ORIE 5741 Final Report

Bitcoin Price Prediction - Using Twitter Comments and Commodity Price

1 Problem Identification

Cryptocurrency has become an emerging market with rocketed prices and trading volume globally. Among the market, Bitcoin is the leader and the representative.The total exchange trading volume reached 1.244B USD in Dec 2021 while the price maintained high at 49.25k USD. How the prices of Bitcoin fluctuate could have a big influence on the investor's portfolio. Thus, this report aims to investigate possible factors that could have an impact on Bitcoin price. We also explore possible separate and integrated regression model to predict the bitcoin price.

In this report, we present how we pre-processed the dataset and perform fundamental data analysis on them. Techniques including ARIMA time series, VAR model, Support Vector Regression and Random Forest are applied to approach the problem identified. At the end, models are evaluated and reflected on its fairness and limitations.

2 Dataset Description

1. **Cryptocurrency Historical Performance**

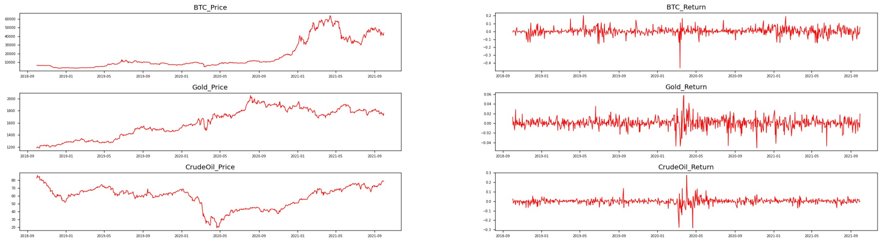
   For the Bitcoin price prediction, its own historical performance are considered to be important feature. The historical performance indicators consist of Bitcoin daily historical price, daily total circulating amount, daily total trading volume.

   Given the industry Bitcoin is in, we find it meaningful to include other cryptocurrency performance as features for Bitcoin price prediction. The currency included in this report are daily Ethereum (ETH) price and Dogecoin price.

2. **Commodity Price: Gold & Crude Oil Historical Price**

   The Gold & Crude Oil Price dataset consists of daily US dollar price series of Bitcoin, gold,and crude oil from Yahoo Finance. 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 below represents 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) Price Series of Three Assets

(b) Return Series of Three Assets

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

3. **Elon Musk's Tweets Data on Bitcoin**

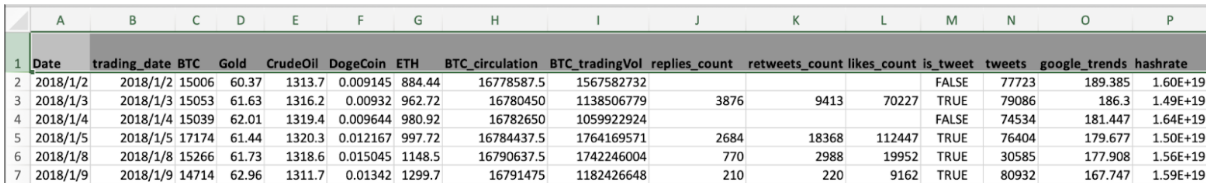
   Elon Musk is a hot celebrity on Bitcoin, which has a considerable influence on Bitcoin's market. The Elon Musk's tweets data on Bitcoin consists of features: For each of the trading day, whether Elon Musk tweeted on bitcoin topic, number of replies, number of retweets and number of likes. If Elon Musk did not tweet, the following features' values would be missing.

4. **Daily Tweets Data on Bitcoin**

   Aside from Elon Musk's influence, there are also institutional and individual investors on bitcoin market. Market attention can manifest in tweet popularity. Daily tweets data includes total amount of tweets, daily google trends and hashrate on bitcoin topic for each trading day.

