Bitcoin Price Prediction

-Agnibha Burman Roy

Abstract - This paper reports my experience with performing Bitcoin Price Prediction. I have trained multiple Machine Learning Models and tried to capture various trends in the past prices of Bitcoin and then used those models to predict the future price of bitcoin.

Introduction

Bitcoin is a cryptocurrency which has the largest market capitalization out of all the existing cryptocurrencies. Since the prices of such currencies vary a lot even in short durations of time, it is very difficult to predict the future values of the prices.
If we can build a nearly accurate model, then this can be used to forecast future values of bitcoin, and its use-case can be in cryptocurrency trading.

Datasets

Link to the dataset used: [Dataset]
The dataset contains 7 columns and 1556 rows.
The columns are:

Date	Open	High	Low	Close	Volume	Market Cap

```
Starting Date 2013-04-28 00:00:00
Ending Date 2017-07-31 00:00:00
```

Methodology Overview

Initially I performed EDA, then converted the dataset into a dataset fit for Supervised Machine Learning, and then tried various models to predict the target variable. The algorithms which I have used are mentioned below:

- Linear Regressor
- XGB Regressor
- LightGBM Regressor
- Ridge regression: L2 regularization variation of Linear Regression
- Lasso Regressor: L1 regularization variation of Linear Regression
- Elastic-Net Regressor: Combination of L1 and L2 regularization
- MultiLayerPerceptron regressor: The activation function used in the last layer is unity, thus making it a regression model instead of a classification model.
- SVR
- LSTM(RNN): Neural Networks designed specifically for Time series forecasting. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. Also tried ARIMA models, but since

data was not stationary, ARIMA models did not give significant results.

Exploratory Data Analysis

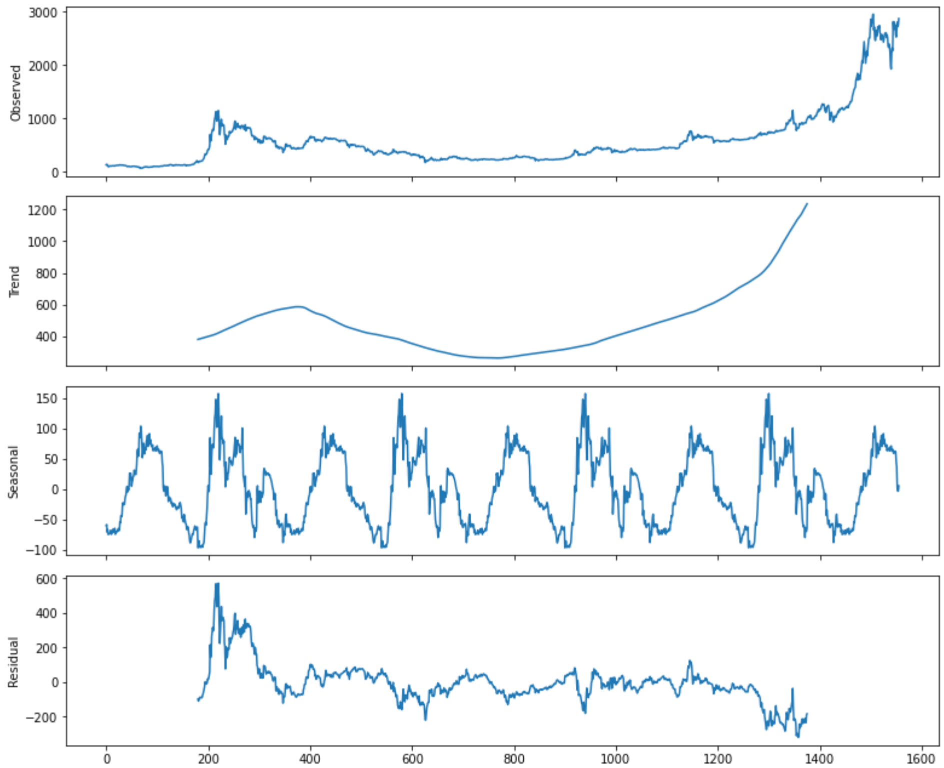
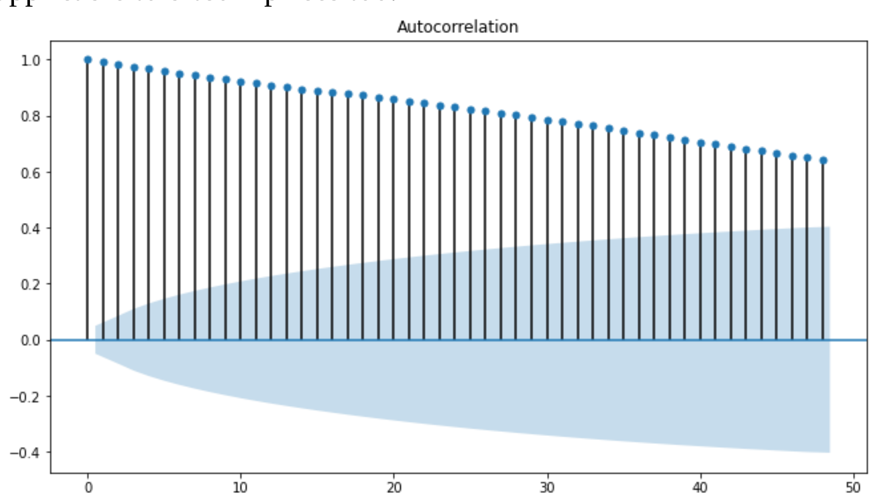


