Forecasting Sustainability of the Bitcoin Market Using Cryptocurrency Market Data

Malvika Singh MALVIKAS@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Prakhar Mishra PRAKHARM@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Simran Handa SHANDA@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract

This paper implements different machine learning time-series forecasting models for discretely timed Bitcoin market data. The work evaluates the prediction performance of the Bitcoin price to cost ratio per unit of Bitcoin using three methods: Gradient Boosting for stochastic data, Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM). We discuss feature selection and the relevance of our analysis, its possible use cases, limitations and future work. This work will aid financial institutions like Goldman Sachs, JP Morgan Chase and others who wish to diversify their reach and gain a strong foothold in the popularly growing Bitcoin market. It will allow them to view the financial viability of trading certain volumes of Bitcoin for the 30 to 50 days in the future assuming all past data for at least the previous year is available. Experimental results demonstrate the performance of the model for time-series prediction. We were able to reduce error significantly by 34.8% by using LSTM over RNN.

Keywords: Bitcoins, Time Series Forecasting, Recurrent Neural Networks, Long Short Term Memory

1. Introduction

We aim to help financial institutions such as Goldman Sachs, Merrill Lynch and JP Morgan predict the sustainability of Bitcoin in order to examine whether it is useful to add to trading activities of the bank. Here, we use the ratio of Bitcoin's price relative to its cost of mining. This lets us know whether the future price-cost ratio of Bitcoin allows it to be sustainable and prevent a bust in the market. This is a measure to analyse the health of bitcoin as a cryptocurrency in the online market.

Works on Bitcoin have focused on a host of issues, from looking at the risk of cybersecurity breaches using a Cox Proportional Hazards Model, to anonymity in Bitcoin (Moore and Christin (2013); Herrera-Joancomartí (2014)).

However, most works related to Bitcoin finance focus on predicting the price of Bitcoin, using several methods ranging from classification to RNN. These papers measure the change in price against a wide variety of parameters, such as the number of Google searches for 'Bitcoin', the number of transaction accounts, and wallets.

The papers mentioned above take a wide range of price types into account, ranging from the weighted price to the opening and closing prices of Bitcoin (Struga and Qirici (2018); McNally et al. (2018); Valencia et al. (2019)). However, in contrast to their work, we focus on the predicting the sustainability of Bitcoin as a cryptocurrency. This is especially important for old and more traditional banks, which must maintain their large profits, customer base, and reputation in the market, all of which will be severely hampered by missteps in transaction timings in volatile industries such as Bitcoin.

Moreover, other papers, focusing only on the price, have not taken the cost of producing Bitcoins into account, which prevents them from studying the maintainability of Bitcoin as a large part of the cryptocurrency industry. Predicting merely on price also does not account for the fact that the cost of producing Bitcoins has also increased over the years and thus price cannot be the only determinant of Bitcoin's sustainability.

Building upon the techniques and parameters used by older papers, we use machine learning methods to determine the possibility of profit for investment banks looking to set up trading desks in this realm. Here, we use machine learning techniques in three different ways: We use XGBoost, LSTM, and RNN, in order to predict the price/cost ratio, using volume, volatility, and competing cryptocurrencies.

1.1 Background on Bitcoin and Institutions

Bitcoin, a cryptocurrency launched in 2008, by a person or persons using the pseudonym Satoshi Nakamoto, has long been seen as volatile and scandal-tainted by traditional banking enterprises and institutions, as well as their customers (Chiu and Koeppl (2017); Arnold; Marr). However, in 2017, this began to change as old and storied names in finance, such as Goldman Sachs, and JP Morgan, began to set up Bitcoin trading operations on Wall Street. While these plans have currently been put on hold due to an uncertain regulatory scene in the market, Goldman executive Rana Yared says the bank is definite about building up a presence in this market (Rogojanu and Badea (2014); Fernández-Villaverde and Sanches (2019); Herrera-Joancomartí (2014); McCann; Rooney; Kharif; Popper; Palmer and DiCamillo; Merced and Popper).

This paper aims to build a model that can provide guidance to investment banks such as Goldman Sachs and JP Morgan on whether the market is profitable or bearish for Bitcoin for these banks in the next month and a half, a useful forecasting time for the fast moving financial markets. Our model centers around competition provided by other cryptocurrencies, the volume of bitcoin traded , and the volatility of Bitcoin pricing.