5. **Overview of Overall Dataset**

   Below is an overview of the combined row data table. It contains 14 features from the first trading day of 2018 to last trading day of 2020. It ought to be notated that we will set up model with features in the overall dataset and suppose label to be the latter day's BTC price. That is, we regard all features in certain cross-section to be in t-1 period and label to be t period.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Date	trading_date	BTC	Gold	CrudeOil	DogeCoin	ETH	BTC_circulation	BTC_tradingVol	replies_count	retweets_count	likes_count	is_tweet	tweets	google_trends	hashrate
2	2018/1/2	2018/1/2	15006	60.37	1313.7	0.009145	884.44	16778587.5	1567582732				FALSE	77723	189.385	1.60E+19
3	2018/1/3	2018/1/3	15053	61.63	1316.2	0.00932	962.72	16780450	1138506779	3876	9413	70227	TRUE	79086	186.3	1.49E+19
4	2018/1/4	2018/1/4	15039	62.01	1319.4	0.009644	980.92	16782650	1059922924				FALSE	74534	181.447	1.64E+19
5	2018/1/5	2018/1/5	17174	61.44	1320.3	0.012167	997.72	16784437.5	1764169571	2684	18368	112447	TRUE	76404	179.677	1.50E+19
6	2018/1/8	2018/1/8	15266	61.73	1318.6	0.015045	1148.5	16790637.5	1742246004	770	2988	19952	TRUE	30585	177.908	1.56E+19
7	2018/1/9	2018/1/9	14714	62.96	1311.7	0.01342	1299.7	16791475	1182426648	210	220	9162	TRUE	80932	167.747	1.59E+19

Figure 2: Overall Dataset Summary

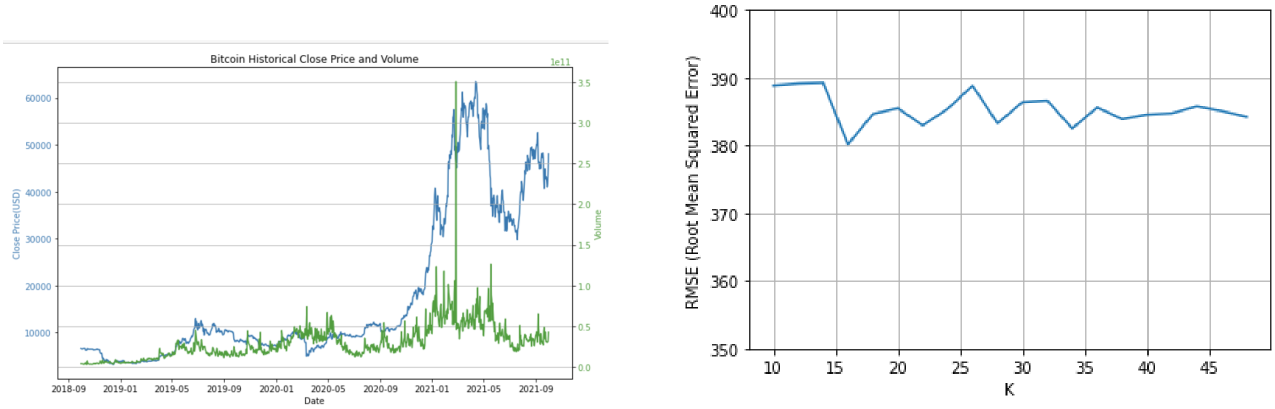
6. **Missing Value Imputation**

Missing value problem is encountered during our data-preprocessing process. There are three categories of missing value:

1. Missing Price Data: We download our data from public website such as Yahoo Finance. There are missing value due to website error. There are 4 days prices missing for ETH and Dogecoin in October 2020. We insert the previous trading price as the missing day value.

2. Missing Elon Musk Twitter Counts: This missing component is actually meaningful. Missing counts of Elon Musk's twitter regarding Bitcoin means that he did not tweetoi on that da. Therefore, we input 0 for this column's missing value.

3. Missing Value for Daily Total Tweets Amount: This feature represents the popularity of Bitcoin on social media. We applied K-means algorithm on this kind of missing value. The result is as below. The best number of neighbours of K is 32 given the Root Squared of Mean Error (RSME) output.



