ORIE 5741 Midterm Report

Bitcoin Price Prediction - Using Twitter Comments and Commodity Price

1 Summary

The objective of this project is to make prediction on Bitcoin Price, utilizing its own features, the commodity prices, and the social media comment. In this mid-term report, we focus on the data set cleaning and processing, visualization, and correlation analysis.

2 Dataset Processing

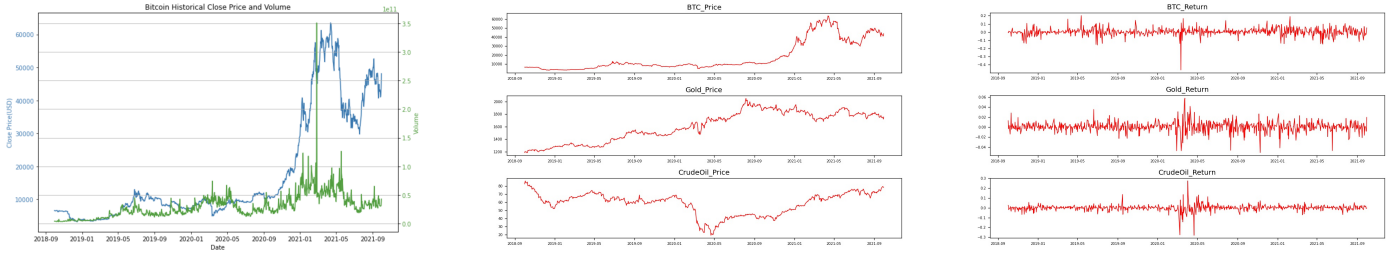
1. **Visualization of Bitcoin Historical Data**

   For the Bitcoin Price dataset, the yahoo finance package is utilized to extract the data from 17 September 2014 (earliest available date)to 30 September 2021. The dataset contains date, open, high, low, close, volume, dividends and stock splits. Given the time series dataset feature, we try to understand whether the historical price of Bitcoin could provide a reliable prediction of its future prices, by plotting the daily trading volume and the daily closing price of Bitcoin as below.

2. **Visualization of Gold and Crude Oil Data**

   The Gold & Crude Oil Price dataset consists of daily US dollar price series of bitcoin, gold,and crude oil from Yahoo Finance. The daily series ranged from 1 October 2018 to 30 September 2021. Since gold and crude oil do not have weekend values, bitcoin's price values for Saturdays and Sundays were removed from the data set. Moreover, if there were any other unmatched days in the data sets, they were also removed. There were 742 data points in total.

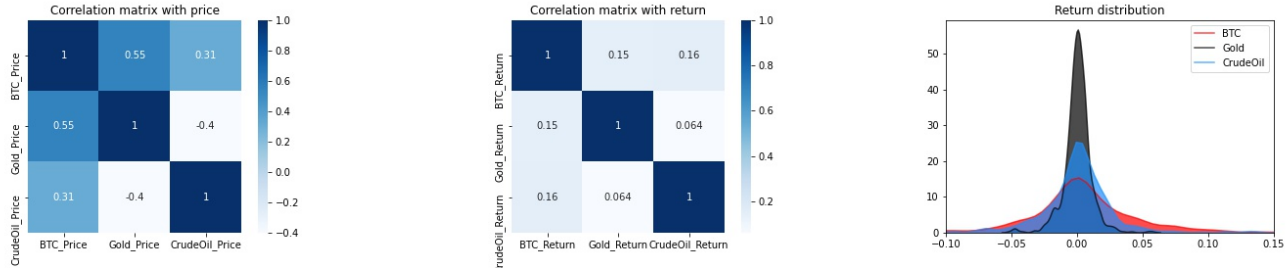
   Figure 1 and Figure 2 below represent the price and return series of bitcoin, gold, and crude oil, respectively. In terms of the returns series, logarithmic returns were computed for all three series.



(a) Bitcoin Daily Price & Volume (b) Price Series of Three Assets (c) Return Series of Three Assets

Figure 1: Data Visulization Bitcoin, Gold, and Crude Oil

We can also get the correlation matrices and price distribution as follows:



(a) Correlation Matrix (Price) (b) Correlation Matrix (Return) (c) Price Distribution'

Figure 2: Correlation Matrices and Price Distribution

From the above, we can see that Bitcoin has the largest price range and variance. Besides, the price of Bitcoin is positively correlated with that of gold and crude oil, and the correlation coefficients are 0.55 and 0.31 respectively.

