Thesis Title — Forecasting in Cryptocurrencies

Student ID	—	31125301
Full Name	—	Simran Singh Gulati
Course	—	FIT5128
Supervisor	—	Dr. Christoph Bergmeir
Word Count	—	5857

Table of Contents

Abstract

In the last decade there has been a tremendous increase in Bitcoin's market value and the public attention around Cryptocurrencies. Its highly volatile nature has ignited a great buzz among machine learning practitioners to predict its prices in real time. In this study we have scraped over 30 time series encompassing global macroeconomic factors, blockchain attributes and public opinion to forecast the Bitcoin price. The economic features include influential stocks, fiat currencies, gold and crude oil prices. The technological features include hash rate, mining difficulty, block size, etc. Lastly, the public sentiment was drawn using number of Tweets and Google Trends. We deployed a Random Forest to draw non-linear relationships among the mentioned feature set. The model predicted log-returns which were further used to compute the actual price of the Bitcoin for the upcoming day. The study focussed on last 7 years of data which had the most fluctuating trends. Thereby the dataset was split into different periods to account for potential market cycles. Finally, the model predictions were compared against the naive forecasts. Although the algorithm had at par performance with the naive forecasts, a Diebold-Mariano test revealed that the forecasts were statistically insignificant.

Introduction

Cryptocurrency refers to a virtual asset functioning as a medium of exchange over the internet. Bitcoin, the first cryptocurrency was devised during The Great Recession (2009) as a response to the failure of traditional financial systems. It was used as a reward for developing distributed applications with an intent to promote a decentralised payment network. This digital ledger was termed Blockchain and it aimed to reduce the reliance on financial institutions acting as intermediaries in a monetary transaction. Thereby the Blockchain acts as a peer-to-peer network wherein each transactional information is stored on multiple sources instead of a central server. Initially, developing these distributed systems demanded intense computational power but with improvements in technology we now have access to frameworks that can do this smoothly.

With over 6000 different types of the cryptocurrencies available in the market, Bitcoin holds the largest market capitalisation and remains the most valuable one. It was effectively worth $0 in 2009 and maintained a steady rise and fall up till the year 2017. In 2017, it grew exponentially from $1,000 to $20,000 and followed a constant decline till it stabilised around $4,000. In the year 2021, where COVID-19 has disrupted even the strongest economies worldwide, Bitcoin proved to be virus-immune and crossed a whopping $60,000. This is the highest ever exchange-price for any cryptocurrency. Considering the growing awareness about Cryptocurrencies and Bitcoin's volatility, especially in recent years, there has been great enthusiasm in the forecasting community to decode the underlying trends and predict its prices in real time. While majority of studies focus solely on historic pricing for predicting returns, a part of those investigations are also trying to unveil the latent factors that may affect the Bitcoin volatility. These latent factors include the blockchain attributes, the media sentiment and the global macroeconomic factors.

Global economies like stock market, exchange rates and commodity prices have been forecasted a lot in the past. The general consensus is that returns are not predictable but volatility is, and these markets are interdependent. A financial shock to any of the three definitely affects the monetary standing of the other two. Our research is targeted at scrutinising if this hold true for cryptocurrencies as well. We use time-series forecasting for Bitcoin to compare how predictable they were in the past and how predictable they are in 2021. We also test which features or markets influence Bitcoin price the most.

Unlike fiat currencies, a cryptocurrency is built on the concept of decentralisation wherein its production and trade entirely depends on the creation of blocks in a blockchain. Thereby the technological features such as block size, hash rate, mining difficulty, etc. have shown some sort of influence on the Bitcoin supply and hence its exchange price. Additionally, cryptocurrency trade is highly dependent on investor's belief in the blockchain technology. The public interest unquestionably drives the transaction flow and consequently its market price. The more the people learn about it, the more likely they are to engage with it. This internet popularity or media sentiment can be measured via the number of social media mentions like Tweets and Google Searches. Bitcoin has also been compared with Gold and the U.S. Dollar to draw conclusions about its volatility. In an attempt to classify Bitcoin as an asset or a medium-of-exchange, it has been proven that global commodities like Gold and Crude Oil, and some fiat currencies (EUR, GBP, JPY, CNY, etc.) can hugely impact the Bitcoin's market value. Considering that most of the Bitcoin users are not miners but investors, it is crucial to include the macroeconomic features for forecasting.

In nutshell, the Bitcoin functions on a distributed network. Thereby *network complexity* and public's interest in this decentralised system influences Bitcoin volatility. The *public interest* is not just limited to media awareness but public's financial stability as well. Therefore it makes sense to also include the *global economies* for Forecasting in Cryptocurrencies.

To summarise, cryptocurrencies are dynamic in nature which makes it hard to predict their prices using just the historic trends. So we included a wide variety of feature sets that potentially influence the Bitcoin price. Our work produces results that are at par with Naive Forecasts, the most reliable technique for forecasting the dynamic trades (like stock market). We split the data into notable periods of price gain and the model predictions were statistically tested against naive-forecasts using a Diebold Mariano Test. The study reveals that the forecasts across all periods, despite a considerably low error in each, were insignificant. Thereby we reject the hypothesis that Bitcoin prices are predictable.

I would conclude by saying that neither bitcoin was predictable in the past, nor is it now.

Background

This section scrutinises the bitcoin-forecasting studies to look for potential determinants of the Bitcoin price.

Unlike fiat currencies, Bitcoin trade is entirely based on mining new blocks in a blockchain. [1] further extends this belief to confirm that Bitcoin volatility is highly influenced by the blockchain attributes such as block-size, hash-rate, miner's revenue, etc. Although, they also confirm that number of transactions and mining difficulty do not affect the pricing in any way.

Being built on a distributed system, each of the network nodes has information that contributes in a transaction flow. This creates a weighted network for the blockchain. [2] evaluates the information held by each node using some statistical measures. Thereby proving that network complexity is highly significant for predicting bitcoin volatility and returns.

[3] compares the involvement of transaction-volume in determining Bitcoin returns. The results imply that transaction volume can accurately forecast the pricing if the market is stable and functioning around the median. Consequently, it can not predict it very well if the market is performing extremely good or extremely bad.

A fiat currency's global standing is subject to its country's macroeconomic status which is governed by criterions like GDP, inflation, unemployment, etc. [4] investigates this concept to affirm that Bitcoin market does not align with traditional economies, instead the investor's sentiment plays a great role in driving Bitcoin returns. His research encompasses Bitcoin's internet popularity via the search queries on Google and confirms that Google Trends is pivotal in determining the Bitcoin's exchange price.

Frequent mentions about Bitcoin on news channels positively affects their daily returns. [5] explores this viewpoint in greater depths to conclude that public awareness unquestionably drives the transaction flow. The more people hear about it, the more likely they are to invest in it. Hence greater media presence leads to higher exchange rates.

