Bitcoin Price Prediction using Machine Learning in Python

[Data Analysis]

By – Mahak Mishra

Bitcoin Price Prediction

AIM - To predict a signal that indicates whether buying a particular stock will be helpful or not by using ML.

Data Set

	Date	Open	High	Low	Close	Adj Close	Volume
0	17-09-2014	465.864014	468.174011	452.421997	457.334015	457.334015	21056800.0
1	18-09-2014	456.859985	456.859985	413.104004	424.440002	424.440002	34483200.0
2	19-09-2014	424.102997	427.834991	384.532013	394.795990	394.795990	37919700.0
3	20-09-2014	394.673004	423.295990	389.882996	408.903992	408.903992	36863600.0
4	21-09-2014	408.084991	412.425995	393.181000	398.821014	398.821014	26580100.0

Head of dataset

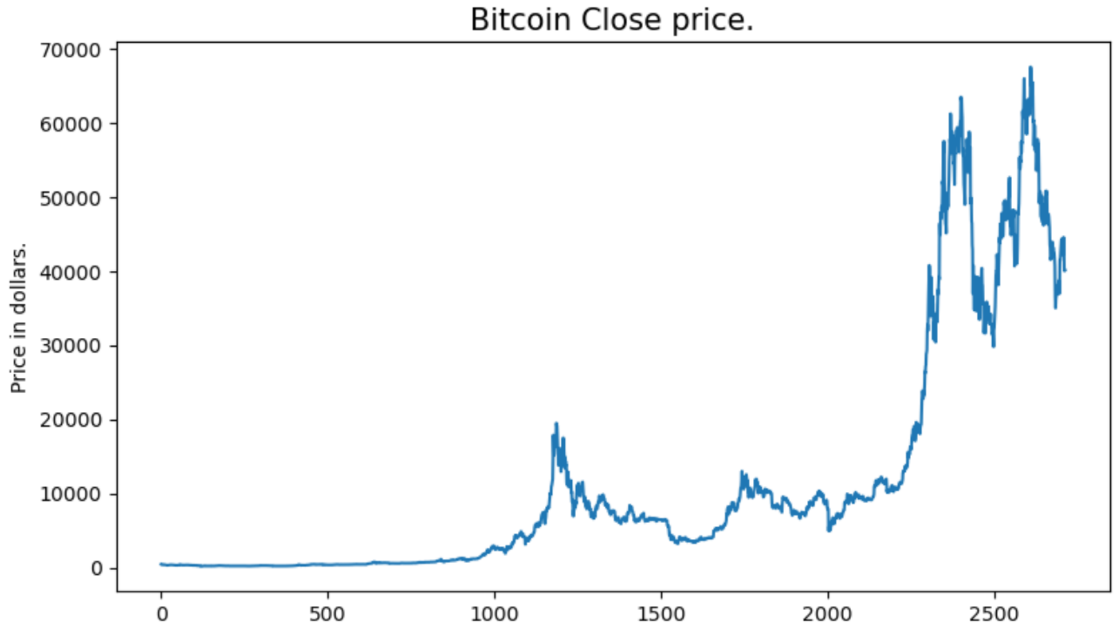
Here we have2904 rows of data available and for each row, we have 7 different features or columns.

	Open	High	Low	Close	Adj Close	Volume
count	2713.000000	2713.000000	2713.000000	2713.000000	2713.000000	2.713000e+03
mean	11311.041069	11614.292482	10975.555058	11323.914637	11323.914637	1.470462e+10
std	16106.428892	16537.390649	15608.572561	16110.365010	16110.365010	2.001627e+10
min	176.897003	211.731003	171.509995	178.102997	178.102997	5.914570e+06
25%	606.396973	609.260986	604.109985	606.718994	606.718994	7.991080e+07
50%	6301.569824	6434.617676	6214.220215	6317.609863	6317.609863	5.098183e+09
75%	10452.399410	10762.644530	10202.387700	10462.259770	10462.259770	2.456992e+10
max	67549.734380	68789.625000	66382.062500	67566.828130	67566.828130	3.509680e+11

