CS6220 Spring 2018

Project: Cryptocurrency Analysis and Prediction

CCReco: A Recommendation System for Selling Cryptocurrency

Team Members

Sabbir Ahmad
Raghavendra Venkatesh
Jatin Taneja

Table of Contents

Table of Tables

Table of Figures

Abstract

This report focuses on the problem of recommending a cryptocurrency as well as date on which a user should sell for a better profit. First, we formulate the problem formally. We define three types of problem, (i) Find Day (ii) Find Coin and (iii) Find Coin and Day. The first and second ones finds the best day and coin, respectively, to sell for better profit, whereas, the third one finds the best coin and day to sell it. We propose algorithms to solve these three problems. We use the cryptocurrency dataset to for the experimental evaluation and also to explore the nature of the data we visualize it using different graphs and charts. For predicting the future values, we used deep learning models and gaussian process models for regression. Finally, we show the result of our experiments on the dataset and discuss it.

1 Introduction

Cryptocurrency is a type of digital or virtual currency. It uses decentralized control as opposed to centralized banking systems. It uses Blockchain which is a public distributed transaction ledger. Blockchain technology was first implemented by a group of people named as Satoshi Nakamoto in 2009 as a core component of Bitcoin, 1st cryptocurrency introduced in the market. Its usage allows secure peer to peer communication by linking blocks containing hash pointers to a previous block, a timestamp and transaction data. By design, blockchain are hard to modify as a copy of main ledger is distributed among many.

In a recent market trend, we have seen remarkable fluctuations in price trend of Bitcoin and other cryptocurrencies. Many wise people who had already invested in cryptocurrencies have made a good profit out their investment. This is a great area for an opportunist to make money but due to lack of knowledge and historical market understanding about cryptocurrencies, many people are reluctant to invest money in it.

There have been thousands of cryptocurrencies introduced in the market since the birth of Bitcoin, mainly to cap the money in market circulation, which is why cryptocurrencies are also known as AltcCins**.** However, not every cryptocurrency sustained in the market such coins are called DeadCoins**.** Most dead coins are the result of scam where organization tried to fool users by a fake cryptocurrency or because of security loopholes.

In this article, we are aiming to bridge this gap and suggest a recommendation system, *Cryptocurrency Recommendation (CCReco)*, where our aim is to recommend a new user a cryptocurrency to invest based on various parameters like his investment budget and percentage of risk he is ready to take, whether he wants long term investment or short term investment and in which period of an year he should invest to make maximum profit out of his investment or in which period of a month a user should buy or sell cryptocurrencies. The problem statement can be formally stated as follows. *Given the investment budget, risk percentage, and term (long/short) of investment, the CCReco system recommends a user a cryptocurrency to sell at a time that will give her the maximum profit. This problem can be easily extended where the system returns top-k cryptocurrencies.*

Making money is easy if you are making calculated investments at right time. Nowadays, many people are investing in cryptocurrencies and most of the people have very little knowledge about it. Suppose, a person is planning to invest in cryptocurrency and due to their volatile nature, he is confused on which one to invest and is finding a reliable system to help him decide. Cryptocurrencies are volatile and hence need a reliable system to predict their future price to help investors make an informed decision to invest in it. This can also be extended to provide market value prediction to back up his decisions. Price history charts will provide them more insight into how a currency was behaving over last few years. The main motivation behind the system is to help them choose a currency that can make their investment profitable.

There are certain studies in the field of cryptocurrency and most of them are related to how the system of cryptocurrency and blockchain work. How this system of currency is being a significant part of the economy and what are the long and short-term effects are the focus of research at this moment. Regarding the technical analysis of different cryptocurrencies available in the market, there is almost no published literature. On the other hand, many people have analyzed the cryptocurrency data to predict the prices and some descriptive analysis on the dataset. To our knowledge, there is no study of recommending a cryptocurrency that is stable and at the same time should be profitable to invest.

Analysis of different cryptocurrencies is done in [1]. Price trend, the correlations between different attributes, plotting with time series are shown. Similar types of analysis are done in [6, 7, 9, 10]. In all the articles only, a small number of cryptocurrencies are analyzed; for example, [6] only analyzes Bitcoin. [5] visualizes different characteristics of cryptocurrencies by plotting the price over years and, market cap over years. This only handled Bitcoin and Ethereum for the visualization and mainly shows the comparison of different attributes between the two specified coins. Time series visualization is also shown in [2]. Prices and average values with respect to time is visualized in this study. D. Sheehan in [3] predicted the prices of Bitcoin using deep learning methods. In this article, he used different models to predict prices for Bitcoin and Ethereum. Price prediction for Bitcoin is studied in [8] with Auto-Regressive Integrated Moving Average (ARIMA) Model. Changing different combinations of parameters of the model, linear regression is constructed through this model to predict the output. In [4] trading strategy simulation for Bitcoin is shown, which actually states that there is no straightforward trading strategy for Bitcoin mining.

All these studies are different from our work. In this report, we explore the dataset by doing some analyses that have already been done in different datasets, however, to address the problem that is mentioned the analysis is necessary to understand the nature of the data. After the exploration of the data, we apply our methodology to address the problem.
In summary the contributions of this report are as follows.

- We define three problems formally that a user can have while selling a cryptocurrency.
- We propose algorithms to solve all the three problems.
- Exploration on cryptocurrency data to show the trends in dataset.

- We visualize different properties of the dataset with graphs and plots.
- We run our algorithm on the dataset to recommend coins that ensures better profit.

The rest of the report is organized as follows. Section 2 describes the problem formally. In Section 3, the methodology to solve each problem is explained. Section 4 contains the code and related sources for the experiments done. Section 5 contains the experimental evaluation, where the used dataset, exploration of data and prediction results and methods are shown. Section 6 shows the results of our algorithm and the discussion is done in Section 7. Section 8 describes the future aspects of the work. The report concludes in Section 9.

2 Problem Formulation

We formally formulate the problem we propose to solve here. Our system basically works on three problems. The user of the system gives input to the system. The inputs are a set of coins, and a range of days. Depending on the inputs three problems can be formulated. The problems are formally formulated below.

Definition 1. (Find Day) Given a specific coin C and a range of days D, find out the best day $d \in D$ to sell the coin C that will ensure the maximum profit.

Definition 2. (Find Coin) Given a set of coins X_c and a specific day d, find out the best coin $C \in X_c$ that will ensure the maximum profit.

Definition 3. (Find Coin and Day) Given a set of coins X_c and a range of days D, find out the best coin $C \in X_c$ and best day $d \in D$ that will ensure maximum profit for selling coin C on day d.

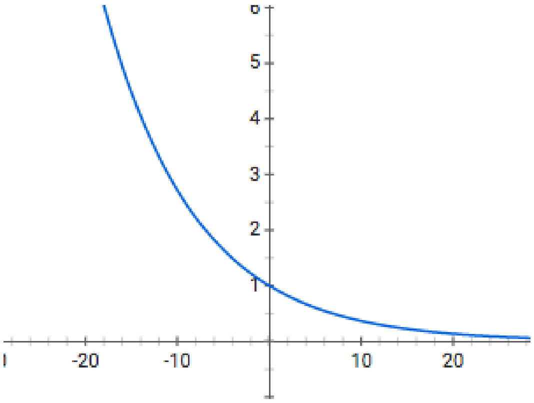
These ideas can be easily extended to rank, where each of the solutions of the problems return a list containing the ranks of the coins or days instead of just one best coin or day. Please note that the algorithms described in Section 3 returns the ranks.

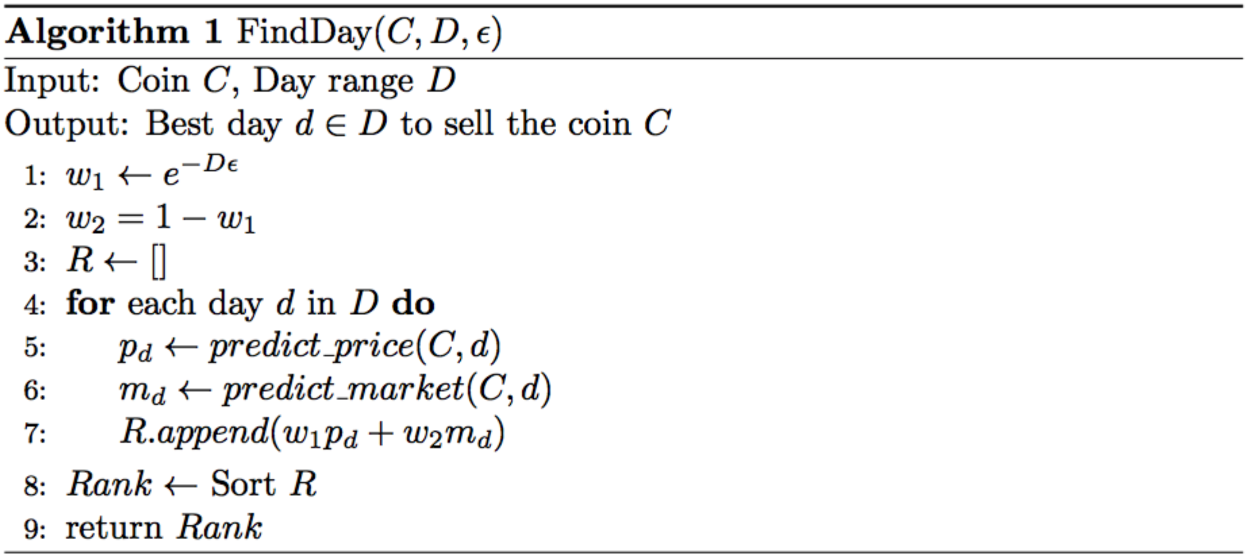
3 Methodology

We propose solutions for all the three problems we mentioned in Section 2. The subsections in this section describe the solutions for the problems.

3.1 Solution for Find Day

To get the best day for selling the coin, the future price needs to be calculated. As the cryptocurrency data is too much random, we consider the predicted market share as well as the predicted price for future. We use a technique with weighted sum taking these two predicted values for each of the days to rank the days given in the range. As different cryptocurrencies have different price ranges, we use the rate of price changes and rate of market share change, which enables the system to compare between different coins. The weights w1 and w2 are calculated in such a way so that $w1 + w2 = 1$. We calculate $w1 = exp\{-D\varepsilon\}$. Note that, as the value of D gets larger, the weight w1 gets smaller. This is because, as the number of days to consider is large, the price is given lower weight. As the number of days gets larger, the user might be thinking for long term investment, and that is why the market share is a good indicator for a cryptocurrency. On the other hand, if D is small, then the user wants to gain profit in short term, then only the price is given high priority and market share is given lower priority. It is worth mentioning that ε is a hyperparameter, and for the experiments we use 0.03 for this. We can see the graph of $exp\{-x\varepsilon\}$ in figure here. As D is always positive, the maximum value of w1 is 1, and as D increases, the weight w1 decreases and w2 increases. Algorithm 1 shows the solution for the problem Find Day. The predict_price(C, d) returns the price of coin C on day d, and the predict_market(C, d) returns the market value of coin C on day d. The days are sorted with the weighted sum value and the sorted list is returned from the algorithm.

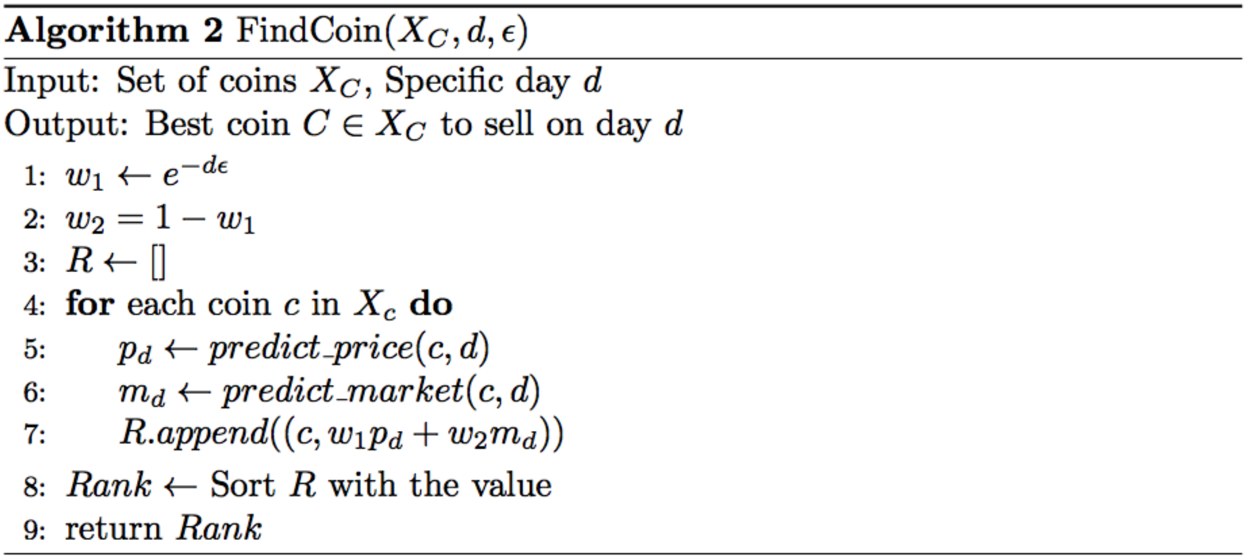




```
Algorithm 1 FindDay($C, D, \epsilon$)
Input: Coin $C$, Day range $D$
Output: Best day $d \in D$ to sell the coin $C$
$w_1 \leftarrow e^{-D\epsilon}$
$w_2 = 1 - w_1$
$R \leftarrow []$
for each day $d$ in $D$ do
    $p_d \leftarrow predict\_price(C, d)$
    $m_d \leftarrow predict\_market(C, d)$
    $R.append(w_1 p_d + w_2 m_d)$
$Rank \leftarrow$ Sort $R$
return $Rank$
```

