CRYPTOCURRENCY AND ITS IMPACT

CSP 571 Final Report

Anitya Kumar Gupta : A20428839 : agupta118@hawk.iit.edu

Kusuma Goli : A20453281 : kgoli1@hawk.iit.edu

Dharamveer Kumar Yadav : A20444272 : dyadav4@hawk..it.edu

Anisha Farzana Shajahan : A20422605 : ashajahan@hawk.iit.edu

Abstract

Cryptocurrency, an encrypted, peer-to-peer network for facilitating digital barter, is a technology developed eight years ago. Bitcoin, the first and most popular cryptocurrency, is paving the way as a disruptive technology to long standing and unchanged financial payment systems that have been in place for many decades. While cryptocurrencies are not likely to replace traditional fiat currency, they could change the way Internet-connected global markets interact with each other, clearing away barriers surrounding normative national currencies and exchange rates. Technology advances at a rapid rate, and the success of a given technology is almost solely dictated by the market upon which it seeks to improve. Cryptocurrencies may revolutionize digital trade markets by creating a free-flowing trading system without fees. A SWOT analysis of different cryptocurrencies with main stream on BITCOIN is presented, which illuminates some of the recent events and movements that could influence whether Bitcoin contributes to a shift in economic paradigms and in the end we also mention that all cryptocurrencies behaves mostly similar and if the rates of cryptocurrencies falls down than other cryptocurrencies with respect to market cap will also drops down.

1.Overview

There are different kinds of cryptocurrencies and when we see the analysis we can clearly observe that Bitcoin, the world's most common and well known cryptocurrency, has been increasing in popularity. It has the same basic structure as it did when created in 2008, but repeat instances of the world market changing has created a new demand for cryptocurrencies much greater than its initial showing. By using a cryptocurrency, users are able to exchange value digitally without third party oversight. Cryptocurrency works on the theory of solving encryption algorithms to create unique hashes that are finite in number. Combined with a network of computers verifying transactions, users are able to exchange hashes as if exchanging physical currency. There is a finite number of bitcoin that will ever be generated, preventing an overabundance and ensuring its rarity. Water, despite its requirement as a life giving material, is generally accepted as being free or of little cost because it is so abundant. If water was rare, it would be more valuable than diamonds. Value exists for bitcoin because its users have trust that if they accept it as payment, they would could use it elsewhere to purchase something they want or need.

Current legal and financial structures are not designed with a technology like this in mind. Financial institutions are built off of much older forms of currency. In some ways, it is comparative to the computing industry. Yet all of our current technology uses this technologically archaic system due to adoption, cultivation, and lack of need for newer systems. If cryptocurrencies became the global norm for transactions, long standing systems for trade would need to be completely reformed to deal with this type of competition.[2] For this reason, cryptocurrencies could possibly be the single most disruptive technology to global financial and economic systems.

Transaction increase is an indicator of user acceptance growing. The conditions for Bitcoin‟s widespread adoption could be described as a “fire triangle”. Where fire needs fuel, oxygen, and heat to exist; Bitcoin needs user acceptance, vendor acceptance, and innovation to ignite. Without all three aspects, bitcoin may not truly become a legitimized mainstream currency. Bitcoin is currently experiencing an increase in user acceptance and use, which is driving the other two aspects of the “fire triangle”.[3] Cryptocurrency’s adoption will be an important subject to watch in the future, as it could be a truly transformative technology that alters the way money is exchanged worldwide. Bitcoin’s increased adoption has been integrally tied to global market shifts. The current Internet fueled global market is very much entangled. If one regional market begins to plummet, it can easily drag the others with it. Bitcoin, like the Euro, can freely move across many national borders, creating an environment that promotes global trade, mutual prosperity, and even peace.

2.Strengths

South America has seen a huge increase in bitcoin transactions, increasing 510% till 2019 and being dropped with 250% Bitcoin: A New Global Economy, 2015). In the past, different countries would convert their currency into US dollars to preserve their value. However, India has recently put restrictions on how many US dollars its citizens can convert. As a result, both a black market for purchasing USD at a higher price and increased bitcoin adoption has arisen. The demand for Indians, Arabians and different citizens to keep their currency value has made itself very apparent, and cryptocurrencies are prominent legal vehicles to meet that demand.

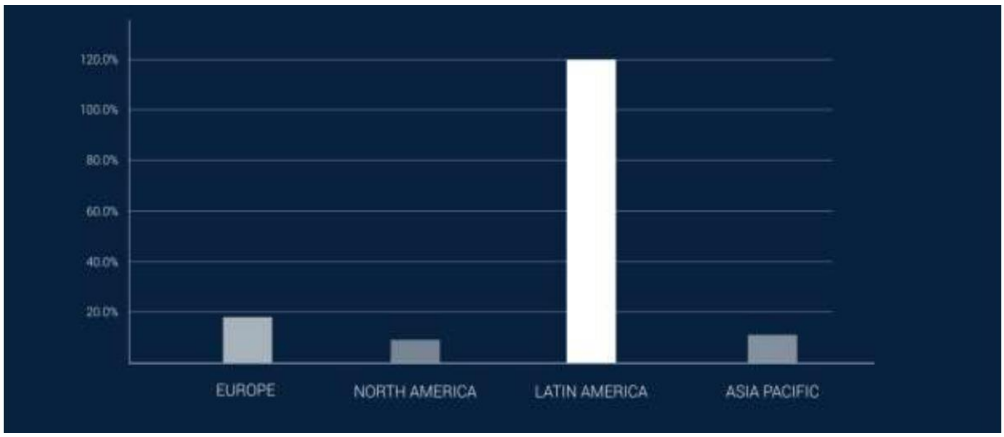

Fig 1. Bitcoin Transaction Value Growth

To purchase bitcoin, one only needs to set up an online account with an online exchange, make their request, and the transaction is usually completed in minutes. Once the bitcoin is in their digital wallet, they would be able to make purchases from thousands of vendors worldwide. In this example, Bitcoin is the more viable solution as quick entry and exit for a currency that can quickly gain value. [5]Other fiat currencies may become stronger and be more desired, but they cannot compete with cryptocurrencies‟ agility. Cryptocurrency is the disruptive technology that could be pushed into acceptance by investors who simply want a refuge from sinking global markets.

3. Business Benefits

Exploratory and Descriptive Analytics: Based on analysing the historical prices of different cryptocurrencies, we can predict the trends for the same, which will help potential investors make informed decisions.

Classification: Identifying fraudulent transactions will help distinguish between legitimate and illegitimate transactions, preventing the case of a dishonest network and avoiding usage of the network for running scams.

Clustering: Highlighting anomalies in the network of users and/or transactions can be done by using various clustering methodologies.

Association rules: There can be established definite correlation between various factors and prices of the cryptocurrencies. This means the data can be analysed for frequent if-then relationships using the criteria of support and confidence to identify the most important relationships. This eventually leads to predict blockchain behaviour.

4. Data Preparation

Our Data Set consists of date wise data, while online references used a 24-hour data,taken from **Coin Exchange and Crypto download**. We leveraged these features in developing a binary and a ternary classification algorithm, to predict the sign change in the Bitcoin price, based on daily data points. Both the algorithms take a manually created label depicting two classes in case of binary classification and three in ternary classification. The binary classification algorithm predicts price positive and no change as 1 and negative price change as -1. The ternary classification algorithm predicts positive price change as 0, negative price change as -1, and no change as 0.

From Coin Exchange We have considered the Bitcoin Data Set which has 24 features or attributes in it. We have extracted 16 important features and build a subset of the data as per the binary and ternary classification mentioned above. We have further classified our data into training and test data. 75 percentage of data is classified as training and the rest as testing data. We have used the different algorithms mentioned below to build the model based on the training data and predicted the Market Price Label based on Model built and Test Data.

From Crypto download the data is already been cleaned and all the 5 features are most important. The data is been update every week and we took for 2018 -2019 evaluation analysis from this data. 60 percentage od the data is classified as training and the rest as

testing data. For the analysis of the current data we use the algorithms and mechanism like ARIMA , AIC , ACF. There was conversion required to convert symbol and date into numerical format.

