PRML Bonus Project

Bitcoin Price Prediction

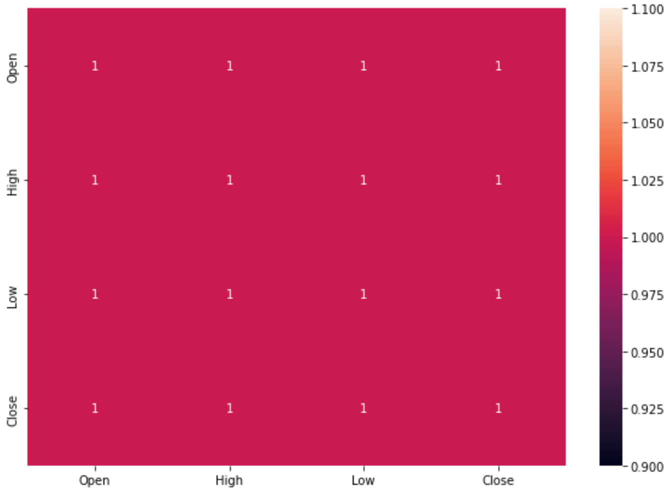
Priyank Mandal (B20AI055)

Introduction:

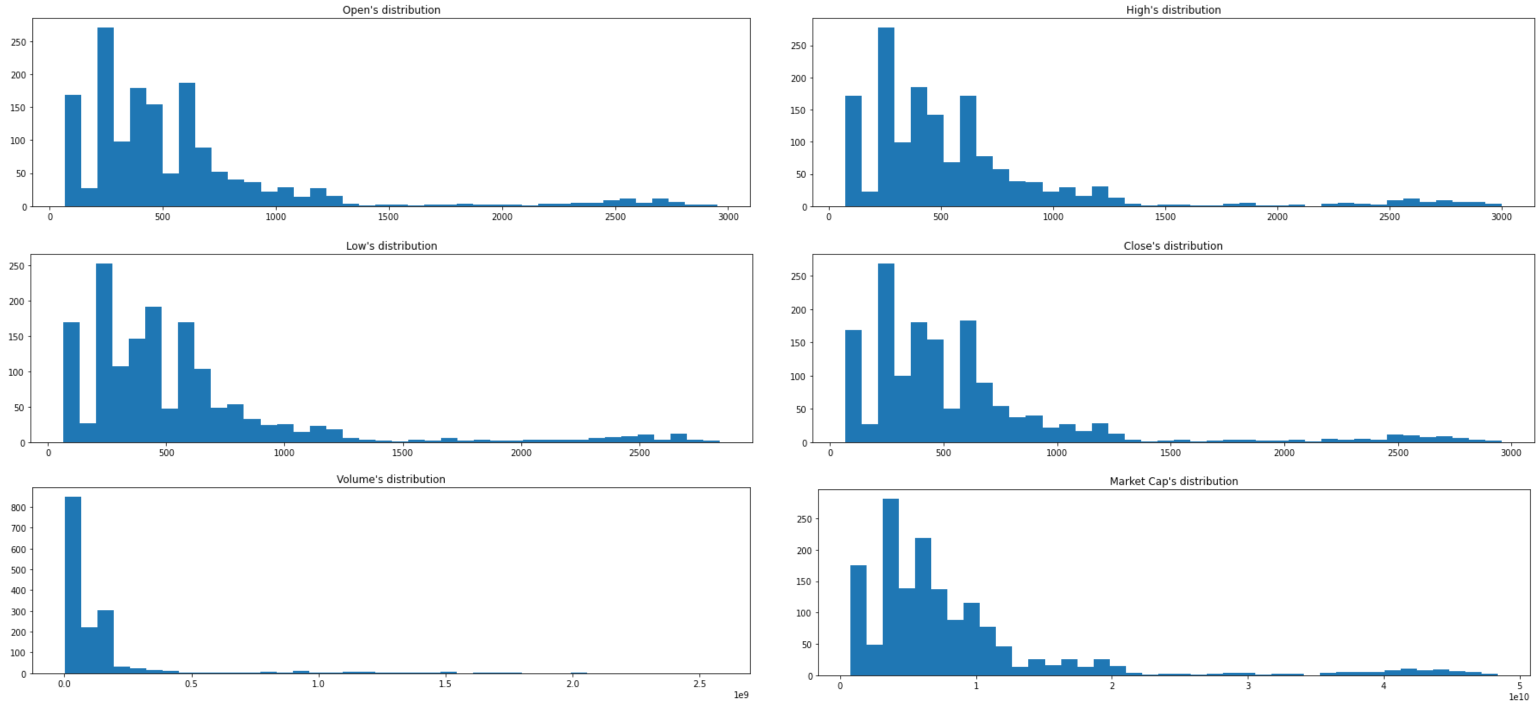
The dataset has data on the Bitcoin prices from the year of Apr,2013 to Jul,2017.

