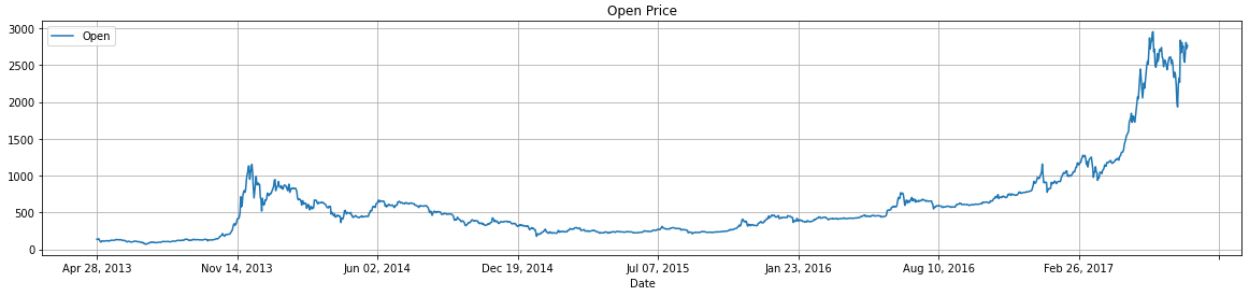
Pattern Recognition and Machine Learning (CSL2050)
Bonus Project Report

Project Title:
Bitcoin Price Prediction

ABSTRACT
Blockchain technology is increasing and there are many digital currencies rising. Bitcoin is a decentralized digital currency, without a central bank or single administrator, that can be sent from user to user on the peer-to-peer bitcoin network without the need for intermediaries.

DATASET DESCRIPTION AND PREPROCESSING
The Bitcoin historical data set consists of 1556 rows and 7 columns. The features of this data set are Date, Open, High, Low, Close, Volume, and Market Cap. Since the dataset is decreasing I reversed the dataset so as to convert the dataset to an increasing form. I checked for any null values present in the dataset, it appears that there are no null values present in the dataset. Then, I checked for the data type of values present in the dataset. Since the "Volume" column and "Market Cap" Column consist of "object" datatype, so I converted them to the "float" data type. Also, the "Volume" column consists of some entries as "-" so for this, I just dropped them only while predicting the column "Volume". Finally, I used the Standard Scaler from sklearn.preprocessing package and scaled the features of the dataset.
I plot the graphs for each feature v/s Date to visualize how the values changes. I got the plot as:

Open v/s Date

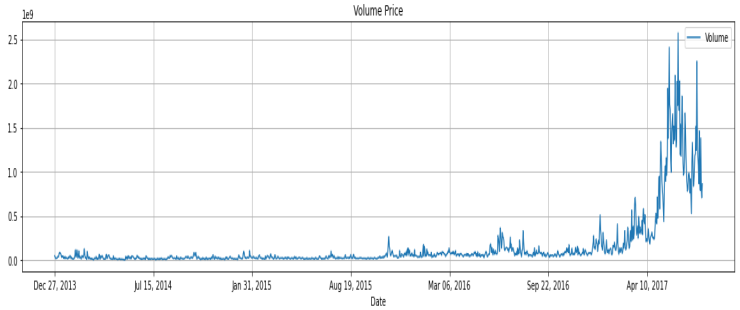


High v/s Date

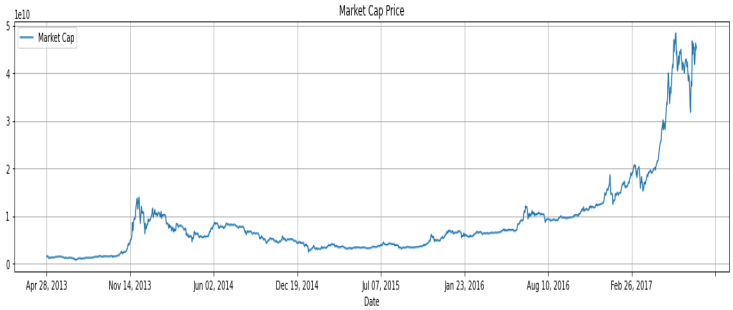
Low v/s Date

Close v/s Date

Volume v/s Date



Market Cap v/s Date



MACHINE LEARNING MODELS

I applied the regression machine learning models because regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. It's used as a method for predictive modelling in machine learning, in which an algorithm is used to predict continuous outcomes.