(a) Bitcoin Historical Price and Trading Volume

(b) Missing Value Imputation - K-means Result

Figure 3: Bitcoin Price Prediction With ARIMA Model

3 Model Construction

1. **ARIMA Model : Could Bitcoin Historical price influence its future performance?** ARIMA model stands for Auto - Regressive Integrated Moving Average. It is commonly used for time series data. Forecast could be made based on series past value. The ARMA process satisfies following formula:

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + ... + \phi_p(Y_{t-p} - \mu) + ... +_1 \epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_p\epsilon_{t-q} + \epsilon_t \tag{1}$$

Y_t follows an ARIMA(p,d,q) model if $\Delta^d Y_t$ follows an ARMA(p,q)model. In this report, we choose $d = 1$, which is to conduct first time differentiation to obtain a stationary series. Then we apply auto.arima in python to automatically select the parameter. From the result, the parameter is selected as $SARIMAX(1, 0, 0) * (2, 1, 0, 12)$.

2. **VAR Model: How Will Commodity Price Influence Bitcoin Performence?**

From literature like Okorie and Lin (2020) and Kyriazis (2020), VAR model which uses the lag term of all endogenous variables in the system is a good way to analyze the dynamic relationship among the returns of Bitcoin, gold, and crude oil. We can represent the moving average version of the VAR process, assuming covariance stationarity, by Equation (1). The Fourier transformation to θ_h coefficients is done using Equation (2).

$$X_t = \theta(L)t = \sum_{h=1}^{\infty} \theta_h \epsilon_{t-h} + \epsilon_t \tag{2}$$

$$\theta(e^{-ih\varpi}) = \sum_{h=0}^{\infty} e^{-ih\varpi} \theta_h \tag{3}$$

3. **SVM (SVR) Model**

We gather all the dataset (Bitcoin price, commodity price, Bitcoin popularity on twitter, and Elon Musk twitter comments) and use SVR model to make a prediction of bitcoin price trend. SVR model uses the same algorithm as SVM but on regression tasks, which constructs a hyperplane in a high dimensional space. The objective function is:

$$\frac{1}{2}||w||^2 + C \sum_{i=1}^{N} (\xi + \xi^*) \tag{4}$$

As SVR model requires data normalization, we preprocess the dataset X into a normalized one using standardization and apply SVR.

4. **Random Forest**

In order to conclude with the optimal tree-based model for our dataset, we applied both K-means algorithm and random forest to impute missing data and search for best random forest model. The following steps are taken to achieve our construction:

1. Grid Search on Hyperparameter: We define possible range of K, even integers in [10,50].

2. Iteration with K-means Algorithm: Perform the imputation on original dataset with the current K value in the loop.

3. Application on Random Forest: First split the dataset into training and testing subsets. Then fit the Random Forests model and predict on the test set.

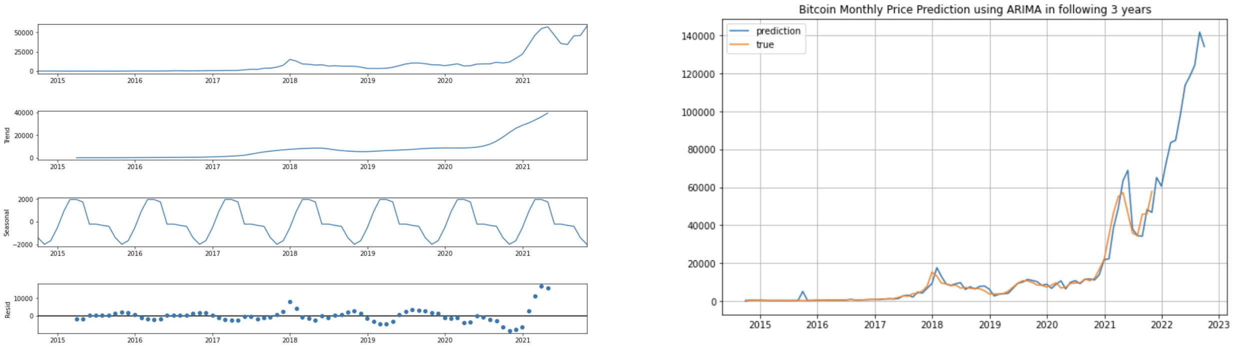
4. Evaluation: Evaluate models using RMSE (Root Mean Square Error), and conclude the number of neighbours with lowest RMSE (K=32).