First five row of the data

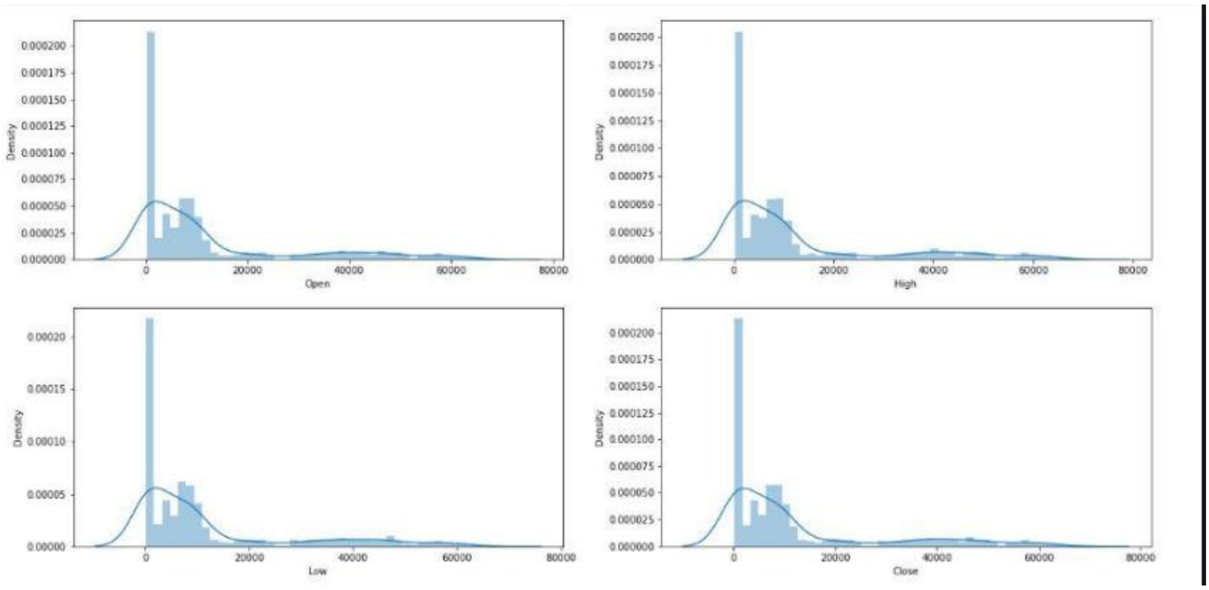
Exploratory Data Analysis

While performing the EDA of the Bitcoin Price data we will analyze how prices of the cryptocurrency have moved over the period of time and how the end of the quarters affects the prices of the currency.