5. Data Cleaning and its Quality

5.1. Cleaning

Data has been verified to identify: Data is divided into 2 segments. First segment 2009 – 2017 and second segment 2018 – 2019.

Missing data: We identified that bitcoin price dataset has missing values for Volume for 7 months of Year 2019. It amounts to around 15% of the data. We also found that bitcoin dataset contains around 27 missing values, which was only around 0.92% of the total dataset. Most important thing data of the second segment hasn't has any missing value.

Data Errors: The dataset does not have any numeric errors. Further, there is no text or factor data, so there are no typographical errors. We convert the symbol and date to the numerical data which is there in second segment 2018 – 2019.

Measurement Errors: There is single source of data and is based on single measurement scheme, thereby no measurement errors recorded.

Coding inconsistencies: Since the data is from single source, there are no coding inconsistencies.

Bad Metadata: Metadata is from standard terminology, hence no bad metadata issues. We observed that the Price data is to see the volatility trend of crypto currency. Volume and Market capitalization data will help in creating models for predicting future prices. Since the data in different attributes are of different units, it needs normalization. The data is clean and consistent, however one of the attributes in every _le has missing values. So, we employed mice package to predict the missing values. We did not encounter any data errors, spelling inconsistencies or bad metadata in the dataset.

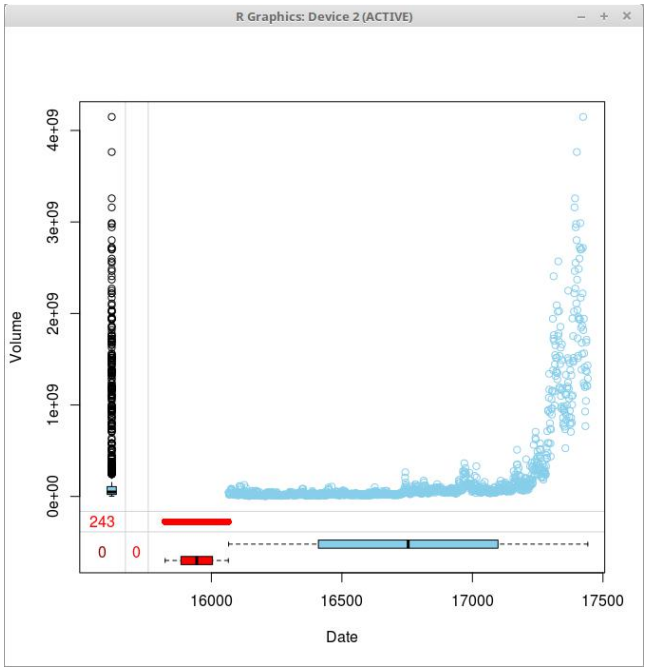

Fig 2. Data value missing 2009 – 2017

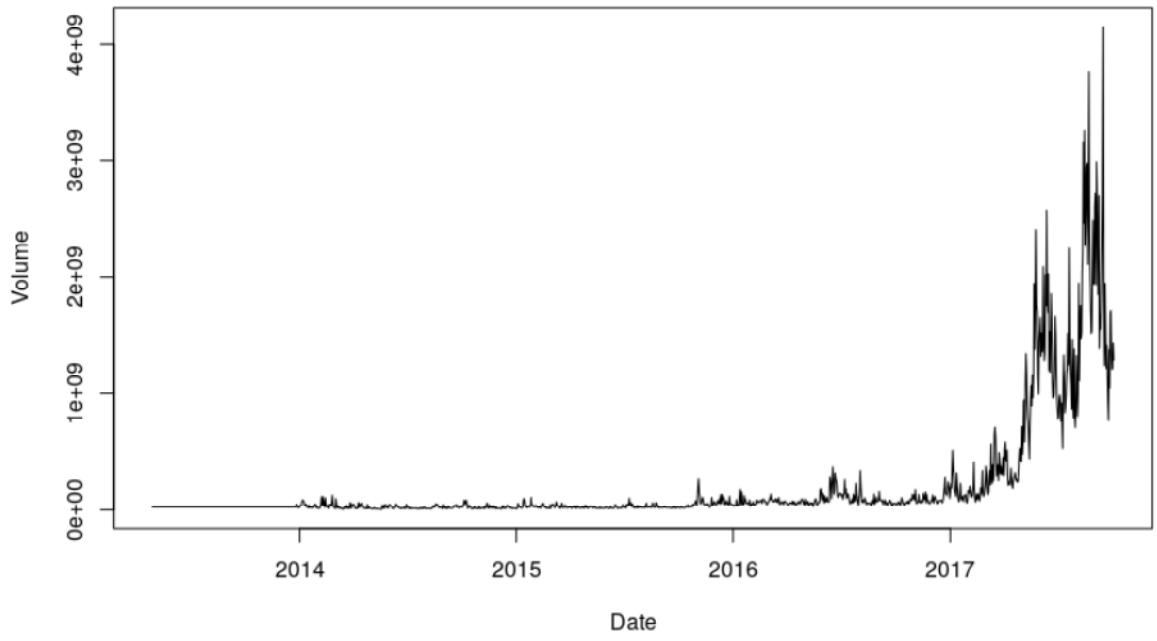

Fig 3. Missing Values Against Date after Filling

From our domain information, we consider that volume of the price might get impacted based on its average daily price. We calculate the average daily price by taking average of daily High and Low.

5.2 Transformation [2nd Segment Refinement]

With the end goal of forecasting, the first step after looking at the data is to transform it to make it stationary, i.e. make it so the future behaves as the past. In order to determine the best way to do that, we should look at the ACF (auto-covariance) and PACF (partial auto-covariance) plots, which depict how data are affected by the observations that precede them. This will also give us an idea of the model for the data.

AR (auto-regressive) models depend on their past values, an error term, and sometimes a constant. MA (moving average) models depend linearly on the current and past values of white noise error terms (which is random). The combination of them, together with differencing, constitute the ARIMA model.

We would like to begin by using the **Box-Cox** power transformation, which is a method that helps to Normalize data. This method finds a value λ which maximizes the log-likelihood and then the data is raised to the power of λ. The higher the likelihood, the better the model parameter. Power transformations such as the Box-Cox help to stabilize the variance which is imperative for Box-Jenkins modelling.[9]

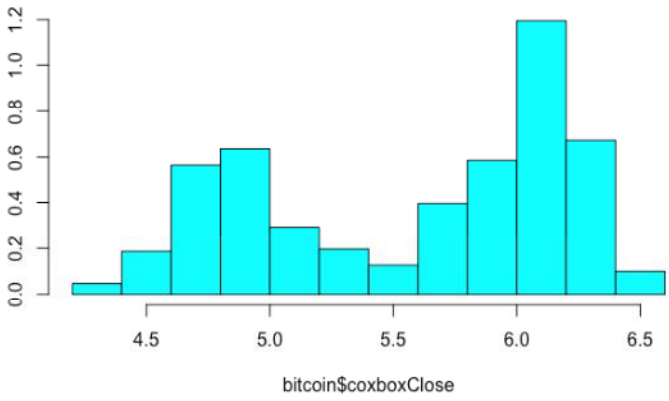

Fig 4. Transforming the 2 features according to close price

Here p value is greater that 0.01 which is 0.83. This makes us to go more deeper and move towards another method called Differencing.