Fig 1.1 The above graph shows the closing price of bitcoin over the years 2013-2017. Also, shows the **Trend, Seasonality, and Residual.** As we can see, there are many fluctuations, and it will be difficult to predict such data.

Converting Time Series Prediction Task into Supervised Learning task

In the case of time series forecasting, we can always see a high correlation between the target variable and the value of the target variable a few days back. Thus, checking if this trend is applicable to bitcoin prices too.



As seen in the graph, the present-day value of the close price is correlated with the previous day's values. (Up to lag of 90-100 days)
Thus, I converted the problem into the following supervised task:
Close-Price(day) = F[Close-Price(day-1), Close-Price(day-2), Close-Price(day-3)]
(To test the hypothesis that the current price depends on the previous prices)

When the hypothesis was true, I tried taking larger dataset and assumed that the current price depends on past 30 days prices.
=> **Close-Price(day) = F[Close-Price(day-1), Close-Price(day-2), Close-Price(day-3),..... Close-Price(day-30)]**

Implementation and Evaluation of Models

Initially the models were trained on the training data(13th April 2013 to 9th Jan 2017). Each of the above-mentioned models was fine-tuned and after finding the best hyperparameters, the models were tested on the last 100 days of the dataset(10th Jan 2017 to 31st Apr 2017).
The models were evaluated on the basis of multiple scores, such as:
R Squared Score: It is the proportion of the variation in the dependent variable that is predictable from the independent variable
Mean Squared Error: It is the mean of the square of all the differences between the predicted value and the true value.

When I trained the models using the features 'Open', 'High', 'Low', 'Market Cap', 'Volume' and the past days of closing price, not much change was observed in the model performance, thus I have dropped those columns. And only used the past values of Closing Price.
Also, only the past value of the Closing Price can be used to forecast the future values.

Results and Analysis

The below results were obtained after training the models:

Model Name	R2_Score	Mean_squared_error
RandomForestRegressor(Using only past 3 days as the features)	-100.49	1.5e6
Linear Regression(Using only past 3 days as the features)	0.458	0.4e6
Linear_Regression(Using past 30 days as the features)	0.99659857	765.198

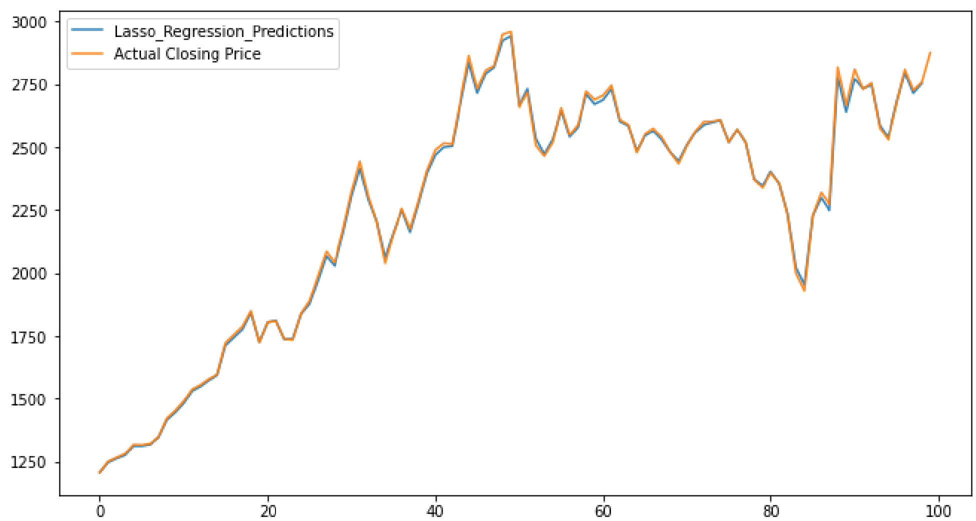
Below table is when I used the past 30 days of data as the feature space.