3.2 Solution for Find Coin

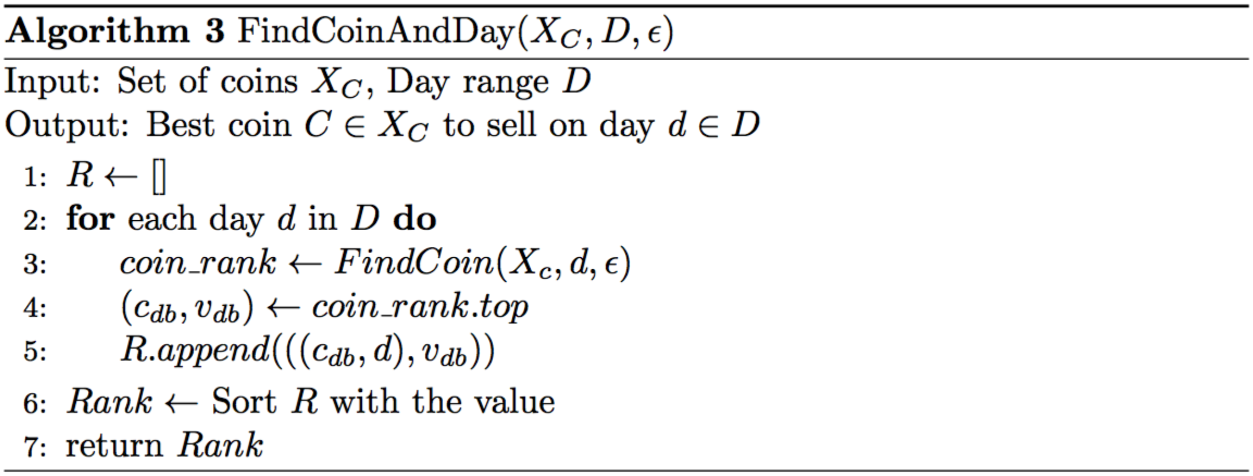
The solution for this problem is done in a similar way with the first problem. Here we predict prices and market value for each coin in the given coin set on the specific day, and take the weighted sum as done previously. In this algorithm the coin price with the weighted sum value is saved as a tuple in the array, and the array is sorted using the weighted sum. Note that, here day d means the d-th day from the current day. After that the sorted list is returned. The solution is shown in Algorithm 2.



```
Algorithm 2 FindCoin($X_C, d, \epsilon$)
Input: Set of coins $X_C$, Specific day $d$
Output: Best coin $C \in X_C$ to sell on day $d$
$w_1 \leftarrow e^{-d\epsilon}$
$w_2 = 1 - w_1$
$R \leftarrow []$
for each coin $c$ in $X_c$ do
    $p_d \leftarrow predict\_price(c, d)$
    $m_d \leftarrow predict\_market(c, d)$
    $R.append((c, w_1 p_d + w_2 m_d))$
$Rank \leftarrow$ Sort $R$ with the value
return $Rank$
```

3.3 Solution for Find Coin and Day

Given the solution to problem 2, we can easily extend and use it to solve the problem 3. The solution is shown in Algorithm 3. For each day in the given day range D, we calculate the coin rank for the given set of coins. At each iteration we get the rank of coins for a specific day. We take the top coin from the rank at each day. This is a tuple with the coin name and the associated weighted sum. We take this tuple and make another tuple with the specific day and make a list (Line 4-5). The list is sorted using the value that is associated with each coin. This gives us the rank of coins and the corresponding day.



```
Algorithm 3 FindCoinAndDay($X_C, D, \epsilon$)
Input: Set of coins $X_C$, Day range $D$
Output: Best coin $C \in X_C$ to sell on day $d \in D$
1: $R \leftarrow []$
2: for each day $d$ in $D$ do
3:     $coin\_rank \leftarrow FindCoin(X_c, d, \epsilon)$
4:     $(c_{db}, v_{db}) \leftarrow coin\_rank.top$
5:     $R.append(((c_{db}, d), v_{db}))$
6: $Rank \leftarrow$ Sort $R$ with the value
7: return $Rank$
```

4 Code

The code to generate the analysis and visualization has been placed at following GitHub repository. Please refer to this repository for detailed code structure.

GitHub Repository: https://github.com/sabbirahmad/cryptocurrency-analysis

5 Experimental Evaluation

5.1 Dataset

We use a cryptocurrency dataset [11] for the experiments in this article. The dataset contains 702,166 instances of 1,516 types of Cryptocurrencies. The instances are recorded each day and the attributes of the instances contain the date of the record, opening and closing price of the respective day, and also the high and low prices for the particular day. Each instance also contains the volume and market share of the coin on that day.

In the dataset we have different cryptocurrencies data for different period of time so when we are comparing different cryptocurrencies, we are dropping a record wherever the data is missing. This has improved the visualization results.

5.2 Dataset Statistics

5.2.1 Properties

- Data size **69.4 MB**
- **1,516** types of Cryptocurrencies
- Number of instances: **702,166**
- Number of attributes: **13**
- Instances are records of different Cryptocurrencies at each day
- Time range of collected data: **04-28-2013** to **02-21-2018**

5.2.2 Attributes

- **Coin name:** identifies the coin type
- **Date:** date of the instance
- **Open price:** opening price of the coin on that date in dollar
- **High price:** highest price of the coin on that date in dollar
- **Low price:** lowest price of the coin on that date in dollar
- **Close price:** closing price of the coin on that date in dollar
- **Volume:** The amount of coin traded on that date
- **Market cap:** total market value of a company's outstanding shares in dollar
- **Ranknow:** Rank of the cryptocurrencies depending on the current market share
- **Close Ratio:** daily close rate, min-maxed with the high and low values for the day
- **Spread:** The price range of the coin on that day

5.3 Data Cleaning

We explored the dataset for missing values. We saw that many coins don't have market share values and those are represented as 0, while they do have the everyday prices and volumes. As a result, we consider those as missing values and exclude them from the dataset. Figure 2 shows the percentage of the used data and data with missing values in the pie chart.

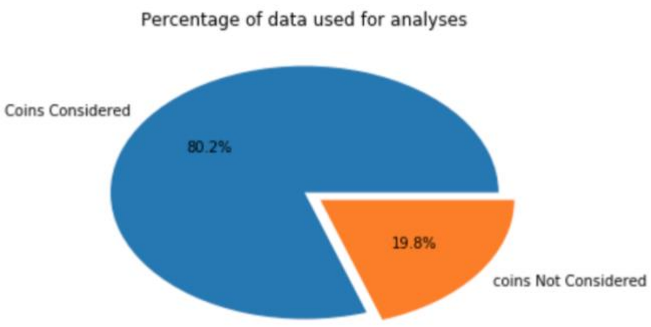


Figure 2: Data cleaning for missing values

5.4 Data Exploration

We explore the dataset by plotting different types of graphs for visualizing the nature of the dataset.

5.4.1 Percentage of Cryptocurrencies

The percentage of different cryptocurrency data in the dataset are shown in the following pie chart. As there are 1516 cryptocurrencies, the chart does not show the amounts clearly. We took the top 50 currencies to demonstrate their data quantity in the overall data.

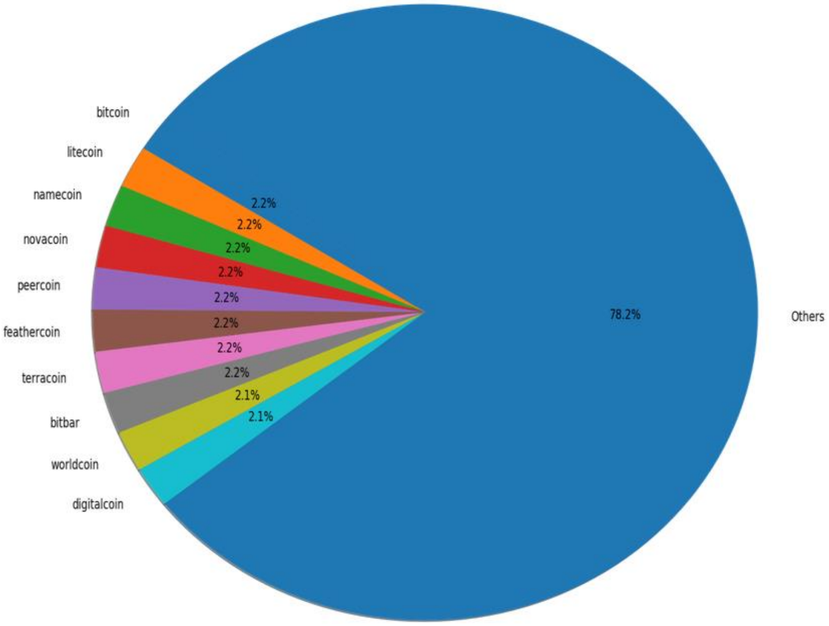


Figure 3: Amount of data for top 50 currencies

5.4.2 Price Distribution

To show the price distribution of different cryptocurrencies, we plot the opening, closing, high and low prices in boxplot format. As there are 1,216 types of cryptocurrencies, we take the first few and last few cryptocurrencies ranked by their market share to show the price nature. We can see from Figure 1 and 2, the first five currencies show high values in their price, whereas the last five currencies show low prices, respectively. The outliers are also plotted in a different graph. From the outliers we can see that some coins (e.g. Bitcoin, Ripple, etc) have long ranges of outliers that can randomize the prices of the coins too much to predict.

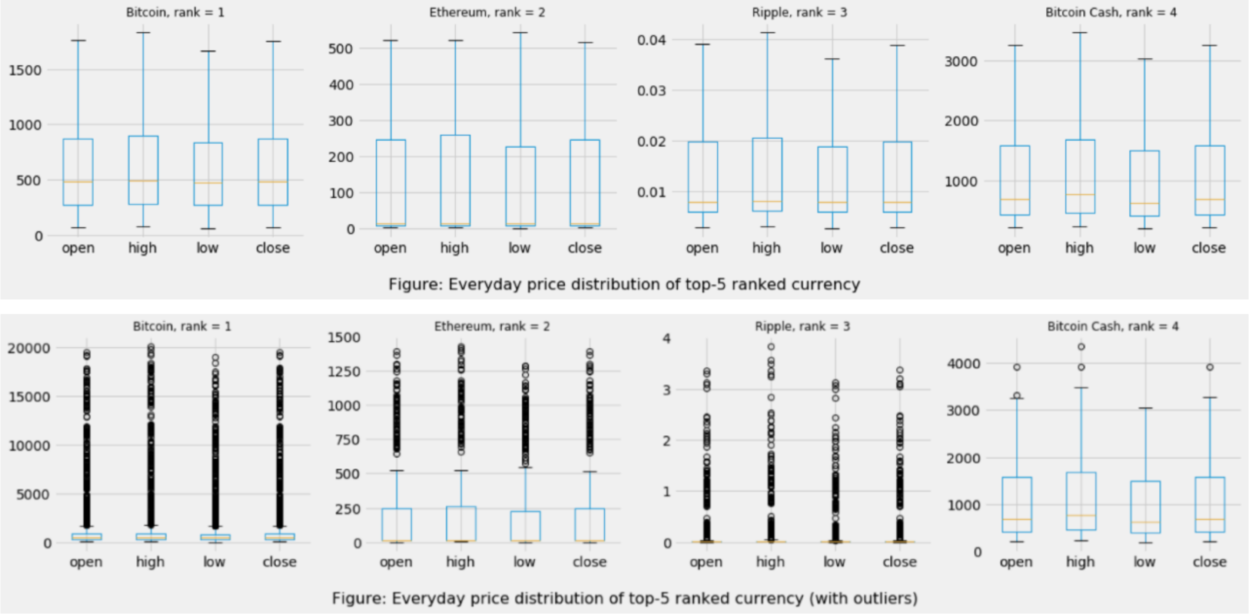


Figure: Everyday price distribution of top-5 ranked currency

Figure: Everyday price distribution of top-5 ranked currency (with outliers)

Figure 4: Everyday price distribution of first five ranked coins

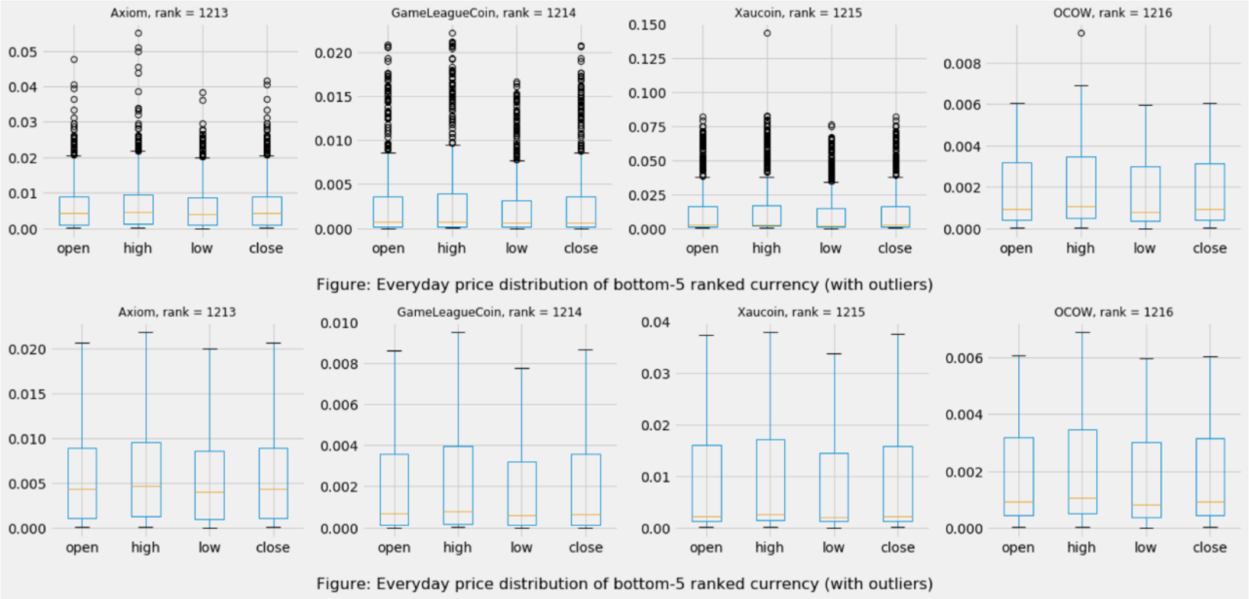


Figure: Everyday price distribution of bottom-5 ranked currency (with outliers)

Figure: Everyday price distribution of bottom-5 ranked currency (with outliers)

Figure 5: Everyday price distribution of last five ranked coins

We also plot the price distributions considering all the coins for the last day of the recorded dataset. The last date on which the data was recorded is February 21, 2018. We again consider the distributions for first ten ranked, first 100 ranked and for currencies ranked more than 1450. The boxplots are shown in Figure 3. We can see that the first ten ranked currencies show high prices, and the price reduces as more currencies are considered.

