International Journal of Engineering & Technology, 7 (4.38) (2018) 1070-1073

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Regression based Analysis for Bitcoin Price Prediction

Azim Muhammad Fahmi, Noor Azah Samsudin, Aida Mustapha, Nazim Razali, Shamsul Kamal Ahmad Khalid

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, 86400, Malaysia
**Corresponding author E-mail: di160049@siswa.uthm.edu.my*

Abstract

In 2017, a significant number of individuals profited from the staggering growth of the price of Bitcoin from $800 USD in January to almost $20,000 USD in December. Because the cryptocurrency market being relatively new when compared to traditional markets such as stocks, foreign exchange, and gold, there is a significant lack of studies in regard to predicting its price behavior. This research is interested in evaluating a number of regression-based algorithms in predicting the price of the Bitcoin (BTC) against United States Dollar (USD). Among the algorithms that will be investigated include the Linear Regression (LR), Neural Network Regression (NNR), Bayesian Linear Regression (BLR), and Boosted Decision Tree Regression (BDTR). By applying such regression-based analysis algorithms, the findings f should further help document the behavior of such a brand new, challenging yet extremely lucrative market.

Keywords: *Bitcoin, Cryptocurrency, Price prediction, Data mining.*

1. Introduction

Bitcoin is simply described as a non-printed form of monetary currency, also known as digital currency held electronically [1]. Such digital currency allows decentralized peer-to-peer network or online transactions carried out by ‘miners’ within a network trading. As a result, there is no centralized authority or any third-party financial institutions that has control of the Bitcoin network. All transactions on the Bitcoin network are embedded in blocks to the open ledger that is known as the blockchain to be verified by the miners using cryptographic proof-of-work.

The Bitcoin blockchain had a split on August 1st2017 that established two blockchains Bitcoin Core (BTC) and Bitcoin Cash (BCH) but for the purpose of this research, the prices of Bitcoin Core (BTC) in US Dollars are taken from the Coindesk BTC Price Index which consists of the average prices from major exchanges. Bitcoin is currently being traded on a global scale over online-exchanges worldwide such as Binance, against dozens of different currency pairings and has a current market capitalization of more than 110 billion dollars with its all-time high reaching more than 300 billion dollars.

Bitcoin has been garnering attention on an unprecedented scale most notable in 2017 when the currency had hit its all-time high price of just below $20,000 in December while starting off the year with a price of $1,000. It is considered an attractive yet highly risky option to investors because of its significant potential when compared to mature financial markets because of the price volatility of Bitcoin which is substantially higher than that of traditional currencies.

This research paper is primarily interested in attempting to predict the price direction of Bitcoin through the use of data mining methods. Bitcoin utilizes blockchain technology to be a type of digital currency commonly referred to as cryptocurrencies. These cryptocurrencies are globally traded by thousands of people from most countries on a myriad of online international exchanges such as Binance and LUNO. There have been a few attempts at predicting the direction of the price of Bitcoin but with varying accuracy and methodologies. For example, the study by [2] achieved an accuracy of 52%, this could be due to lack of exploring other algorithm options. Because of this, the accuracy of past models may still be considered to be low compared to traditional trading markets.

This research is interested in testing new models of regression-based models because of the need to predict continuous numerical values which are the price of Bitcoin (BTC) against United States Dollar (USD). Among the algorithms that will be investigated include the Linear Regression (LR), Neural Network Regression (NNR), Bayesian Linear Regression (BLR), and Boosted Decision Tree Regression (BDTR). This research will also analyse the performance of various data mining algorithms in determining the best accuracy of predicting the price direction. The data will consist of the daily prices of Bitcoin which are available from multiple sources that closely observe the price of the emerging market.

The rest of the paper is organised as follows, the second section will describe the related works on Bitcoin and price prediction. Lastly, the third and final section defines the methodology chosen for the research.

At present, the prices of these cryptocurrencies do not have a significant amount of studies and research as compared to traditional trading markets. However, the number of studies is steadily increasing as the popularity of Bitcoin is surging. [3] used time-series analysis along with sentiment analysis using Support Vector Machines (SVM) to study the relationship of Bitcoin prices and have found that the mining difficulty, number of Wikipedia search queries about Bitcoin, and a sentiment ratio from Twitter are positively correlated with the prices of Bitcoin. This shows that the prices of Bitcoin, although volatile, still has a certain amount of predictability given the right parameters.

In [4], the chosen method was Bayesian regression and its efficacy were tested for predicting price variation of Bitcoin. The paper was said to be able to nearly double their investment in less than a 60-

day period. However, we were not able to replicate the experiment and concluded that original authors had hand-selected 20 patterns observed in their clusters. Given this, it shows the risk of cherry-picking data solely to obtain good results.