Some common steps were taken for all the models before using them to predict each feature that is I created a variable prediction_days for predicting 'n' days out into the future, and a column called 'prediction' that will contain the price of Bitcoin 'n' days from the current price.

1.) Support Vector Regression (RBF)

Support Vector regression is a type of Support vector machine that supports linear and non-linear regression. RBF Kernel is popular because of its similarity to K-Nearest Neighborhood Algorithm. It has the advantages of K-NN and overcomes the space complexity problem as RBF Kernel Support Vector Machines just need to store the support vectors during training and not the entire dataset.

2.) Linear Regression

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables.

3.) Random Forest Regression

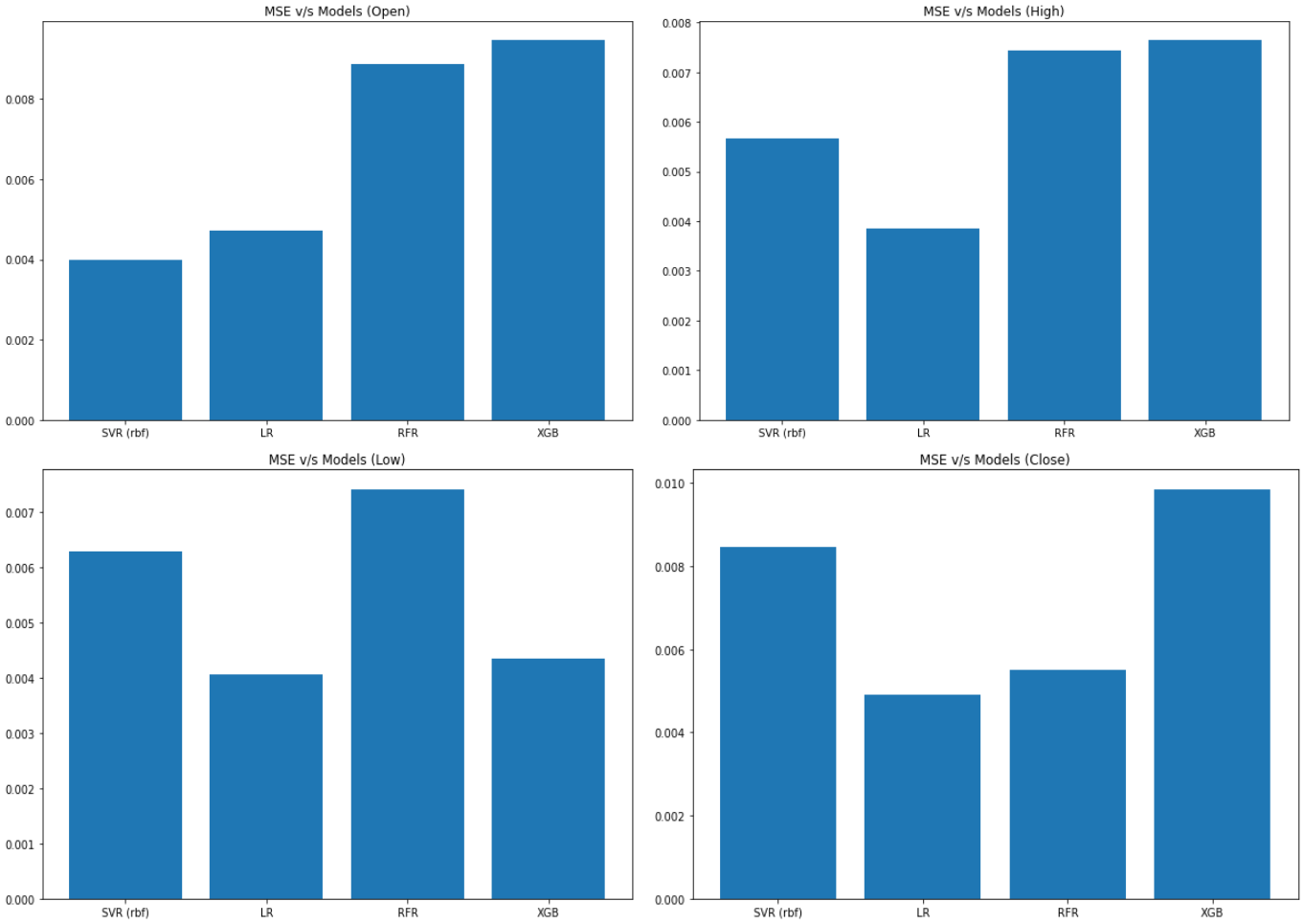
A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging.