Including new information in a trade directly influences its returns and volume. While price-shift reflects the market's response to the new information, volume-drift reflects the investor's understanding of this information. Examining price-volume correlation can expose rate, extent, retention of market information and the degree to which prices indicate market information. Small fluctuations in price are more persistent than those in volume. On the other hand, large fluctuations in volume are anti-persistent. [6]

According to [7], Bitcoin's characteristics can be described as a mix of Gold and U.S. Dollar. Firstly, Bitcoin is limited and very expensive to mine just like Gold which also makes it very easy to trade as it can simply be exchanged for money. On the other hand, it does not possess any intrinsic value and acts just as medium of exchange. Thereby its volatility stands somewhere in between the Gold and the U.S. Dollar. This claim was further investigated by [8] and their research produced quite the opposite results. It says that Bitcoin is a much riskier asset than Gold. Its volatility is one of its kind and cannot be associated with gold, stock exchange or even fiat currencies.

Yet another study by [9] tries to decode the effect of global commodities on Bitcoin. It shows that there is strong positive correlation between Gold and Bitcoin. Gold prices can highly influence Bitcoin returns and have great influence on its volatility. Contrastingly, fiat currencies like Euro, Pound or Yen are negatively impacted by the changes in global assets like Gold or Crude Oil.

As Bitcoin works on a distributed system, its trade hugely depends on the cumulative power produced by all the computers involved in the mining process. Conversely, majority of the users are not miners (or blockchain developers) but people who have invested money in Bitcoin. Thus [10] produced a report to compare the effect of technological and economical features on Bitcoin returns. Surprisingly, the study exhibits that technological attributes have no impact on the bitcoin price in the long run, rather the economic variables tend to show great influence on its returns. Mining difficulty used to have significant control in driving the exchange price but due to the advancements in technology its hold tends to decrease over the years.

According to [11], the last decade can be split into four periods of significant price gain for the Bitcoin. It originates from Information Share (IS) and Activity Share (AS) for each of the leading exchange operators. While there are plenty of operators dealing with Bitcoin exchange, Mtgox and Btce remain the principal catalysts for determining daily returns. Beginning with the steadier of the two, Btce holds a strong control over the trade and remains the most influential exchange throughout. On the other hand, Mtgox greatly influenced the price-discovery in the early years but later went onto a downward trend and lost its impact eventually. Other operators, especially the smaller ones, simply follow the market with a lag and do not actually affect the pricing. In totality, the time period and corresponding IS/AS ratio have significant effect on the Bitcoin's market capitalisation.

According to the literature covered here, we can see that a lot of these studies are focussing either on historic pricing or on a potential feature-set. This potential feature-set is often one of the following — blockchain's complexity and other network attributes, public opinion or social popularity, and finally the global macroeconomic factors like stock market, fiat currencies, gold price, crude oil price, etc. None of the current methods have combined all these features to predict the Bitcoin's exchange price. It would be wise to say that researchers are either using just the price or a few other variables to forecast upcoming trends. As evident from the current works, these features as independent studies have shown some sort of influence in the exchange price. Thus their collective contribution should definitely produce forecasts with lowest error rate. Hence we can also compare which variables affect our forecasts the most.

We would like to combine 30 potential determinants (explained in the next section) into one huge feature set and then try to predict the bitcoin price. Once we have reasonable forecasts we would like to see what factors affect bitcoin volatility. Most importantly, we would like to split our data based on the time-period and see *if and when was bitcoin most predictable*. Or we can say, is bitcoin as predictable today as it was in the past? Was there a time in history when we could confidently predict the bitcoin volatility? If it is, *then what features affect bitcoin's price the most*. Do these dependencies change over time?

Methodology

In this section we will look at data preparation, forecasting techniques and statistical testing for our forecasts.

1. Data Preparation

As discussed in literature [1,2,3] blockchain attributes seem to be a strong feature-set for determining the bitcoin price. The website bitinfocharts.com includes historic information about the technical data associated with bitcoin. It presents the statistics in the form of dynamic web-charts. Thus the below listed *daily time-series* about the blockchain data were scraped from the above mentioned website:

1.1 Blockchain attributes

- **Mining Profitability** : Profit in US dollar for mining 1 THash per second
- **Transactions** : Total number of transactions
- **Market Capitalisation** : Number of Bitcoins sent in US dollar.
- **Transaction Fee** : Handling fee in US dollars for all the transactions in a day.
  - Average
  - Median
- **Transaction Value** : The economic value between two parties for bitcoin exchange.
  - Average
  - Median
- **Confirmation Time** : Average time in minutes to complete a decentralised transaction.
- **Block Size** : The size of data (in Megabytes) that is permanently recorded.
- **Fee Reward** : Average fee percentage in Total Block Reward.
- **Hash Rate** : Cumulative computing power used to mine.
- **Active-addresses** : Number of unique addresses each day.
- **Bitcoins sent** : Total number of Bitcoins sent in US dollar.
- **Mining Difficulty** : Average difficulty to mine a block in blockchain.
- **Top Addresses** : Percentage of total coins held by the richest 100 addresses.

Similarly according to [8] and [9], frequent media mentions can also influence the bitcoin price. So we include two new variables encompassing internet popularity:

1.2 Media Sentiment

- **Tweets** : Total number of tweets about Bitcoin in a day
- **Google Trends** : Total number of Google Searches mentioning the keyword “Bitcoin”

While tweets indicate the ongoing discussion about exchange price, Google tends to show the public’s interest in learning about bitcoin and the underlying network technology.

Unlike most data-repositories with a standard tabular-representation, this website contains interactive graphs for each attribute. While this visual-representation allows better understanding of the trends, we need numerical data for producing a model. The tutorial available at [21], shares the code to extract numeric values from JavaScript component of the graph. The reference is concentrated on getting the tweets-data from the website, so we automated a script to follow the same steps for all the relevant attributes. Thereby producing a table with 4657 rows and 18 columns, containing daily records for each of the 18 attributes.

As mentioned by [10], [8] and [9] — global commodities and macroeconomic features affect the upcoming markets and newer economies. We refer Yahoo Finance website to download historical data about the below listed features. All the exchange prices obtained are in US dollar.

1.3 Macro-Economical Factors

- Gold Price
- Crude Oil (WTI)
- Influential Stocks
  - S&P500
  - DOW30
  - FTSE
  - NASDAQ
  - Volatility Index
- Fiat Currencies
  - Euro
  - Pound
  - Japanbese Yen
  - Chinese Yen
  - Swiss Franc

Yahoo Finance enables its users to download financial data using a Python Package [22]. Although they do not permit redistribution of data, it can be used for research purposes. Thus an API call enabled us to download historic data about the above-mentioned 12 macroeconomic factors. The data collected was in a multi-level format, with 6 features for each of the macroeconomic attribute:

- **Open** : The price of the stock when the market opens.
- **High** : The highest price of the stock in a day.
- **Low** : The lowest price of the stock in a day.
- **Close** : The price of the stock when the market closes.
- **Adjusted Close** : The close price adjusted after using splits and dividends.
- **Volume** : The number of shares for the stock that were traded in a day.

