Bitcoin Price Prediction using Machine Learning

Siddhi Velankar*, Sakshi Valecha*, Shreya Maji*

*Department of Electronics & Telecommunication, Pune Institute of Computer Technology, Pune, Maharashtra, India

velankar.siddhi@gmail.com, sakshivalecha.96@gmail.com, shreyamaji50@gmail.com

Abstract— **In this paper, we attempt to predict the Bitcoin price accurately taking into consideration various parameters that affect the Bitcoin value. For the first phase of our investigation, we aim to understand and identify daily trends in the Bitcoin market while gaining insight into optimal features surrounding Bitcoin price. Our data set consists of various features relating to the Bitcoin price and payment network over the course of five years, recorded daily. For the second phase of our investigation, using the available information, we will predict the sign of the daily price change with highest possible accuracy.**

Keywords— **Bayesian regression, Bitcoin, Bitcoin prediction, Blockchain, crypto currency, generalized linear model (GLM), machine learning.**

I. INTRODUCTION

A. *Bitcoin:*

Bitcoin is a crypto currency which is used worldwide for digital payment or simply for investment purposes. Bitcoin is decentralized i.e. it is not owned by anyone. Transactions made by Bitcoins are easy as they are not tied to any country. Investment can be done through various marketplaces known as "bitcoin exchanges". These allow people to sell/buy Bitcoins using different currencies. The largest Bitcoin exchange is Mt Gox. Bitcoins are stored in a digital wallet which is basically like a virtual bank account. The record of all the transactions, the timestamp data is stored in a place called Blockchain. Each record in a blockchain is called a block. Each block contains a pointer to a previous block of data. The data on blockchain is encrypted. During transactions the user's name is not revealed, but only their wallet ID is made public.

B. *Prediction:*

The Bitcoin's value varies just like a stock albeit differently. There are a number of algorithms used on stock market data for price prediction. However, the parameters affecting Bitcoin are different. Therefore it is necessary to predict the value of Bitcoin so that correct investment decisions can be made. The price of Bitcoin does not depend on the business events or intervening government unlike stock market. Thus, to predict the value we feel it is necessary to leverage machine learning technology to predict the price of Bitcoin.

II. LITERAURE SURVEY

Bitcoin is a new technology hence currently there are few price prediction models available. [1] deals with daily time series data, 10-minute and 10-second time-interval data. They have created three time series data sets for 30, 60 and 120 minutes followed by performing GLM/Random Forest on the datasets which produces three linear models. These three models are linearly combined to predict the price of Bitcoin.
According to [2] the author is analysing what has been done to predict the U.S. stock market. The conclusion of his work is the mean square error of the prediction network was as large as the standard deviation of the excess return. However, the author is providing evidence that several basic financial and economic factors have predictive power for the market excess return.

In [3], instead of directly forecasting the future price of the stock, the authors predict trend of the stock. The trend can be considered as a pattern. They perform both short term predictions (day or week predictions) and also long-term predictions (months). They found that the latter produced better results with 79% accuracy. Another interesting approach the paper reflects is the performance evaluation criteria of the network. Based on the predicted output the performance evaluation algorithm decides to either buy, sell or hold the stock.

From [4], a comparison between Multi-Layer Perceptron (MLP) and Non-linear autoregressive exogenous (NARX) model is made. They conclude that MLP can also be used for stock market prediction even though it does not outperform NARX model in price prediction. The authors made use of MATLAB's neural network toolbox to build and evaluate the performance of the network.

III. FLOW OF PAPER

The first part of the paper is database collection. We have acquired bitcoin values from two different databases namely: Quandl and CoinmarketCap. After acquiring this time-series data recorded daily for five years at different time instances, it must be normalized and smoothened. For this, we have implemented different normalization techniques like log transformation, z-score normalization, boxcox normalization, and so on. After this, data is smoothened over the complete time period.

The next step is to select parameters that will be fed to the predictive network. From an array of available features, some are mentioned below:

TABLE 1. BITCOIN FEATURES AND THEIR EQUATIONS

S.no.	Features	Equations/Definitions
1.	Block Size	Average block size in MB
2.	Total bitcoins	Total number of bitcoins mined
3.	Day high, day low	Highest and lowest values of different days
4.	Number of transactions	Total number of unique Bitcoin transactions per day
5.	Trade Volume	USD trade volumes from the top exchanges

After feature selection, the sample inputs will be fed to the model. The variation in the bitcoin values can be considered as a pattern. The pattern can be either going up, down or staying within a certain margin of the previous day's price. The next choice that is available is the number of layers and the number of neurons per layer. Hence, the model will perform a pattern aided regression algorithm and artificial neural networks to correctly predict the bitcoin value. The accuracy can be compared with different models after the final prediction.