Random forest model does not require standardization of dataset, which differs from SVR model. We use the parameter, max depth of trees, to be 5, according to manual attempts. According to the model results, we can give out estimated feature importance in the below part.

4 Model Evaluation

1. **ARIMA Model**

The regression and forecast result of Bitcoin price is as below in Figure 3. The fitting of the Bitcoin price is generally well. From the plot, the prediction shows that the bitcoin price will continue to roar after going back up from mid of 2021 all the way up until 2023.



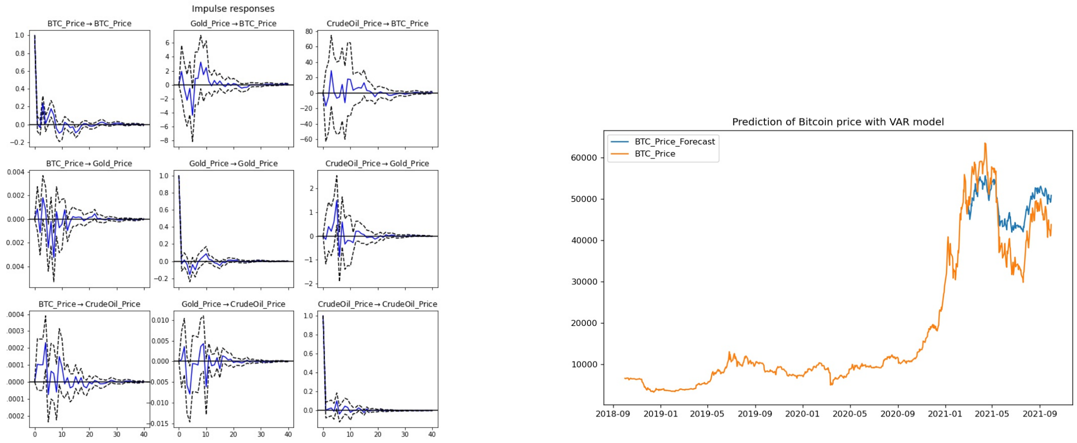
(a) Seasonal Decompose

(b) ARIMA Prediction Result

Figure 4: Bitcoin Price Prediction With ARIMA Model

2. **VAR Model**

Due to ADF test, we construct a stationary sequence by first-order difference for regression. The prediction result of VAR model and impulse response analysis are shown in Figure 2 below.



(a) Impulse Response Analysis

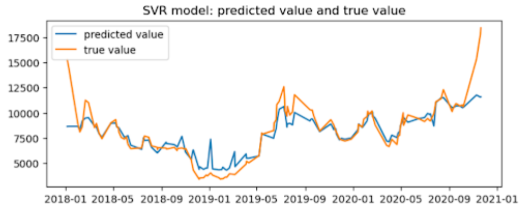
(b) Prediction Result

Figure 5: Prediction With VAR Model

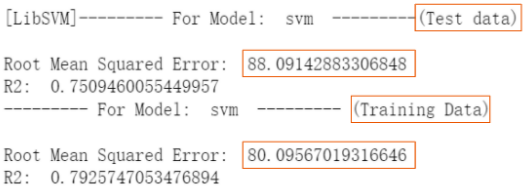
We can see that in the test set, although there are certain errors regarding the predicted values, VAR model fits the trend of Bitcoin price well. From the impulse response diagram, the short-term impact of gold price on Bitcoin price is positive, while the short-term impact of crude oil price on Bitcoin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

3. **SVR Model**

The result of the trained SVR model on test set is showing in Figure 5 below. We can see that the predicted value is very close to the trained value, and the model fits well in the test set for both the trend and the value. The test RMSE of SVR model is 88.09 and the R-square is 0.75, while the training RMSE is 80.10 and the R-square is 0.79. Therefore, SVR model predicts well on Bitcoin price and there's no overfitting problems.