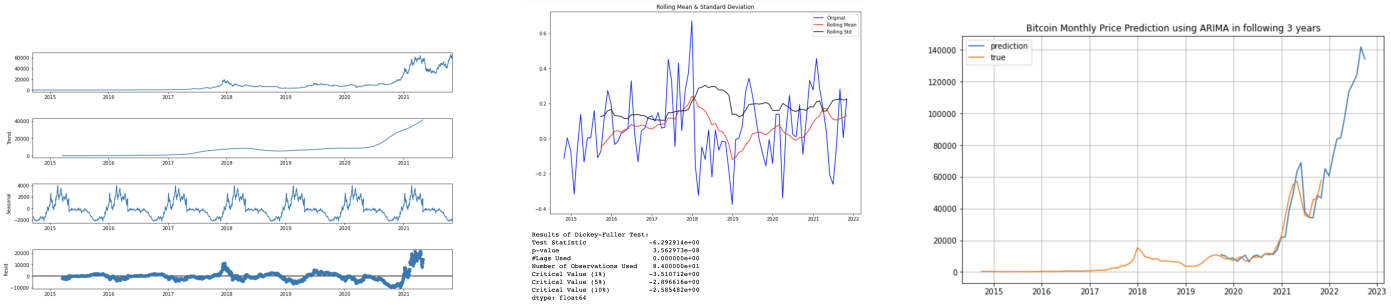
3 Data Analysis

1. **ARIMA**

   To begin, we would like to explore the seasonality of the bitcoin price.

   From above plot, we can see that there exists seasonality in the Bitcoin Price. We aim to eliminate the seasonality of the data. The approach we take here is to take the logarithm, than difference for 1 time of the BTC price, in order to obtain a stationary time series. The ADF test is applied to examine the stationary of the data.

   To apply seasonal ARIMA model on the data, we select the parameters by Auto_ARIMA, the best model selected is SARIMAX$(1,0,0)(2,1,0,12)$. The true and predicted Bitcoin price are as below.



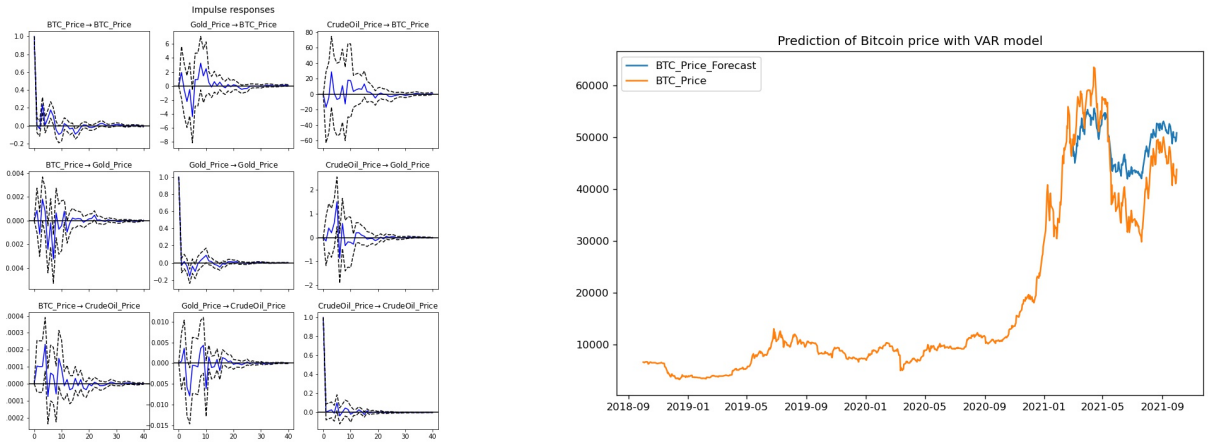
(a) Bitcoin Log price (b) Bitcoin Price (c) Bitcoin Price

Figure 3: Bitcoin Price prediction using ARIMA

2. **Prediction With VAR Model**

   We divide the dataset into train set and test set, and try to provide a reliable prediction model of Bitcoin prices. Here we choose VAR model which uses the lag term of all endogenous variables in the system to construct the model.

   Due to ADF test, we construct a stationary sequence by first-order difference for regression. The impulse response analysis and prediction result are shown in the figure below.