In Section 2, we provide background on the topic of Bitcoin, in order to understand the context of the transactions. Moreover, we discuss the XGboost, LSTM and RNN models. In Section 3, we describe the model we have used and our sources for creating our model. Section 4 discusses our experimental setup, our data preprocessing, our study design, and evaluation criteria. Section 5 presents our results and our analysis thereof.

2. Background

2.1 Method Background

We use gradient boosted decision trees, pushed to the extent of our computational resources. This is done in order to predict our price/cost ratio by fitting each new model on a modified version of our original data set.

Recurrent Neural Networks (RNN) are a class of neural networks that have multiple hidden states that allow previous outputs to be used as inputs, as shown in Figure 1. LSTM networks have been outperforming traditional neural networks ever since they were discovered. They can be thought of as a special type of RNN that use special memory units to maintain information in memory for a long amount of time. LSTM's use a special set of gates that moderate when information enters the cell, when its output and when its forgotten which, unlike traditional RNN's, allows them to learn long term dependencies. We see this in Figure 2.

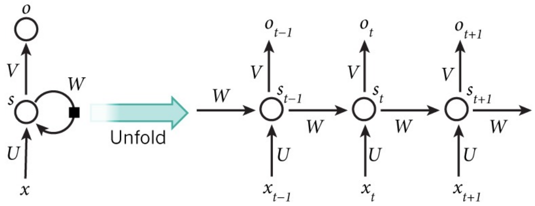
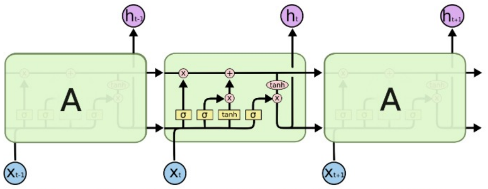


Figure 1: RNN, Colah (2015)

2.2 Bitcoin Price Analysis

Bitcoin transactions take place all day long and there are several methods of transactions, from extremely short term transactions to swing trading, where traders aim to take advantage of an upswing in the price and trade at the first sign of an uptick.

There exist two types of analyses that help predict the price of bitcoin: fundamental analysis and technical analysis. Fundamental analysis aims to predict the price by evaluating the cryptocurrency industry, technical developments (such as a reduction in cost required to produce electricity), world regulations, and the number of competitors in the industry.



The repeating module in an LSTM contains four interacting layers.

Figure 2: LSTM, Colah (2015)

Technical analysis tries to predict the price by studying the market, rather than the industry. It aims to study past statistics, such as the movements of price and volume traded. Through this, it aims to identify if patterns and trends exist in the determination of price and predicts price based solely on market factors.

As with stocks, Bitcoin's price cannot be determined by fundamental or technical analysis alone. A mix of both types of analyses is required in order to prevent mispredicting factors. For example, it is possible that a fundamental analysis predicts that an increase in competitors will cause the price of Bitcoin to fall. However, it is possible that the new cryptocurrencies have almost no traders, leading to barely any movement in price. In the same vein, predicting on basis of price points alone cannot control for a fall in price caused by any new regulations that ban the use of Bitcoin in another country, such as China (Thum (2018); Hayes (2019); Fernández-Villaverde and Sanches (2019); Ametrano (2016); Rogojanu and Badea (2014)).

It is due to the above reason that we have taken both factors into account when building our model, by including the cost of mining bitcoins which is dependent on electricity costs, as well as the number of competing currencies, and the volatility of prices as well as the volume of Bitcoin traded. Due to the lack of information on regulation and the lack of implementation of regulations regarding cryptocurrencies, we have not included regulations as a parameter in our analysis.

3. Methods Used

The methods that we will be using are in the form of models which are capable of dealing with non-i.i.d (independently and identically distributed) data. Since we have time-series data, the unit of analysis is a point in time (every one minute interval recorded) and the corresponding Bitcoin parameter values like volume of Bitcoin currency traded, Open Price, Close Price etc. at that point of time. There appears to be a correlation in the time-series data which is common in stochastic time series analysis. However, neural network methods like RNN and LSTM are capable of dealing with non-i.i.d data as indicated by Nejadettehad et al. (2019). This is due to the fact that RNNs and LSTMs take inputs from previous time steps in earlier layers to make predictions. Since these models are inherently dealing with correlations in time-series in order to make predictions, it is appropriate to use RNNs and LSTMs for our time-series forecast of bitcoin price ratios.