Model Name	R2_Score	Mean_squared_error
Linear Regressor	0.996598	765.198
XGB Regressor	0.996215	780.863
LightGBM Regressor	0.996467	773.898
Ridge regression	0.996592	765.378
Lasso Regressor	0.999224	174.066
Elastic-Net Regressor	0.999193	180.960
MultiLayerPerceptron regressor	0.9879	807.992
SVR	0.876588	5057.598
LSTM(RNN)	0.986834	1754.833

The models arranged according to the best R2_score are as follows:

Model	R2_Score
Lasso Regressor	0.999224
Elastic-Net Regressor	0.999193
Linear Regressor	0.996598
Ridge regression	0.996592
LightGBM Regressor	0.996467
XGB Regressor	0.996215
MultiLayerPerceptron regressor	0.987900
LSTM(RNN)	0.986834
SVR	0.876588

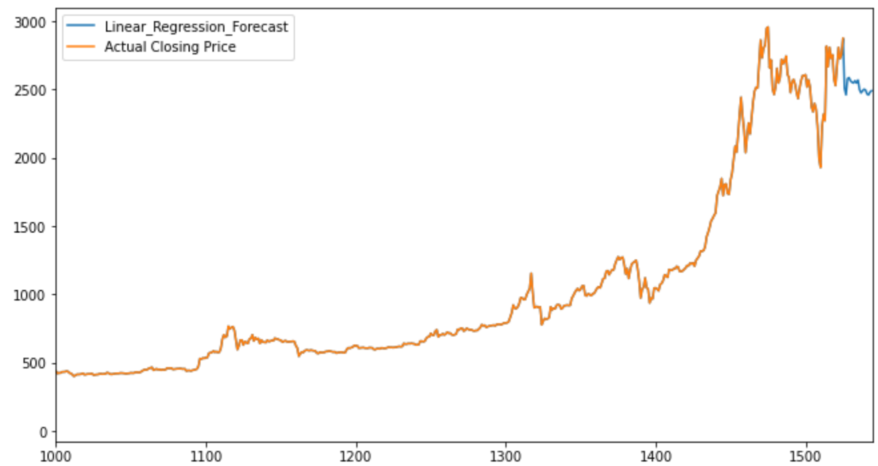
The Outputs of the Lasso regression model are as follows:

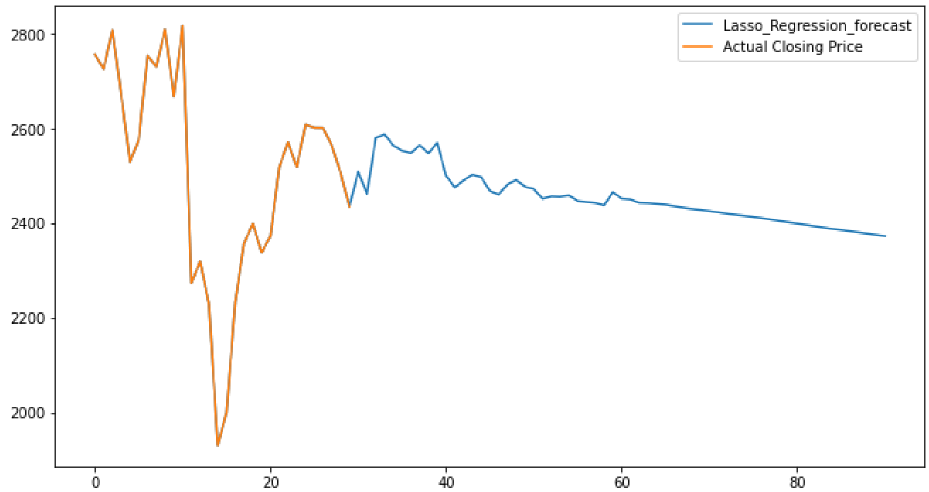


- **This is the task of only predicting the closing price 1 day into the future, (as we are using the data of the previous 30 days while predicting)**
- **As observed, we can almost accurately predict the value of Closing price for the next day.**
- **But for forecasting the future prices, we need to re-predict the values of the next day recursively, taking into account the predictions made by our model.**

Forecasting Future Values

For forecasting the values of the future, we initially predict the value of the immediate next day, then assume that our prediction was correct and append that value into the main dataset, now predict the value for the later day. And perform the above step recursively.





X-axis denoted the day number

The above graph denotes the forecasted values of closing price, as we can see, the model is only able to forecast the values up to 5-10 days into the future and after that, the forecast is not appropriate. Thus, there is still a long way to go in developing models that predict values into the future. But using our model, we can only predict the value up to 5-10 next days.

References

[1] Pattern Classification -Book by David G. Stork, Peter E. Hart, and Richard O. Duda

[2] Guide to Time Series Analysis | Analytics Vaidya