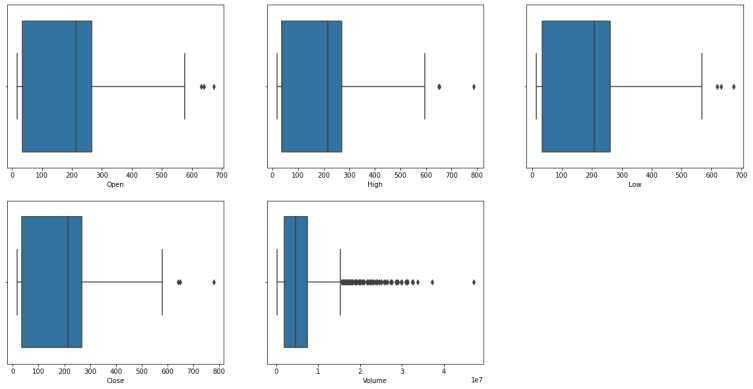
1. While performing the EDA of the Tesla Stock Price data I will analyze how prices of the stock have moved over the period of time and how the end of the quarters affects the prices of the stock.



* The prices of tesla stocks are showing an upward trend as depicted by the plot of the closing price of the stocks.
* From around 19th June 2013, a substantial increase is noticed.
* From around 24th January 2020, another leap is noticed.



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

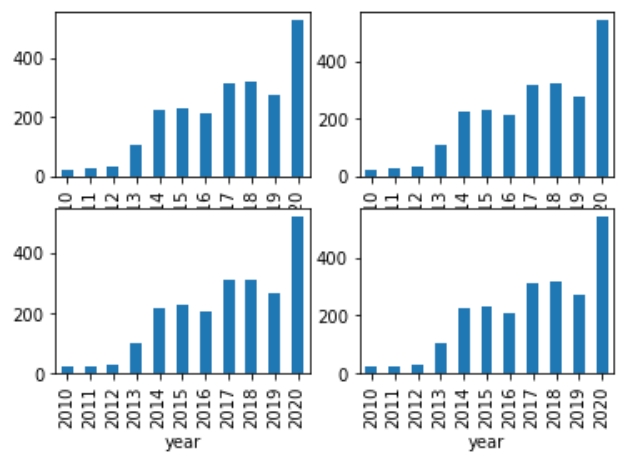


1. From the boxplot I can say

* The volume column contains many outliers and the rest of the columns contain very few outliers.
* The distributions of all variables are positively skeId.
* Except volume, the median value for OHLC price is a slight more than 200.

Feature Engineering :

1. From existing date column, I extracted 3 more columns day, month and year.
2. A quarter is defined as a group of three months. Every company prepares its quarterly results and publishes them publicly so, that people can analyze the company’s performance. These quarterly results affect the stock prices heavily which is why I have added this feature because this can be a helpful feature for the learning model.



1. The bar chart shows –
   * OHLC Stock price has continuously increased over the years but stock price has magnificently increased in 2020.
   * The stock price increased at a more rapid rate from 2013 and it got doubled in 2014.
2. day month year open-close low-high target

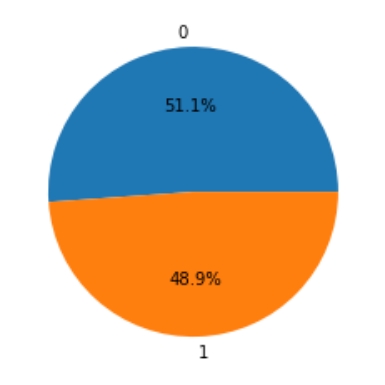
is\_quarter\_end

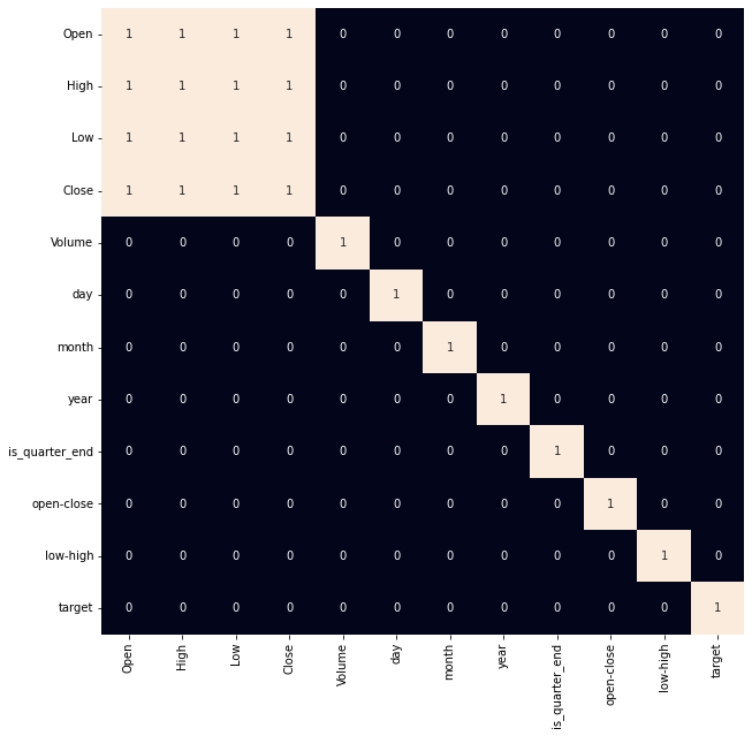
0 6.635860 15.457794 2014.771411 -0.183247 -6.624233 0.497227

1 6.664565 16.343001 2014.788146 -0.028650 -6.738033 0.539723

* Prices are higher in the months which are quarter end as compared to that of the non-quarter end months.
* The volume of trades is lower in the months which are quarter end.

1. Above I have added some more columns which will help in the training of our model. I have added the target feature which is a signal whether to buy or not I will train our model to predict this only. But before proceeding let’s check whether the target is balanced or not using a pie chart.





1. From the above heatmap, I can say that there is a high correlation between OHLC that 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.

1. After selecting the features to train the model on I normalized the data 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 I can evaluate the performance of our model on unseen data.

1. Training set contains 2174 rows and test set contains 242 rows.

Model :

I have applied Logistic regression, Support vector machine and XGBClassifier. The response variable is the target and the features are (open-close), (low-high) and quarter\_end(binary 0 or 1).

LogisticRegression() :

Training Accuracy : 0.5181036927909507

Validation Accuracy : 0.5318828680897646

SVC() :

Training Accuracy : 0.4721484905828115

Validation Accuracy : 0.48741105637657356

XGBClassifier() :

Training Accuracy : 0.9603851543284354

Validation Accuracy : 0.455836070060208

Among the three models, I 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 the Logistic Regression, this is not the case.

Evaluation :

For the evaluation metric, I will use the ROC-AUC curve but why this is because instead of predicting the hard probability that is 0 or 1 I would like it to predict soft probabilities that are continuous values between 0 to 1. And with soft probabilities, the ROC-AUC curve is generally used to measure the accuracy of the predictions.

The confusion matrix for validation data -