Differencing: We only want to difference once and at lag 1 because we did not observe any seasonality when we first plotted the original time series. The differenced series will represent the *change* between observations that are dd units of time apart, in this case, d=1d=1, which is to say one day apart. So our differenced data will represent the change from each value to the next, one day at a time. [9]

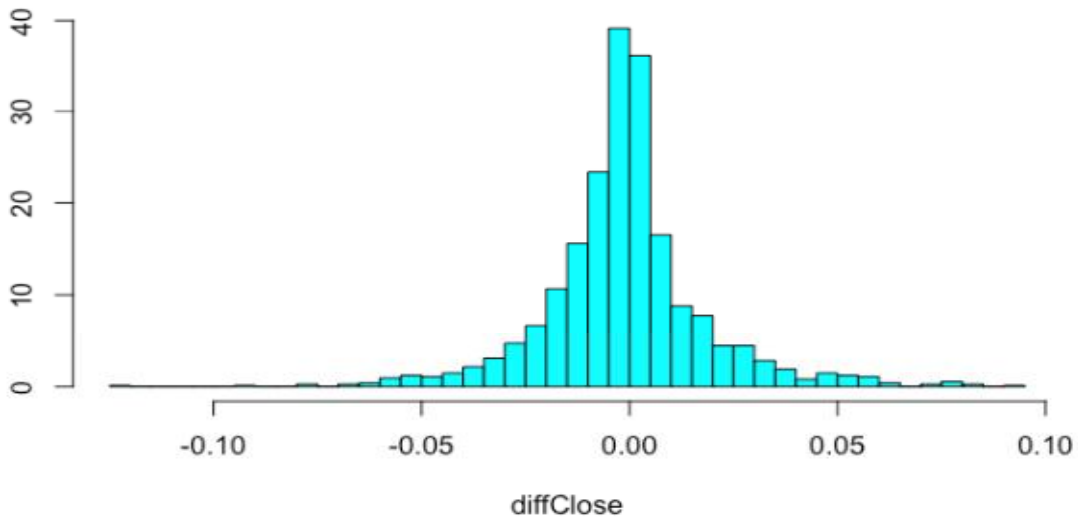

Fig 5. Execution of Differencing

Our p-value of 0.010.01 being lower than confidence level α=0.05 gives us significant evidence to reject the null hypothesis that the data is not stationary, thereby concluding that it is. Satisfied by this, we can now move on with estimating the parameters of the model.

6. Methods and Algorithms

6.1 Random Forests

Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Evaluation of the model: Data set has accuracy of maximum of 61.12% which is slightly greater than the Support Vector Model Classification.

Pros:

- It runs efficiently on large data bases.
- It can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- It has methods for balancing error in class population unbalanced data sets.
- It computes proximity's between pairs of cases that can be used in clustering, locating outliers, or (by scaling) give interesting views of the data.

The capabilities of the above can be extended to unlabelled data, leading to unsupervised clustering, data views and outlier detection.

Benefits for the Target Users: Using the Random Forest Model we have predicted the Market Price of the Coin Exchange Data Set with accuracy of maximum of 61.12% which is slightly greater than the Support Vector Model Classification. Consumers and Merchants who accept and deal with the cryptocurrency would like to know the stability of it.[6] We have other statistical measures like Kappa, p-value apart from accuracy to help us in evaluating the Classification Model.

6.2 Linear Regression

Linear Regression is used for predictive analysis. In this case, there is a response variable whose outcome has to be predicted based on the input variables which are also called as dependent variables. Linear Regression is used with continuous type of data.

Pros:

- Useful based on relationships between two quantitative continuous variables.

Evaluation of the Model: The mean square error which we derived is 0.2127196982.Figure 1 below shows the predicted vs actual value.

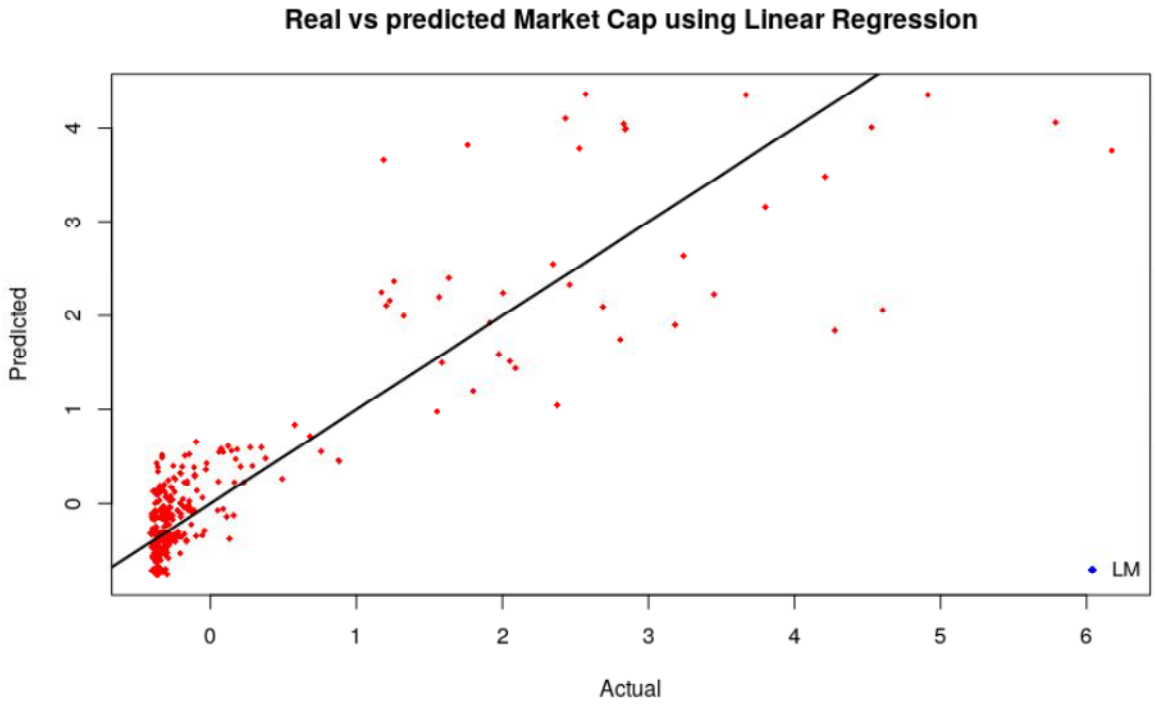

Fig 6. Bitcoin Market Cap prediction based on Open Price

Benefit for the target users : Trading of cryptocurrency is similar to stock market trading. Investors want to predict the prices so that they can plan their strategy. Open Price is one such parameter which help investors to decide to purchase or to sell. Further it helps them to plan their day strategy. If the Open Price predicted is much lower, then they can plan to

purchase and wait for later in the day to sell at higher price. If the predicted price comes higher than the expected, they can plan to sell or short sell. Thus, predicting the opening prices helps investors to plan their day trading strategy.

6.3 ARIMA

Relatively basic Time Series model that we will be coding out and explaining the components when necessary. Facebook Prophet uses an additive model for forecasting time series data that is fast and tuneable. After modelling, we will compare the results from each model's unique insights into Bitcoin's future.

Steps involved in framing :

1. Gather, explore, and visualize the data.
2. Difference the data and check for stationarity.
3. Plot the ACF and PACF for the differenced data.
4. Start modelling by searching for the best parameters.
5. Train and test the model with the optimized parameters.
6. Forecast the future!

Optimising parameter: In order to get the best performance out of the model, we must find the optimum parameters.[10] We do this by trying many different combinations of the parameters and selecting the one with the relatively lowest **AIC score**.

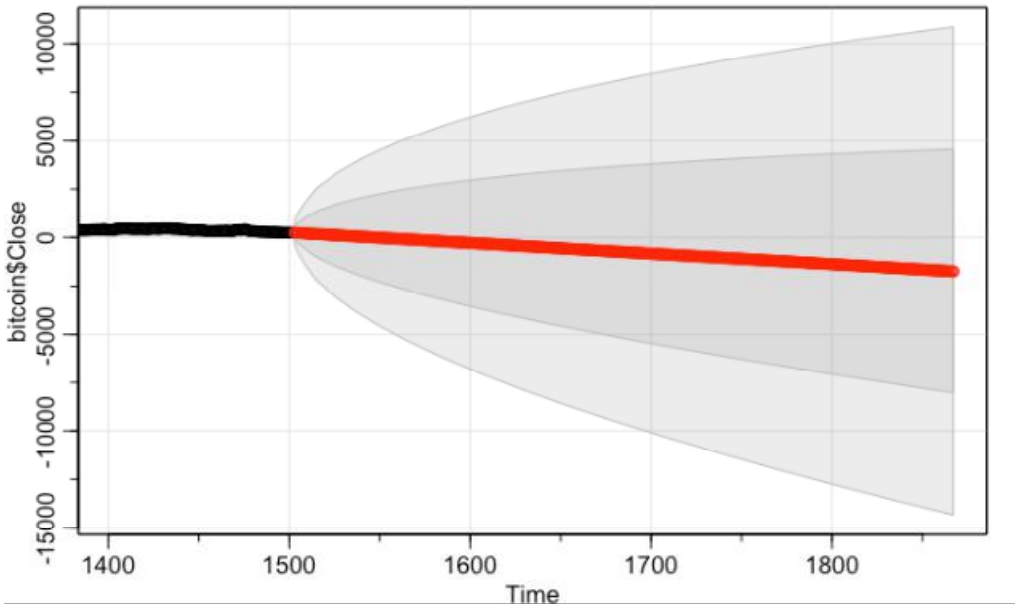

Fig 7. Indicating ARIMA Forecasting

We can observe that there is drop down to the price in future from 90 percentile to 75 percentiles.