Extreme Gradient Boosting(XGBoost) is a powerful ensemble boosting tree based method and is widely used in stochastic problems. We use XGBoost implementation from the XGBoost python package.

There are sources that say that several non-i.i.d. datasets are often assumed to be i.i.d in order to carry out a larger breadth of machine learning analysis. In order to test the same, we have used XGBoost, a well known technique that carries out its analysis with the assumption that the data it is trained on and will predict is i.i.d. This is later contrasted with Simple RNN and LSTM.

3.1 Simple RNN

A recurrent neuron has backward connections with previous layers of neurons. It has internal memory due to which it is used for predicting time-series trends. This attribute made recurrent neural networks very popular for cases of sentence completion, handwriting detection (Nejadettehad et al. (2019)) and time series forecasting. Given a sequence $X = X_1, X_2, X_3....., X_t$ as input, RNN computes the hidden state sequence $H = H_1, H_2, H_3....., H_t$ and output sequence $Y = Y_1, Y_2, Y_3....., Y_t$

$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h)$$

$$y_t = g(W_{yt}h_t + b_y)$$

Here, we make use of an equation and analysis by Nejadettehad et al. (2019).

Where W_{hx}, W_{hy}, W_{yt} denote the input-to-hidden, hidden-to-hidden and hidden-to-output weight matrices, respectively. b_h and b_y are hidden layer bias and output layer bias vectors and $f(.)$ and $g(.)$ are the activation functions for the hidden layers and the output layer respectively (Nejadettehad et al. (2019)).

3.2 LSTM

LSTMs are an advanced and complex version of RNN. They are capable of learning long-short term dependencies. Such networks remember information for long periods of time. That is why, for our case, LSTMs will produce the best results because of their ability to retain important information from the past data and not just the recent-most data like in the case of RNNs. LSTMs are able to apply context to predict values (Nejadettehad et al. (2019); McCann). So, they are better suited for time-series prediction for bitcoin price ratio as these networks will tend to capture important trends from the past and use it to depict what the future values of the ratio will look like.

4. Experimental Setup

The following pipeline indicates the process that was followed for running our experiment.

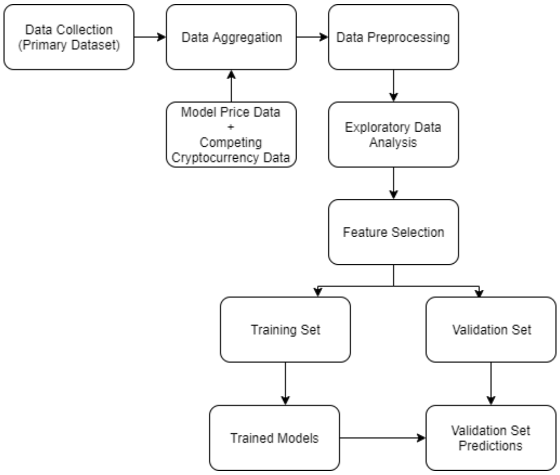

Data Collection (Primary Dataset)
Data Aggregation
Data Preprocessing
Model Price Data + Competing Cryptocurrency Data
Exploratory Data Analysis
Feature Selection
Training Set
Validation Set
Trained Models
Validation Set Predictions

Figure 3: The Execution Pipeline

Dataset Collection and Description: The primary dataset is taken from Zielak(2019). It is present in the form of csv for the time period of December 2011 to May 2017 with one minute updates of Open, High, Low and Close, Volume in Bitcoin Transaction, Volume in Currency, and Weighted Bitcoin Price per unit. Timestamps here are UNIX timestamps, which were converted to date and time stamps.

The code is uploaded on the submitted github link which has the procedure for reproducing the results on local machines.

Next, we describe the computing resources as a part of experimental setup which is required to execute our analysis. The hardware and software resources required are as follows:

- Hardware:
  Processor Intel(R) Core(TM) i5-7200 CPU @2.50 GHz 2.70 GHz and an installed RAM of 8.00 GB (7.73 GB Usable). The system is a 64-bit operating system, x64-based processor.