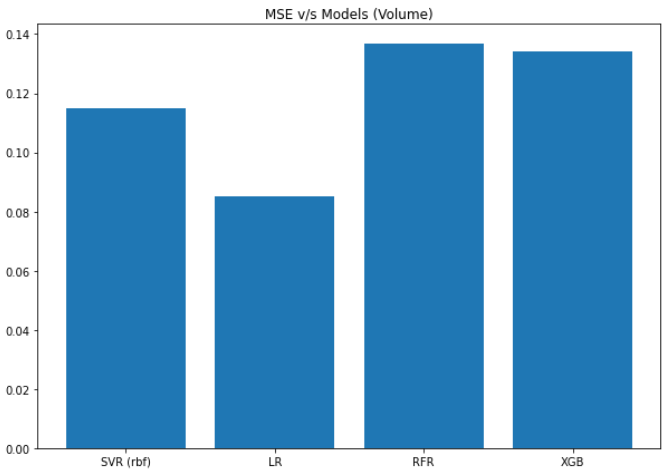
4.) XGBoost Regression

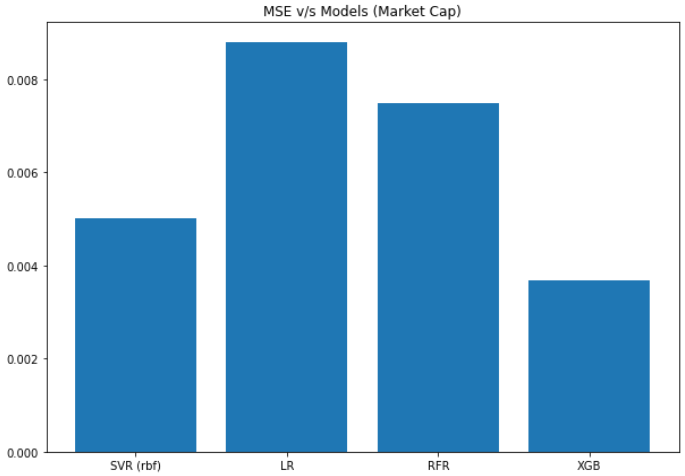
Extreme Gradient Boosting, or XGBoost for short, is an efficient open-source implementation of the gradient boosting algorithm. XGBoost is an efficient implementation of gradient boosting that can be used for regression predictive modelling.

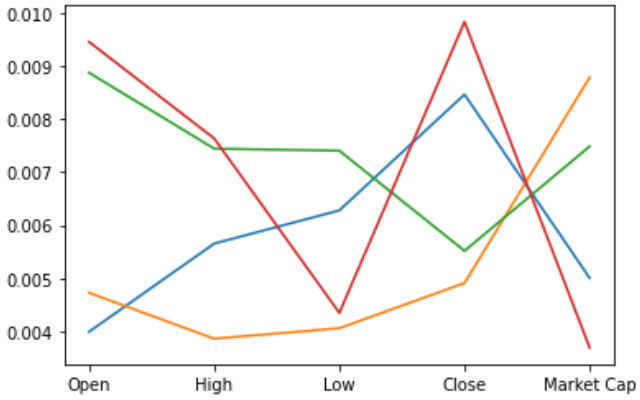
RESULT AND ANALYSIS

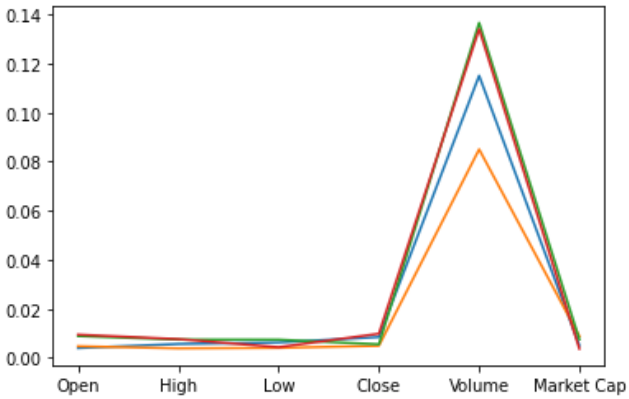
For each of the models specified above I made graphs/plots for actual value and predicted value for each feature. Also, I varied the prediction_days from 1 to 30 and visualized how the MSE varies with increasing the prediction_days.