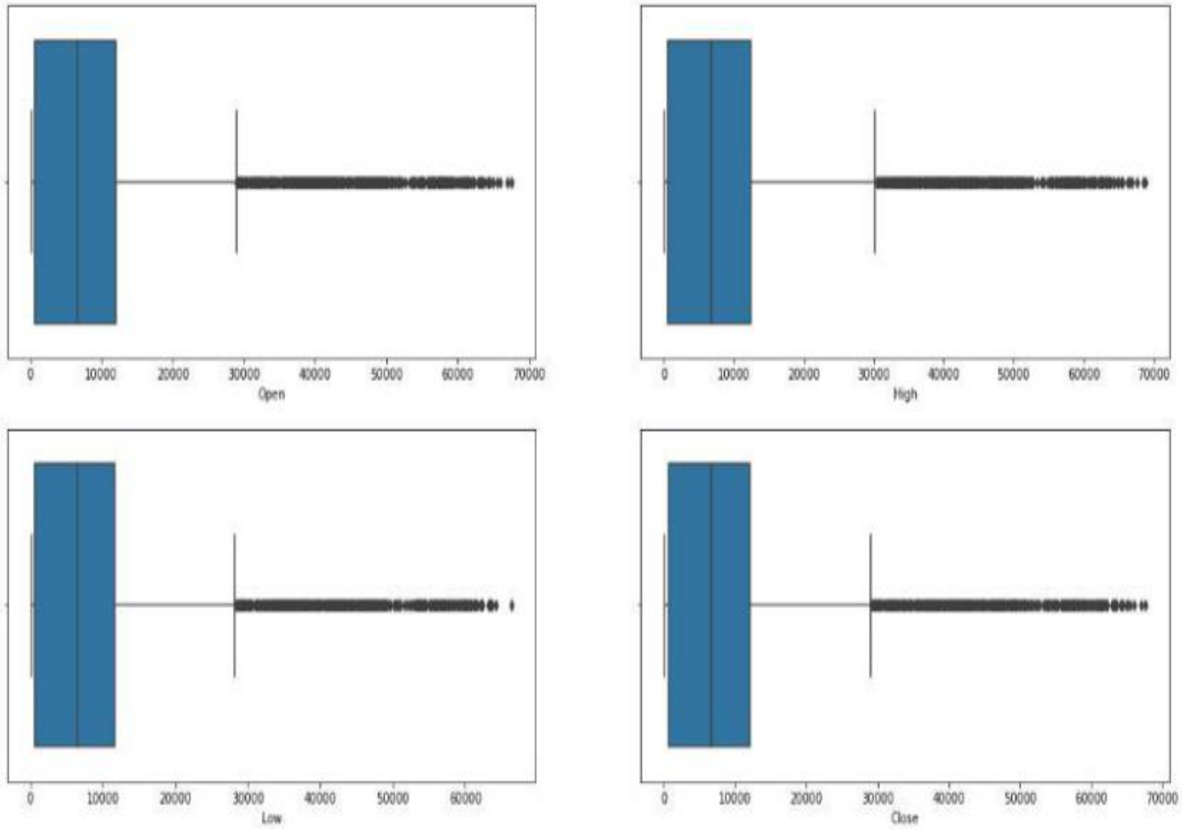


Variation in the price of cryptocurrency

The prices of the Bitcoin stocks are showing an upward trend as depicted by the plot of the closing price of the stocks.



Distribution plot of the OHLC data

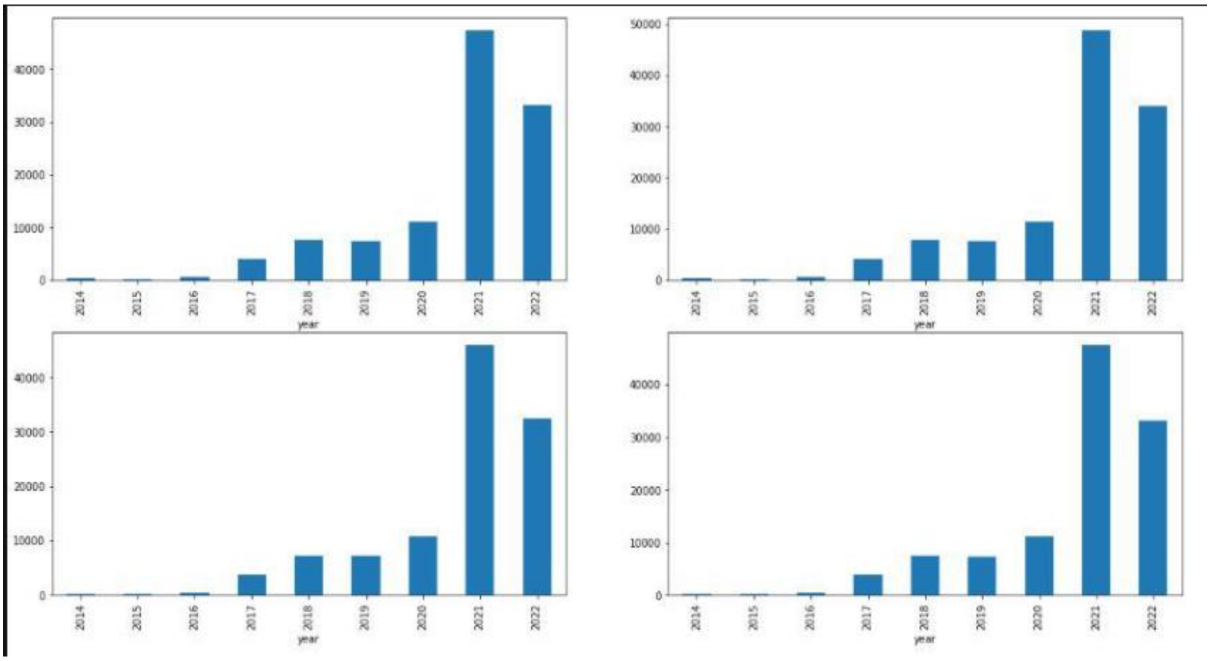


Boxplot of the OHLC data

There are so many outliers in the data which means that the prices of the stock have varied hugely in a very short period of time. Let's check this with the help of a barplot .