(a) Prediction Result



```
[LibSVM]--------- For Model:  svm  ---------(Test data)

Root Mean Squared Error:  88.09142883306848
R2:  0.7509460055449957
--------- For Model:  svm  --------- (Training Data)

Root Mean Squared Error:  80.09567019316646
R2:  0.7925747053476894
```

(b) The RMSE and R-Square

Figure 6: Prediction With SVR Model

4. **Random Forest**

The prediction result of random forest is shown in Figure 6 below. We can infer that with the feature t-1 period BTC price, random forest predicts well in test set, with RMSE 360.50. When we drop the autocorrelation feature from our model, our prediction is worse in the test set, with RMSE 763.08. Without autocorrelation, the model prediction becomes less alert to large fluctuation and local high or low peaks, thus presents inaccuracy in spite of smoothness.

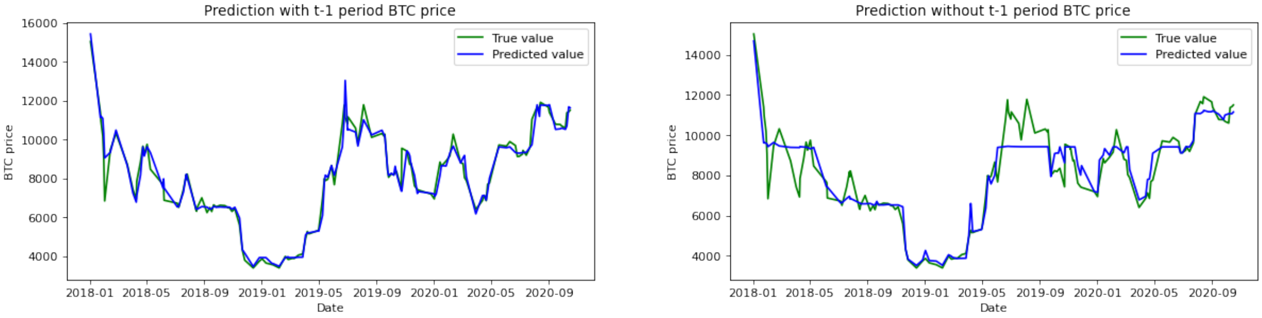


Figure 7: Prediction With Random Forest

The feature importances using Mean decrease impurity is shown in Figure 7 below. According to our results, we can conclude that BTC price's historical performance, which refers to the t-1 period, has the biggest impact on future BTC prices (t period). And it seems to be most related to the t period price, with coefficient 0.98, while the coefficients for other features near 0. To develop impacting factors aside from autocorrelation, we also drop t-1 period BTC price and fit the model again to evaluate feature importance. The important features follow by crude oil's daily price, BTC total circulation amount, ETH price and Dogecoin price. Thus, we conclude that feature's impact on BTC price from high to low is: BTC's historical price, crude oil price, other cryptocurrency's price, and other features including tweets' data, celebrity's impact, gold price, etc.