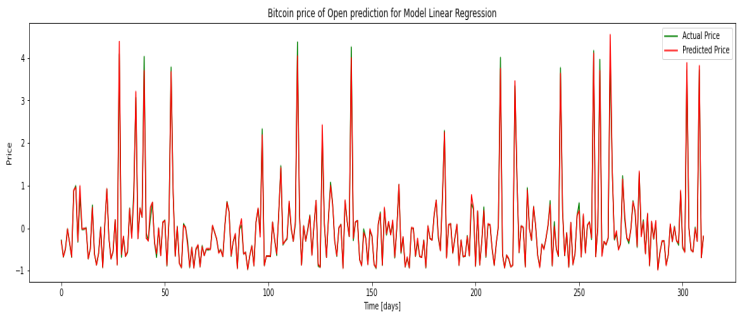


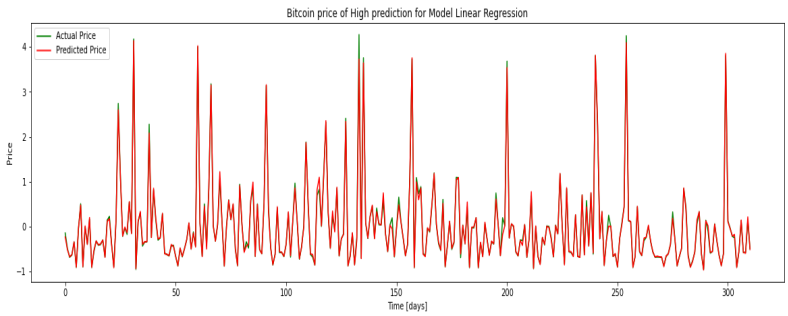


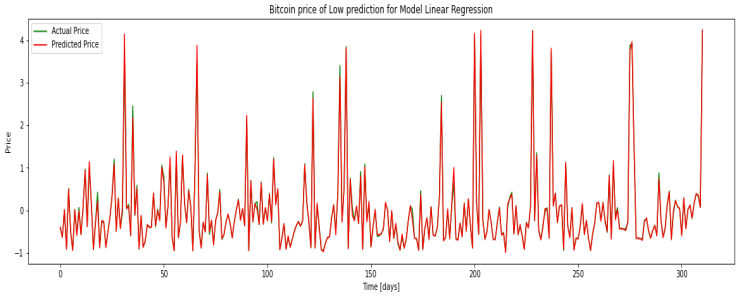
	Models	Open Price MSE	High Price MSE	Low Price MSE	Close Price MSE	Volume Price MSE	Market Cap Price MSE
0	SVR (rbf)	0.003992	0.005654	0.006280	0.008465	0.115070	0.005004
1	LR	0.004727	0.003861	0.004058	0.004911	0.085030	0.008789
2	RFR	0.008875	0.007444	0.007405	0.005516	0.136638	0.007487
3	XGB	0.009458	0.007636	0.004347	0.009836	0.134080	0.003689

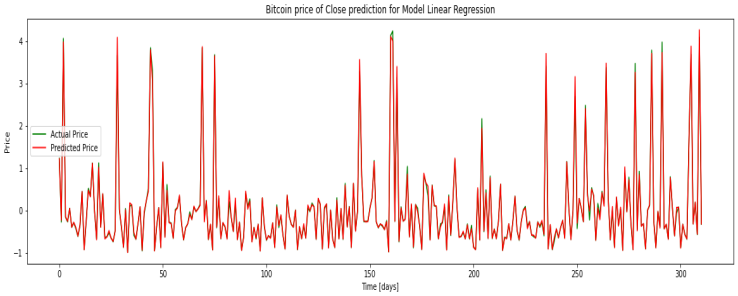
On a comparison note, LR performed the best among the other 4 models.