In a more recent work, [2] was able to achieve a classification accuracy of 52% through the use of a Long Short-Term Memory (LSTM) network of a Bayesian-optimized recurrent neural network (RNN). The findings of this study are quite closely related to this proposed research as it dealt with the same dataset which were both obtained from the Coindesk website (https://www.coindesk.com/price/bitcoin). More methods and techniques used in studying the behavior and movement of the price of Bitcoin are summarized in Table 1.

Based on the table, it can be ascertained that there exist multiple varying techniques that could be used in order to approach predicting the price of Bitcoin.

2. Materials and Methods

This research will be based on the CRISP-DM model or also known as the ‘Cross-Industry Process for Data Mining’ which is one of the most frequently used analytics models as reported by [15]. Originally proposed by [16], this standard has been in use for a significant period of time to which it is widely regarded to be a very robust, well-proven and overall quite popular model for data analytics. Figure 1 shows the adapted CRISP-DM model.

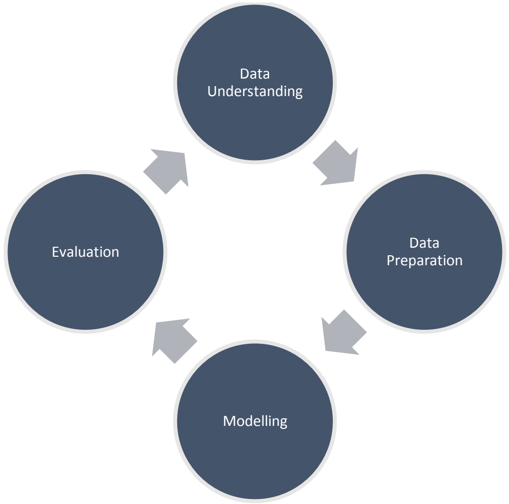


Fig. 1: CRISP-DM Model adapted from [16]

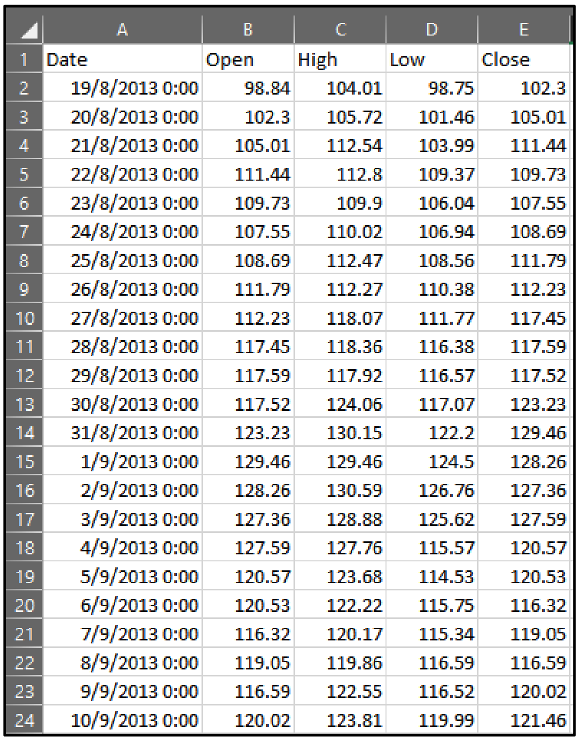
Fig. 1 shows the four phases of CRISP-DM Model: data understanding, data preparation, modelling, and evaluation. The cycle may be repetitive, that is the potential of going through a next iteration of the same cycle with the newly evaluated results from the past iteration. In the data understanding phase, the features of the dataset were considered with reasoning from existing literature. As for data preparation, the dataset is obtained and processed in order to proceed with the modelling phase which would give results that would be evaluated and lead to more understanding of the data and whether improvements could be made to the dataset for a better result.