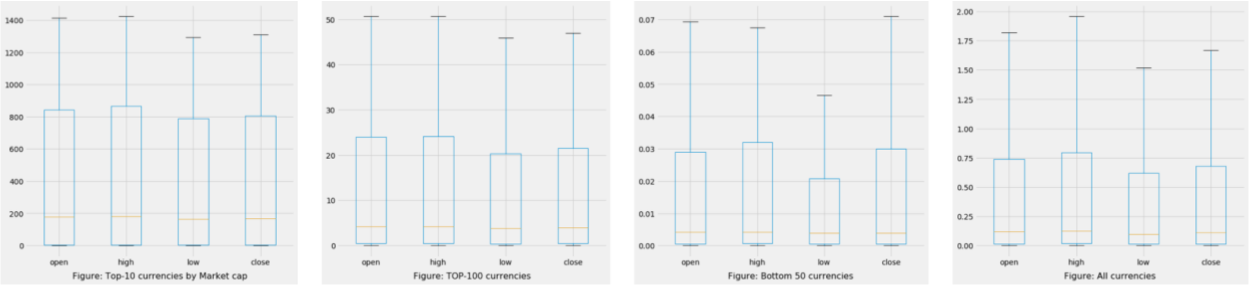


Figure 6: Price distribution on a specific day (February 21, 2018) for a number of coins

5.4.3 Comparison on Different Aspects

For comparison between different cryptocurrencies, we are using line graph to visualize and analyze the differences between cryptocurrencies in various aspects.

Figure 7 shows the comparison of top 10 cryptocurrencies in respect of their market value over the period of time.

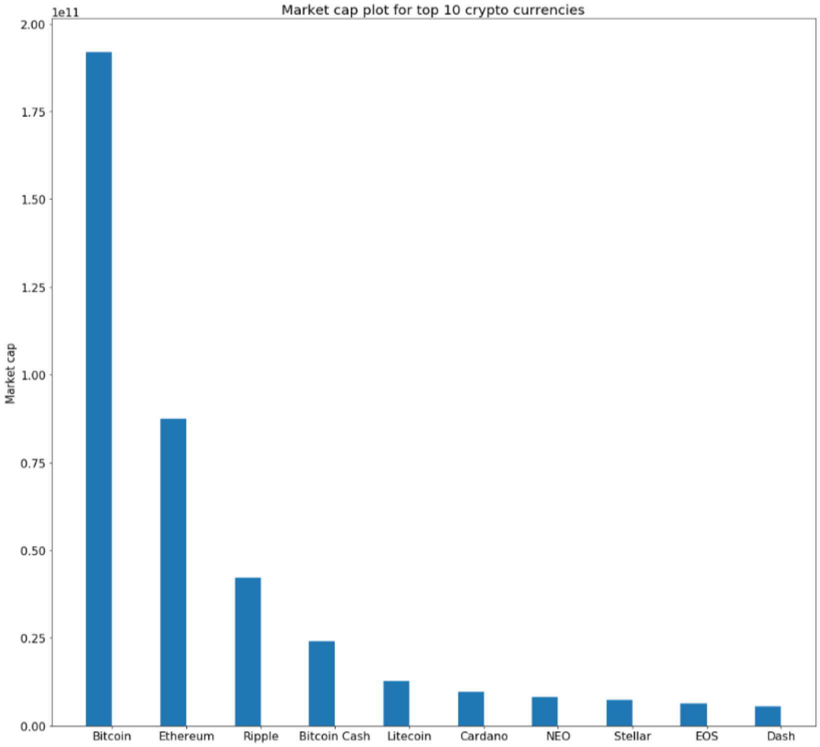


Figure 7: Comparison of top 10 cryptocurrencies with respect of their market value

Figure 8 represents the comparison of top 5 cryptocurrencies based on their market value.

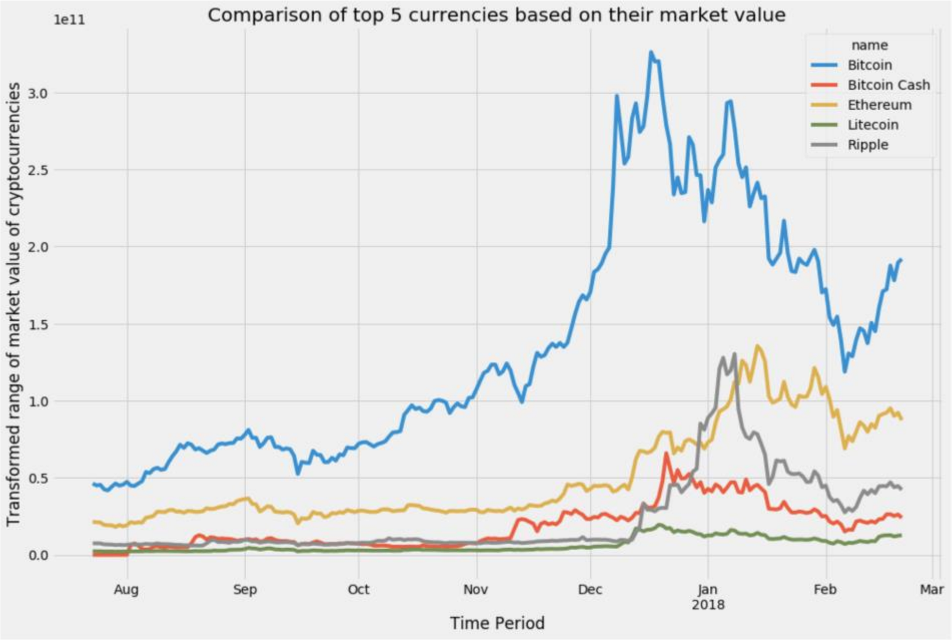


Figure 8: Comparison of top 5 cryptocurrencies with respect of their market value

Figure 9 depicts the comparison of top 5 cryptocurrencies in respect of their volume over the period of time.

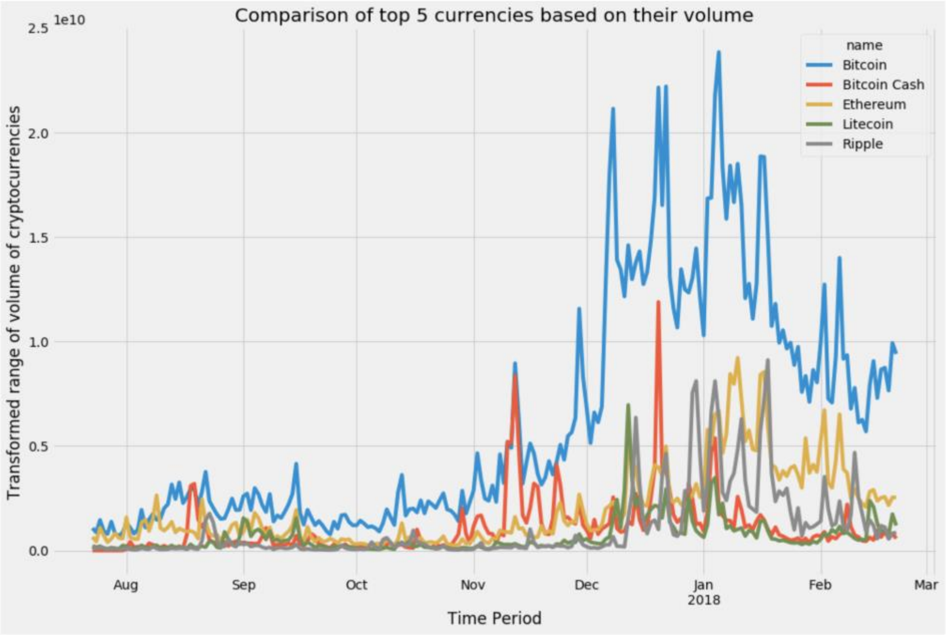


Figure 9: Comparison of top 5 cryptocurrencies with respect of their volume

Figure 10 describes the comparison of top 5 cryptocurrencies in respect of their closing values over the period of time.

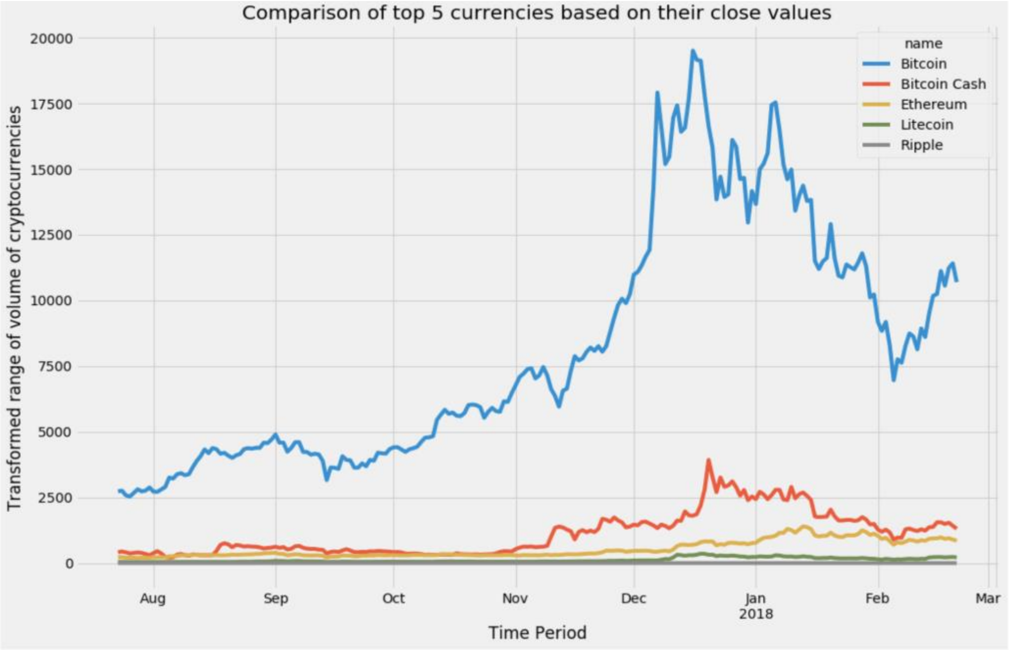


Figure 10: Comparison of top 5 cryptocurrencies with respect of their closing values

Figure 11 illustrates the comparison of top 5 cryptocurrencies in respect of their daily high and daily low values over the period of time. We can easily observe here that Bitcoin's low was higher than high of other top cryptocurrencies for most of the time.

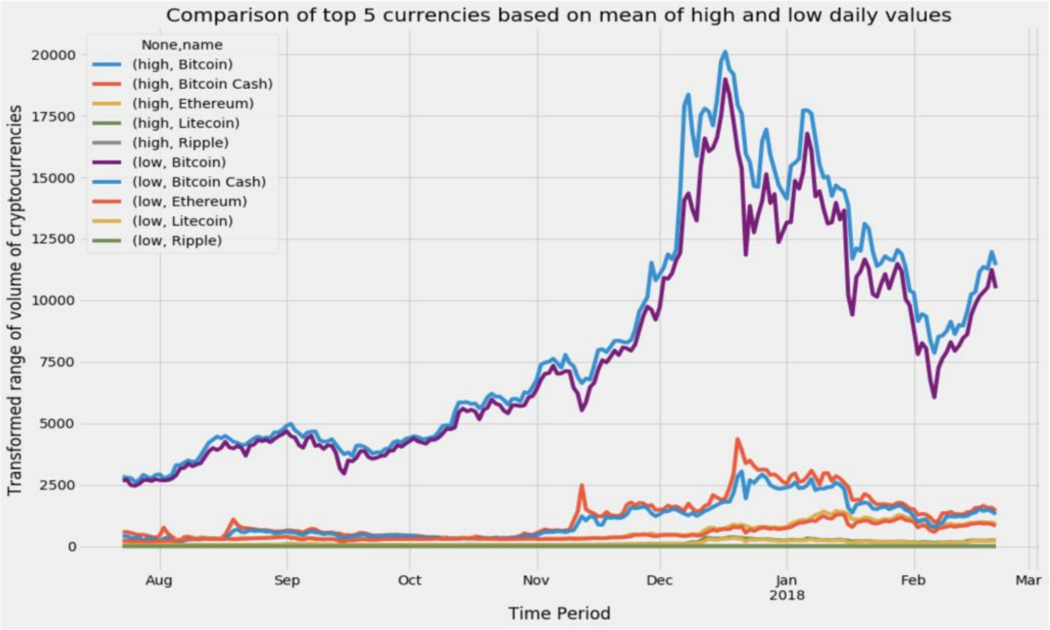


Figure 11: Comparison of top 5 cryptocurrencies with respect of their daily high and daily low values

5.4.4 Fluctuations

Trend of dataset properties can be seen from the fluctuations over a certain time range. Fluctuation means of change of a value over a range of times. For example, the market share for a coin changes every day as we can see from the dataset. As different coins have emerged different times and remain in the market for varying times, their market share depends on that. To see the trend of change and to compare them we need to figure out something that can be used to calculate the fluctuations. To do this, we came out with the idea of rate of change in market share, or price of any other attributes of the coin that is continuous in nature.