While traders are keen on opening price as it reflects a stock's demand-and-supply mechanism, it is the closing price that reflects a stock's performance for the day. [12] analyses the above mentioned features to study the informational-content stored within them. Their investigation reveals that close-price contains information that is useful for forecasting.

Thus we discard the other features and compose a new dataset with closing prices. The data from Bitinfocharts and Yahoo Finance is merged using an Outer Join, thereby creating a table with 4584 rows and 30 columns. Each column represents the 30 separate features that we have listed above and the dataset is indexed on date. Finally we have 12 years worth of financial data which can potentially determine the bitcoin's exchange price.

After procuring all the time series, we plot line-charts to visualise their trends along the timeline. From graphs in Figure 1, we can observe that along with the bitcoin's exchange price — its market capitalisation, number of transactions, block size, mining difficulty, hash rate as well as number of active addresses have gone up as well. Similarly the influential stocks including S&P500, NASDAQ, DOWJONES and FTSE have also experienced a steady gain over the years.

Contrastingly, mining difficulty has significantly gone down, presumably due to advancements in technology and upcoming blockchain frameworks that simplify the development process.

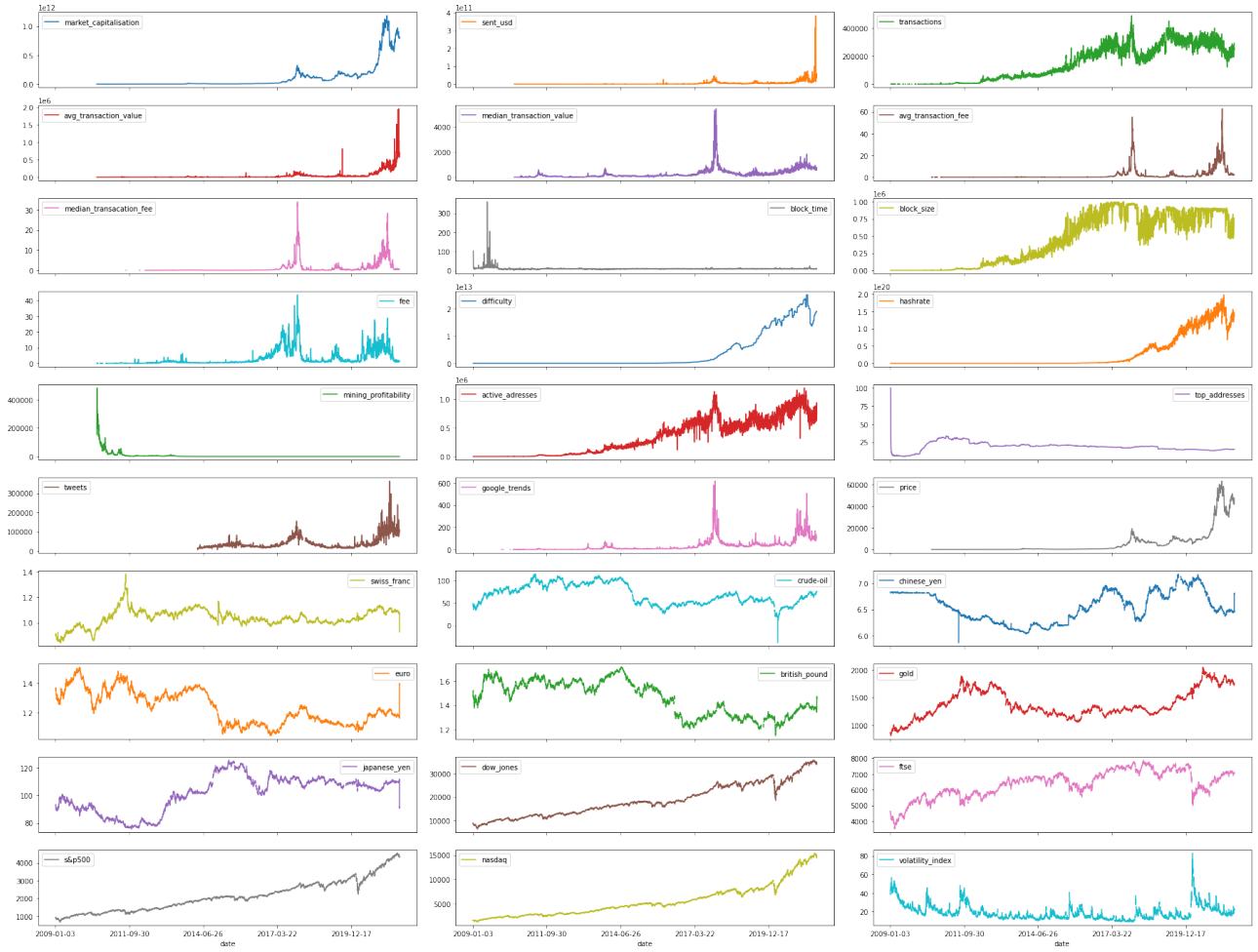


Figure 1: Price trends for all the potential determinants of the bitcoin price

Looking at rest of the series — transaction fee for mining, block time, media mentions through tweets and google have experienced occasional peaks during the last decade but eventually come down to the same value. On a whole, despite the sudden rise and fall, media mentions have definitely risen over the years.

Similarly, there's no definitive trend with fiat currencies. They do fluctuate rapidly but can not be associated with a conclusive gain or loss.

A lot of the columns have empty values in the initial rows. On comparing the technological features with the economic ones, we can see that majority of blockchain attributes have absolutely no records for the first 2 years of the bitcoin's launch. Likewise, the data from Yahoo Finance has no record for the weekends. The latter is understandable because the stock market is open for trade only from Monday to Friday.

As evident in Figure 2, most of the technological attributes (market capitalisation, bitcoins sent, transaction value, transaction fee, mining profitability) have exactly 563 rows with no data. This implies that data is entirely missing for a period of 18 months. Mostly importantly, even the dependent variable "bitcoin's exchange price" has no account for these months. Likewise we do not have any record about the tweets on Bitcoin before the year 2014. These records can be classified as Missing Completely at Random (MCAR) [13].

market_capitalisation	563
sent_usd	563
transactions	263
avg_transaction_value	563
median_transaction_value	563
avg_transaction_fee	643
median_transacation_fee	1207
block_time	10
block_size	8
fee	643
difficulty	8
hashrate	9
mining_profitability	563
active_adresses	30
top_addresses	14
tweets	1980
google_trends	569
price	562
swiss_franc	1358
crude-oil	1472
chinese_yen	1486
euro	1355
british_pound	1355
gold	1473
japanese_yen	1355
dow_jones	1447
ftse	1444
s&p500	1447
nasdaq	1447
volatility_index	1447

Figure 2 : Number of empty records in each

While there's no rational approach to impute the records for early years, weekend data can easily be imputed. Considering that stock market works Monday through Friday, we can simply use the closing price on Friday as that on Saturday and Sunday as well. Instead of deleting the other attributes' weekend values, it would be logical to estimate weekend prices for the stock exchange. Thus we use a forward-fill methodology which simply propagates the last observed entry for the missing one.