AIC : Formula based on the number of estimated parameters as well as the maximum likelihood estimator. The smaller the AIC, the stronger the model. The AIC corrects for finite sample sizes, addressing an issue with the AIC which can result in overfitting models, so we prefer it when fitting for an ARIMA model.

It is however important to note that the AIC depend on a model being univariate, linear, and with normally distributed residuals. We know the first two are true from what we have done so far but are unsure about the latter condition. It is something that we will diagnose in the model diagnostics.

```
> AIC.mat
          q = 0    q = 1    q = 2    q = 3    q = 4    q = 5    q = 6    q = 7    q = 8    q = 9   q = 10
p = 0  14.31064 14.31175 14.31281 14.31287 14.31174 14.30515 14.30421 14.30482 14.30473 14.30584 14.28771
p = 1  14.31176 14.31309 14.30931 14.30912 14.30829 14.30539 14.30529 14.30084 14.30134 14.30203 14.28856
p = 2  14.31286 14.30918 14.31160 14.31124 14.29569 14.29806 14.30139 14.29505 14.30158 14.30290 14.28789
p = 3  14.31342 14.30826 14.31094 14.29599 14.29693 14.29801 14.30071 14.30357 14.28269 14.27949 14.28279
p = 4  14.31262 14.30629 14.29556 14.29693 14.28481 14.28171 14.28718 14.28412 14.28186 14.27501 14.27669
p = 5  14.30309 14.30151 14.29538 14.29795 14.29062 14.28970 14.29325 14.29279 14.27716 14.27631 14.27931
p = 6  14.30030 14.30164 14.30293 14.30426 14.28502 14.29321 14.29452 14.29354 14.26800 14.27027 14.27774
p = 7  14.30163 14.30019 14.29514 14.29565 14.28086 14.29227 14.29394 14.27907 14.27841 14.28191 14.27260
p = 8  14.30278 14.30412 14.30104 14.28439 14.28223 14.27927 14.27084 14.28552 14.26851 14.27784 14.26330
p = 9  14.30403 14.30223 14.30237 14.28571 14.28080 14.28067 14.28078 14.28328 14.25965 14.27232 14.27078
p = 10 14.28878 14.29008 14.28933 14.28155 14.27834 14.28083 14.27878 14.27387 14.27158 14.27288 14.26268
>
```

Fig 8. AIC matrix when value of p=q

We want to estimate as few parameters as possible due to the principle of parsimony (to select as few parameters as possible). In our case, the more parameters there are, the weaker the model. We can fit a couple of different models and compare. It is generally advisable in fact to stick to models in which at least one of p and q is no larger than 1, since this is likely to lead to overfitting.[10]

6.4 K Means clustering

It is a centroid based partitioning technique that uses the centroid of a cluster, Ci to represent the cluster. Conceptually, the centroid of a cluster is its centre point. This algorithm requires to specify the number of clusters (k) be-forehand. This method is not guaranteed to converge to the global optimum and often terminates to a local optimum. We have applied this algorithm on Open/High attributes of the daily prices. And with some experimentation, of taking no of clusters values from 3, 4, 5, we found that between 4 and 5 here is not much difference. And so we decided to take the k value as 4. We attempt to identify the outliers based on these attributes.

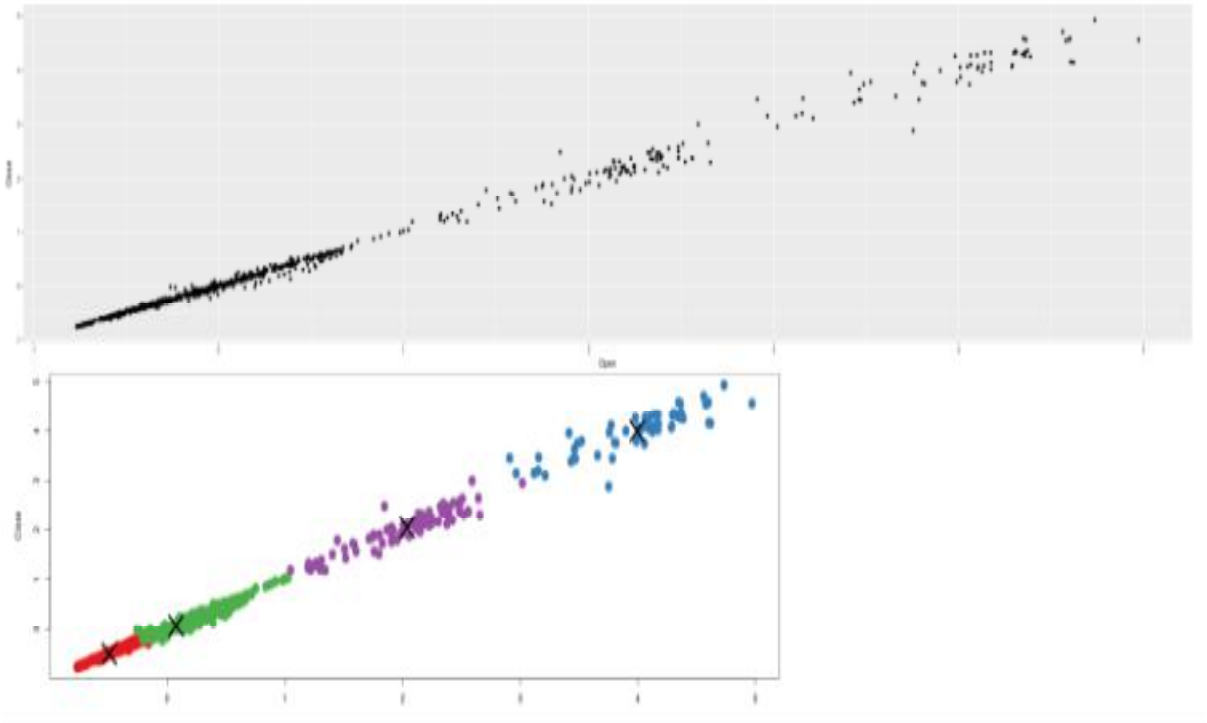

Fig 9. Before and After K Means Clustering

7.Data Visualization

7.1. Inference visualised on different variables executed

The hash rate of a cryptocurrency is a measure of how many proofs of work calculations are being performed by all the miners participating in the network. It is measured in terms of number hashes per second, but because such a large number of these calculations are performed, it is more usual to see values in larger units such as Giga hashes per second (GHS) or Terra hashes per second (THS). That means, if a cryptocurrency network has a low hash rate, then the cost for an attacker wanting to purchase enough hashing power to attack the network would be relatively low. As the hash rate goes up, so does the cost of attacking the network. Knowing this correlation would help a network of miners avoid any fraudulent transaction by a malicious miner.[8]