To compare the fluctuations of the values for different coins, we use the concept of different measures of rate of change specified below.

a. **Rate of Change:** Rate of change implies the change in everyday market share with respect to the current share.
b. **Absolute Rate of Change:** Absolute value of the rate of change.
c. **Rate of Increment:** This implies the rate of changes of values while the value is increasing.
d. **Rate of Decrement:** This implies the rate of changes of values while the value is decreasing.

We calculate mean rate for each of the four types of rates. And with that we can interpret the performances of the coins. Four types of rate for Bitcoin is shown in Figure 12. We can see the rate of changes in the graphs.

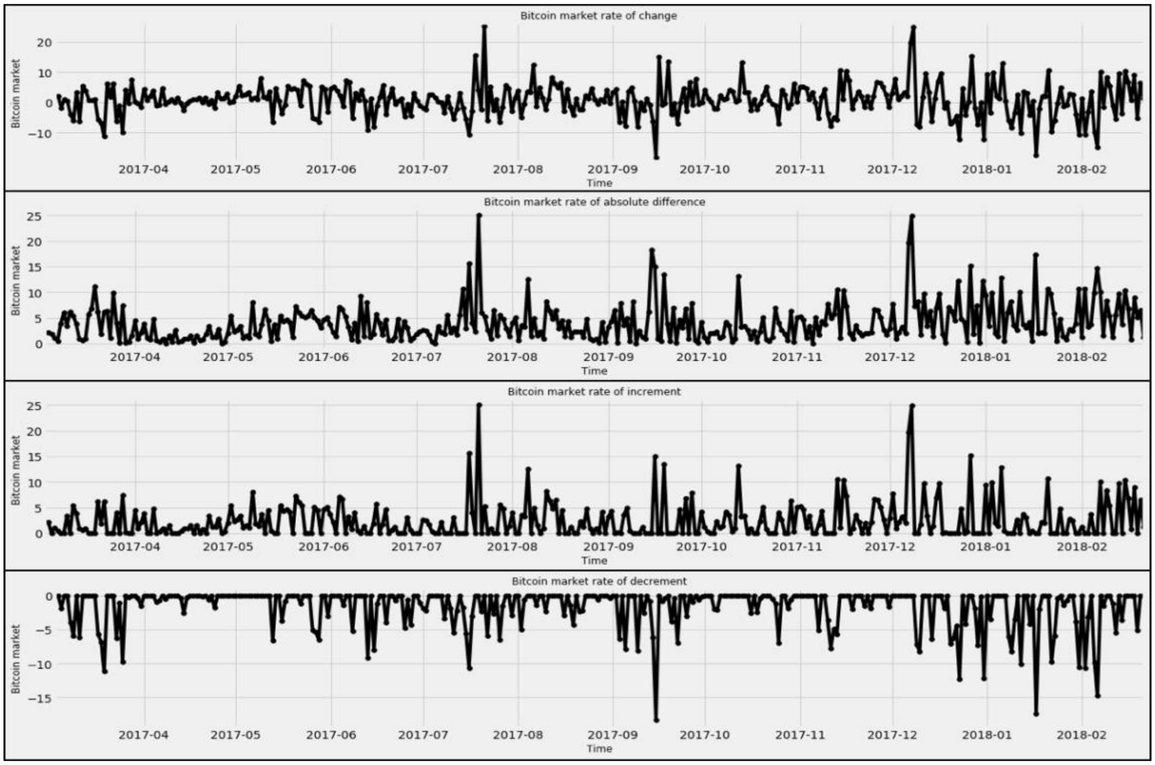


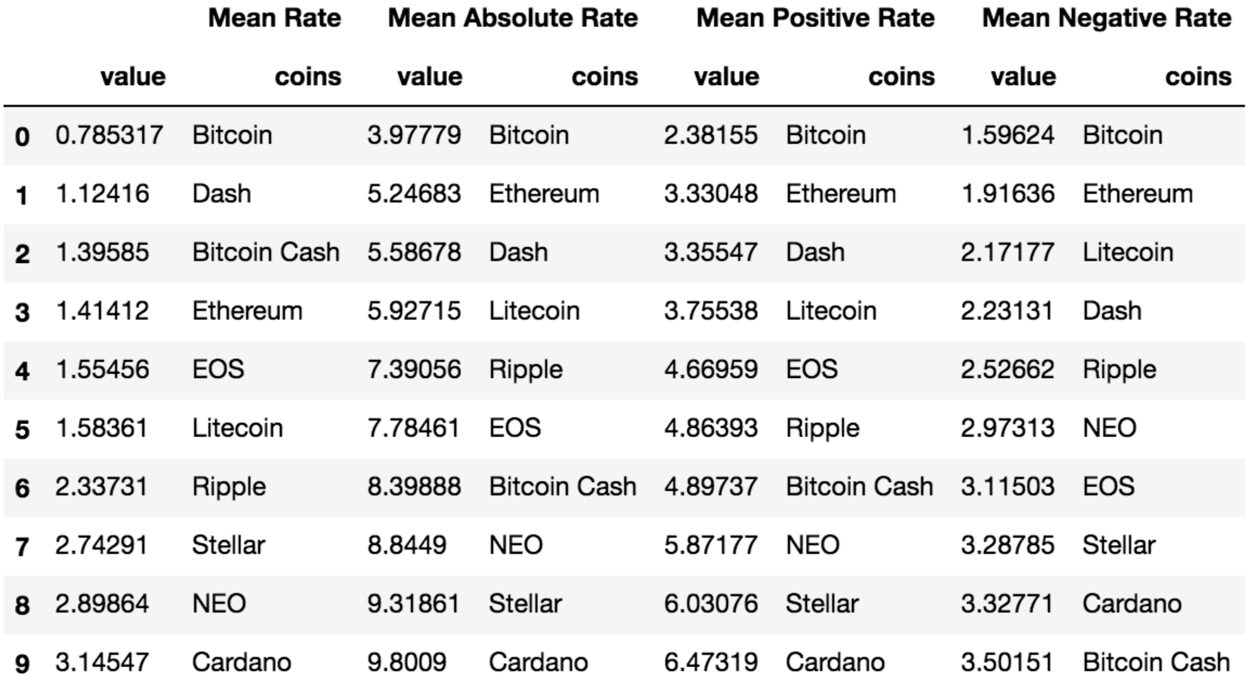
Figure 12: Market share fluctuation of Bitcoin

To compare the performances of each coin, we take mean for each types of fluctuations. The mean rate implies the average change in the rate of the values, whereas, the absolute mean rate implies all the changes, which is a measure of volatility. Lower value of absolute mean rate means the changes in values is low. The mean rate of increment (mean positive rate) and mean rate of decrement (mean negative rate) imply the average increment and decrement of the rates, respectively.

Table 1 shows the four different types of mean rates of changes for ten different coins. From here, we can see that Bitcoin has minimum fluctuation in the in rates. Bitcoin Cash has low mean rate change, but high Absolute rate change, which means it has similar changes in both decrement and increment rate and the changes are higher compared to other coins.

Mean rate implies another important aspect in determining the performance of the coins. Whenever a coin has negative mean rate, it means that the rate of decrement of the coin is larger than the rate of increment of the values over the specified amount of time. That means, the price of market share of the value is consideration has gone lower most of the times when it changed, which indicates a bad performance for the coin. In Figure 13 we can see 82 coins (7.28%) have mean rate less than zero. That tells us that while recommending, we should never consider those coins because their rate of decrement is higher than the rate of increment.

Table 1: Comparison of price fluctuations for ten coins



	Mean Rate		Mean Absolute Rate		Mean Positive Rate		Mean Negative Rate	
	value	**coins**	**value**	**coins**	**value**	**coins**	**value**	**coins**
0	0.785317	Bitcoin	3.97779	Bitcoin	2.38155	Bitcoin	1.59624	Bitcoin
1	1.12416	Dash	5.24683	Ethereum	3.33048	Ethereum	1.91636	Ethereum
2	1.39585	Bitcoin Cash	5.58678	Dash	3.35547	Dash	2.17177	Litecoin
3	1.41412	Ethereum	5.92715	Litecoin	3.75538	Litecoin	2.23131	Dash
4	1.55456	EOS	7.39056	Ripple	4.66959	EOS	2.52662	Ripple
5	1.58361	Litecoin	7.78461	EOS	4.86393	Ripple	2.97313	NEO
6	2.33731	Ripple	8.39888	Bitcoin Cash	4.89737	Bitcoin Cash	3.11503	EOS
7	2.74291	Stellar	8.8449	NEO	5.87177	NEO	3.28785	Stellar
8	2.89864	NEO	9.31861	Stellar	6.03076	Stellar	3.32771	Cardano
9	3.14547	Cardano	9.8009	Cardano	6.47319	Cardano	3.50151	Bitcoin Cash

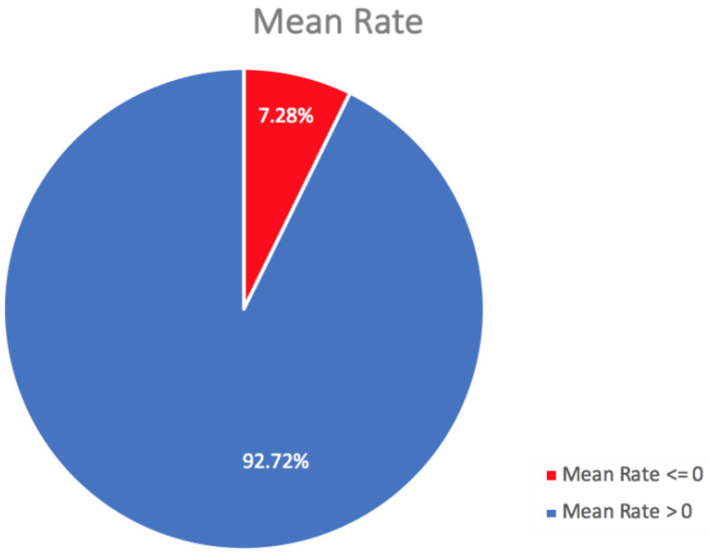


Figure 13: Comparison of Mean Rate fluctuations

5.4.5 Clustering of Coins

Gap statistics is used to determine the elbow point which is optimal number of clusters for the data we are using. We can see that elbow point is at 4 hence optimal number of clusters is 4.

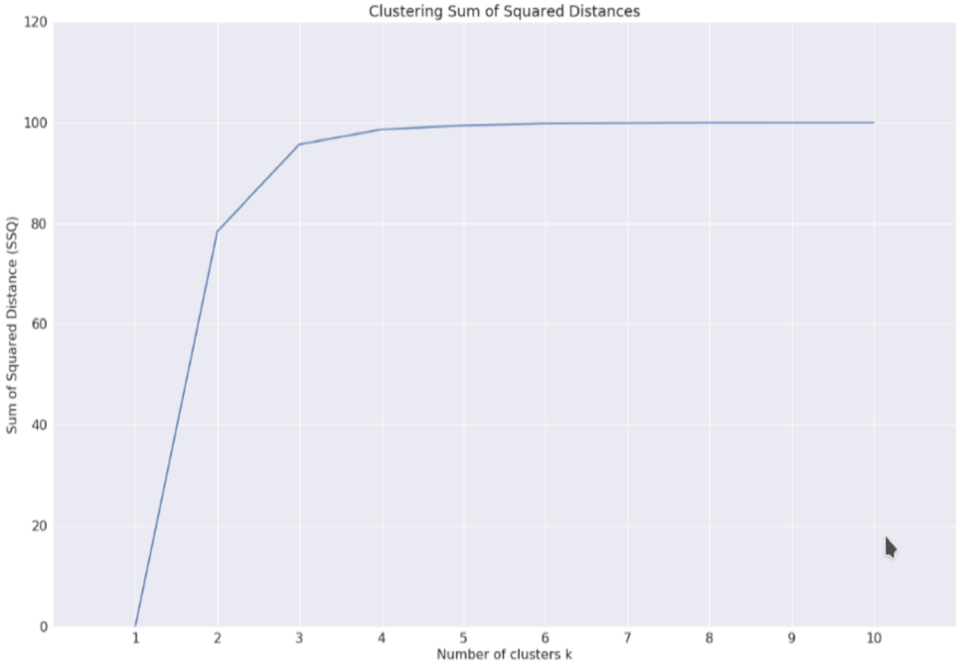


Figure 14: Clustering by sum of squared distances

We have used k means clustering to cluster the coins based on their market cap value and close price using their most recent data. This will provide insight on coins groups based on their worth. we can observer from the graph most of coins worth less and one coin is worth the most.

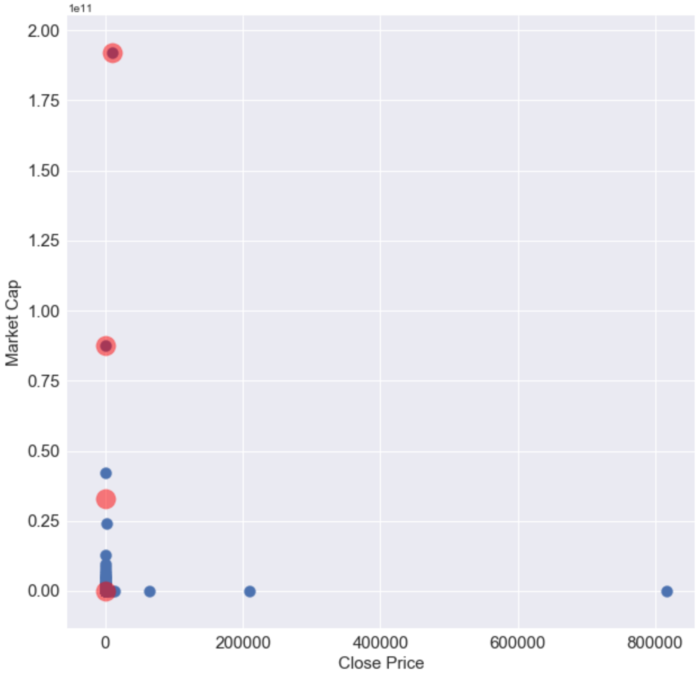


Figure 15: Clustering by Market cap and volume

Similarly, we have clustered using market cap, close price and volume. We can see the graph is similar to the above. Hence volume is directly proportional to market cap and close price.