Coming back to the missing values for initial years, another observation from the trends is that bitcoin's exchange price did not change much until 2017. Although it increased from mere $0 in 2009 to a whopping $1000 in 2016 but its volatility post 2017 was exponentially high and way riskier in comparison. Thus we decided to keep data only 2015 onwards.

The final dataset comprises of 2400 records from January 2015 to September 2021. At the moment, each row comprises of 30 features for the current day's value which would be used to forecast the upcoming day's exchange price. Thus we create a matrix type representation wherein each row is fed in with values for current day as well as those for the last 2 days.

Ultimately we have 2400 rows and 100 columns with each row containing values for all the features for current day, the day before and the day before yesterday. We also include information about day of the week using one hot encoding.

2. Modelling

Random forests is an ensemble modelling technique in which a large number of decision trees are constructed on different data samples from the same distribution. The predictions from each decision tree are considered as a vote and the majority constitutes the final prediction. The key advantage of using random forests is the avoidance of overfitting due to randomised feature selection and the data sampling methodology. [14]

According to [15], Random Forest is robust to noisy data and is better than a Single Regression Tree. The non linearity of cryptocurrency time-series makes Random Forest the optimal choice for bitcoin forecasting. Gradient Boosted Tree (GBT) is another comparable ensemble technique [18] and may even have better performance than Random Forest but we are inclined towards optimal feature selection. More importantly, GBT does not cope well with data involving high noise. Therefore Random Forest is a good representative method, especially when our work focusses on variable importance.

We deployed a Random Forest regressor with 100 trees on the first 90% records. It utilises the squared-error for splitting criteria and was tested on the remaining 10% records. As evident from the graph in Figure 3, there's huge gap between predictions and the true price.

```
MAE:  10381.769028340083
RMSE: 12968.248325363511
```

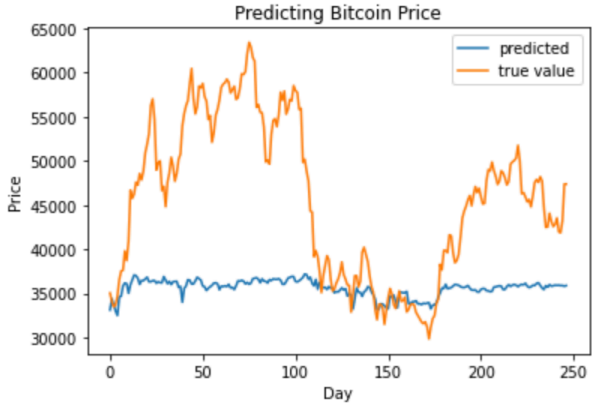


Figure 3 : Random Forest predictions on entire dataset

A reasonable explanation for the huge error gap would be — *the training set never observed the bitcoin price higher than 20,000 US dollars*. In this particular scenario, the model was trained on data from early 2015 to mid 2020. However, the bitcoin price sky-rocketed late 2020 onwards and hit an exorbitantly high price of 60,000 US dollars in April 2021. The model being trained on a notably smaller price range, could not comprehend the unprecedented 300% price jump in 2021.

The series looks non-stationary with stochastic trends and this needs to be taken into account for better modelling.

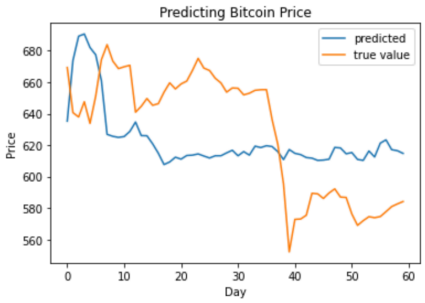
It makes sense to train the model on shorter periods (and test for near future) as their dynamics are changing drastically. The model that worked a few years ago might not have the same predictive power for the current scenario. So we train and test on smaller chunks. Moreover, [11] shares that there are three relevant periods of price gain and each one has different catalysts for that gain:

- January 2013 to December 2014
- July 2014 to December 2017
- July 2015 to July 2018

We split the dataset into 4 sequential subsets of 600 days each and train the model on the first 540 days. The one-step ahead predictions on the test set (60 days) are as follows:

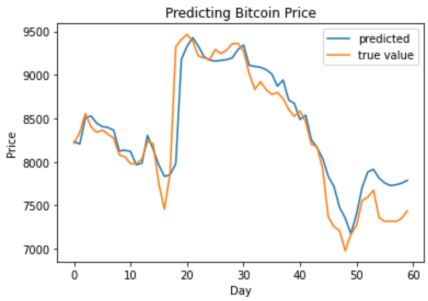
```
Test Set : 1 of 4

MAE:  37.302633666666694
RMSE: 39.435971700684895
```



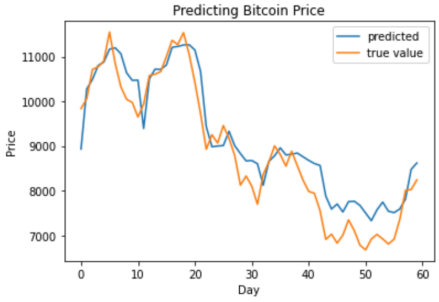

```
Test Set : 3 of 4

MAE:  182.90266666666668
RMSE: 271.9983090756264
```




```
Test Set : 2 of 4

MAE:  441.7508333333334
RMSE: 526.8977925951737
```




```
Test Set : 4 of 4

MAE:  1410.1621666666667
RMSE: 1795.4452286016228
```

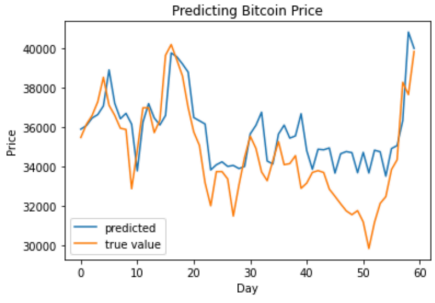


Figure 4 : Random Forest predictions on test-sets with narrower price-range.

While the forecasts on smaller test-sets are much better than those on the entire dataset, the results are still not reliable as there are huge errors. Considering the scale of these price fluctuations, we need a better approach for forecasting. Thus we tried slight changes in the way the forecasts were computed.