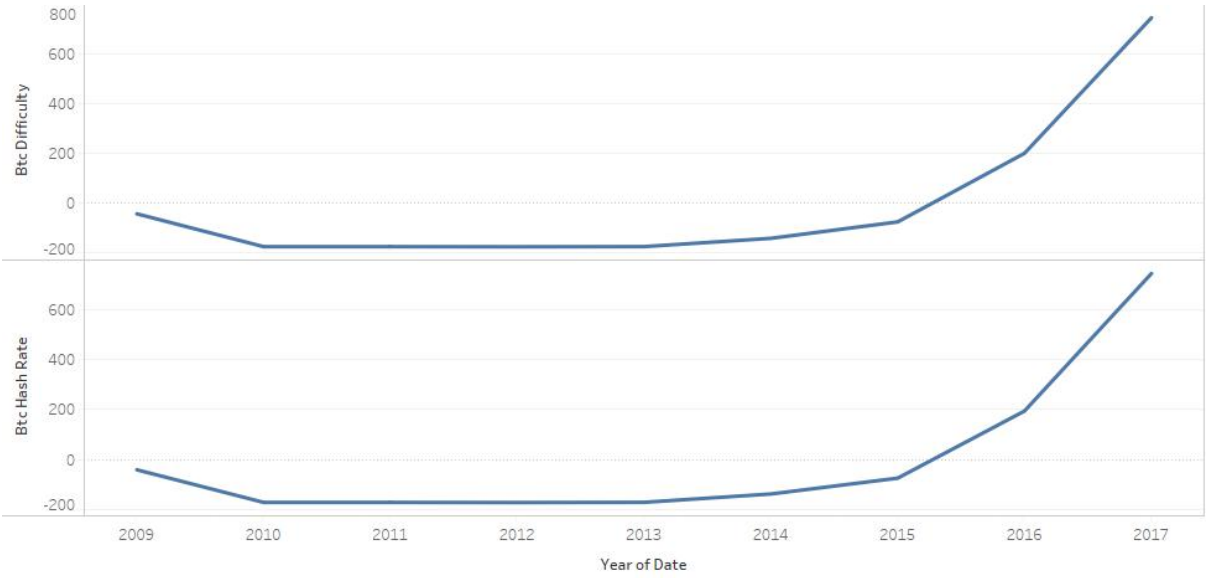

Fig 10. Variations of Bitcoin's Difficulty and Hash rate attributes with time

When the number of transaction increases, more blocks are produced. Since there is a cap of average 1 block per 10 minutes, the difficulty of the network increases. With the difficulty increase, the hash rate has to increase to keep generating blocks on time. Hence, it can be inferred that with increase in hash rate, the difficulty increases as well because with number of transactions, the hash rate has to increase to keep processing them. With the hash rate increased blocks are generated faster and thus difficulty is increased. This knowledge helps maintain a balance between number of transactions and difficulty, so that neither is the number of transactions too high, leading to high difficulty, neither is it too low, leading to an inverse causation on market price.

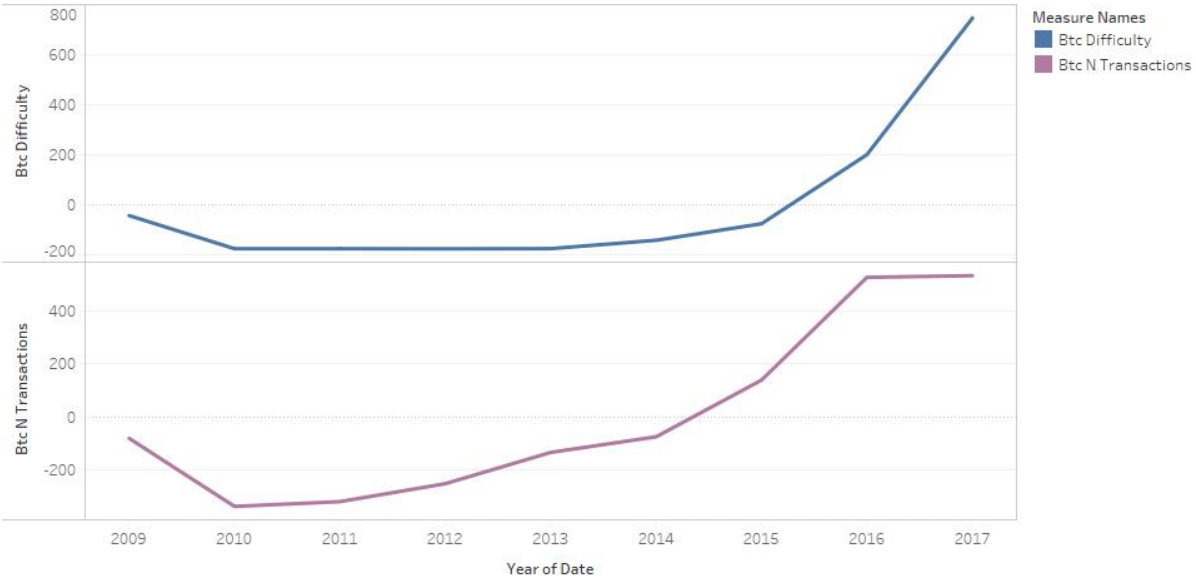

Fig 11. Variation of Bitcoin's Number of Transactions and Difficulty attributes with time

As expected, it turns out that the top 5 features affecting the price are:

1. Eth Hash rate: Hash rate in Giga hashes per second.

2. Eth Difficulty: Difficulty level

3. Eth address: Cumulative address growth

4. Eth Block size: Average block size in bytes

5. Eth Market cap: Market Capitalization in USD

Above realisation helps in more accurate predictions about the cryptocurrencies' price, now that we know the major price drivers.

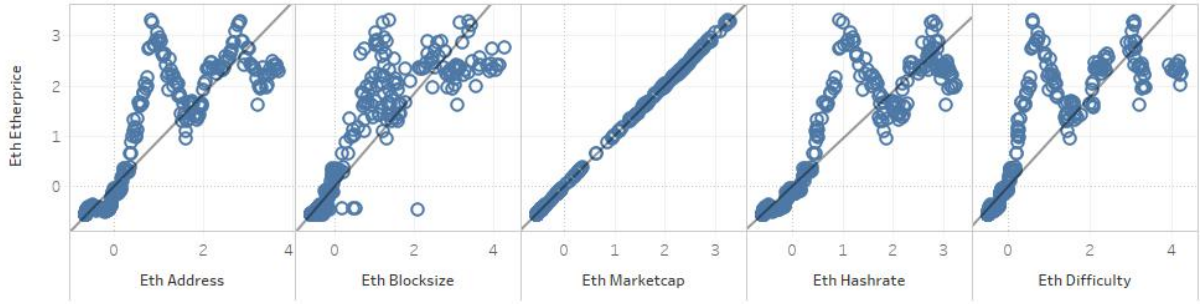

Fig 12. Correlation of attributes with the Market Price

From the historical behaviour of the cryptocurrencies, we know that hash rate increases with the number of transactions, and in turn Difficulty also increases. As seen from the plot below between the transaction fees paid to miners (validating the transactions) and the median time for a transaction to be accepted into a mined block, it can be inferred that with increase in the transaction fees, the median time also seems to be increasing, however not with similar rate. Thus, the visualizations below confirm our hypothesis that with increase in number of transactions, transaction fees increase, and hence average confirmation time also rises.[5] This would help miners verify if the transactions fees paid to them is fair according to the time spent on a block.

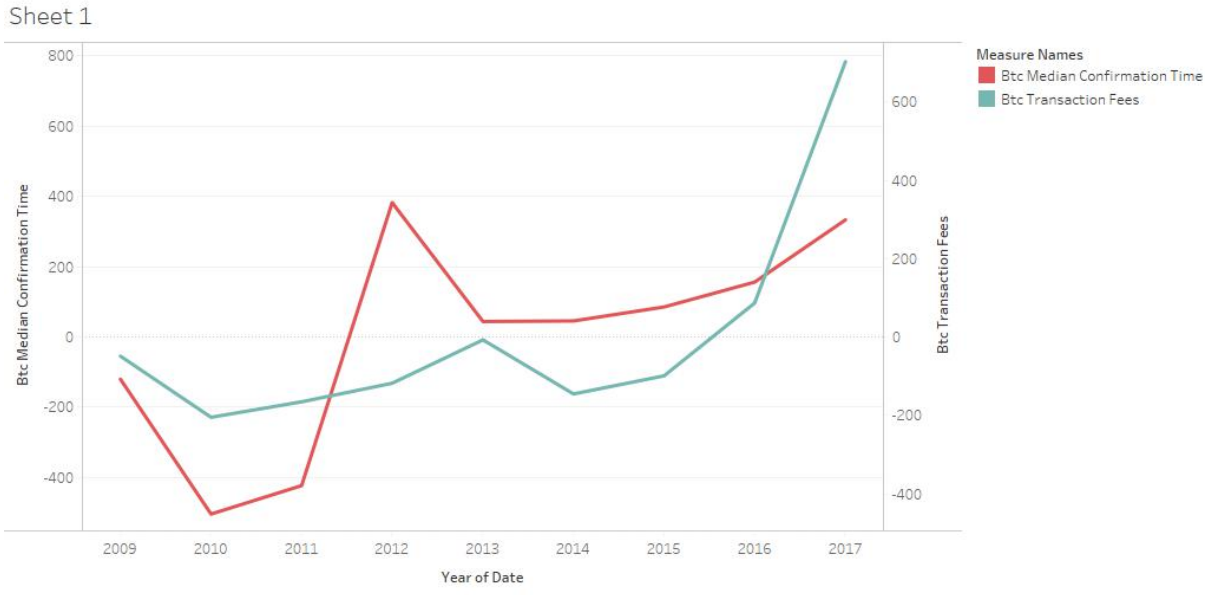

Fig 13. Variation of Bitcoin's Transaction Fees attribute with Median Confirmation Time attribute

8.Data Processing

8.1. Stationary:

Next step identifies if the data is stationary or not, by applying Dickey Fuller Test and verifying with Plot.

