BIRLA INSTITUTE OF TECHNOLOGY, MESRA RANCHI INDIA

MASSIVE OPEN ONLINE COURSE (CA590)

Topic: Bitcoin Price Prediction and Analyzation using Machine Learning

Name: Rishabh Anand
Roll: MCA/10007/2019
Branch: MCA
Semester: 4th
Submitted to: Rashmi Rathi Upadhyay

Index

Abstract

Cryptocurrency, especially Bitcoin, is one of the most volatile markets today and has gained a lot of attention from investors across the globe. Cryptocurrency, being a novel technique for transaction system, has led to a lot of confusion among the investors and any rumours or news on social media has been claimed to significantly affect the prices of cryptocurrencies. The goal of this study is to predict prices for Bitcoin using Machine Learning Techniques for the next day and prepare a strategy to maximize gains for investors. We also aim to find out if there is a co-relation between fluctuating Bitcoin prices and related news.

Introduction

Researchers have been long trying to predict the stock market and any breakthrough in this field would result in, literally, the people being able to mint money. Cryptocurrencies, to be specific, has gained a lot of traction in the recent years from investors across the globe. Cryptocurrency being a novel technique for transaction system has led to a lot of confusion among the investors and any rumours or news on social media has been claimed to significantly affect the prices of cryptocurrencies.

Bitcoin has a market share of more than 55% as compared to other cryptocurrencies, being followed by Ethereum at 8.57%.

Bitcoin has seen its highest price around 61000$ and lowest price around 5200$ in last 2 years. Therefore, it is very sporadic analysing it would have been interesting.

Related work

ARIMA:

In Forecasting, we predict the future values given the past data. Time-series forecasting has been performed predominantly using statistical based methods, for example, the linear autoregressive (AR) models because of their flexibility to model many stationary processes. These include the well-known ARMA (autoregressive moving average) model (Fan and Yao, 2003) and its extensions by Weron and Misiorek (2008) for short term time series forecasting. Also, ARIMA models are largely limited to capturing the first-order non-stationarity in a time series data. However, most of these methods tend to be limited for nonlinear and stationary time series forecasting by the local linearity assumption implicit with an AR-type structure.

Stanford:

A student at Stanford tried to predict the price direction for the next 3 days based on past data. He performed comparative study of Logistic Regression, Naive Bayes and Deep Neural Networks and finds that Artificial Neural networks can classify the directions with an accuracy of 72 percent. They also confirm that spikes in bitcoin prices are also co-related to sentiments of people in social media forum about bitcoin.

Kaggle:

Various Kaggle Competitions and online forums report better accuracy for stock market predictions using RNN, but there have been rare or no such formal investigation on predicting Spikes.

Method and Model

To solve the problem using Machine learning, we first tried to categorize the problem and tried to find previous solutions on how they solved it. We quickly learned that, since, the problem involves prices which are changing with time, this could be modelled as a Series prediction problem.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:
For b = 1, B:
Sample, with replacement, n training examples from X, Y; call these Xb, Yb.

Train a classification or regression tree fb on Xb, Yb. After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x' or by taking the majority vote in the case of classification trees.

This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated.

Experiments

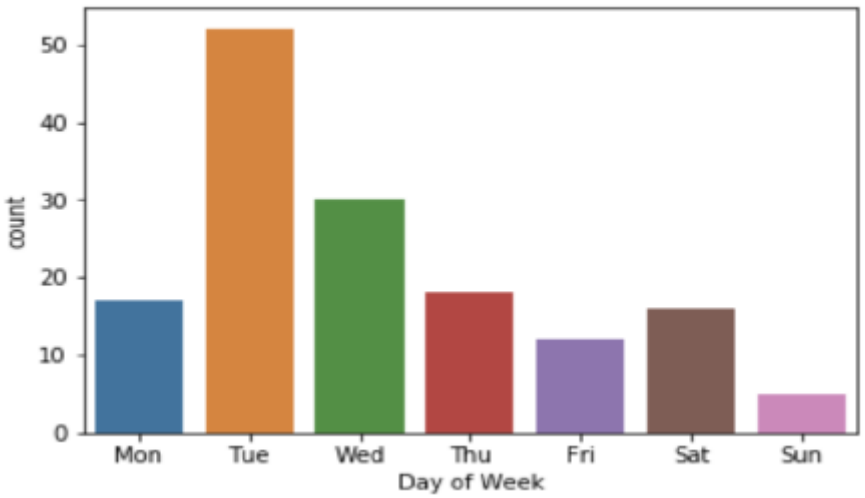
Data:

The database is extracted from Yahoo Finance, where a csv file has been generated which includes the closing price, opening price, Day high, Day Low, Volume on each date. This data file is later cleaned for all the null values or similar values. Our horizon forecast is increase in price of bitcoin every day, therefore we kept only the data of closing price for each day and deleted rest of the data.

NaN Values:

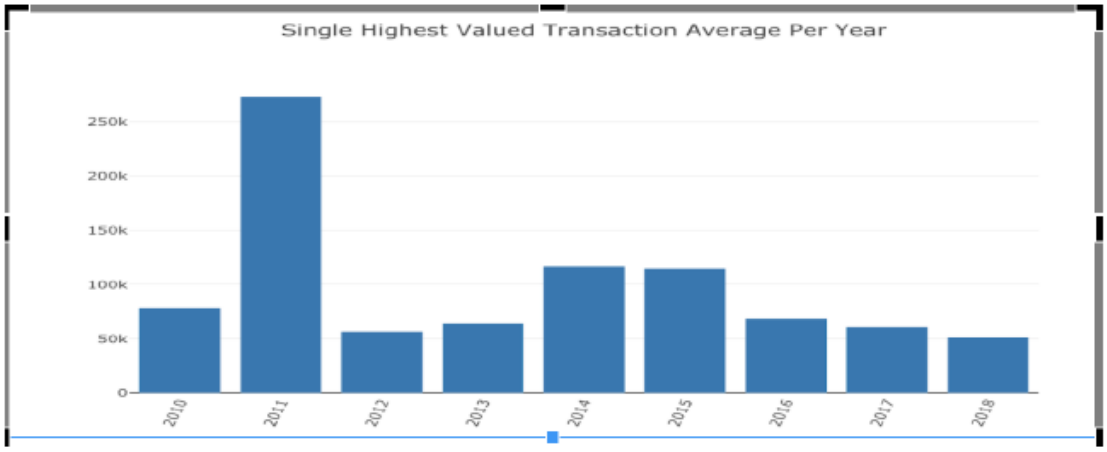
A NaN value is handled by the method of forward fill. In some models, backward fill method is also used but since we don't want to predict the past given the future, forward filled method is used.

Trends



Observation:

There is highest count of bitcoin transactions on Tuesdays and lowest on Sunday which was somehow expected but this confirms that there is some seasonality in data.



Observation:

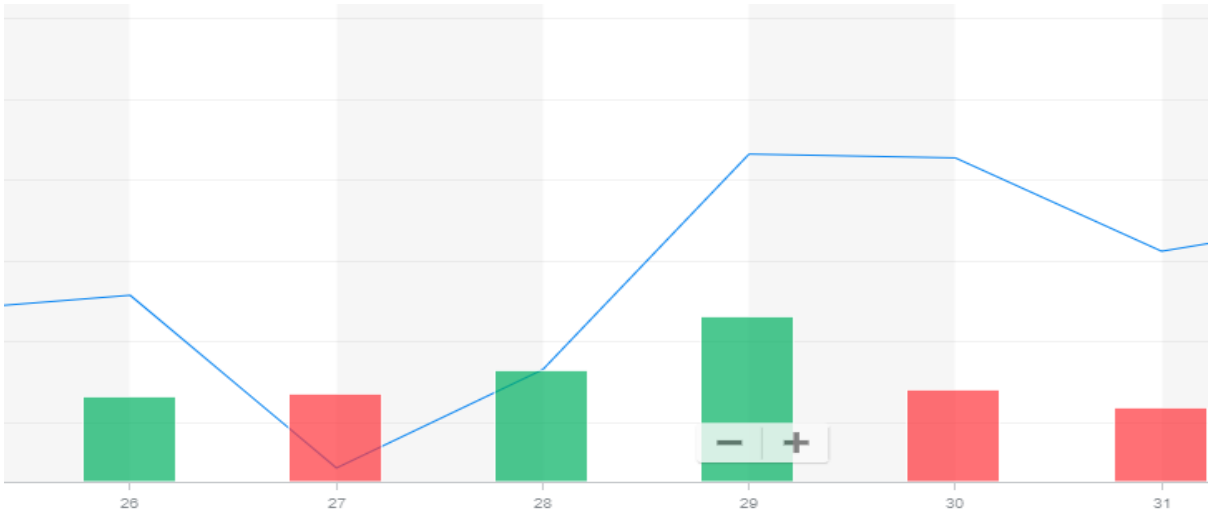
It is observed that the highest valued transaction per year was actually during 2011 and not 2020 or 2019 where the prices shoot up massively.

News Data:

Now, to investigate the claim of bitcoin related news and tweets affecting the bitcoin prices, Doc2Vec utility (based off of Word2Vec) from the gensim library was used. e.g.: A one word tweet by Elon Musk or Government of India's trying to control cryptocurrency transactions or GameStonks.

e.g: Elon Musk (CEO, Tesla) tweeted "**#bitcoin**" on Jan 29, 2021.

Result: Price of bitcoin shoot up 14%


26
27
28
29
30
31

Experiments and Evaluation:

Random Forest:

Random Forest performed in a pretty standard way when the window length was 3 and forecasted a single day's value (present). The RMSE with Random forest was decent and increasing the max depth up to 16 gave the best results. Increasing the max depth up to 50 gave us the best results when considering the news data. It is observed that Random forests started to over fit when we decreased the number of trees, and this was expected.

Tools Used

The following tools are used in creation of this project:

1. Programming Language:
   Python: The whole project is coded in Python language.

2. Libraries used:
   a. Numpy: It provides the basic data structure for the entire project. It is used to store and pre-process data.

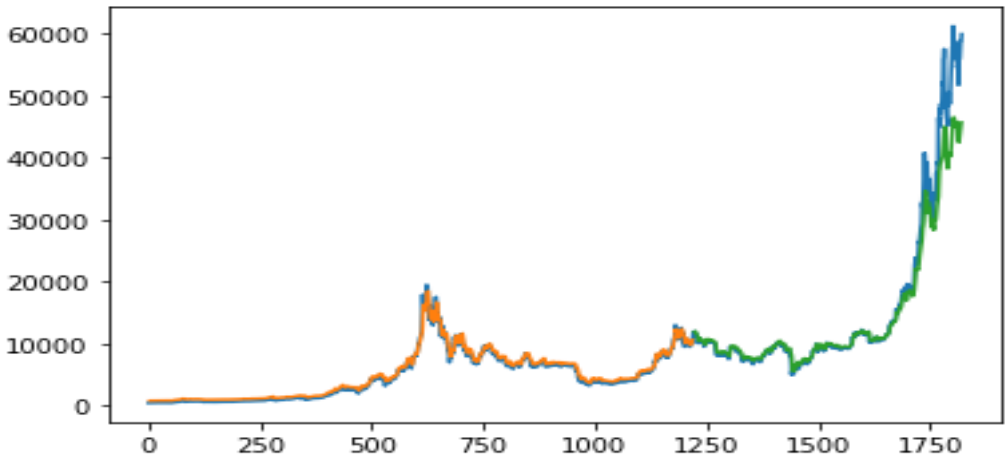
   b. Scikit learn: It provides the necessary tools needed to construct the CNN model. Along with the construction tools it also provides tools for training and testing the data.

   c. Matplotlib: It is used for drawing the graph that is displayed in the GUI interface.

   d. Pandas: It is used for data analysis and manipulation tool.