[16] states that bitcoin's daily returns are predictable, given they are tested on a narrow range. Since we have already split the data into smaller periods it adheres to the narrowed range constraint. The previous technique was directly predicting the tomorrow's price which fails to account for unprecedented rise or fall. So instead of predicting the price, we now forecast tomorrow's returns and add them to current day's price. This way we aim to reduce the gap between forecasted and actual price. This better utilises the current day information which should be greatly for one-step ahead predictions. We also propose the inclusion of naive forecasts wherein instead of adding the returns, we simply use the current day price as tomorrow's prediction. The naive-forecast follows the market with a lag but should explain if adding returns to current day price is useful or not.

In case these returns are not predictable, it makes our series a random walk and then naive forecast is probably the best method [19]. So we need to ensure that our model performs at par with the naive, if not better.

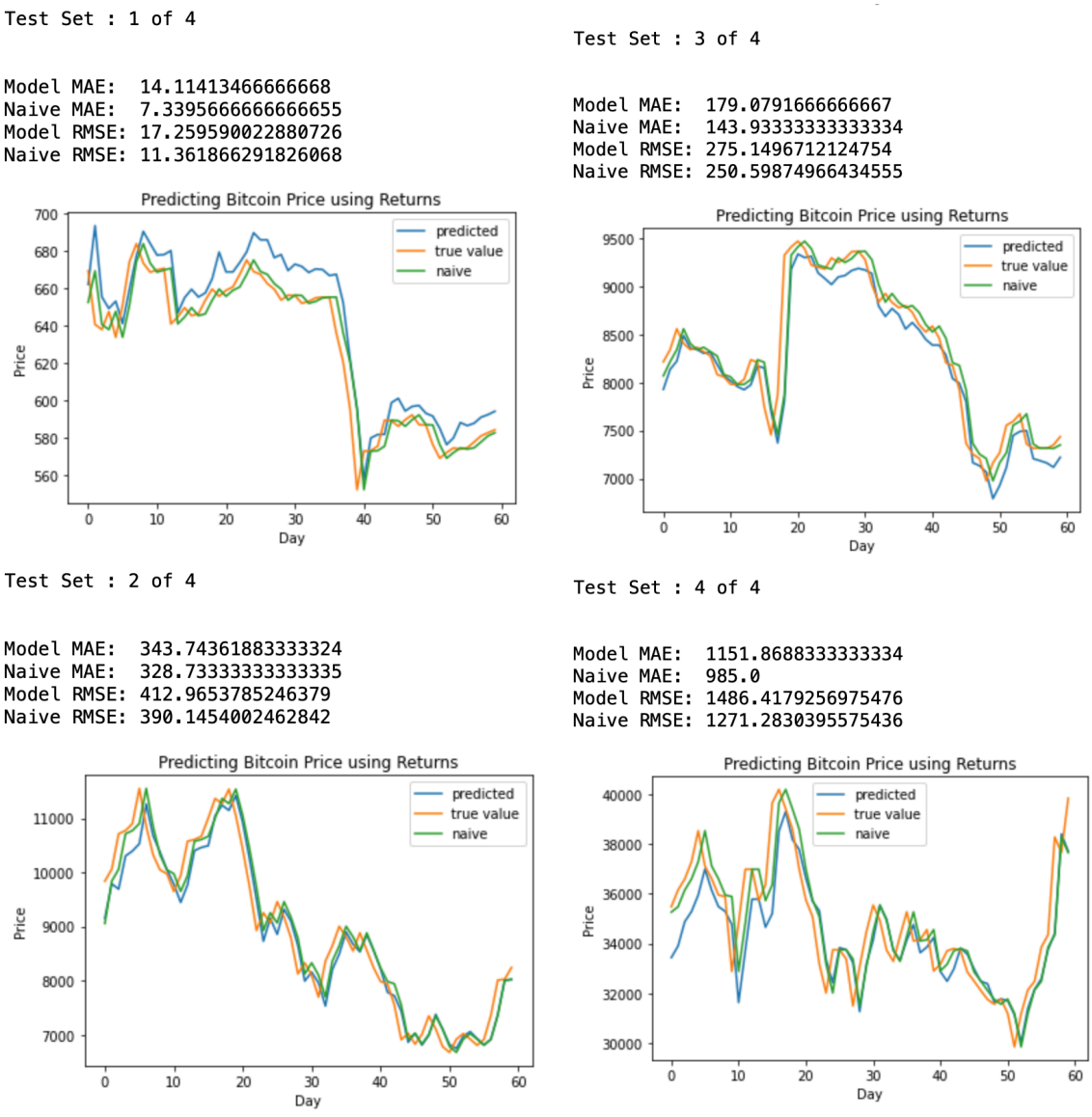


Figure 5 : Random Forest predictions using Returns and performance comparison with Naive Forecasts

Looking at the results in Figure 5, all test sets have lower errors than those in the previous technique. Therefore predicting returns definitely seems to be better approach. Another notable observation from the error-comparison is, that naive forecasts have much lower error when compared to our model. Although the predictions look better across these different datasets, it is of no use if the person simply trusts the current day price to be the actual market trend. Thus the aim is not just to reduce the error but also beat the naive forecast.

According to [17], using log-returns is advantageous in comparison to raw pricing. They have explained that it creates a symmetric representation of price-change and thus the time-series for logarithmic returns can be approximated as stationary.

We try another technique by normalising returns into log-returns using the formula:

$$R_i = \log (B_i/B_{i-1})$$

Where R_i represents log-returns, B_i signifies current day price and B_{i-1} signifies yesterday's price.

So the model would now use the same 4 datasets to predict the transformed log-returns.