- Software:
  Experiment is performed on Google CoLab having its own RAM and disk space on Chrome. Python 3 was used for building the notebook. Laptop we used had Windows 10 OS.

4.1 Cohort Selection

The following table presents the summary statistics of the variables that were used in our analysis:

Table 4.1 Aggregated Data Statistics.

Variable Name	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Open	3.8	239.9	420.0	495.9	641.0	2754.7
High	3.8	240.0	420.0	496.2	641.7	2760.1
Low	1.5	239.8	419.9	495.5	640.2	2752.0
Close	1.5	239.9	420.0	495.9	641.0	2754.8
Bitcoin Volume	0.0	0.3	1.8	11.8	8.0	5853.8
Currency Volume	0.0	124.0	614.6	5316.3	3107.9	1865888.5
Weighted Price	3.8	239.9	420.0	495.9	641.0	2754.5
Time (converted from UNIX)	2011-12-31	2014-02-23	2015-03-21	2015-03-30	2016-04-09	2017-05-31
Cost of Mining	32.0	254.7	411.6	524.7	684.7	1681.7
Competition	4.0	506.0	562.0	545.9	644.0	1335.0
Price Ratio	0.05	0.87	0.97	1.01	1.07	3.92
Status	0.0	0.0	0.0	0.4	1.0	1.0
Volatility	0.0	0.0	0.05	0.68	0.74	587.03

4.2 Data Aggregation and Pre-Processing

Carrying out data preprocessing for this project involved a multitude of steps from choosing which data points to include to creating new parameters based on the number of competing cryptocurrencies.In addition, we included volatility, as well as the cost of mining Bitcoin in our determined time frame using the equation from Hayes (2015).

Dropping Rows: There were a significant number of bitcoin transactions from the year 2011 in our dataset that contained missing values that were crucial for our analysis. The values seemed to be missing completely at random and therefore we decided to drop those cases instead of imputing values into it.

Including Competing Cryptocurrencies: We added the number of competing cryptocurrencies that were present in the market at the time each transaction was carried out as a separate parameter to our data. This was done by referring to several sources listing out the dates of creation for cryptocurrencies and adding them to our column (Fernández-Villaverde and Sanches (2019); Ametrano (2016); Bigmore; Chan; Liquid).

Cost of Production: For each transaction, we also added a model price - the cost of mining when that particular amount of bitcoin at the time the transaction was carried out. This should ideally be the floor of the market price, though we find that this is not always the case. This was done by using the following equation by Hayes (2015):

$$E_{day} = (priceperkWh * 24hrday * WperGH/s)(GH/1,000)$$

Here, E(day) is the cost of mining bitcoins everyday, kWh stands for Kilowatt-hours, W is Watts, while GH is the time required to create 1 billion hashes per second. The price of kWh was provided by Hayes, in a dataset provided with his paper and the cost (or model price) was calculated accordingly using Excel. Further analysis is provided by Thum (2018).

The model price and the weighted price of the bitcoin at that point in time was then used to calculate the "Price Cost Ratio", which we then use as an indicator for calculating the profitability of the bitcoin.

Volatility of Bitcoin: We created a new column, subtracting the lowest price of Bitcoin from the highest price in a minute. This allowed us to measure the volatility of Bitcoin prices (Abdul and Rajabu. (2017)).