3. IDE Used: Google Collab

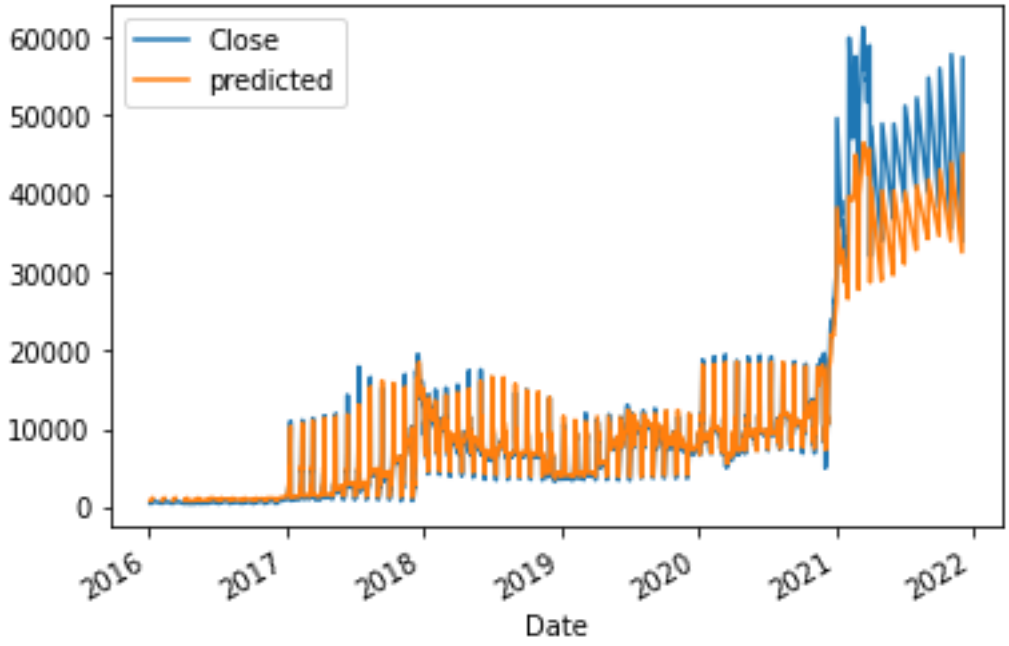
Screenshot



The price fluctuation can be seen from the above graph.

```
Epoch 41/50
25/25 [==============================] - 0s 16ms/step - loss: 3.5937e-05
Epoch 42/50
25/25 [==============================] - 0s 16ms/step - loss: 3.6993e-05
Epoch 43/50
25/25 [==============================] - 0s 15ms/step - loss: 3.8135e-05
Epoch 44/50
25/25 [==============================] - 0s 15ms/step - loss: 3.9326e-05
Epoch 45/50
25/25 [==============================] - 0s 15ms/step - loss: 4.0515e-05
Epoch 46/50
25/25 [==============================] - 0s 16ms/step - loss: 4.1640e-05
Epoch 47/50
25/25 [==============================] - 0s 14ms/step - loss: 4.2621e-05
Epoch 48/50
25/25 [==============================] - 0s 15ms/step - loss: 4.3393e-05
Epoch 49/50
25/25 [==============================] - 0s 16ms/step - loss: 4.3865e-05
Epoch 50/50
25/25 [==============================] - 0s 15ms/step - loss: 4.4040e-05
Train Score: 514.87 RMSE
Test Score: 3486.61 RMSE
                 Close     predicted
Date
2016-02-04    420.872986           NaN
2016-03-04    420.903992           NaN
2016-04-04    421.444000           NaN
2016-05-04    424.029999    792.804932
2016-06-04    423.412994    794.276733
...                  ...           ...
2021-03-28  55950.746090  43949.523438
2021-03-29  57750.199220  44402.660156
2021-03-30  58917.691410  45045.828125
2021-03-31  58918.832030  45693.210938
2021-02-04  59926.535160           NaN
```

A total of 50 epochs are done, and the predicted price is compared to the close price.