Table 1: The summary of the related work

Study	Features	Findings
Bayesian Regression and Bitcoin [4]	Prices obtained from Okcoin.com between February 2014 – July 2014	Correlation found
Using Time-Series and Sentiment Analysis to Detect the Determinants of Bitcoin Prices [3]	Mining difficulty, Wikipedia search queries, Sentiment ratio of Bitcoin, Price movement of Bitcoin	89% investment return in 50 days
Predicting the Price of Bitcoin Using Machine Learning [2]	Prices from obtained from CoinDesk.com between the 19th August 2013 – 19th July 2016	Classification accuracy of 52%
Bitcoin Spread Prediction Using Social and Web Search Media [5]	Automated Sentiment Analysis on tweets about Bitcoin	Positive sentiments may contribute to predict the movement of Bitcoin’s price
What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis [6]	Uses wavelet coherence to localize correlations between series and evolution in time and across scales.	Usage in trade, money supply, price level, and interest of investors drives the price of Bitcoin
The Fractal Nature of Bitcoin: Evidence from Wavelet Power Spectra [7]	Continuous wavelet transforms on historical returns of Bitcoin	Clear dominance of specific investment horizons during periods of high volatility
The Predictor Impact of Web Search Media on Bitcoin Trading Volumes [8]	Sentiment analysis on search queries from Google Trends	High cross-correlation value showing Google Trends as a good predictor
Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin [9]	SVM and ANN analysis of the blockchain of Bitcoin	Price direction accuracy of 55% with a regular ANN
Automated Bitcoin Trading Via Machine Learning Algorithms [10]	SVM, GLM/Random Forest on a dataset of 26 features from the Bitcoin Blockchain	50% accuracy in predicting the future price change on a 10-minute interval
Comparative Performance of Machine Learning Algorithms for Cryptocurrency Forecasting [11]	SVM, ANN, DL and Boosted NN analysis on various cryptocurrency prices	SVM obtained highest accuracy and lowest mean absolute percentage error
Predicting Short-Term Bitcoin Price Fluctuations from Buy and Sell Orders [12]	Dataset comprising of realized volatility observations and order book snapshots of different time scales and data types	Ensemble method XGT and regularized regression ENET outperformed other methods in predicting short-term fluctuations
An Empirical Study on Modeling and Prediction of Bitcoin Prices with Bayesian Neural Networks Based on Blockchain Information. [13]	Analyses the time series of Bitcoin by using Bayesian neural networks (BNN)	BNN model succeeded in relatively accurate direction prediction
Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies [14]	Analyses user comments in online communities	Identified comments that significantly influence fluctuations of price

3.1. Data Understanding

One of the collected datasets consists of the daily opening price, highest price, lowest price, and closing price of Bitcoin (BTC) in United State Dollars (USD) from 19/8/2013 until 19/7/2016 sourced from www.coindesk.com. This dataset contains 1,066 days’ worth of data about the price of Bitcoin and was utilized by [2]. In addition, data such as transaction count, on-chain transaction volume, value of created coins, market cap, and exchange volume of the Bitcoin blockchain is obtained from coinmetrics.io. Fig. 2 shows the excerpt of the dataset sourced from coindesk.com.



	A	B	C	D	E
1	Date	Open	High	Low	Close
2	19/8/2013 0:00	98.84	104.01	98.75	102.3
3	20/8/2013 0:00	102.3	105.72	101.46	105.01
4	21/8/2013 0:00	105.01	112.54	103.99	111.44
5	22/8/2013 0:00	111.44	112.8	109.37	109.73
6	23/8/2013 0:00	109.73	109.9	106.04	107.55
7	24/8/2013 0:00	107.55	110.02	106.94	108.69
8	25/8/2013 0:00	108.69	112.47	108.56	111.79
9	26/8/2013 0:00	111.79	112.27	110.38	112.23
10	27/8/2013 0:00	112.23	118.07	111.77	117.45
11	28/8/2013 0:00	117.45	118.36	116.38	117.59
12	29/8/2013 0:00	117.59	117.92	116.57	117.52
13	30/8/2013 0:00	117.52	124.06	117.07	123.23
14	31/8/2013 0:00	123.23	130.15	122.2	129.46
15	1/9/2013 0:00	129.46	129.46	124.5	128.26
16	2/9/2013 0:00	128.26	130.59	126.76	127.36
17	3/9/2013 0:00	127.36	128.88	125.62	127.59
18	4/9/2013 0:00	127.59	127.76	115.57	120.57
19	5/9/2013 0:00	120.57	123.68	114.53	120.53
20	6/9/2013 0:00	120.53	122.22	115.75	116.32
21	7/9/2013 0:00	116.32	120.17	115.34	119.05
22	8/9/2013 0:00	119.05	119.86	116.59	116.59
23	9/9/2013 0:00	116.59	122.55	116.52	120.02
24	10/9/2013 0:00	120.02	123.81	119.99	121.46

Fig. 2: Small sample of the dataset (USD Prices of Bitcoin)

3.2. Data Preparation

The collected data will need to go through cleaning to ensure that there the findings are not skewered by errors such as missing values. The dataset of the daily opening price, highest price, lowest price, and closing price of Bitcoin is formatted in a .CSV file and does not require much cleaning as there were no missing or unreliable entries of the data. However, the data from coinmetrics.io too is in a .CSV file but is in a raw state given it has daily data entries all the way from January 2009 until November 2018. Data cleaning is required as only the entries after April 2013 were there actually data for transaction volume, transaction count, market cap, price in USD, and exchange volume in USD.