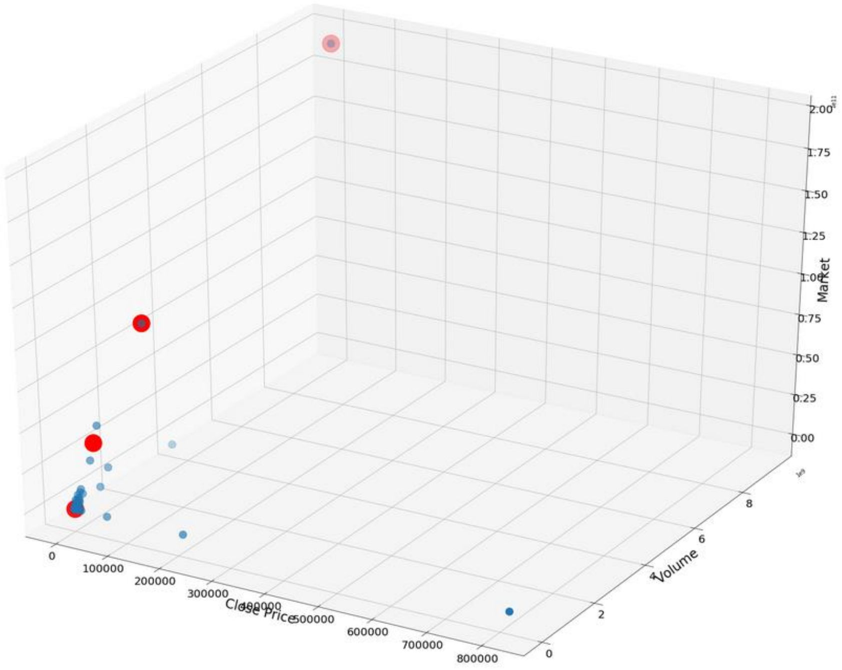


Figure 16: Clustering by Market cap, close price and volume

5.4.6 Seasonality in time Series

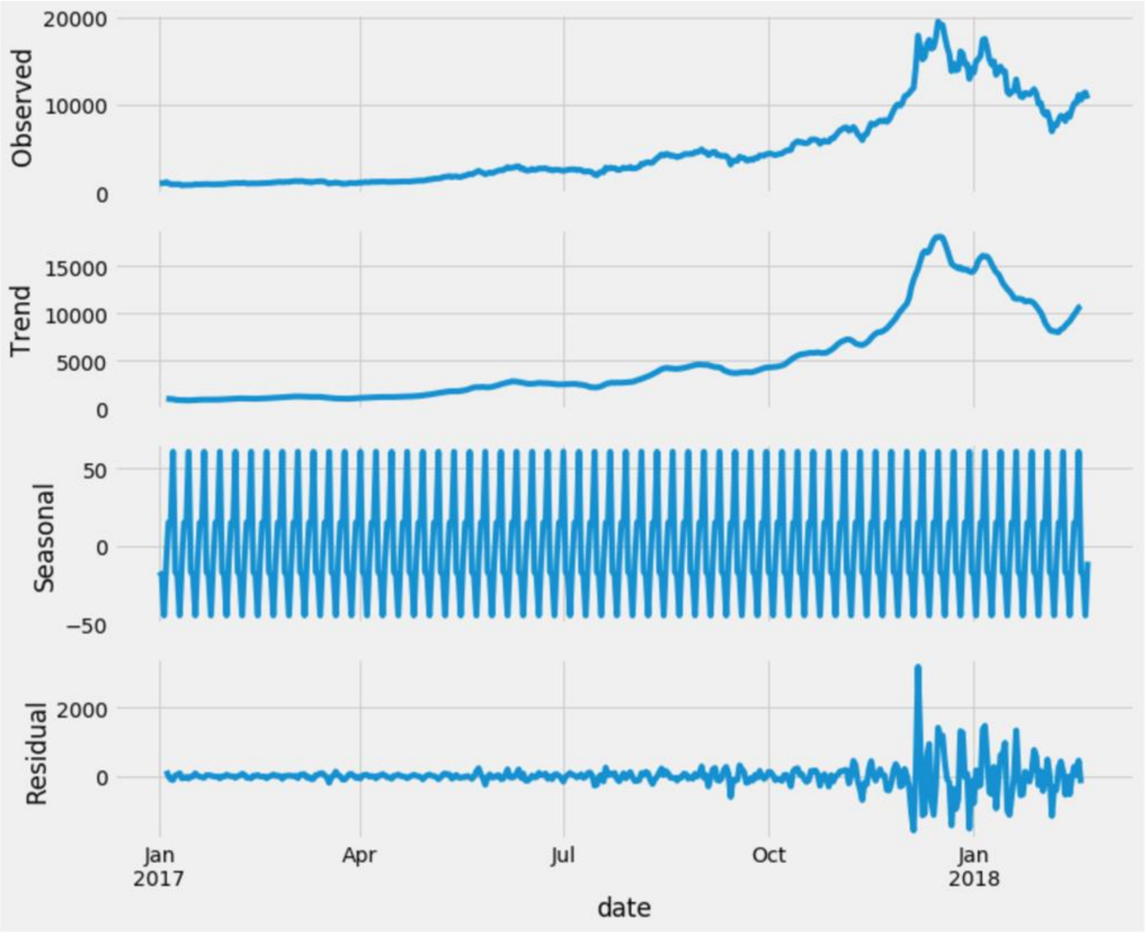


Figure 17: Seasonality of data

ARIMA analysis method requires a series with no trend in it. However, the data set close column values have trend and it requires a transformation in data series if we want to feed it to ARIMA model. Or other option to deal with seasonal series is to use SARIMAX model which implicitly transforms the value based on value of parameter.

5.4.7 Volume Trend of coins

Graphs below shows us volume trend of top 5 coins as per market value. First graph shows daily trend where we can observe that bitcoin's high volume is on Wednesday whereas Ethereum's high volume is on Tuesday and Ripple high volume is on Thursday. This depicts that user make transactions in different coins on different days.

Similarly, monthly, we can see that all 5 coins are not performing well between Feb to Mar end and from Apr the number of transactions is increasing again. While recommending a coin to a user, we can take this information into consideration.

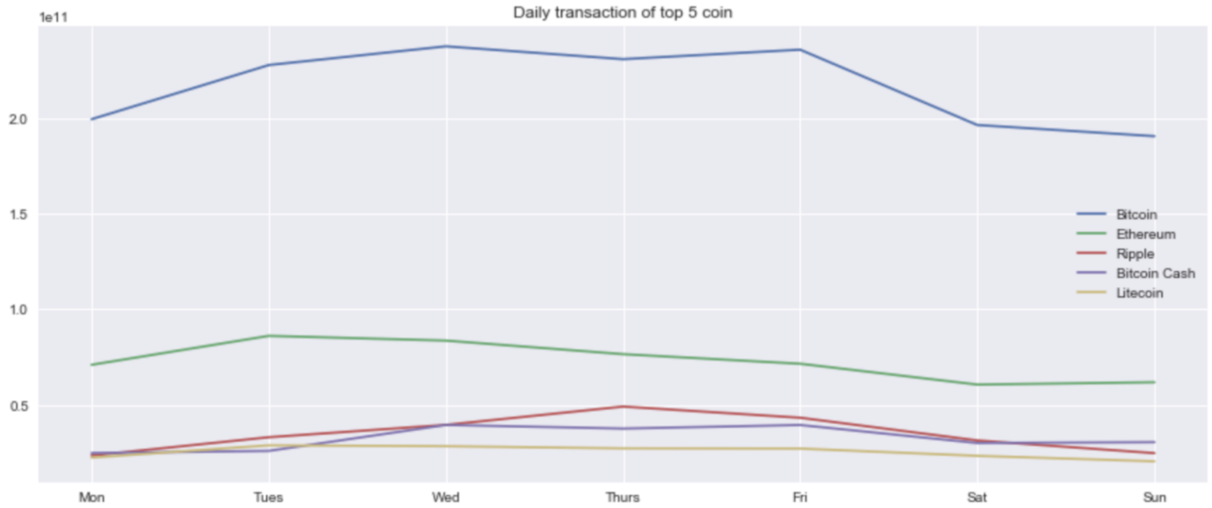


Figure 18: Volume trend of coins (Daily basis)

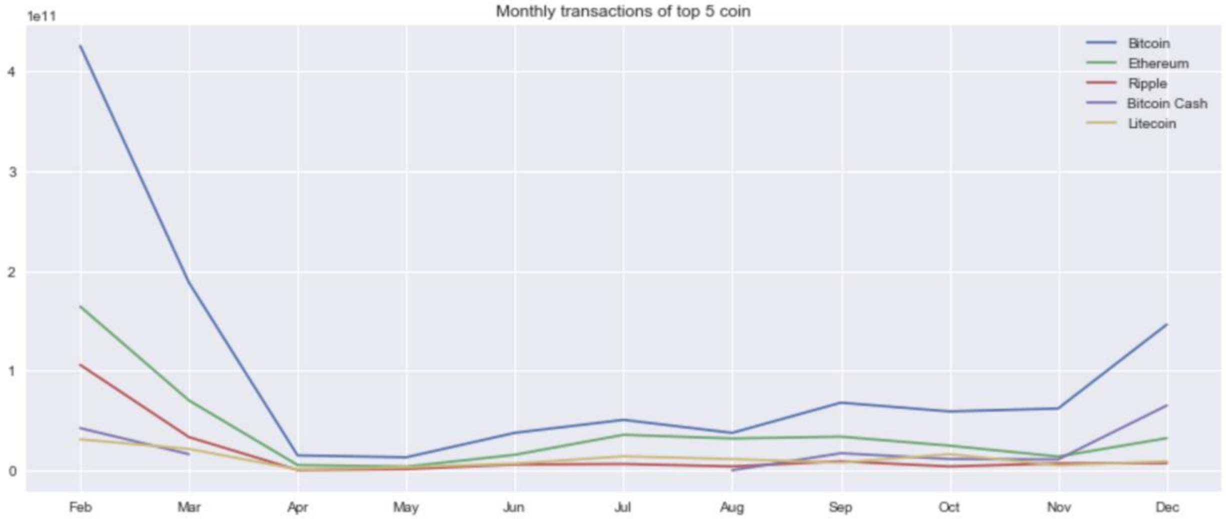


Figure 19: Volume trend of coins (Monthly basis)

5.4.8 Price value trend of coins

Like volume data, we analyze coins data by close values and found that close values is almost constant on daily basis, however, on monthly basis close value decreases which is opposite to our expectations. Figure 21 also shows us that May is the best month to invest in Bitcoin.

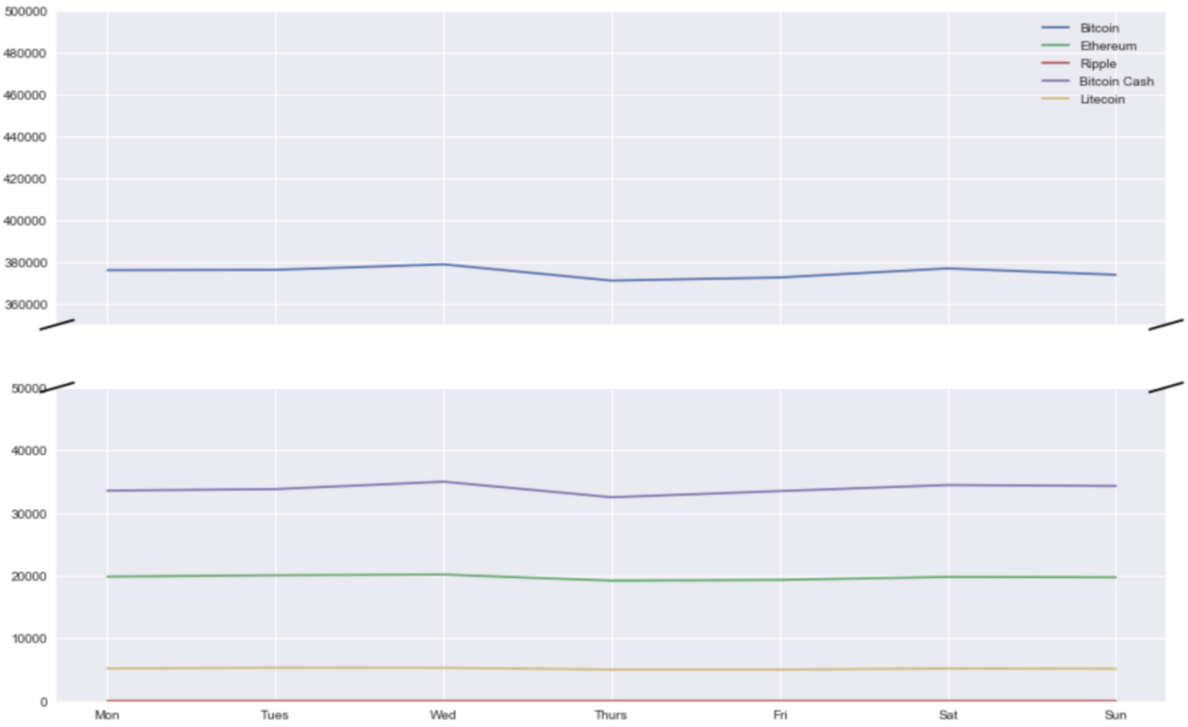


Figure 20: Price value trend of coins (Daily basis)