Differencing 0 Without any differencing, the result of Dickery Fuller Test, showed following value p-value = 0.99, Alternative Hypothesis: stationary

Since p-value > 0.05, it implies that there is a need of differencing.

Differencing 1 Differencing value as 1, the result of Dickery Fuller Test, showed following value p-value = 0.01, Alternative Hypothesis: stationary Since p-value < 0.05, it implies that data is now stationary, also seen from Figure 23 for differenced Closed Price Data.

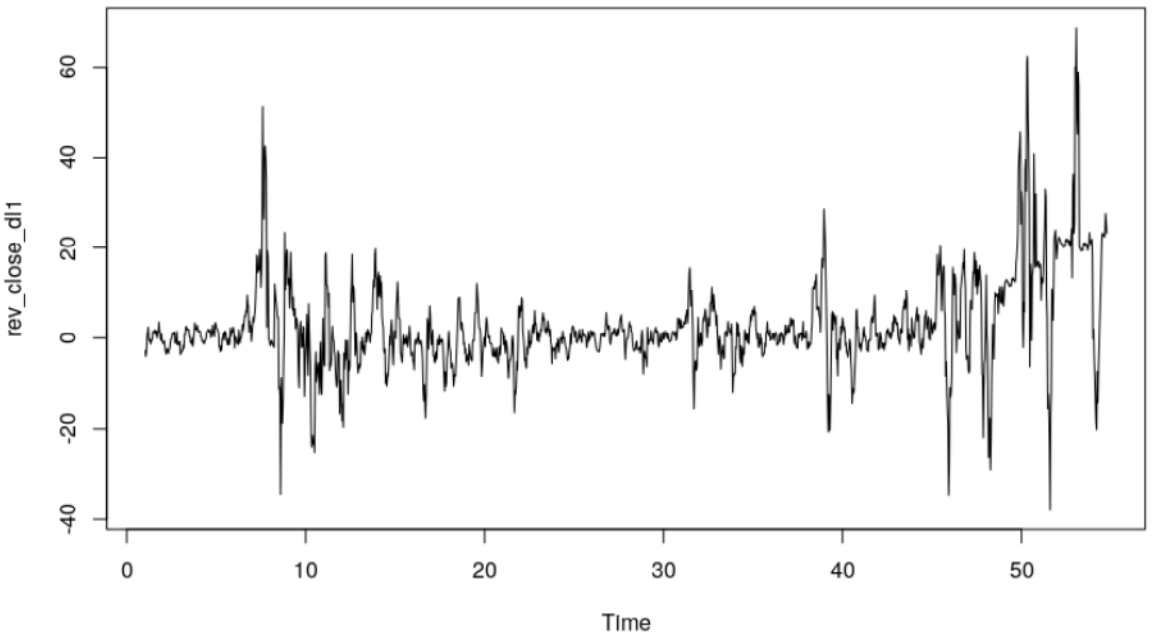

Fig 14. Differencing with 1

Differencing 2 Differencing value as 2, the result of Dickery Fuller Test, showed following value p-value = 0.01, Alternative Hypothesis: stationary Since p-value < 0.05, it implies that data is now stationary, also seen from Figure 24 for differenced Close Price Data.