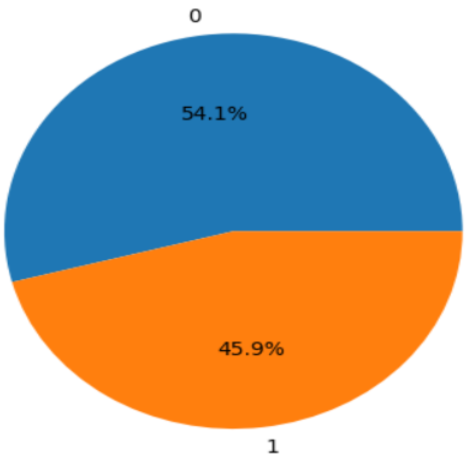
Feature Engineering

Feature Engineering helps to derive some valuable features from the existing ones.



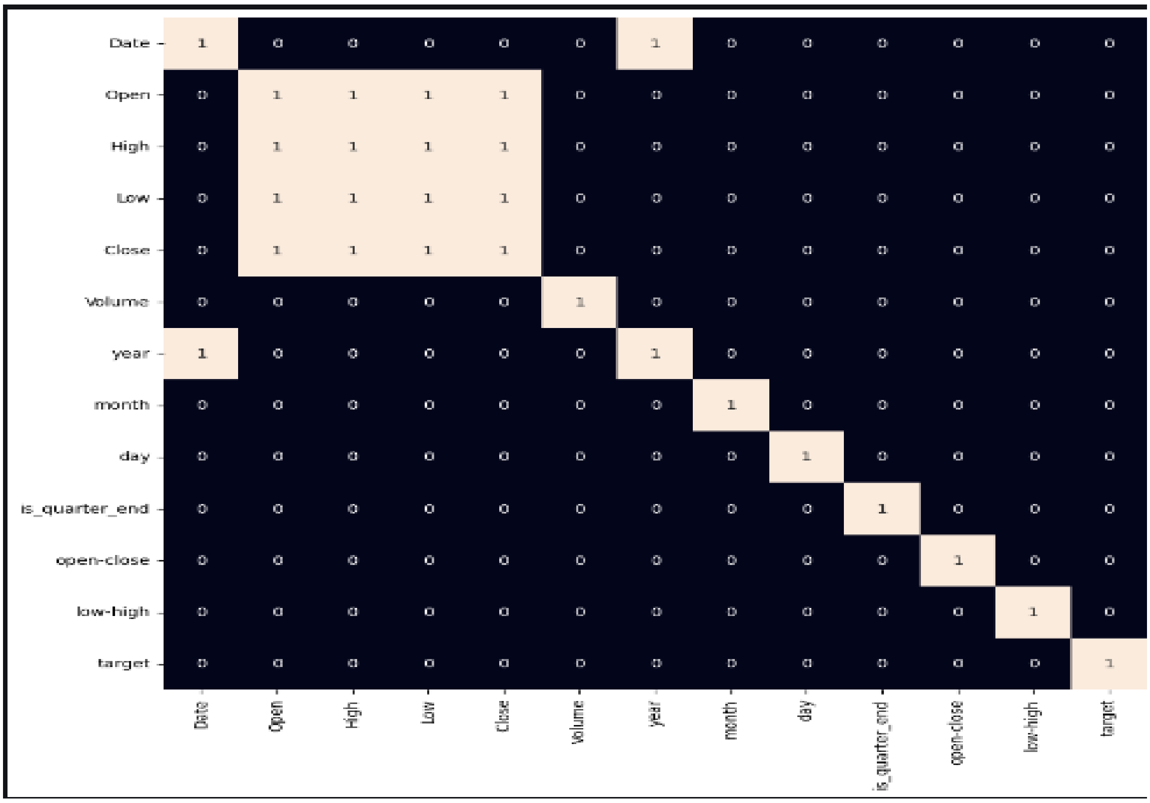
Barplot of the mean price of the bitcoin year wise

Here we can observe why there are so many outliers in the data as the prices of bitcoin have exploded in the year 2021.