4.3 Feature Selection and Choices

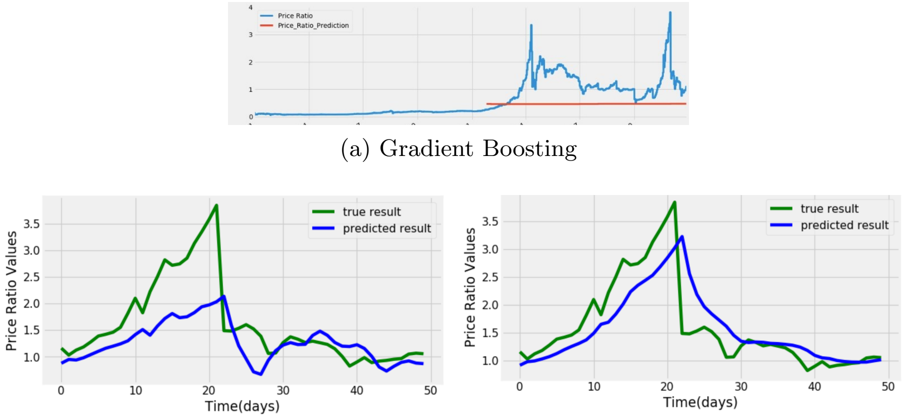
We used a correlation heatmap to analyze the correlation between different variables and drop one or more variables that were highly correlated with each other. Features that have high correlation are more linearly dependent on each other and have almost the same effect on the dependent variable and as a result we drop one of them. We set the correlation threshold at 0.9 i.e. we drop one of two variables that have a correlation coefficient that is greater than 0.9 and keep the other for our analysis. We found that the variables that showed the Open, High, Close and Low variables were highly correlated to one another and therefore we dropped the Open, High and Low variables from our analysis.

4.4 Comparison Methods

In the papers by McNally et al. (2018) and Struga and Qirici (2018), mentioned in Section 1, LSTM was used to predict prices of bitcoin. This is in contrast to our paper which uses the price-cost ratio as our dependent variable. There are differing features in the two papers; our paper included the impact of competing cryptocurrencies, which the other paper did not, as well as volatility. However, we did not include popularity of bitcoin on Google Trends as a sign of its actual demand and we did not include block size, transaction amounts and mining revenue because of data constraints.

4.5 Evaluation Criteria

For this project, we use the root mean squared error(RMSE) between the values of our predictions of profitability and the actual observed values from the dataset for evaluating



(a) Gradient Boosting

(b) RNN Predictions

(c) LSTM Predictions

Figure 4: Model Performance on Testing Data

the fitted models. We believe this is the best way to evaluate our models as RMSE is highly sensitive to the error distribution of samples, which helps us gain a better result than using Mean Absolute Error (MAE). It also enables us to measure the gradient better. We do not believe it is best to use accuracy, as a close prediction is good enough for us and we might not need an exact prediction in order to satisfy our requirements, like in classification models.

5. Results

The results can be divided into three categories based upon the three models that we have used for price ratio prediction. These are as follows:

- XGBoost for Stochastic Data
- Recurrent Neural Networks
- Long Short Term Memory

We started with the preliminary predictions using gradient boosting and moved onward to more precise analysis using neural networks. We obtain the predictive graphs as shown in the figures depicting the predictive performances of the XGBoost, RNN and LSTM models.

The RMSE values for the performance of the 3 models on the testing data are summarized as follows:

Table 5.1 Results from 3 models.

Model	RMSE
XGBoost	0.8427
RNN	0.6658
LSTM	0.4610

6. Discussion and Related Work

6.1 Comparison of Models

From the results, we observe that LSTM performs the best in terms of predictive performance, producing an RMSE of 0.43 as compared to RNN that produces RMSE of 0.66. Clearly, Long-Short-Term-Memory is best suited for this kind of bitcoin sustainability ratio prediction due to the fact that it makes use of 'context' from long duration in the past and effectively uses it to forecast future values. As mentioned in Section 3, we carried out XGBoost in order to test the idea that non i.i.d. data can be treated as i.i.d. data. We can clearly see here that that is incorrect, due to our high RMSE, and that methods specific to non-i.i.d. have more and better predictive power.

6.2 Implications of Analysis For Financial Institutions

We can clearly see here the rise of Bitcoin's price-cost ratio in 2013, towards the end of the recession that lasted from 2008-2012 across the world. This is in line with our hypothesis that macroeconomic factors as well as technical factors should be taken into account when measuring the viability of investing in Bitcoin. Bitcoin was most profitable to invest in in late 2013, when we begin to see an upswing in the price, which continues to grow, so that we can buy low and sell high.

Since we see that our models are in line with the true prices of the remaining years as well, we can use our model to predict when it is most viable to invest a certain volume into Bitcoin for 30 to 50 days in the future, a result that is highly useful for financial institutions looking to set up trading desks in Bitcoin in these times as costs and prices fluctuate. Our model also allows banks to engage in Bitcoin transactions without having to monitor the prices constantly, which was previously necessary for a volatile currency such as Bitcoin.