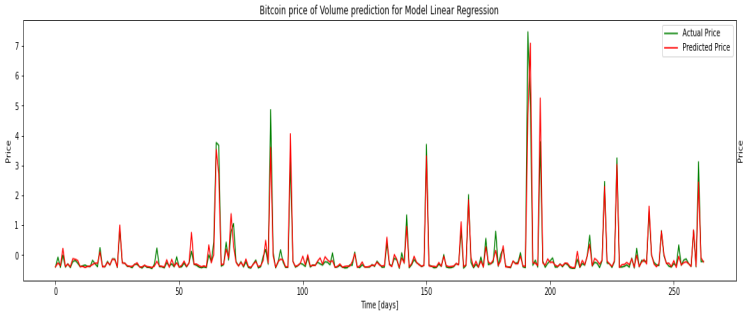
So, each of the graphs that I got for Linear Regression (LR) are:

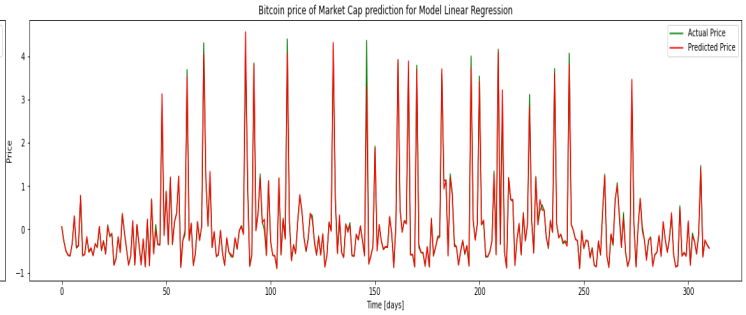


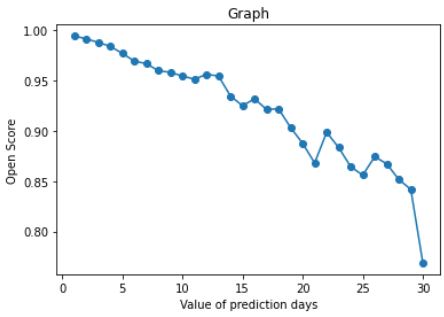


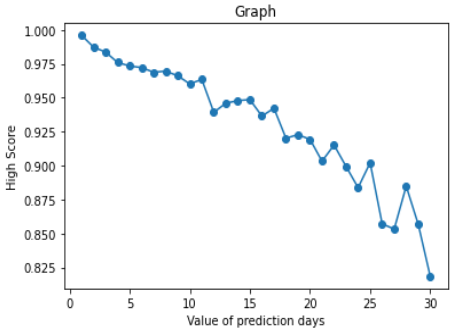


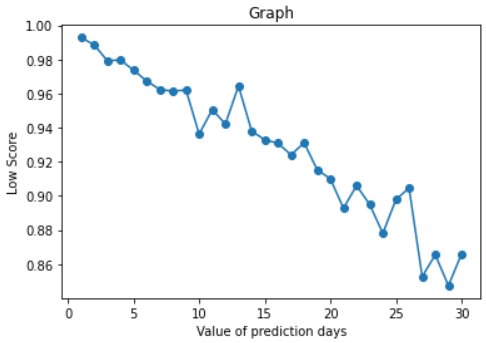


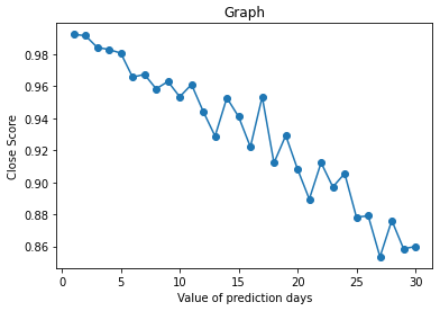


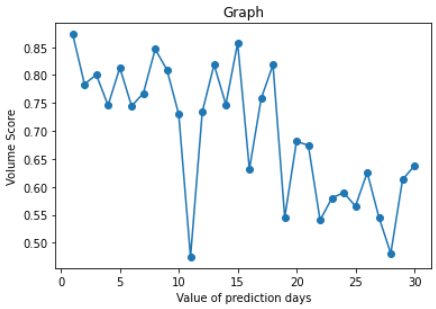


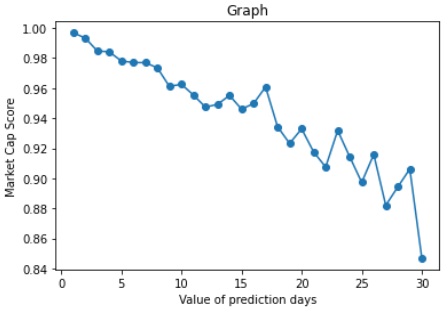












Similar to these graphs for Linear Regression I have also plotted the graphs for each model. The above graphs represent the difference between actual and predicted values. The below graphs shows the variation of the values in the range of 30 days.

REFERENCES

- https://scikit-learn.org/stable/supervised_learning.html#supervised-learning