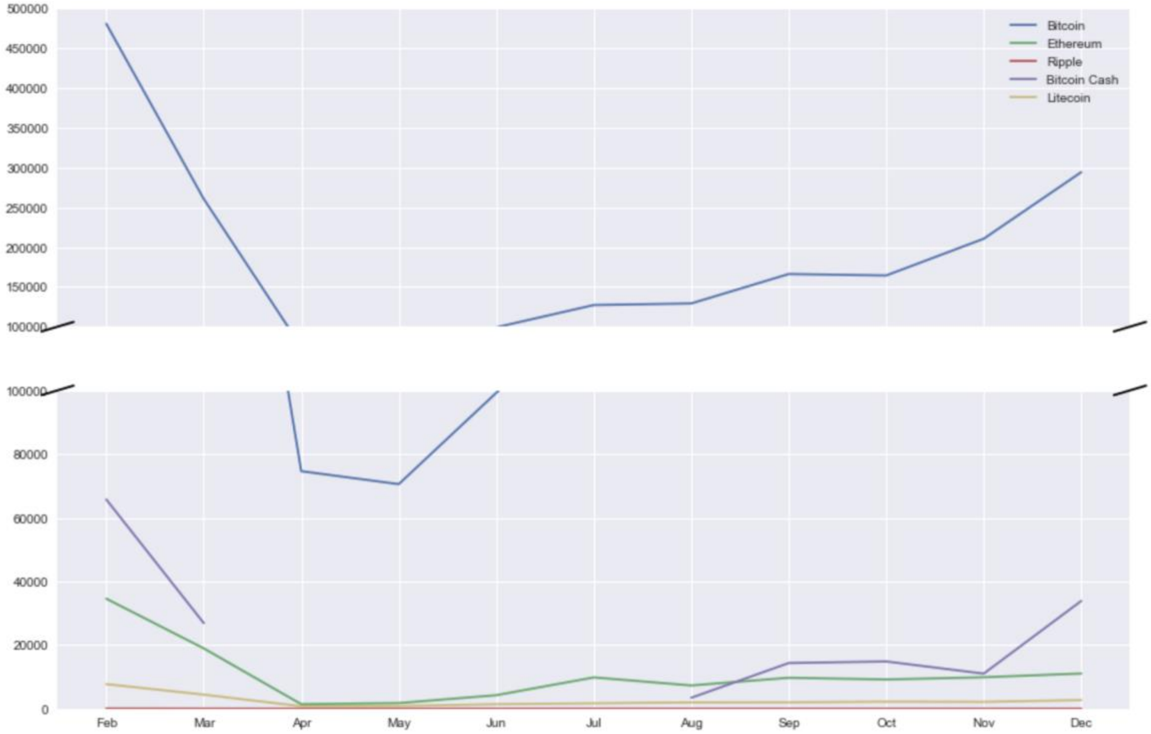


Figure 21: Price value trend of coins (Monthly basis)

5.5 Prediction

5.5.1 ARIMA Method

Figure 22 represents the Stationary Analyses of Time series.

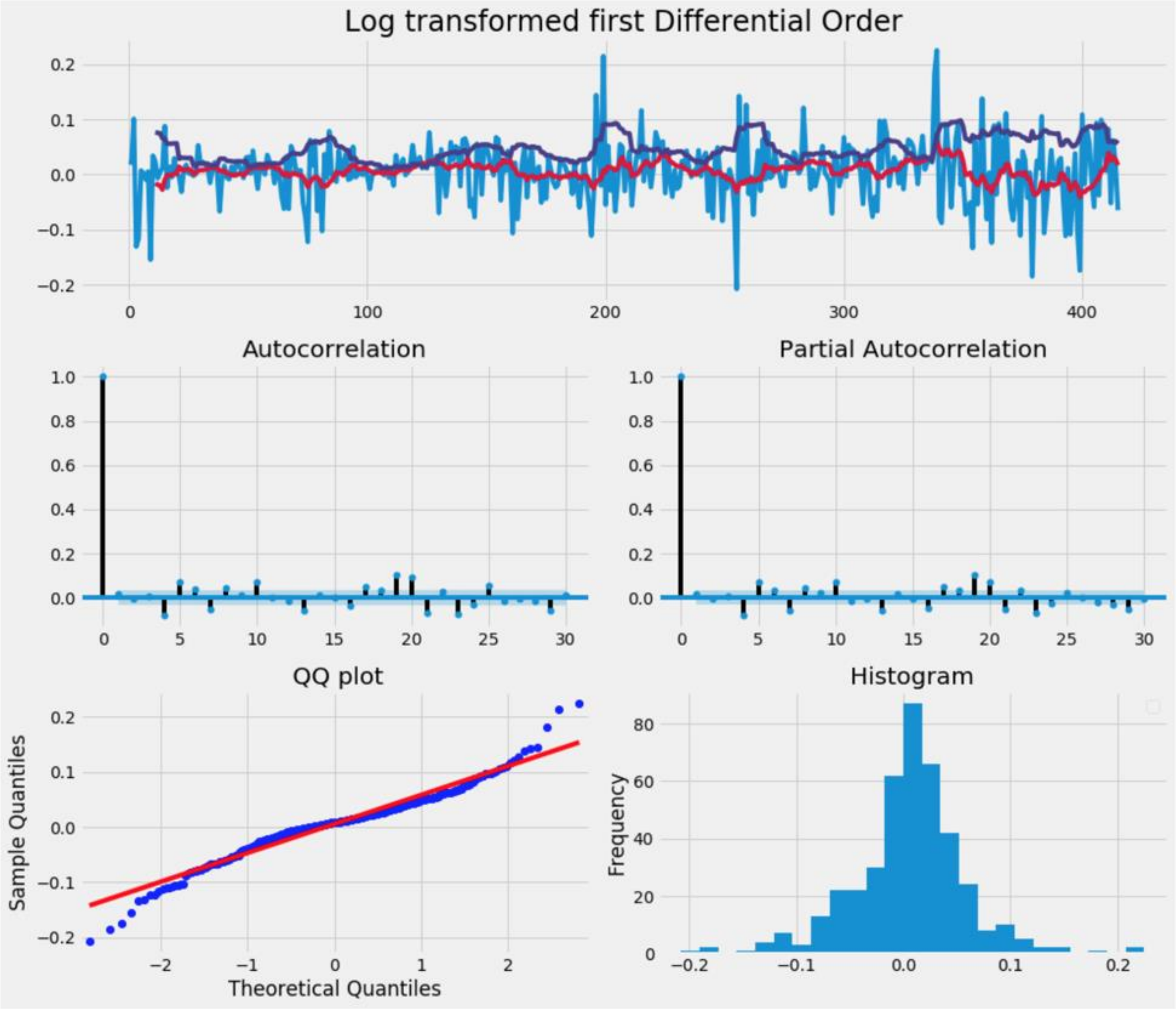


Figure 22: Stationary Analyses of Time series with ARIMA

To remove seasonality from time series we have tried to convert it non-trending series by taking log of each value and differentiating x th value by x+1 th value. This graph shows that 5 subplots. 1st subplot shows transformed series. 2nd and 3rd subplot shows Autocorrelation and Partial Autocorrelation of the transformed series and Histogram shows that the new series depicts binomial distribution characteristics.

There is another method that we have used to deal with the trend data in give time series. We have resampled our original daily data into monthly data series. Now we are using Dickey-Fuller test to check the trend and seasonal value in monthly data series. Figure 23 illustrates the original monthly data series.

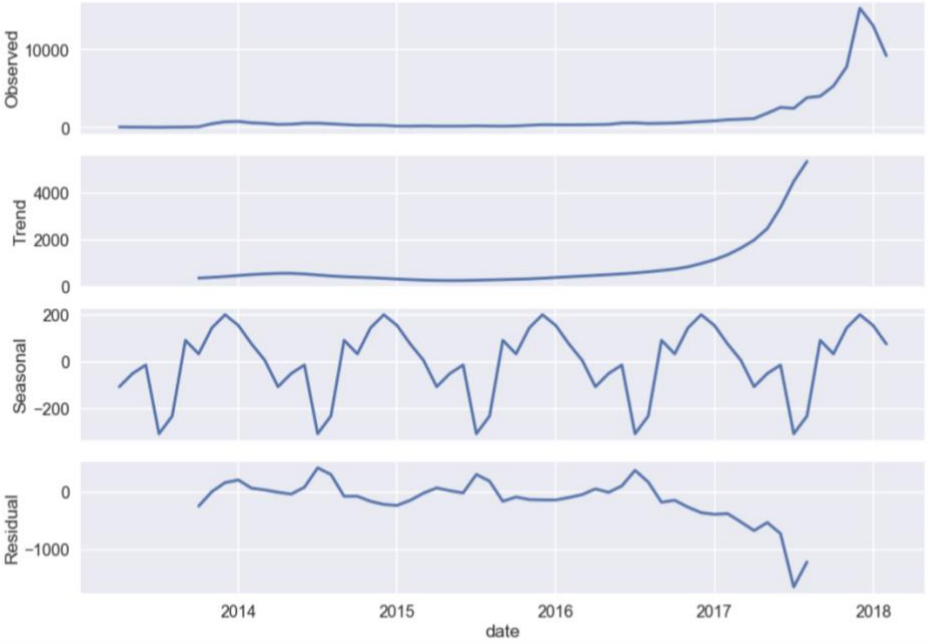
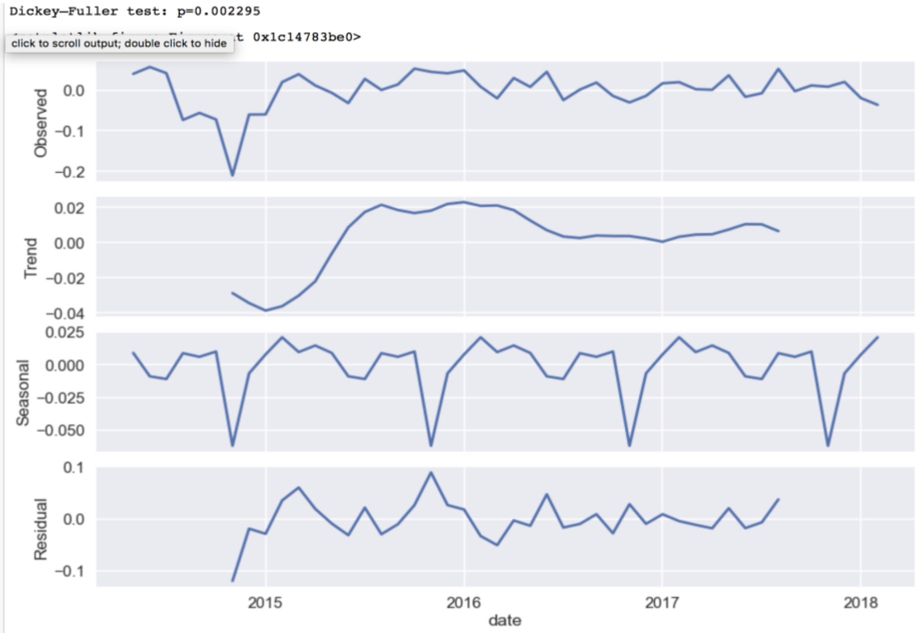


Figure 23: Original monthly data Series

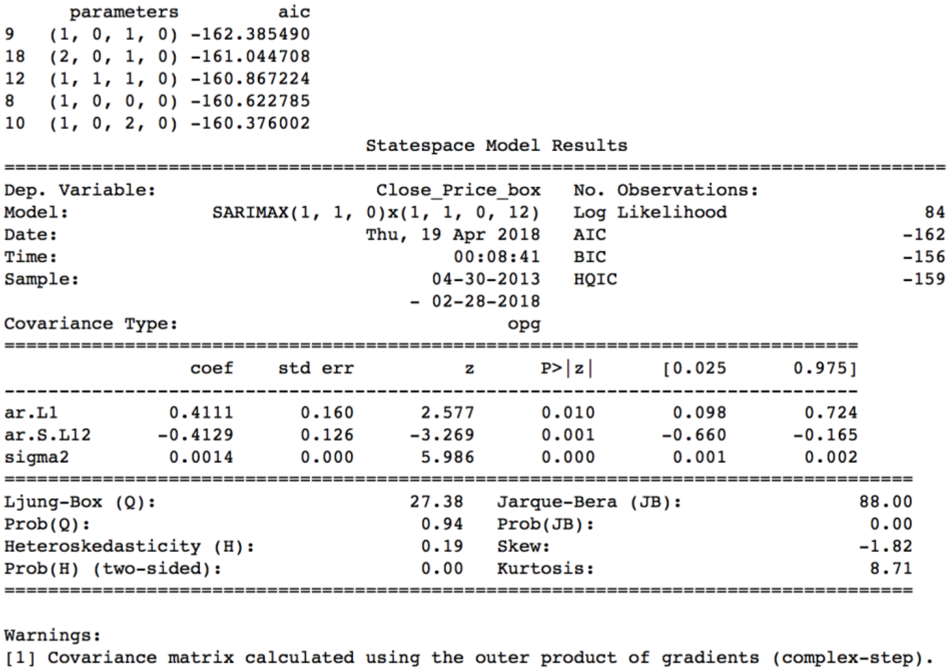
After taking two differentiations, figure 24 shows Dickey Fuller results for transformed values.



```
Dickey—Fuller test: p=0.002295
t 0x1c14783be0>
```