IV. ONGOING WORK AND ACHIEVED RESULTS

The first step towards Bitcoin prediction is database collection. For our paper we have collected database from the following sources:

A. Quandl

Quandl holds databases related to financial, economic, and social background from over 500 publishers. Data available on Quandl can be used on different platforms such as Python, MATLAB, Maple and Strata. We were able to procure datasets for Bitcoin for up to 5 years of timestamp data with specifications such as –Data high, Data low, Open, Close, volume of transaction, weighted price.

B. CoinMarketCap

CoinMarketCap keeps a track of all the cryptocurrencies available in the market. They keep a record of all the transactions by recording the amount of coins in circulation and the volume of coins traded in the last 24-hours. They continuously update their records as they receive feeds from various cryptocurrency exchanges. CoinMarketCap provides with historical data for Bitcoin price changes.

The next step is database normalization. We basically perform this step to achieve consistency i.e. reduce or eliminate duplicate data, insignificant points and other redundancies. For normalizing our data, we have used five different techniques.

1) ***Log Normalization:*** In this method, the range is compressed and we get the values that were close to zero before normalization. The function is: A'= log(A)/log(max)

2) ***In built MATLAB method***: The function used is 'normc' to normalize database columns. It compresses the range to the best possible extent as compared to other methods.

3) ***Standard deviation normalization***: Here, we take into consideration the difference of every value with respect to the mean value. The advantage of this technique is that we get the negative values as well due to proper compression of the Y axis. The formula is $z = (x - \mu) / \sigma$.

4) ***Z score normalization***: This method uses technique similar to standard deviation method by considering the mean value.

5) ***Boxcox normalization:*** The function used is:-

$$data(\lambda)=(data^{\lambda}-1)/\lambda \quad \ldots.\lambda \text{ is not } = 0$$
$$data(\lambda)=\log(data) \ldots.\lambda \text{ is } = 0$$

The sudden changes in data are observed significantly in this type of normalization, so that the data can be processed more accurately.

Here are the results obtained after implementing various normalization techniques:

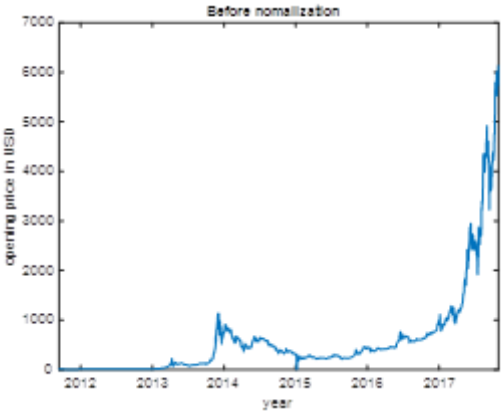

Figure 1. Graph of data before normalisation

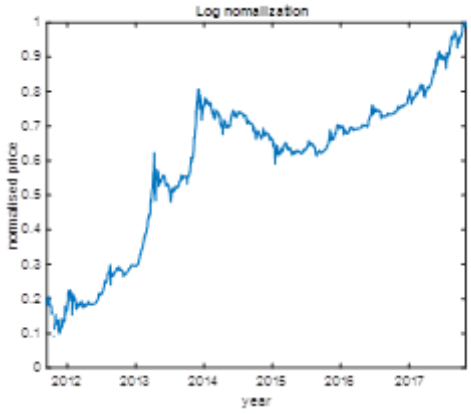

Figure 2. Graph of data using Log Transform normalisation technique

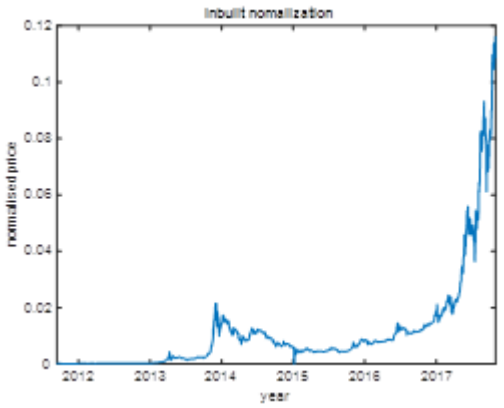

Figure 3. Graph of data after using matlab in-built function

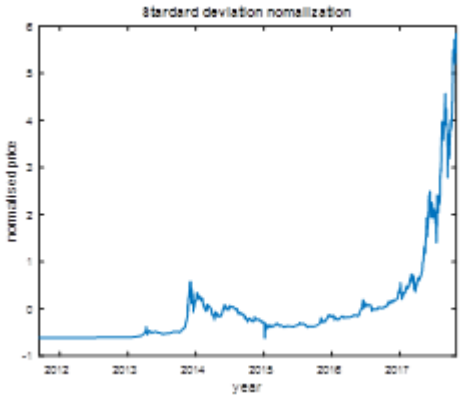

Figure 4. Graph of data after standard deviation normalization technique

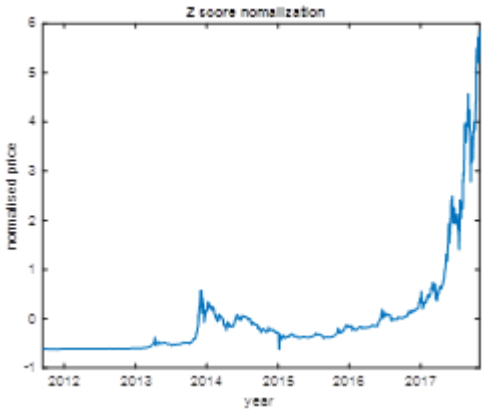

Figure 5. Graph of data after z-score normalization technique

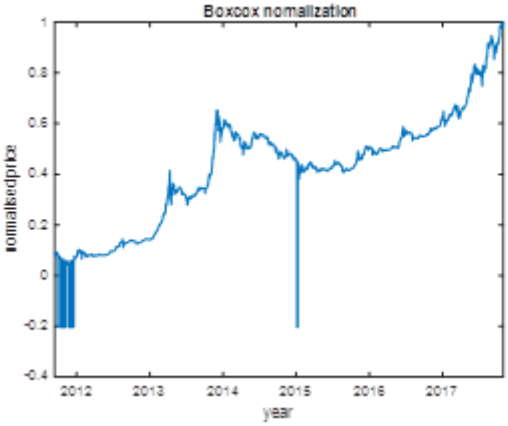

Figure 6. Graph of data after boxcox normalisation technique

V. PROPOSED WORK

A .Bayesian Regression

1. Break the first third of the data into all possible consecutive intervals of sizes 180s, 360s and 720s. Apply k-means clustering to retrieve 100 cluster centers for each interval size, and then use sample Entropy to narrow these down to the 20 best/most varied and hopefully most effective clusters.
2. Use the second set of prices to calculate the corresponding weights of features found using the Bayesian regression method. The regression works as follows –

- At time t, evaluate three vectors of past prices of different time intervals (180s, 360s and 720s).
- For each time interval, calculate the similarity between these vectors and our 20 best kmeans patterns with their known price jump, to find the probabilistic price change dp_i.
- Calculate the weights, w_i for each feature using a Differential Evolution optimization function.

3. The third set of prices is used to evaluate the algorithm, by running the same Bayesian regression to evaluate features, and combining those with the weights calculated in step 2.

B. GLM/Random forest:

1) Construct three-time series data sets for 30, 60, and 120 minutes (180, 360, 720 data points respectively) preceding the current data point at all points in time respectively.
2) Run GLM/Random Forest on each of the two time series data sets separately.
3) We get two separate linear models: M1, M2 corresponding to each of the data sets. From M1, we can predict the price change at t, denoted $\Delta P1$. Similarly, we have $\Delta P2$ for M2.
4) Combine these values to predict the macro price change defined as $\Delta P=W_0+\sum W_j \Delta P_j$, where W_0 is initial market value at t=0, and W_j denotes the weight at the given interval.

5) In addition to using 10-second interval data, we can also use 10-minute interval data to gain a longer-term picture of the price trends.

VI. CONCLUSIONS

After establishing the learning framework and completing the normalization, we intend to use the two methods mentioned above and choose the best method to solve the Bitcoin prediction problem.

ACKNOWLEDGMENT

We would like to thank our project guide Mr. Kaustubh Sakhare for encouraging us to work on this project and our parents for the constant support.

REFERENCES

[1] D. Shah and K. Zhang, "Bayesian regression and Bitcoin," in 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2015, pp. 409-415.

[2] Huisu Jang and Jaewook Lee, "An Empirical Study on Modelling and Prediction of Bitcoin Prices with Bayesian Neural Networks based on Blockchain Information," in IEEE Early Access Articles, 2017, vol. 99, pp. 1-1.

[3] F. Andrade de Oliveira, L. Enrique ZÃ¡rate and M. de Azevedo Reis; C. Neri Nobre, "The use of artificial neural networks in the analysis and prediction of stock prices," in IEEE International Conference on Systems, Man, and Cybernetics, 2011, pp. 2151-2155.

[4] M. Daniela and A. BUTOI, "Data mining on Romanian stock market using neural networks for price prediction". informatica Economica, 17,2013.

AUTHOR BIOGRAPHIES

Siddhi S. Velankar was born in Pune, India on 23rd December 1995. She is currently studying Electronics and Telecommunication engineering in Pune Institute of Computer Technology. She has immense interest in fields like signal processing and machine learning and wishes to do further research in the same. This will be her first IEEE published paper.

Sakshi Valecha was born in Patiala, Punjab, India, in 1996. She is currently pursuing her B.E. from Savitribai Phule Pune University in the field of Electronics and Telecommunication from Pune institute of computer Technology in Pune, Maharashta, India. With your conference Ms. Valecha is attempting to get a place in IEEE publication for the first time.

Shreya Maji, age 21, was born in Pune, India. She has been extremely devoted towards the technical education and is currently studying in final year of the undergraduate engineering course from SPPU, Pune, India. Her area of specialization is electronics and telecommunications. With the current project of Bitcoin Prediction, Ms. Shreya Maji is aiming for a place in IEEE publications for the first time.