Pie chart for data distribution across two labels

When we add features to our dataset we have to ensure that there are no highly correlated features as they do not help in the learning process of the algorithm.

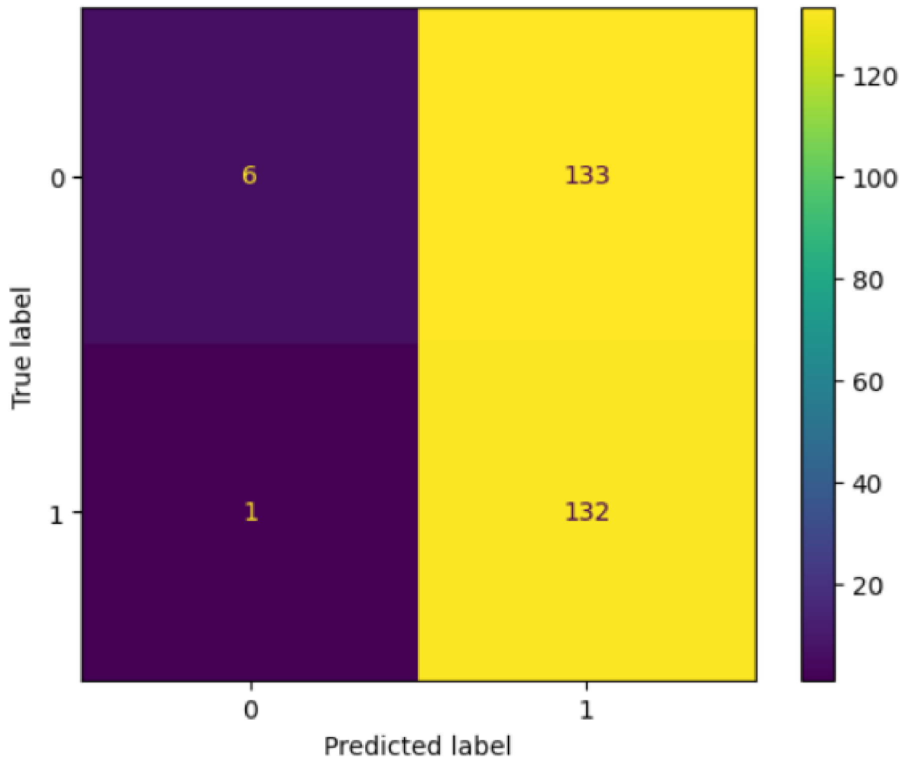


Heatmap to find the highly correlated features

From the above Heatmap , 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.

- 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



Confusion matrix for the validation data

The graph displayed is a **confusion matrix** commonly used in classification problems to evaluate the performance of a model. Here's an analysis of this confusion matrix:

Confusion Matrix Breakdown:

- **True Negatives (Top-Left):** 6
  - The model correctly predicted 6 instances as negative (class 0), which matched the true labels.
- **False Positives (Top-Right):** 133
  - The model incorrectly predicted 133 instances as positive (class 1), although they were actually negative (class 0).
- **False Negatives (Bottom-Left):** 1
  - The model incorrectly predicted 1 instance as negative (class 0), although it was actually positive (class 1).

- **True Positives (Bottom-Right):** 132
    - The model correctly predicted 132 instances as positive (class 1), which matched the true labels.

Analysis:

- **High Recall, Low Precision**: The model performs well at detecting actual positives (high recall of ~99.2%) but has low precision (~49.8%), meaning that a significant portion of its positive predictions were incorrect.
- **Imbalance**: The high number of false positives (133) compared to true negatives (6) suggests a potential issue with the model's threshold, training data balance, or an inherent difficulty in distinguishing between the two classes in this dataset.