Figure 24: Data after Dickey Fuller transformation

When we run our model on given dataset to get optimized parameters, we get following results which is represented in figure 25.



```
      parameters          aic
9    (1, 0, 1, 0) -162.385490
18   (2, 0, 1, 0) -161.044708
12   (1, 1, 1, 0) -160.867224
8    (1, 0, 0, 0) -160.622785
10   (1, 0, 2, 0) -160.376002
                                 Statespace Model Results
==========================================================================================
Dep. Variable:                    Close_Price_box   No. Observations:                   59
Model:             SARIMAX(1, 1, 0)x(1, 1, 0, 12)   Log Likelihood                  84.193
Date:                            Thu, 19 Apr 2018   AIC                           -162.385
Time:                                    00:08:41   BIC                           -156.153
Sample:                                04-30-2013   HQIC                          -159.953
                                     - 02-28-2018
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4111      0.160      2.577      0.010       0.098       0.724
ar.S.L12      -0.4129      0.126     -3.269      0.001      -0.660      -0.165
sigma2         0.0014      0.000      5.986      0.000       0.001       0.002
===================================================================================
Ljung-Box (Q):                       27.38   Jarque-Bera (JB):                 88.00
Prob(Q):                              0.94   Prob(JB):                          0.00
Heteroskedasticity (H):               0.19   Skew:                             -1.82
Prob(H) (two-sided):                  0.00   Kurtosis:                          8.71
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Figure 25: Results from Dickey Fuller transformation

Model with configuration (1,0,1,0) gave the minimum AIC value. Using this model, the predicted Cryptocurrencies value are shown in Figure 26 and 27. It looks like that the predicted values are just increasing in with constant rate and if we look for few more months, prediction will start decreasing at one point. Next graph is showing that trend.

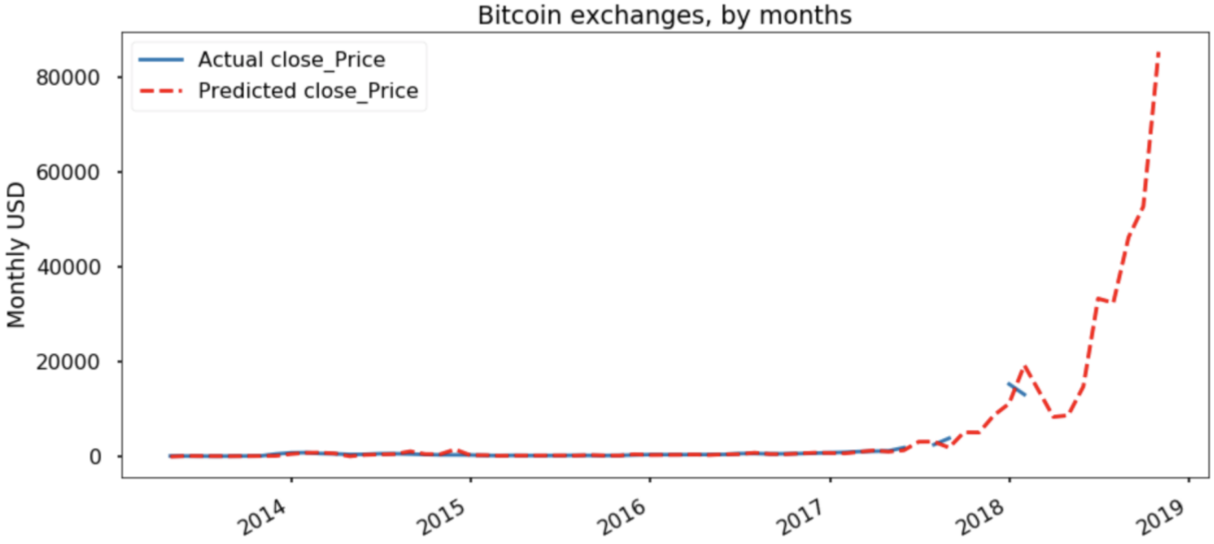


Figure 26: Using ARIMA to predict Bitcoin price

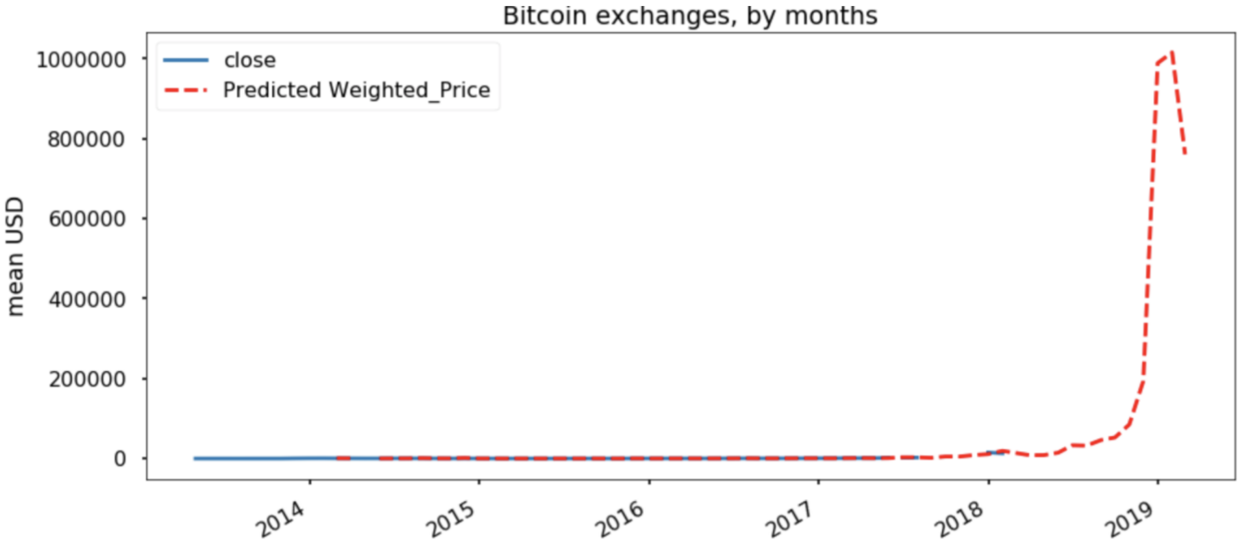


Figure 27: Using ARIMA to predict Bitcoin price. Mean price is shown in this figure

5.5.2 GP Flow Package

GPflow [12, 13] is a package for building Gaussian process models in python, using TensorFlow. We use the Regression model in the package using different types of normalizing and kernels. To use this package, we need to convert the data to zero mean. For normalizing we used two methods.

a. **Normalize Data**

   i. **Log Normalizer:** We take the log values to train the model and generate prediction. The data value a of attribute A is normalized to $\hat{a}$ by computing:

$$\hat{a} = log(a)$$

   ii. **Tanh Estimators:** We use the tanh estimator described in [14], which is a robust and highly efficient method to normalize time series data. The data value a of attribute A is normalized to $\hat{a}$ by computing:

$$\hat{a} = 0.5\left[tanh\left[\frac{0.1(a - \mu)}{\sigma}\right] + 1\right]$$

b. **Kernels**

   The kernels that are used for regression are specified below:

   i. Linear and Constant
   ii. RBF
   iii. Periodic

We used these kernels and also the mixture of the kernels to predict the values. The prediction for Bitcoin is shown using different mixtures of kernels and normalization. We can see from the graphs in figure 28 that the graphs cannot predict the future values (shown in red color) correctly.

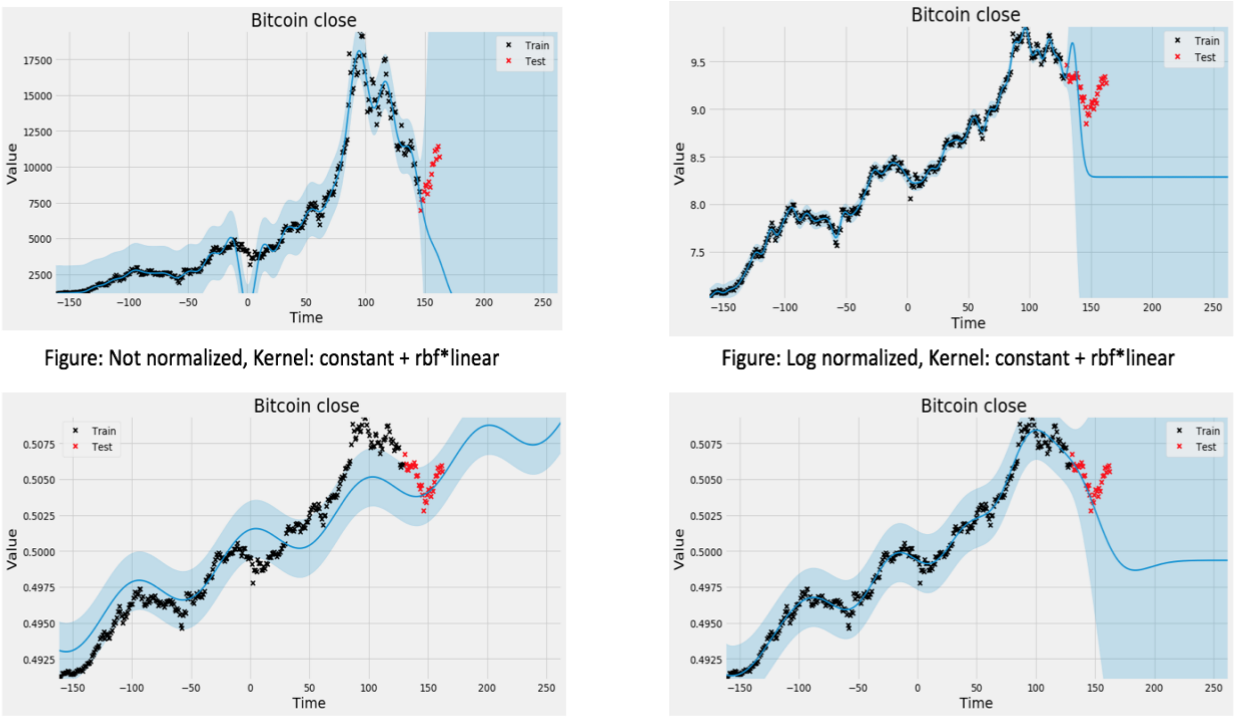


Figure: Not normalized, Kernel: constant + rbf*linear

Figure: Log normalized, Kernel: constant + rbf*linear

Figure: Log and tanh normalized, Kernel: constant + rbf*linear + periodic

Figure: Log and tanh normalized, Kernel: constant + rbf*linear

Figure 28: Prediction for Bitcoin close values using GP flow with different kernels and normalization

5.5.3 LSTM Method

The Long Short-Term Memory(LSTM) network is a recurrent neural network that overcomes the vanishing gradient problem using back propagation.

As such, it can be used to create large recurrent networks that in turn can be used to address difficult sequence problems in machine learning and achieve state-of-the-art results.

A block has components that make it smarter than a classical neuron and a memory for recent sequences. A block contains gates that manage the block’s state and output. A block operates upon an input sequence and each gate within a block uses the sigmoid activation units to control whether they are triggered or not, making the change of state and addition of information flowing through the block conditional.

Implementation

We are using Keras deep learning library to predict the future crypto currency values. LSTMs are sensitive to the scale of the input data, specifically when the sigmoid (default) or tanh activation functions are used. In this implementation we are using tanh activation function. It can be a good practice to rescale the data to the range of 0-to-1, also called normalizing. We can easily normalize the dataset using the MinMaxScaler preprocessing class from the scikit-learn library.

We convert data into two column data set. The first column contains current day value and second column contains next day's value i.e. the value to be predicted.

The LSTM network expects the input data (X) to be provided with a specific array structure in the form of: *[samples, time steps, features].* We can transform the prepared train and test input data into the expected structure using numpy.reshape(). The network has 100 neurons and trained for 25 epochs and batch size of one is use. For predicting we are training the model by taking current prediction into account before predicting the next value.

Once model is trained and values are predicted we inverse transform to reverse the normalization. The error scored are calculated using root mean square error method to evaluate the model's performance.

Close price prediction for BitCoin

Root Mean Square error:
Train Score: 3059.64 RMSE
Test Score: 11129.32 RMSE

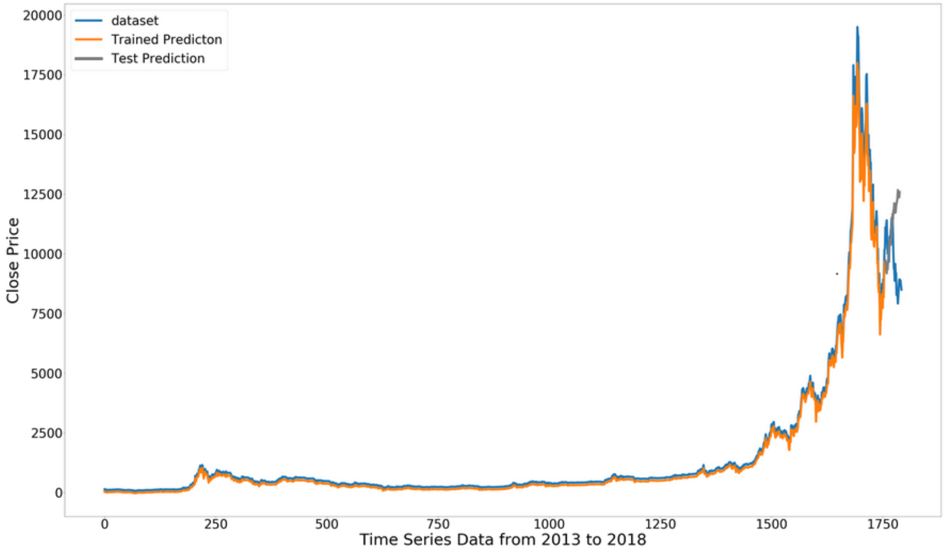


Figure 29: Bitcoin close price prediction by LSTM

Market price prediction for BitCoin:

Root Mean Square error:
Train Score: 53613057137.95 RMSE
Test Score: 200815290625.86 RMSE

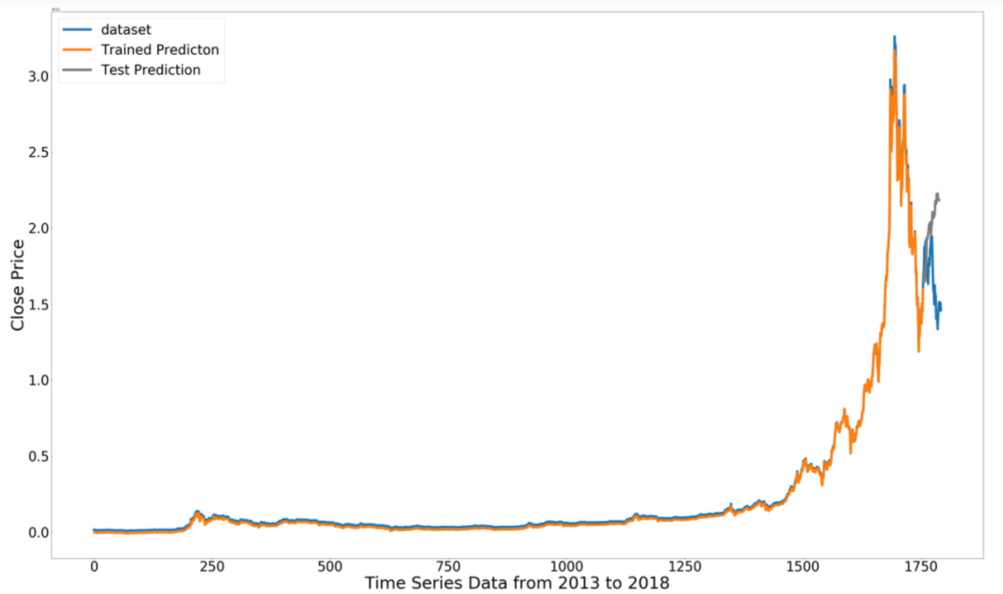


Figure 30: Bitcoin market share prediction by LSTM

LSTM RNNs are implemented in order to estimating the future sequence and predict the trend in the data. It does predict unseen data really well within the range of training data. But outside the boundaries of training data, it does not make the estimation as expected. We can this from our implementation. Further improvement need to be incurred to predict the trend accurately similar to regression techniques.

6 Results

6.1 Descriptive Analysis

As of now, by our descriptive analysis, we have found out top valued cryptocurrencies. We have analyzed following from our analysis.

1) We have calculated the mean, 10 and 90 percentile and outliers of each feature of top 10 and last 10 cryptocurrencies.
2) In the current dataset, out of data for top 50 cryptocurrencies, we just have 22.8% data for top 10 cryptocurrencies.
3) Top 10 cryptocurrencies are Bitcoin, Ethereum, Ripple, BitcoinCash, LiteCoin, Cardano, Neo, Stellar, EOS and Dash.
4) We have compared market value of top 10 cryptocurrencies and found out that Bitcoin's market value is almost double to it's runner up cryptocurrency Ethereum. Similar is the case with Ethereum and its runner up cryptocurrency Ripple. Market value of 5th to 10th position cryptocurrency varies with a difference of 1-2%.
5) We have summed up cryptocurrencies volume on a daily basis and monthly basis and found out following.
   a. Most of the cryptocurrencies transactions happened on Wednesday and Bitcoin transactions are the main contributor this volume.

b. Ethereum transactions are high on Tuesday and for Ripple transactions are high on Thursday.
c. On monthly basis, April and May are lowest in number of transactions, whereas Jan, Feb and Mar are marked highest in number of transactions.

6) We have summed up cryptocurrencies price data on a daily and monthly basis and found out following
   a. For Ripple and Litecoin, price values are almost constant for all days.
   b. For Ethereum, price value on Thursday and Friday are lower than other days
   c. For Bitcoin and Bitcoin Cash, price improves on Wednesday and drops on Thursday.
   d. In case of monthly trend, we have observed that Apr and May are lowest price month and are safe to invest.
7) By applying k Means clustering algorithm, we found out that optimal value of K is 4 and all given coins can be easily clubbed into 4 clusters.
8) The fluctuation results and the Mean values for 4 types of rates shows us the comparison between different coins. We also found out the coins that have decreasing price rates. From that data we the system can avoid those coins from recommending.

6.2 Predictive Analysis

For our predictive analyses, we have tried 3 different algorithms (ARIMA, GP Flow and LSTM) to predict future values for given time series, however, are able to use LSTM for different coins. In case of ARIMA, it's hard to generalize that model for different coins as it requires a lot of manual effort and in case of GP Flow framework, predicted figures were not good at all.