Coming to the EDA and preprocessing of the dataset, firstly we can see that the date column of the dataset isn’t of any meaning to predict the Close variable or the bitcoin price, as a result we can drop this column. Now for the columns Volume and Market Cap the readings are in string format which cannot be used for the prediction of the prices. So the dataset needs to be iterated over these columns and then the commas were removed and then this string was then type casted into integer for the further processing.

Now the data is continuous in nature, we can plot the heatmap of the features of the dataset.



As we can see from the heatmap that the correlation between the features are close to 1 and that means that all the features are highly correlated to each other.



The above plots are the distribution plots of the feature data.

From the above distribution plots and the heatmap we can see that the features are highly correlated and also the data is also continuous hence the regression models will be suitable to predict the Bitcoin Prices.

Hence the models that are chosen for this dataset are as follows:

- **Linear Regression**
- **LightGBM Regressor**
- **Support Vector Regressor**
- **K-Neighbor Regressor**

Experimentation and Hyperparameter Optimization

The dataset was first split into training and testing sets for analyzing the performance of various models.

The base models (without the parameter tuning) were trained on the training dataset and the performance was reported.

Metrics used for the performance evaluation are:

- **R2 - Score**
- **Mean Squared Error**
- **Mean Absolute Error**

The metrics of the base models are as follows:

Models	R2	MSE	MAE
- Linear Regression	0.999	160.146	6.439
- LGBM Regressor	0.997	745.531	11.473
- SVR	0.253	20069	206.286
- KNeighborsRegressor	0.988	3150	30.247

As we can see from the metrics of the base models, they need hyper tuning in order to get better results (especially SVR). The exception here is the Linear Regression model from Sklearn which already has a very high R2 score compared to all the other models. The Hypertuning can't be done on Linear Regression as it doesn't have any customizable parameters for the related problems.

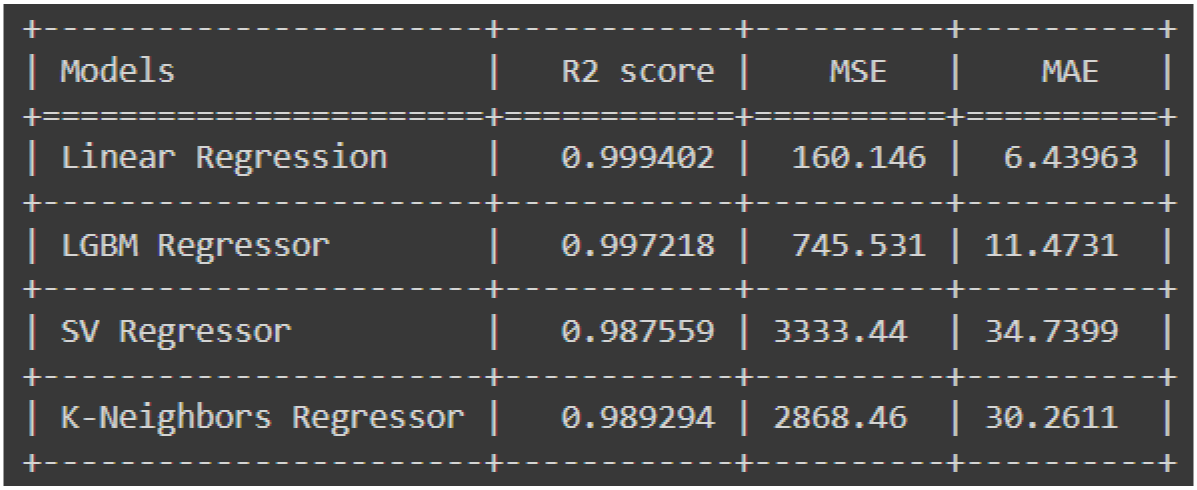
For the other models some of the parameters are chosen (Those relevant to regression problem) and then iterated over to train models with different parameter combinations, Then the R2 score was chosen as a metric to optimize the performance of the models.