6.3 Limitations

Due to data constraints, we were unable to include certain factors in our study, such as regulations in different aspects of the world. Moreover, constrained by the non-i.i.d. nature of our data, we were unable to conduct certain classification analysis and carry out a study of Bitcoin investment survival in the market and the exact risks of investing at a certain time using Cox proportional hazards.

7. Conclusion

In conclusion, this paper uses machine learning techniques based on several aspects of a pipeline extending from data collection to prediction to predict the best time to invest in Bitcoin for financial institutions, such as Goldman Sachs, Merrill Lynch, and JP Morgan, within the next 1.5 months.

Post using XGBoost, RNN, and LSTM, we see that LSTM provides the best results for this use, with a RMSE 0.43.

In the future, we would also like to include the effect of worldwide regulations on the price of Bitcoin, once they are created, and the effects of reputational harm through a loss of privacy and anonymity.

References

Okwuchukwu Abdul and Rajabu. "what is bitcoin mining and is it still profitable in 2020? (complete guide)." 99 bitcoins. 99 bitcoins. accessed december 3, 2019. https://99bitcoins.com/bitcoin-mining/. 2017.

Ferdinando M Ametrano. Hayek money: The cryptocurrency price stability solution. *Available at SSRN 2425270*, 2016.

Andrew Arnold. "how institutional investors are changing the cryptocurrency market.". *Forbes. Forbes Magazine, November 29, 2018.*

Rosemary Bigmore. Cryptocurrencies: a timeline. URL `https://www.telegraph.co.uk/technology/digital-money/the-history-of-cryptocurrency/`.

Christine Chan. Cryptocurrencies: Growing in number but falling in value. URL `https://graphics.reuters.com/CRYPTO-CURRENCIES-CONFLICTS/0100818S2BW/index.html`.

Jonathan Chiu and Thorsten V Koeppl. The economics of cryptocurrencies–bitcoin and beyond. *Available at SSRN 3048124*, 2017.

Colah. Understanding lstm networks. URL `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`.

Jesús Fernández-Villaverde and Daniel Sanches. Can currency competition work? *Journal of Monetary Economics*, 106:1–15, 2019.

Adam Hayes. The socio-technological lives of bitcoin. *Theory, Culture & Society*, 36(4): 49–72, 2019.

Jordi Herrera-Joancomartí. Research and challenges on bitcoin anonymity. In *Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance*, pages 3–16. Springer, 2014.

Olga Kharif. Goldman sachs analyst is bullish. URL `https://www.bloomberg.com/news/articles/2019-08-12/a-goldman-sachs-analyst-is-bullish-on-bitcoin-but-is-goldman`.

Liquid. How many cryptocurrencies are there. URL `https://blog.liquid.com/how-many-cryptocurrencies-are-there`.

Bernard Marr. A short history of bitcoin and crypto currency everyone should read. URL `https://www.forbes.com/sites/bernardmarr/2017/12/06/a-short-history-of-bitcoin-and-crypto-currency-everyone-should-read/#51defe333f27`.

Chris McCann. 12 graphs that show just how early the cryptocurrency market is. URL `https://medium.com/@mccannatron/12-graphs-that-show-just-how-early-the-cryptocurrency-market-is-653a4b8b2720`.

Sean McNally, Jason Roche, and Simon Caton. Predicting the price of bitcoin using machine learning. In *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, pages 339–343. IEEE, 2018.

Michael J. De La Merced and Nathaniel Popper. Jpmorgan chase moves to be first big u.s. bank with its own cryptocurrency. URL `https://www.nytimes.com/2019/02/14/business/dealbook/jpmorgan-cryptocurrency-bitcoin.html`.

Tyler Moore and Nicolas Christin. Beware the middleman: Empirical analysis of bitcoin-exchange risk. In *International Conference on Financial Cryptography and Data Security*, pages 25–33. Springer, 2013.

Alireza Nejadettehad, Hamid Mahini, and Behnam Bahrak. Short-term demand forecasting for online car-hailing services using recurrent neural networks. *arXiv preprint arXiv:1901.10821*, 2019.

Daniel Palmer and Nathan DiCamillo. Goldman sachs analysts' slide suggests now's a good time to buy bitcoin. URL `https://www.coindesk.com/goldman-sachs-analysts-note-says-nows-a-good-time-to-buy-bitcoin`.