The predicted values for the actual train and test data are shown for the three types of models in Section 5.5 and from the plotting we can see that the prediction is not so good. The reason behind this is the random nature of the price change of the coins. Cryptocurrency prices change so randomly and, in the plotting, we can see the spikes at certain times that has no pattern. As a result there is almost no seasonality in the data, and that is main reason behind the bad performance of the predictive models.

6.3 Problem Solutions

LSTM was easy to tune and generalize for different coins. By following the strategy that we have described in methodology section, here we are providing the results of all 3 different problems.

Problem 1. (Find Day)

Suppose a user want to invested in Ethereum and wants to sell it in next 5 days. The CCReco system detects this as FindDay problem.

Here, the inputs are: day range, D = 5 and coin, C = Ethereum (ETH). The result we get from the system is given below.

Table 2: Rank of days for selling Ethereum in FindDay problem

Rank	Day	Value
1	3	0.020480660745812392
2	1	0.005410711634362619
3	5	-0.010773787187558241
4	4	-0.020319878633982755
5	2	-0.06882641605153517

What we can interpret from here is that, the user should sell Ethereum on the 3rd day. Table 2 shows the values used for ranking, which is the weighted sum of price and market prediction. The negative values say that the rates decreasing on the corresponding days.

Problem 2. (Find Coin)

Now, suppose a user has invested in three different coins: Bitcoin (BTC), Ethereum (ETC) and Ripple s(XRP). If she wants to sell any of the coin on the 6th day, the system detects as the FindCoin problem.

Here, the inputs are: set of coins, X_c = {BTC, ETC, XRP} and d = 6. The result we get from the system is given below.

Table 3: Rank of coins on day 6 in FindCoin problem

Rank	Coin	Value
1	XRP - Ripple	0. 05080423400580009
2	ETH - Ethereum	-0.013498750822835952
3	BTC – BitCoin	-0.02246938638715267

Problem 3. (Find Coin and Day)

Now suppose, the user invested in in three different coins: Bitcoin (BTC), Ethereum (ETC) and Ripple (XRP). And she is willing to sell one or more coins in the next 10 days. In this case, this is detected as the FindCoinAndDay problem.

Here the inputs are: set of coins, X_c = {BTC, ETC, XRP} and day range, D = 10. The result we get from the system is given below.

Table 4: Rank of coins to cell on corresponding day for FindCoinAndDay problem

Rank	Coin	Day
1	XRP - Ripple	10
2	BTC - BitCoin	4
3	BTC - BitCoin	1
4	XRP - Ripple	6
5	BTC - BitCoin	7
6	BTC - BitCoin	9
7	BTC - BitCoin	5
8	BTC - BitCoin	2
9	BTC - BitCoin	8
10	BTC - BitCoin	3

From the result shown in Table 4, we can see that in the next 10 days, the user would be most profited if she sells Ripple on day 10.

7 Discussion

From our descriptive, predictive and solution of the three problems we discuss the following points that arose while the experiments were being done.

7.1 Descriptive Analysis

1) In recent years, we have seen a great jump in market value and volume of top 3 cryptocurrencies where high value of Bitcoin's reached almost 9 times of Ethereum's high value of that period.
2) We can see lots of spikes in top 2 cryptocurrencies trend which shows their volatility whereas 3rd, 4th and 5th top cryptocurrency' graph is more linear which shows their stability.
3) Bitcoin's high and low value over the period of time is always higher than high and low of any other cryptocurrency which makes it more valuable in the market. But due to spikes in the trend, it's more volatile and risky to invest for a short period of time.
4) We noticed fluctuations in prices, volumes and market shares almost for all coins. The patterns in their rates of changes were also negligible.

7.2 Predictive Analysis

1) Cryptocurrency data is extremely random. The seasonality and pattern of data is almost absent which makes the prediction an extremely hard problem. That is why no existing system can predict the prices correctly.
2) We tried with different models using deep learning models and gaussian process models to predict the prices and market shares. The prediction is not what prediction model would be expected to produce, but the it is reasonable at the same time because of the random spike and changes of the data.
3) The cryptocurrency started in this generation and the data that is available is not much in amount. May be in future, with more data, we would be able to find more patterns that would be able to predict the values more precisely.
4) Of the three models ARIMA, GP Flow and LSTM, the performance of LSTM was significantly better, and that is why we used LSTM to predict the prices and market shares for our algorithm.

7.2 Problem Solutions

1) We have run the algorithms on a small subset of the data. We used 3 coins and day range 10 for the experiments.
2) Any number of coins and day range can be given in our system to get the rank.
3) The calculation of the predictions needs significant amount of time. As a result, we pre-calculate the predicted values for each coin and save it to files.

4) When a user gives a query to the system, it only calculates the ranking for the coins and days. This calculation only depends on addition and multiplication; hence, this can be done in real time.
5) There is no current automatic system that recommends a user to sell a cryptocurrency, and that hindered the way of evaluating the results of the recommendation.
6) In future, collecting user data for selling the coins can be a way of evaluation of the recommendation system.

8 Future Work

We can think of better ranking functions that will take more features of the coins to rank them. At this moment we are using weighted sum of predicted price and market share as the ranking function. More features like fluctuations, volumes, price spread, etc. can be taken into consideration to design a better ranking function. In addition to that deep generative models for prediction might work better to predict the values.

9 Conclusion

In this report, we focus on solving three different problems to recommend a user to sell cryptocurrency for better profit. We formulated three problems formally and propose algorithms to solve them. The cryptocurrency dataset is used in the experiments where we expose the nature of the dataset through different aspects. We used advanced techniques to predict the future trend of the data and used them to solve the problems specified. There is no current automatic system that recommends a coin to a user. This system recommends a user from a user specified coin set and day range for making better profit. With more data and better ranking function the system can be perform better which is an aspect of future work that we look forward to.

References

[1] Wayward Artisan, “Cryptocurrency Market Analysis”, www.kaggle.com, January 2018, https://www.kaggle.com/taniaj/cryptocurrency-market-analysis

[2] Hassan Aftab Mughal, “Time Series Analysis [Closing market]”, www.kaggle.com, August 2017, https://www.kaggle.com/hassanaftab/time-series-analysis-closing-market/notebook

[3] David Sheehan, “Predicting Cryptocurrency Prices With Deep Learning”, dashee87.github.io, November 2017, https://dashee87.github.io/deep%20learning/python/predicting-cryptocurrency-prices-with-deep-learning/

[4] Adam Smith, “Bitcoin Trading Strategy Simulation”, www.kaggle.com, June 2017, https://www.kaggle.com/smitad/bitcoin-trading-strategy-simulation

[5] Sachin Shelar, “Bitcoin vs Ethereum (candlestick-chart at the end)”, www.kaggle.com, January 2018, https://www.kaggle.com/shelars1985/bitcoin-vs-ethereum-candlestick-chart-at-the-end

[6] Lorenzo Pagliaro, “Bitcoin in depth analysis”, www.kaggle.com, February 2018, https://www.kaggle.com/lorenzopagliaro01/bitcoin-in-depth-analysis

[7] Nate, “Nate's Cryptocurrency Analysis”, www.kaggle.com, October 2017, https://www.kaggle.com/natehenderson/nate-s-cryptocurrency-analysis/notebook

[8] Wayward Artisan, “Cryptocurrency Predictions with ARIMA”, www.kaggle.com, February 2018, https://www.kaggle.com/taniaj/cryptocurrency-predictions-with-arima

[9] John Young, “10 Statistical Price Predictions for 10 Cryptocurrencies”, medium.com, January 2018, https://medium.com/@spreadstreet/10-statistical-price-predictions-for-10-cryptocurrencies-january-2018-3dcf04bf9d9a

[10] Vijay Vaidyanathan, “Crypto currency analysis”, www.kaggle.com, January 2018, https://www.kaggle.com/vvijay26/crypto-currency-analysis

[11] “Every Cryptocurrency Daily Market Price”, www.kaggle.com, February 2018, https://www.kaggle.com/jessevent/all-crypto-currencies/data

[12] “GPflow Documentation”, http://gpflow.readthedocs.io, http://gpflow.readthedocs.io/en/latest/intro.html

[13] Matthews, Alexander G. de G., et al. "GPflow: A Gaussian process library using TensorFlow." *Journal of Machine Learning Research* 18.40 (2017): 1-6.

[14] Nayak, S. C., B. B. Misra, and H. S. Behera. "Impact of data normalization on stock index forecasting." *Int. J. Comp. Inf. Syst. Ind. Manag. Appl* 6 (2014): 357-369.

[15] Praneeth, “Bitcoin price prediction with ARIMA”, www.kaggle.com, October 2017, https://www.kaggle.com/praneethji/bitcoin-price-prediction-with-arima

[16] Артём, “Bitcoin Price Prediction by ARIMA”, www.kaggle.com, November 2017, https://www.kaggle.com/myonin/bitcoin-price-prediction-by-arima

[17] Wasim, “Trend Prediction with LSTM RNNs using Keras (Tensorflow) in 3 Steps”, https://www.freelancermap.com, April 2017, https://www.freelancermap.com/freelancer-tips/11865-trend-prediction-with-lstm-rnns-using-keras-tensorflow-in-3-steps