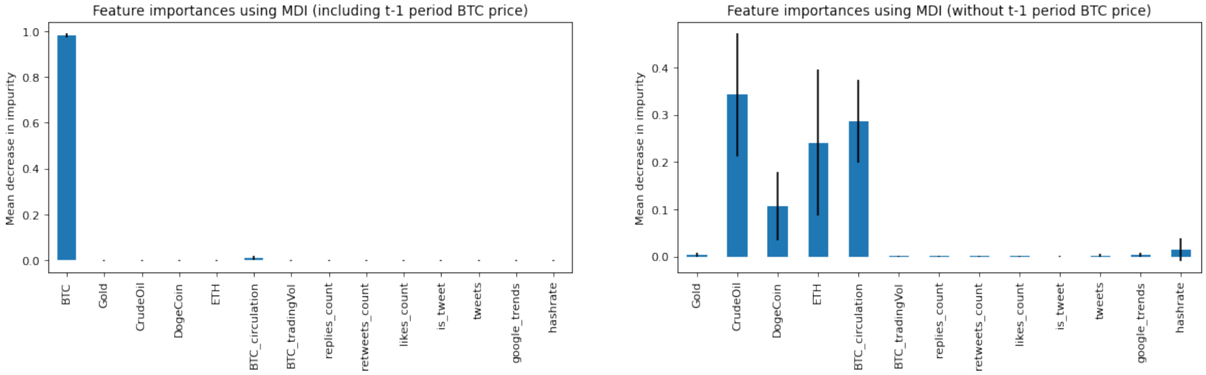


Figure 8: Feature Importances using MDI

5 Fairness, limitation and Conclusion

1. **Fairness and Limitation**

   Fairness of the model is critical, since it could have big impact on the result and bias available information. In this report, we reflected on the limitation and fairness of the model.

   The data used might be limited. The total circulating amount of Bitcoin are underestimated. Some of the Bitcoins in circulation are believed to be lost forever or become unspendable. They are due to lost passwords, wrong output addresses or mistakes in the output scripts. In addition, the total trading volume of Bitcoin are also underestimated. The volume in this report is only from a small proportion of exchanges and the actual total trading volume is much higher. Part of the trading volume is also made outside exchanges, for example in the Over The Counter (OTC) market. For the trading days, this only use trading days in favor of the characteristics of commodity price of gold and crude oil, but bitcoin are traded 24/7. Considering only part of the value might cause unfairness to final result.

   Apart from the data used, the algorithm itself has some limitations. The data we collected related to social media platform has a lot of access restrictions and limitations. We can only get data up to end of 2020, which are not most up to date. Thus, there is rather larger training error, regardless of how we adjust the parameters including max-depth in random forest or number of neighbours choice.

2. **Conclusion**

   To conclude, by applying different method of regression, this report found that the past price of Bitcoin, other commodity price, and a combination of 14 features could have a satisfactory regression performance on the Bitcoin prices.

   SVR model, the regression version of SVM model, predicts well on Bitcoin price with RMSE 88.09 and R-square 0.75. Its similar RMSEs on the test set and training set indicate that the model doesn't have an overfitting problem.

   Compared to SVR model, random forest is not giving out a better prediction. The model predicts with RMSE 360.50 (2.40% of BTC stock price), which may due to high autocorrelation and limited data. Features' importances on BTC price from high to low are: BTC's historical price, crude oil price, other cryptocurrency's price, and other features including tweets' data, celebrity's impact, gold price, etc.

6 Reference

1. https://finance.yahoo.com/

2. https://www.blockchain.com/charts/market-price

3. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

4. https://www.kaggle.com/ayhmrba/elon-musk-tweets-2010-2021

5. https://www.kaggle.com/kaushiksuresh147/bitcoin-tweets

6. https://bitinfocharts.com/comparison/bitcoin-tweets.html3y

7. https://www.machinecurve.com/index.php/2019/09/20/intuitively-understanding-svm-and-svr/summary-support-vector-machines-and-support-vector-regression

8. Bouri, Elie, Mahamitra Das, Rangan Gupta, and David Roubaud. 2018a. Spillovers between bitcoin and other assets during bear and bull markets. Applied Economics 50: 5935–49.

9. Bouri, Elie, Rangan Gupta, Amine Lahiani, and Muhammad Shahbaz. 2018b. Testing for asymmetric nonlinear short-and long-run relationships between bitcoin, aggregate commodity and gold prices. Resources Policy 57: 224–35.

10. Gkillas, Konstantinos, Elie Bouri, Rangan Gupta, and David Roubaud. 2020. Spillovers in Higher-Order Moments of Crude Oil, Gold, and Bitcoin. The Quarterly Review of Economics and Finance.

11. Panagiotidis, Theodore, Thanasis Stengos, and Orestis Vravosinos. 2019. The effects of markets, uncertainty and search intensity on bitcoin returns. International Review of Financial Analysis 63: 220–42.