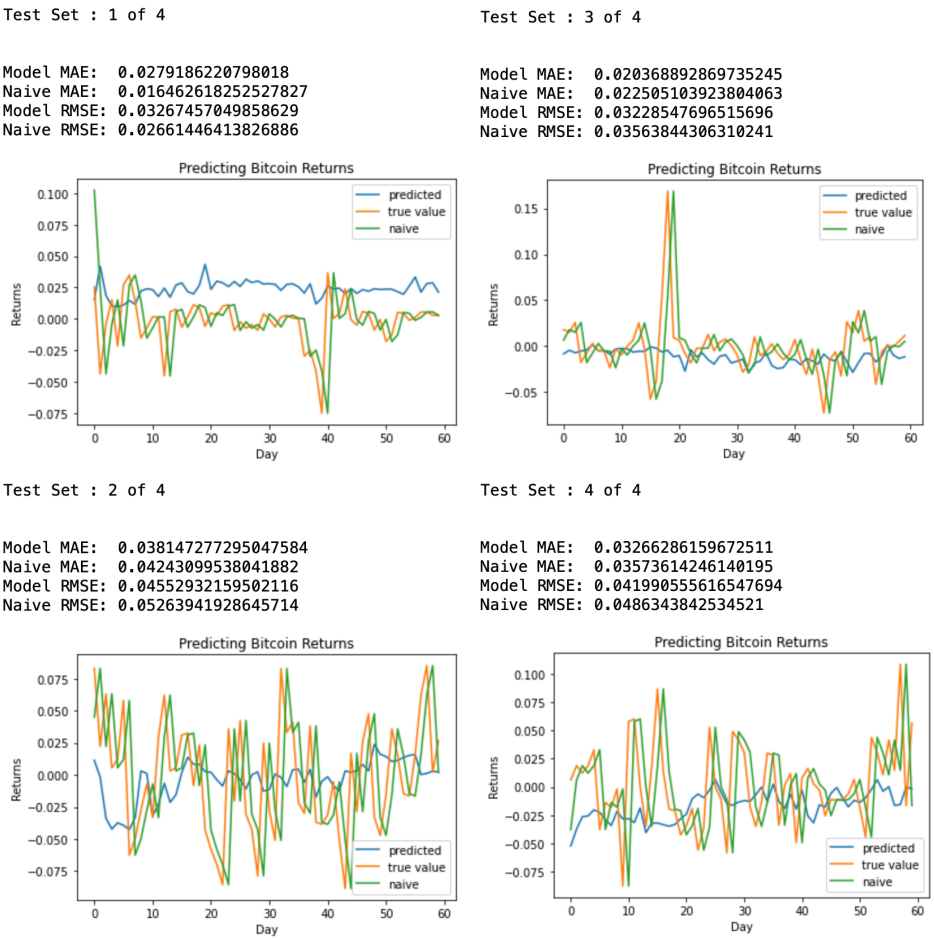


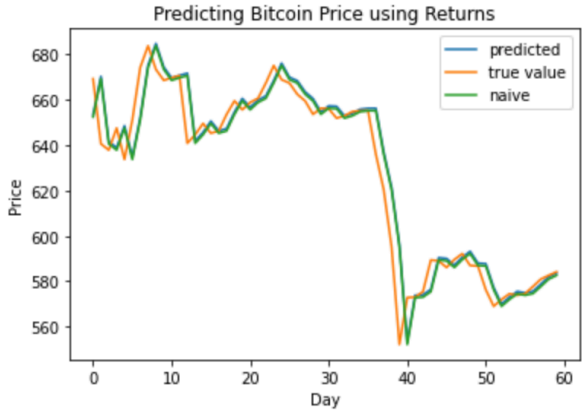
Figure 6 : Random Forest predictions for log-returns

Comparing the results in Figure 6, all datasets except the second one have low errors and our model performs close enough to a constant zero.

This is a validation to further utilise this transformation. Now we predict the log-returns and use an exponential function to convert these back into raw values. Finally these raw-values will added to current day price for predicting tomorrow's price.

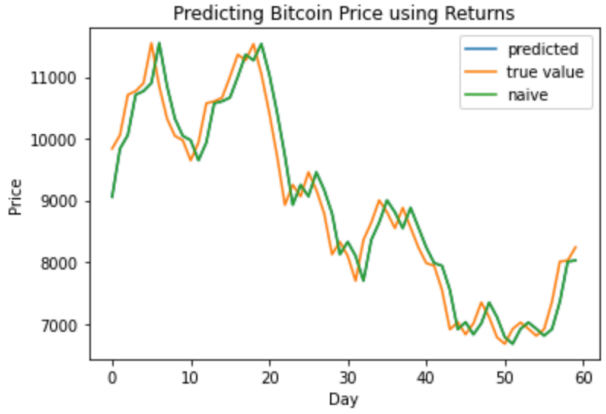
```
Test Set : 1 of 4

Model MAE:  7.2987342165250295
Naive MAE:  7.3395666666666655
Model RMSE: 11.508989975317695
Naive RMSE: 11.361866291826068
```



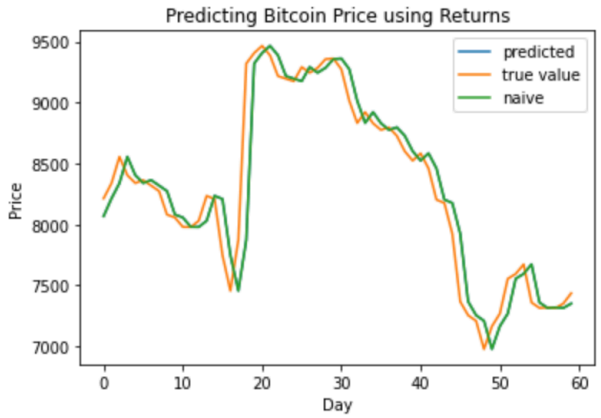

```
Test Set : 2 of 4

Model MAE:  328.73389536194355
Naive MAE:  328.73333333333335
Model RMSE: 390.1801060131269
Naive RMSE: 390.1454002462842
```




```
Test Set : 3 of 4

Model MAE:  144.09783238943874
Naive MAE:  143.93333333333334
Model RMSE: 250.6408360991514
Naive RMSE: 250.59874966434555
```




```
Test Set : 4 of 4

Model MAE:  985.0008774307156
Naive MAE:  985.0
Model RMSE: 1271.2238613882578
Naive RMSE: 1271.2830395575436
```

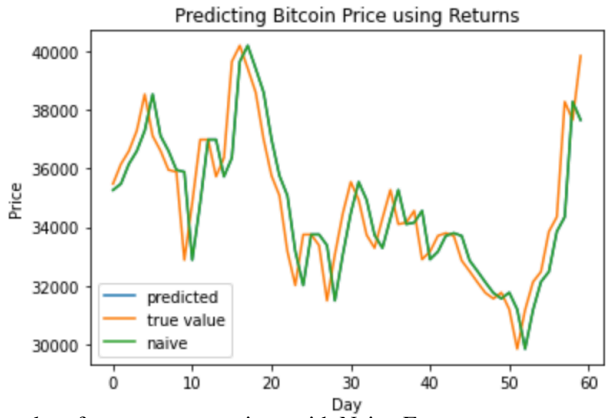


Figure 7 : Random Forest predictions using log-returns and performance comparison with Naive Forecasts

Looking at the graphs in Figure 7, the model predictions are aligned so closely to the naive forecasts that we can not tell the difference visually. This is definitely better than the variance observed in Figure 5 but can not be used as a metric due to high noise in our data. Thus it makes sense to look at Mean Absolute Error and Root Mean Squared Error (RMSE) from Model and Naive forecasts. All the datasets have same RMSE for our model as that of the naive forecast, implying that the model does perform quite well.

3. Statistical Testing

Now that we have a model with comparable forecasts, next step is to ensure that these results are not by-chance. Thus we use a Diebold Mariano Test to see if the results are statistically significant. Diebold Mariano test is a predictive accuracy test to ensure that two sets of forecasts have same forecasting accuracy [20]. We use the Python library developed by [23] to do a DM test on all four test-sets (using Mean Squared Error as test criteria).

Experimental Results

1. Forecasting Techniques

We implemented three forecasting techniques for one-step-ahead predictions. Each regressor model was built in succession, on four different test sets, to account for the dynamic characteristics. In the first approach, we directly predicted the bitcoin's exchange price using 30 independent variables (Forecast 1). We tried to enhance its accuracy by predicting returns and then computing the exchange price from those values (Forecast 2). Finally, we upscaled this by using log-transformation on returns to convert it into a stationary series (Forecast 3).