It can be observed that close and predicted price by our model works well for the data before 2021 but several other factors including tweets, government intervention has caused the model to deviate from the reality.

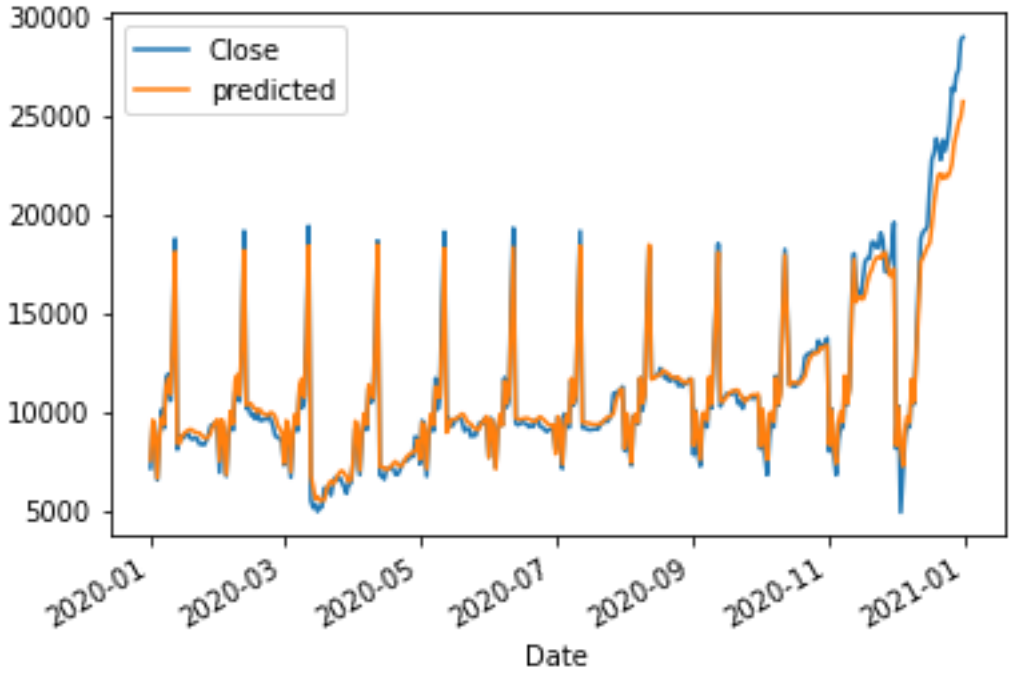


Fig: Price prediction and closing price in 2020

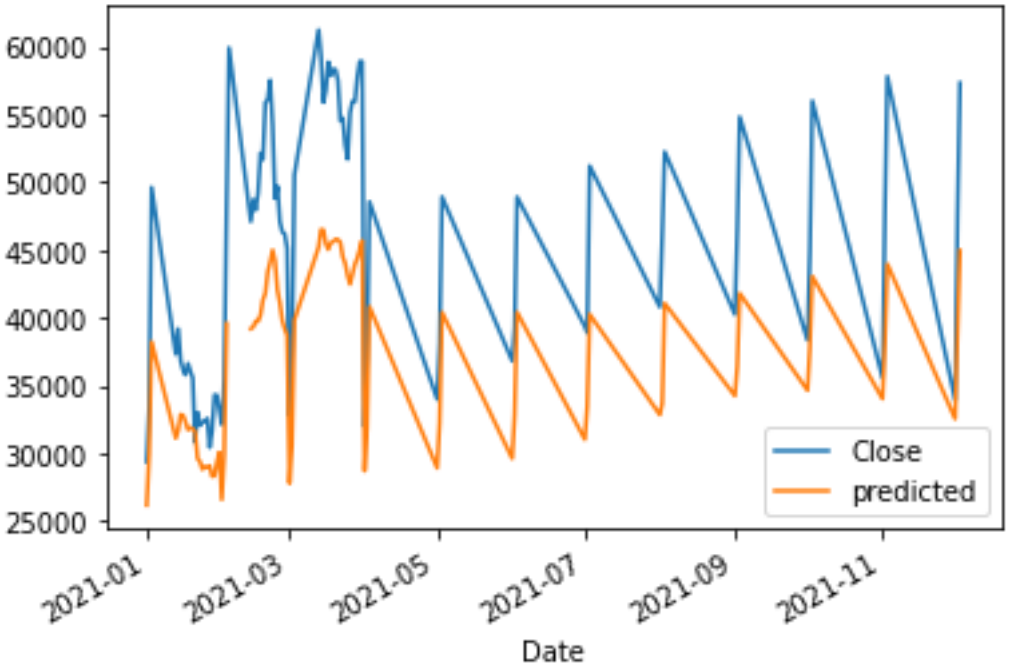


Fig: Price prediction and closing price in 2021

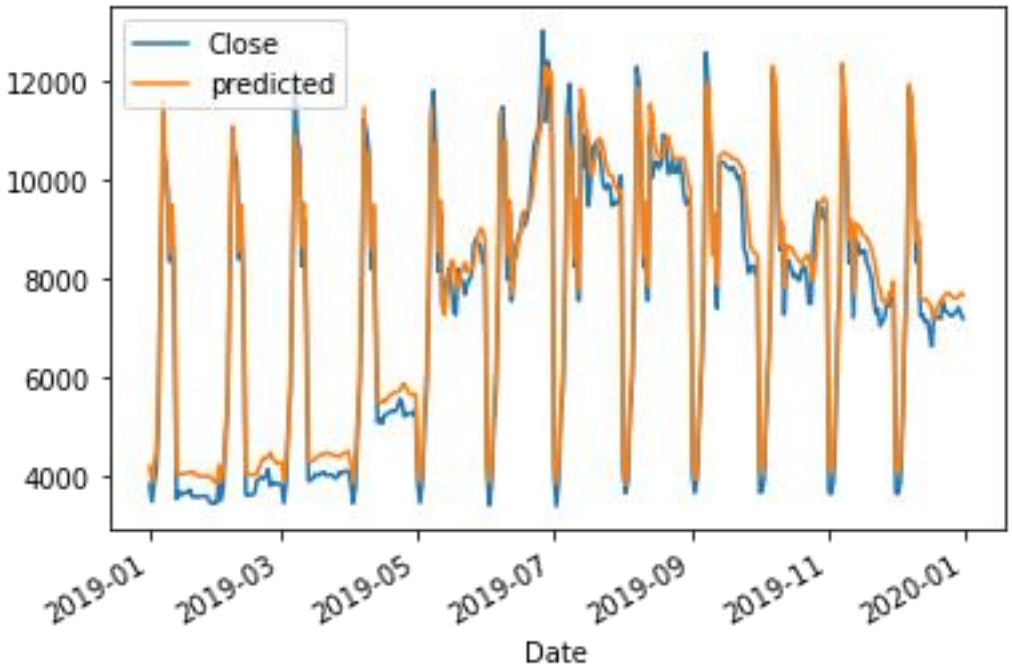


Fig: Price prediction and closing price in 2019

Conclusion

This project uses important concepts that we have learnt from fields like mathematics, biology, and computer science.

We have also learnt how Time series data mining is identified and how different models works on the different types of data.
We also learnt that external factors like news or tweets by influential people also affects the price of bitcoin. This project also demonstrates how stock market is so unpredicted and how it gets affected by news or rumours. It is also observed that while some model is highly effective for the short data, on the other hand they fail hugely when the data is large.

References:

[1] Connor Lamon & Eric Nielsen & Eric Redondo, Cryptocurrency Price Prediction Using News and Social Media Sentiment.

[2] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)

[3] Yahoo Finance.

[4] Case Study done by Manthan Thakker & Bharat Vaidhyanathan, Northeastern University, Boston. 2017

[5] Reading Materials from NY Times, Medium, Economic Times.