(a) Impulse Response Analysis (b) Prediction Result

Figure 4: Prediction With VAR Model

We can see that in the test set, although there are certain errors in the fitting values, the VAR model fits the changing trend of Bitcoin price well. From the impulse response diagram, we can see that the short-term impact of gold price on Bitcoin price is positive, while the short-term impact of crude oil price on Bitcoin price is negative.

3. **Tweets**

   Now we focus on the Tweets factor. We analyze the popularity of Bitcoin-related topics on Twitter. We use the Twitter standard API to capture data before the end of 2019 and visualize Bitcoin's popularity.

   There are few data mentioning Bitcoin before 1 January 2017, thus we drop the data and visualize the popularity from January 2017 to November 2019 monthly.

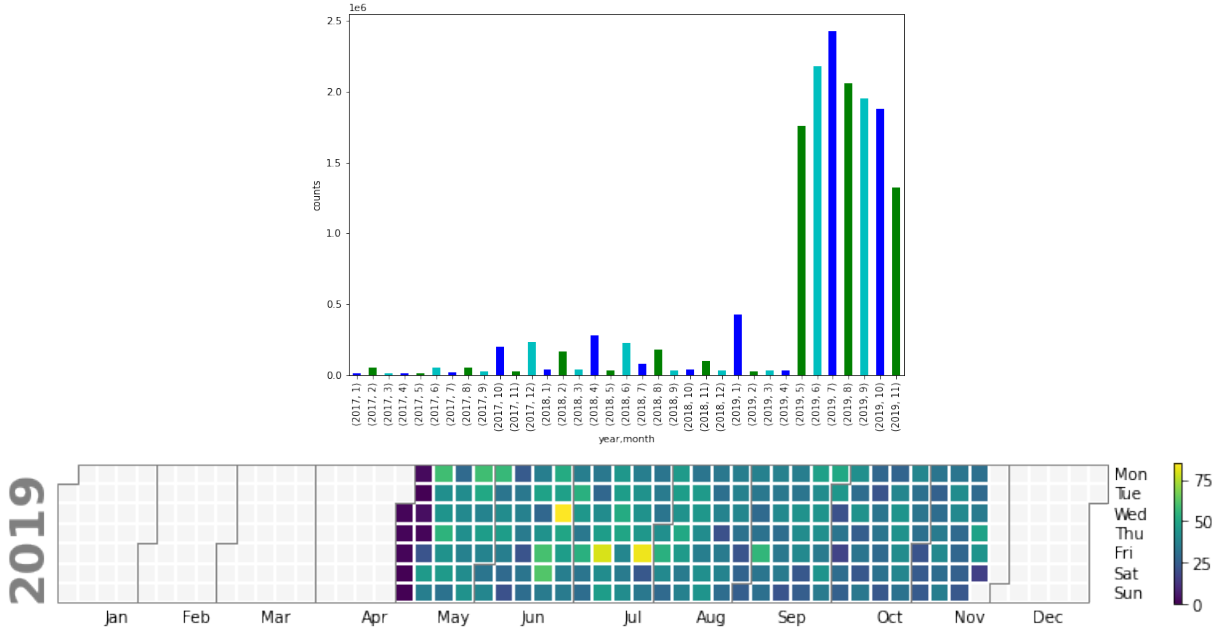


Figure 5: Prediction With VAR Model

From above plot, we can infer that the topic is highly heated from May to November 2019, with 7,718,389 tweets containing hashtags with #BTC or #Bitcoin (topic related directly, not only mentioning), indicating a heated level of comments with individual and institutional investors' attention.

The total tweets is really a large dataset with tick frequency. Among totally 15,996,724 datapoints, 146,759 tweets containing "Follow us" or "subscribe", which indicates the texts are institutional posted. As a result, individual investors and Bitcoin market followers are the main contributors to Bitcoin popularity on Twitter.

4 Future Planning

1. There is a certain gap when predicting Bitcoin price using historical price data of Bitcoin, gold and crude oil. Therefore, we plan to conduct feature engineering which combines tweets and other messy information with historical price data to reduce the error.

2. We encountered difficulties when dealing with the tweets data set due to the large volume and limited data access. We plan to improve this issue by using key opinion leaders tweets comment as reference resource, while trying for more effective methods to process over 10 million comment.