Each approach was built on top of the last method and thus the errors dropped with each iteration. We used RMSE to analyse the predictions and compared each model's performance against the Naive Forecast. Below is a table that compares our forecasts from each technique.

Dataset	Forecast 1	Forecast 2	Forecast 3	Naive Forecast
1	36.9946	26.0767	11.5098	**11.3619**
2	487.6657	404.6762	390.1809	**390.1454**
3	276.6219	286.8624	250.6412	**250.5987**
4	1936.0871	1498.6399	**1271.2241**	1271.283

Table 1 : Error comparison for our forecasts

As marked in bold, in Table 1 — the first three test sets had lowest error rate with the Naive Forecast. However in the last test-set our model did beat the Naive Forecasts. It has to be noted that among all the comparisons between Naive and Forecast 3, the winner beat the other by only a slight margin (as low as 0.02). So even though our model couldn't beat the Naive in first three cases, the performance is still exemplary. This implies that the model must be producing genuine forecasts. So we check the feature importance for each test set, which was the underlying purpose of this research. That is, we now have a look at the most influential input-variables in each time-period.

Referencing the Figure 8 (next page), we observe that in early times (from 2015 to mid 2016) hash-rate, mining profitability, market capitalisation, block time, Euro and google trends were the top 10 estimators, in the order of their importance. It is clearly evident that blockchain attributes had a great influence on Bitcoin's exchange price and macroeconomic features had no role whatsoever. This is probably because Bitcoin did not get as hyped up as it did after 2017 (on hitting $20,000).

Moving on to the next time period, from late 2016 to early 2018 — tweets, google trends, Chinese Yen, crude oil, transaction count, transaction value and block size seem to be quite influential. As per our expectation, the economic involvement is slightly increasing. Moreover, tweets have also shown some influence by this time.

From mid 2018 to late 2019 — google trends, transaction count, block time, bock size, Japanese Yen, transaction fee, active addresses and transaction value have shown to be quite important in determining the bitcoin price. Surprisingly, there is negligible involvement of macroeconomic features while technological attributes continue to have control over the price. Google seems to be present across all the years up till now.

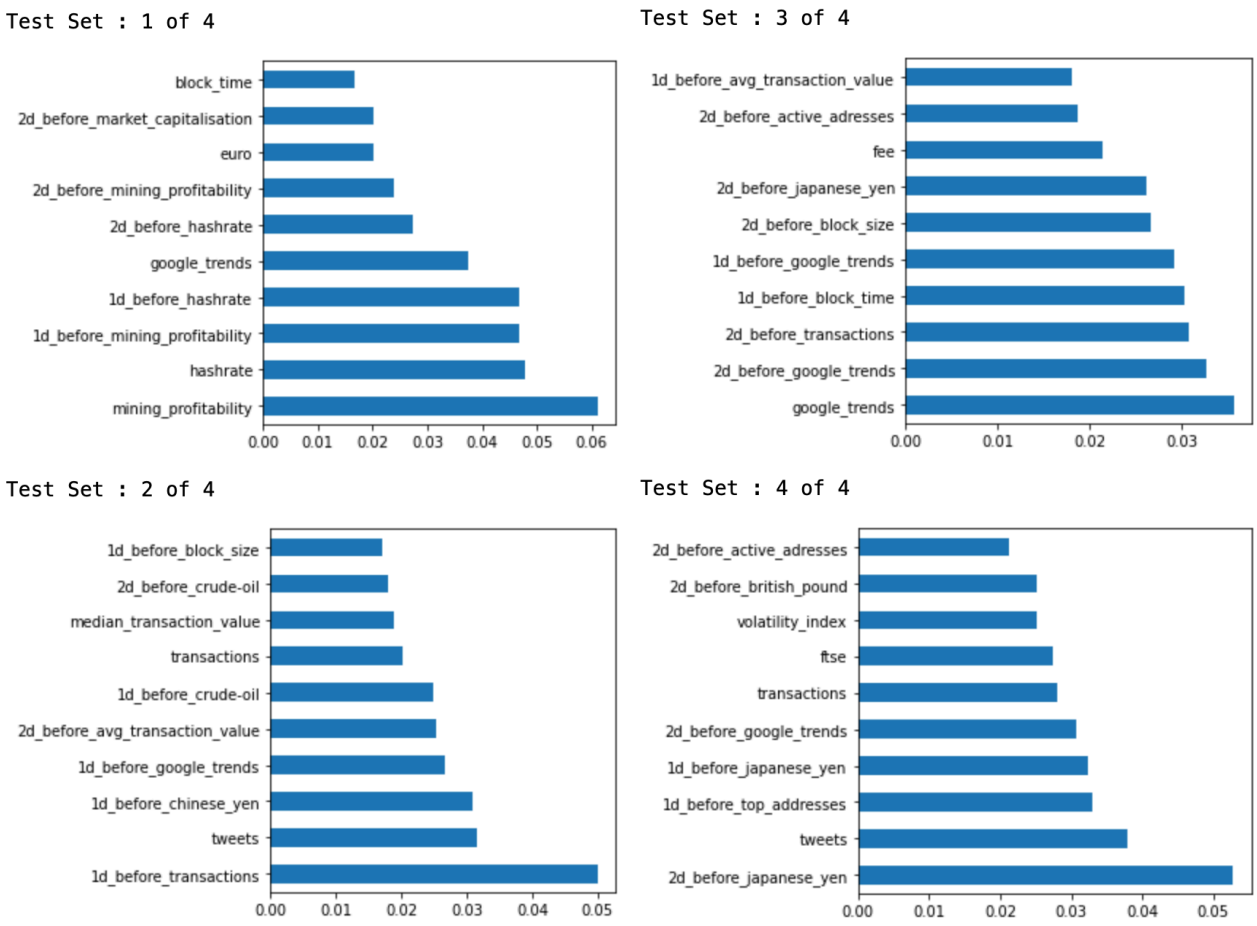


Figure 8 : Feature importance across all test-sets

Finally, we look at the recent months (early 2020 onwards) when the world economy was hit by global pandemic — Japanese Yen, British pound, FTSE, volatility index, top addresses, transaction count, google trends, tweets, constitute the important feature set. Contrastingly, the blockchain involvement has significantly dropped and the economic features have suddenly taken a strong control over bitcoin pricing. Yet again google and tweets are still relevant.

To summarise these trends, there has been a gradual transition from bitcoin's dependence on technological features to that on macroeconomic ones. On the other hand, internet popularity plays a crucial role throughout and is expected to do so in future as well. As blockchain frameworks come up, developing distributed applications is becoming easier than ever. Therefore the technological involvement in price determination is declining with improvements in technology. Similarly, as the awareness about cryptocurrency spreads, more markets and economies are seeking dependence on Bitcoin for financial stability. This also confirms our assumption that determinants of the bitcoin price change over the years.