After the hyper tuning the following parameters were the best for each model:

- **LGBM Regressor : boosting_type = ‘gbdt’ , n_estimators = 100**
- **SVR : C = 10000**
- **KN Regressor : n_neighbors = 14 , weights = ‘distance’ , algorithm = ‘auto’**

Results and Analysis:

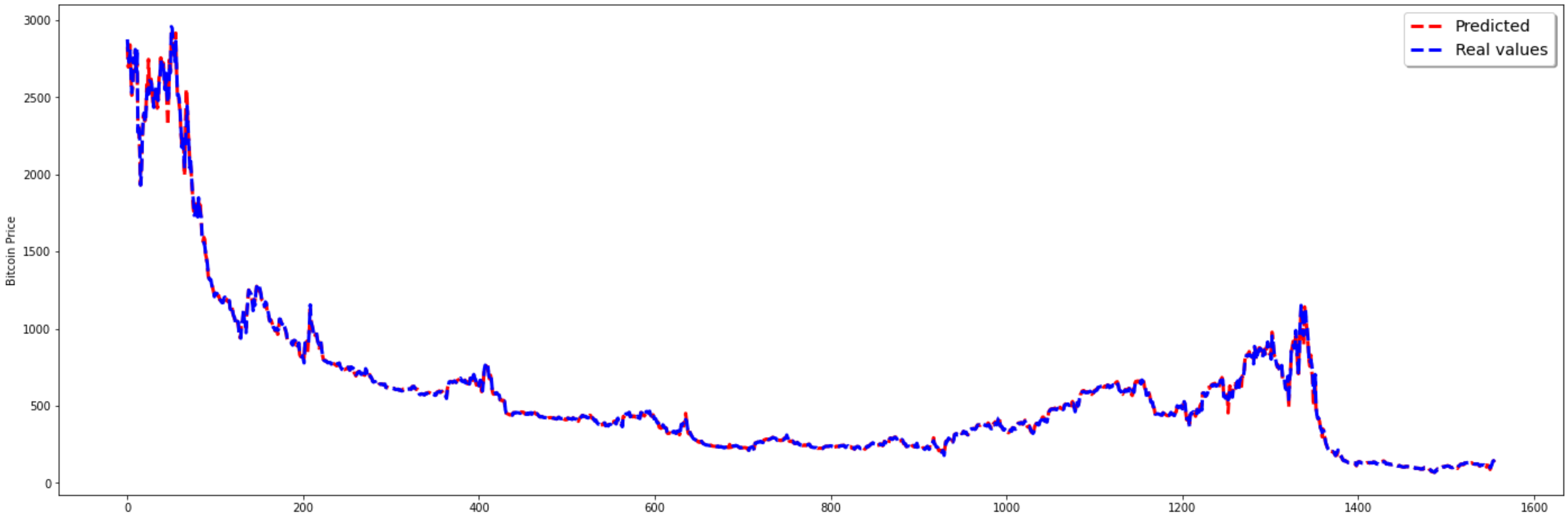
The models with the best parameters were created and then fitted over the training dataset, and then used to predict the testing dataset and then the metrics were reported for comparison.



```
+-----------------------+------------+----------+----------+
| Models                |   R2 score |    MSE   |    MAE   |
+=======================+============+==========+==========+
| Linear Regression     |   0.999402 |  160.146 |  6.43963 |
+-----------------------+------------+----------+----------+
| LGBM Regressor        |   0.997218 |  745.531 | 11.4731  |
+-----------------------+------------+----------+----------+
| SV Regressor          |   0.987559 | 3333.44  | 34.7399  |
+-----------------------+------------+----------+----------+
| K-Neighbors Regressor |   0.989294 | 2868.46  | 30.2611  |
+-----------------------+------------+----------+----------+
```

From this we can see that the Linear Regression has performed better than the other models even after hyper parameter tuning.

The Linear Regression model was fitted with the whole dataset and then plotted the predicted price vs the actual prices.



The final predictions had very little error compared to the actual price.