Nathaniel Popper. Goldman sachs to open a bitcoin trading operation. URL `https://www.nytimes.com/2018/05/02/technology/bitcoin-goldman-sachs.html`.

Angela Rogojanu and Liana Badea. The issue of competing currencies. case study-bitcoin. *Theoretical & Applied Economics*, 21(1), 2014.

Kate Rooney. Goldman sachs cfo says bank is working on bitcoin derivative for clients. URL `https://www.cnbc.com/2018/09/06/goldman-sachs-cfo-calls-reports-of-shutting-down-crypto-desk-fake-news.html`.

Kejsi Struga and Olti Qirici. Bitcoin price prediction with neural networks. In *RTA-CSIT*, pages 41–49, 2018.

Marcel Thum. The economic cost of bitcoin mining. In *CESifo Forum*, volume 19, pages 43–45. München: ifo Institut–Leibniz-Institut für Wirtschaftsforschung an der . . . , 2018.

Franco Valencia, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy*, 21(6):589, 2019.

Zielak. Bitcoin historical data. URL `https://www.kaggle.com/mczielinski/bitcoin-historical-data/`.

Appendix A.

Plot of the correlation plot(Shown in figure 5 below) that we obtained for feature choices. Here, we removed the highly correlated variables as described in section 4.3 to ensure that we have relevant features.

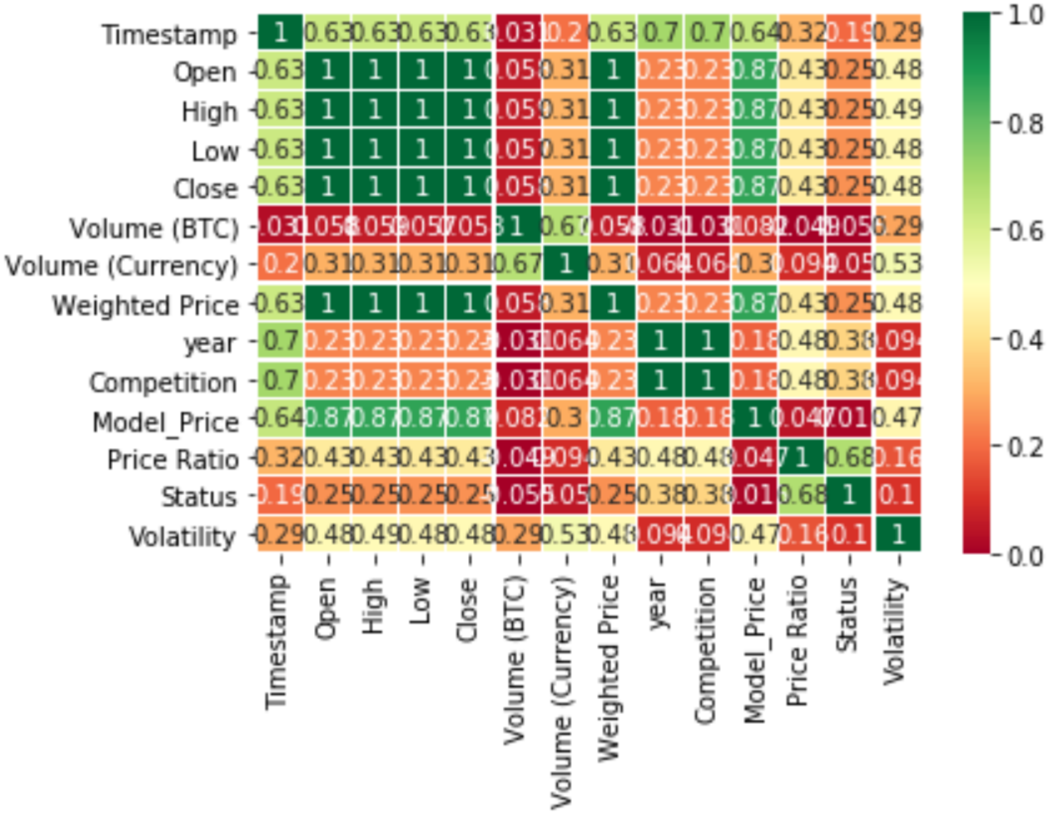


Figure 5: Correlation Plot Obtained