2. Statistical Significance

Diebold-Mariano test was performed to compare the model forecasts against the naive forecasts.

The test results are attached in Table 2. Considering the high p-value for each of the test sets, *we cannot reject the Null Hypothesis that two forecasts have same level of accuracy*. Thus the alternative hypothesis that our model and naive forecast have different level of accuracy is accepted.

Test Set	P Value
1	0.264628565532027
2	0.782360984738494
3	0.733995049092697
4	0.377103085927652

Table 2 : DM Test Results

It is evident from the results that none of the test sets had statistically significant forecasts. That is, none of them had forecasting accuracy similar to that of a Naive Forecast.

This aligns with the general consensus that dynamic trades like stock market and cryptocurrencies are not at all predictable. If they were, every person would potentially exploit the machine-learning algorithms to make money and become rich.

Markets like these, involve high risk and do not follow any particular trends. They are non stationary time-series that follow stochastic trends. These are dynamic trades that are dependent on unlimited number of factors. A simple explanation would be — blockchain attributes were highly influential for our early test-sets but eventually lost their determining power due to advancements in technology. Similarly, Tweets played an important role in alternating test-sets while Google always stayed influential.

Another literature from [7] says that bitcoin characteristics are somewhat similar to that of Gold. On the other hand, [8] claims that Bitcoin is far from Gold and is in fact a much riskier asset. We could not spot Gold as an important estimator in any of the test sets. To conclude, cryptocurrency is not predictable. Neither was it in the past, nor is it now. Its volatility depends on a huge range of attributes which makes predicting the returns impossible.

It would be very easy to cherry pick a dataset and say that our model predicts the bitcoin price with superior accuracy but in reality — their dynamics change so drastically that it may fail any point in time. The risk involved is so high that even slight misinterpretation can cause high monetary losses. Similarly, if one is lucky enough there might be great benefits as well. In any of the two cases, we can not tell how the returns would be performing tomorrow.

Conclusion

Bitcoin is the most valuable cryptocurrency and has been extensively compared to Gold or U.S. dollar to classify it as a medium-of-exchange or an asset. There has been extensive amount of research done to predict the bitcoin price. We refer a widespread literature to shortlist the potential determinants of the bitcoin price. These include the technological attributes, macroeconomic features and public opinion. We scraped over 30 time series from the internet to forecast bitcoin prices. We developed a model using Random Forest that predicts the log-returns and computes the exchange price based on those values. These forecasts were made on different time periods for one-step ahead prediction. We scrutinise the variable importance in each test-set to draw conclusions about bitcoin's determinants. The model forecasts were compared against Naive Forecasts. Despite comparable results and really low error-rate, our forecasts were statistically insignificant on running a Diebold Mariano Test. We conclude by saying that bitcoin is not predictable, and never was in the past. It is a dynamic trade which changes every moment with no logical explanation for its high volatility. High noise and fluctuations in bitcoin time-series make it unpredictable.

References

1. Jang, H., & Lee, J. (2017). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *Ieee Access*, *6*, 5427-5437.
2. Yang, S. Y., & Kim, J. (2015, December). Bitcoin market return and volatility forecasting using transaction network flow properties. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 1778-1785). IEEE.
3. Balcilar, M., Bouri, E., Gupta, R., & Roubaud, D. (2017). Can volume predict Bitcoin returns and volatility? A quantiles-based approach. *Economic Modelling*, *64*, 74-81.
4. Kristoufek, L. (2013). BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific reports*, *3*(1), 1-7.
5. Polasik, M., Piotrowska, A. I., Wisniewski, T. P., Kotkowski, R., & Lightfoot, G. (2015). Price fluctuations and the use of bitcoin: An empirical inquiry. *International Journal of Electronic Commerce*, *20*(1), 9-49.
6. Zhang, W., Wang, P., Li, X., & Shen, D. (2018). Multifractal detrended cross-correlation analysis of the return-volume relationship of Bitcoin market. *Complexity*, *2018*.
7. Dyhrberg, A. H. (2016). Bitcoin, gold and the dollar–A GARCH volatility analysis. *Finance Research Letters*, *16*, 85-92.
8. Baur, D. G., Dimpfl, T., & Kuck, K. (2018). Bitcoin, gold and the US dollar–A replication and extension. *Finance Research Letters*, *25*, 103-110.
9. Gajardo, G., Kristjanpoller, W. D., & Minutolo, M. (2018). Does Bitcoin exhibit the same asymmetric multifractal cross-correlations with crude oil, gold and DJIA as the Euro, Great British Pound and Yen?. *Chaos, Solitons & Fractals*, *109*, 195-205.
10. Li, X., & Wang, C. A. (2017). The technology and economic determinants of cryptocurrency exchange rates: The case of Bitcoin. *Decision Support Systems*, *95*, 49-60.
11. Brandvold, M., Molnár, P., Vagstad, K., & Valstad, O. C. A. (2015). Price discovery on Bitcoin exchanges. *Journal of International Financial Markets, Institutions and Money*, *36*, 18-35.
12. Fiess, N. M., & MacDonald, R. (2002). Towards the fundamentals of technical analysis: analysing the information content of High, Low and Close prices. *Economic Modelling*, *19*(3), 353-374.
13. Pratama, I., Permanasari, A. E., Ardiyanto, I., & Indrayani, R. (2016, October). A review of missing values handling methods on time-series data. In *2016 International Conference on Information Technology Systems and Innovation (ICITSI)* (pp. 1-6). IEEE.
14. Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.
15. Chen, W., Xu, H., Jia, L., & Gao, Y. (2021). Machine learning model for Bitcoin exchange rate prediction using economic and technology determinants. *International Journal of Forecasting*, *37*(1), 28-43.
16. Huang, J. Z., Huang, W., & Ni, J. (2019). Predicting Bitcoin returns using high-dimensional technical indicators. *The Journal of Finance and Data Science*, *5*(3), 140-155.
17. Pichl, L., & Kaizoji, T. (2017). Volatility analysis of bitcoin. *Quantitative Finance and Economics*, *1*(4), 474-485.
18. Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367-378.
19. Aaker, D. A., & Jacobson, R. (1987). The sophistication of 'naive'modeling. *international Journal of Forecasting*, *3*(3-4), 449-451.
20. Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics*, *33*(1), 1-1.
21. How to scrape data from chart on Bitinfocharts. (2019, December 18). Stack Overflow. https://stackoverflow.com/questions/59395294/how-to-scrape-data-from-chart-on-https-bitinfocharts- com
22. Aroussi R. (2017). *yfinance* (Version 0.1.64) [Python Package]. PyPi. https://pypi.org/project/yfinance/
23. Tsang J. (2017). *Diebold-Mariano-Test*. https://github.com/johntwk/Diebold-Mariano-Test