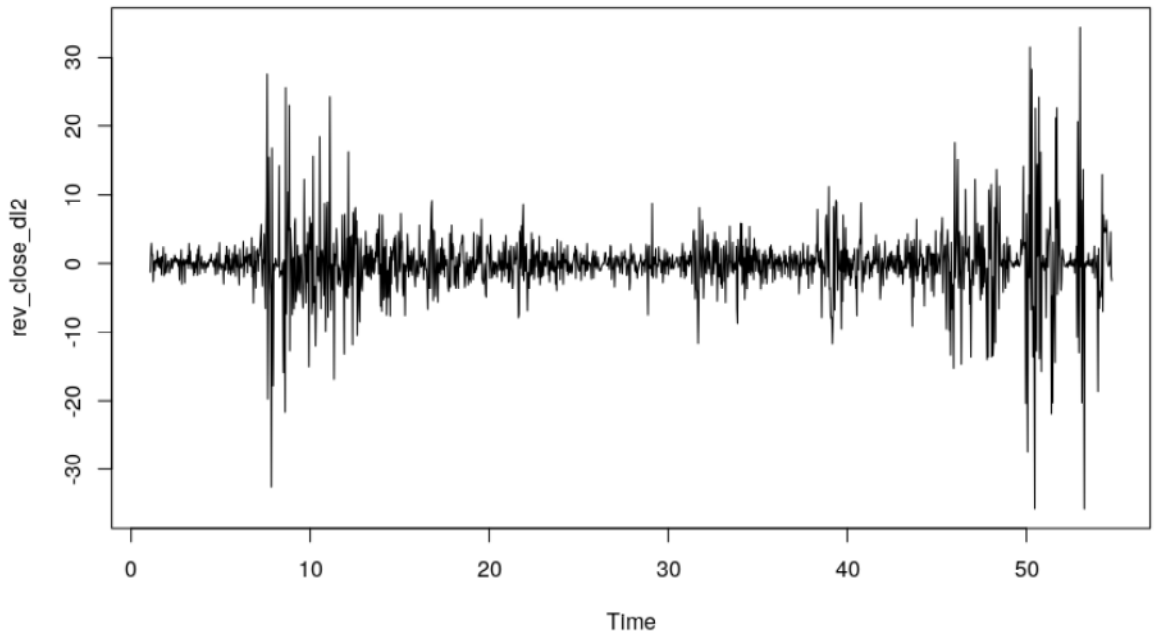

Fig 16. Differencing with 2

8.2 ACF and PACF

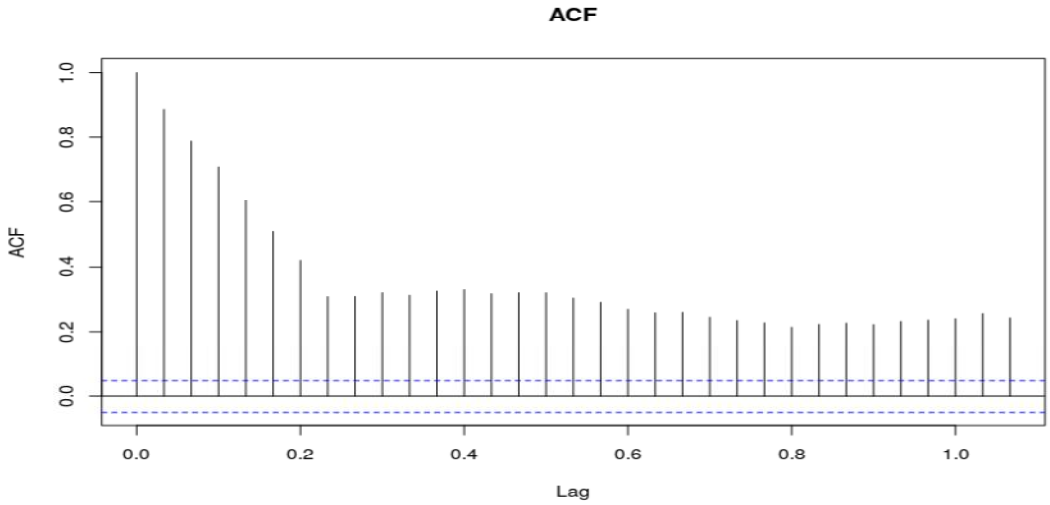

Fig 17. ACF with Differencing 1

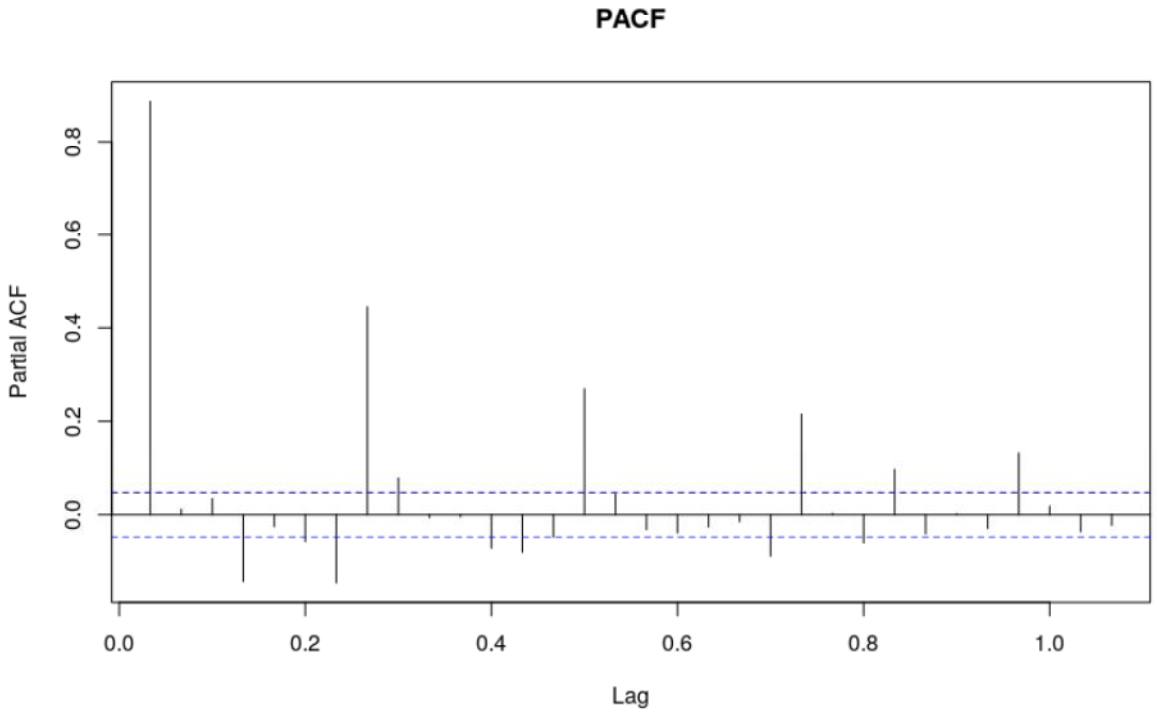

Fig 18. PACF with differencing 1

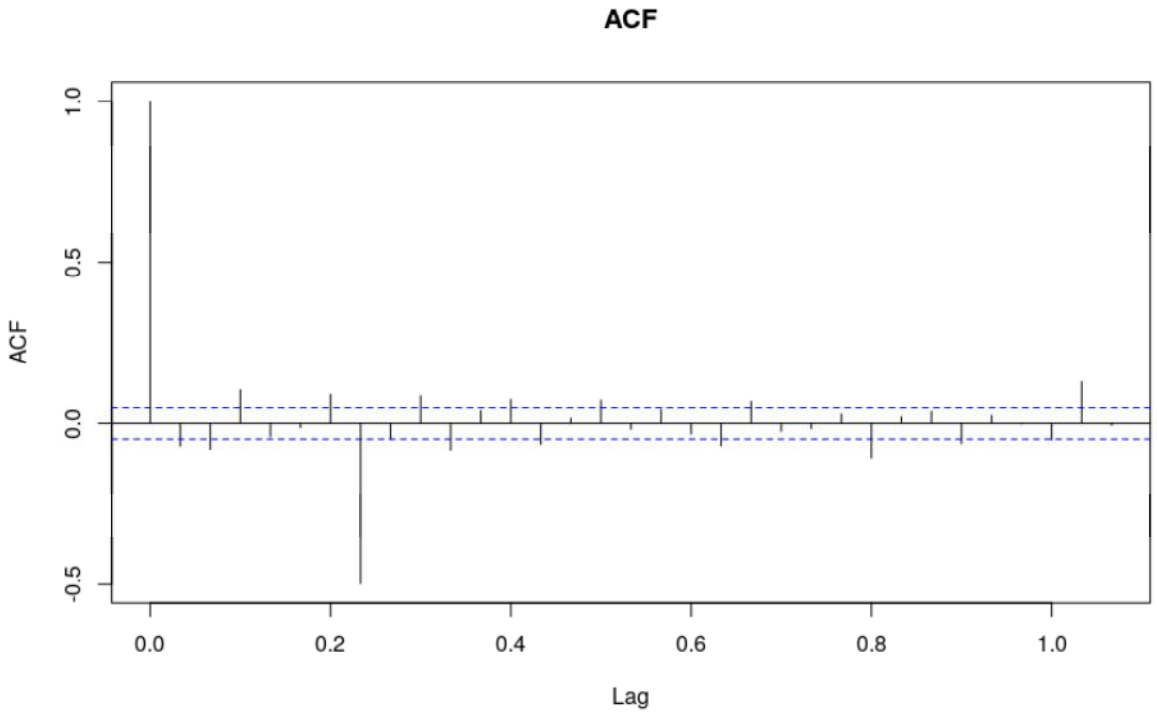

Fig 19. ACF with differencing 2

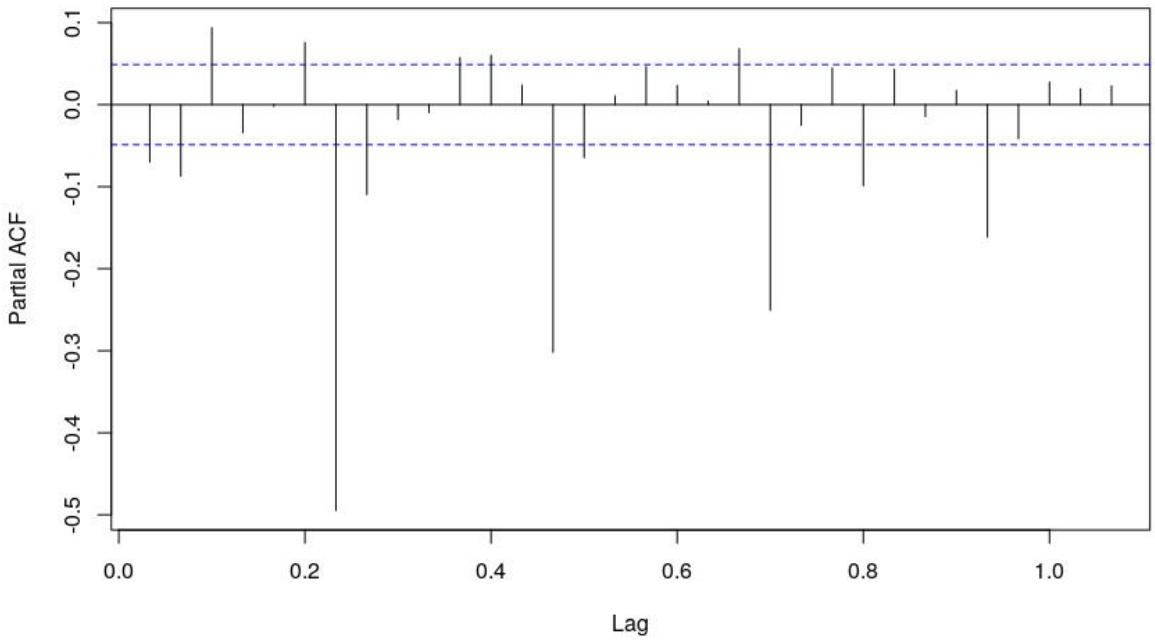

Fig 20. PACF with differencing 2

Conclusion

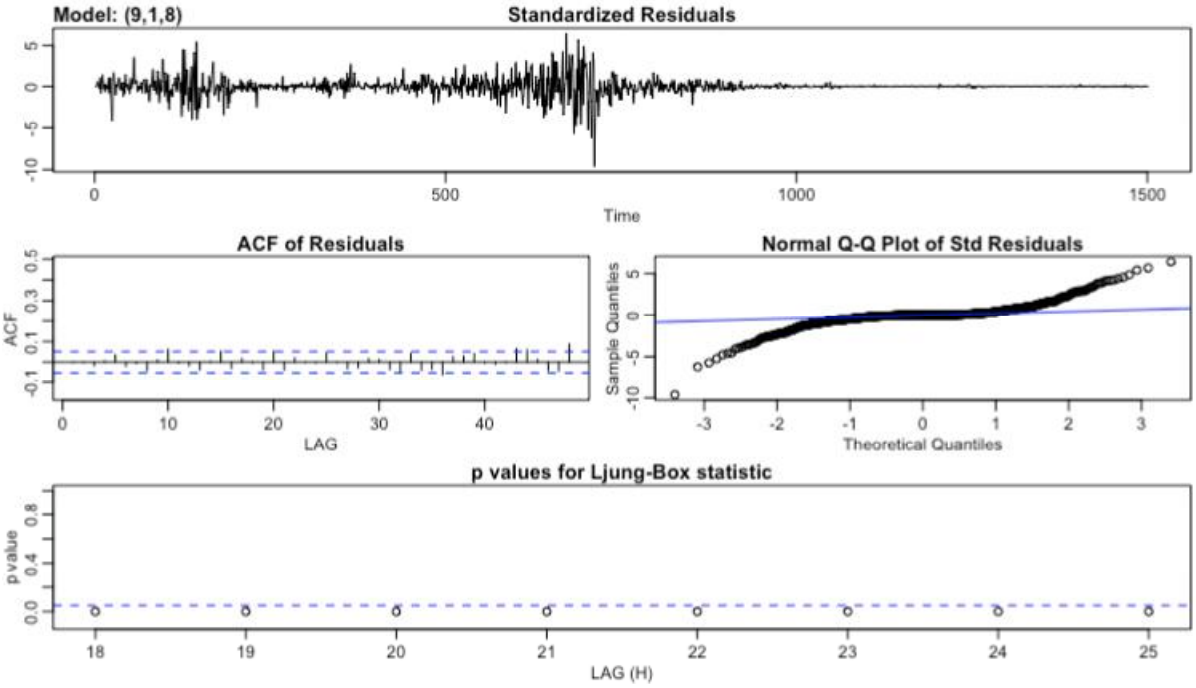

Fig 14. Model for 2018 – 2019

This dataset was difficult to fit with an ARIMA model because of the fact that the variance was non-constant, which results in squared residuals that were dependent on one another.

The residuals were also not Normal, even after transforming with the Box-Cox transformation in the beginning. Both issues easily apparent in our time series plot of the original data which has many sporadic bursts up and down.[9] While the residuals themselves were not autocorrelated, this still does not make for a forecast-ready model, at least not with ARIMA.

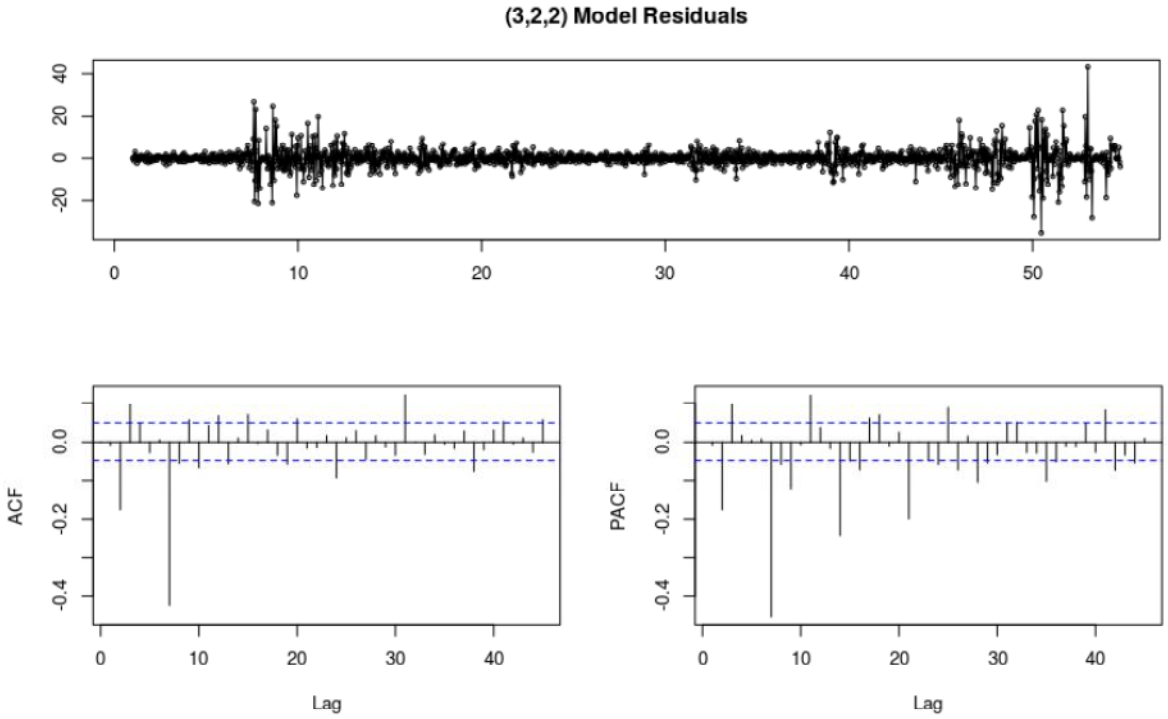

Fig 15. Model for 2009 – 2017

If someone were to accurately forecast on this dataset, it would be prudent to utilize a GARCH model instead, which better considers the non-constant variance. GARCH is a typical model for financial data. Considering our dataset in question is concerning the value of cryptocurrency (which has been extremely variable in the past several years), prospective analysts would be best advised to utilize this model in the future.

If we need to choose between the two models for trade we will go with the 2009-2017 model reason because there are lag values in 2018-2019 model which is more than 0.1 which will have adverse nature in the profit instead the price drops and turns to the lose.

We can see ACF and PACF plots for differencing order 1, 2,as above. And we see that ACF plot shows the lags in the graph with d = 2. So, this again matches with our previous conclusion when we considered the stationary data for d = 2. With this we take p = 3 from ACF and q = 7 from PACF as that is where we see real positive and negative correlation sharp cut-off respectively.[10]

References

[1]Feature Selection and Classification Techniques for Multivariate Time Series Basabi Chakraborty ;Faculty of Software and Information Science, Iwate Prefectural University - 152-52 Sugo, Takizawa-mura, Iwate, 0200193, Japan E-mail: basabi@soft.iwate-pu.ac.jp.

[2]http://www.rdatamining.com/examples/time-series-clustering-classi_cation.

[3]The Application of Machine Learning Techniques to Time-Series Data -Author Scott Mitchell; Master of Computing and Mathematical Sciences at the University of Waikato.

[4]Article by Sunil Ray at Analytics Vidhya Website Predicting the direction of stock market prices using random forest. Luckyson Khaidem Snehanshu Saha Sudeepa Roy Dey. khaidem90@gmail.com, snehanshusaha@pes.edu sudeepar@pes.edu

[5]Stock Price Prediction Using Regression Analysis Dr. P. K. Sahoo, Mr.Krishna charlapally

[6]Random Forests by the founders ; Leo Breiman and Adele Cutler rstudio-pubs website describing about the Random Forest.

[7]Article published at the Tutor website for R tutorial.

[8]Article published at the Institute for Digital Research and Education website.

[9]https://jstevenr.com/bitcoin.html#transforming

[10]https://medium.com/datadriveninvestor/predicting-cryptocurrency-prices-with-machine-learning-1b5a711d3937