The total combination of parameters of the obtained datasets consists of the daily opening price, highest price, lowest price, closing price, transaction volume, adjusted transaction Volume, transaction count, market capitalization, daily average price, exchange volume, generated coins, fees, active addresses, average hash difficulty, payment count, median transaction value, median fee, block size, and block count. Given the number of parameters, significant experimentation and literature review needs to be done in order to determine which parameters would best lead to a higher accuracy in predicting the price direction of Bitcoin.

3.3. Modelling

The four regression-based analysis algorithms used in the experiments include linear regression, neural network regression, Bayesian linear regression and Boosted Decision Tree regression. Note that the conventional regression analysis aims to model a target value based on independent predictions. Such regression-based analysis algorithms are also commonly used in forecasting applications, especially in determining cause and effect between variables or features.

Linear Regression tries to model a relationship between variables, a dependent variable and an explanatory variable, by drawing a linear equation with the data. Equation 1 is the equation for a linear regression line where X is an explanatory variable and Y is a dependent variable and the slope of the line is b with a being the intercept. [17]. The ultimate goal is to figure out the best possible values for a and b, which would provide the best fitting line for a given data points.

$$Y = a + bX \qquad (1)$$

Neural Network Regression uses adaptive weights and are able to approximate non-linear functions about their inputs. It is a feedforward neural network as it responds to an input pattern by processing the input data from one layer to the next with no feedback paths. [18].

Bayesian Linear Regression uses linear regression supplemented by additional information in the form of a prior probability distribution. Prior information about the parameters is combined with a likelihood function to generate estimates for the parameters. [19].

Boosted Decision Tree Regression also known as gradient boosting builds each regression tree in a step-wise fashion, using a predefined loss function to measure the error in each step and correct for it in the next. [20].

The experiments using linear regression, Neural Network regression, Bayesian linear regression, and Boosted Decision Tree regression algorithms are implemented using Microsoft Azure Machine Learning Studio. The dataset consists of the daily opening price is labeled as 'Open', highest price is labeled as 'High, lowest price is labeled as 'Low', and closing price is labeled as 'Close'. The dataset are split randomly, with training set size of 30% and validation set size of 0.70. The data presented shows the daily opening price, highest price, lowest price, and closing price of Bitcoin for 1,066 days.

3.4 Evaluation

The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1 with 1 meaning that it can be predicted without error. Fox et al. (1961) stated that it signifies a way to measure how well the outcomes are replicated by the given model, based on the rate of total variation of outcomes given by the model. The coefficient of determination (R_2) for a linear regression model with one independent variable is shown in Equation 2.

$$R_2 = \left\{ \left(\frac{1}{N}\right) * \sum \frac{(x_i - x) * (y_i - y)}{\sigma_x * \sigma_y} \right\}^2 \qquad (2)$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, x is the mean x value, xy_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

3. Results and Discussions

The purpose of the experiments is to evaluate potential regression-based analysis algorithms available on the Microsoft Azure Machine Learning Studio platform in predicting the price behavior of the Bitcoin. Table 2 shows the results given from running the experiment with Linear Regression (LR), Neural Network Regression (NNR), Bayesian Linear Regression (BLR), and Boosted Decision Tree Regression (BDTR) algorithms.

Table 2: Prediction results

Prediction Algorithm	Price	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Linear Regression (LR)	High	4.876637	9.271067	0.034447	0.002660	0.997340
	Low	5.001090	11.057373	0.037486	0.004408	0.995592
	Open	5.933206	11.64092	0.043219	0.004564	0.995436
	Close	4.095639	8.474145	0.029633	0.002373	0.997627
Neural Network Regression (NNR)	High	9.958728	16.24066	0.070345	0.008162	0.991838
	Low	10.34984	17.098883	0.077578	0.010541	0.989459
	Open	9.959548	14.84553	0.072548	0.007423	0.992577
	Close	9.819348	15.853681	0.071046	0.008306	0.991694
Bayesian Linear Regression (BLR)	High	4.936148	9.290835	0.034867	0.002671	0.997329
	Low	5.054751	11.134975	0.037888	0.004470	0.995530
	Open	5.974957	11.644766	0.043523	0.004567	0.995433
	Close	4.106296	8.470431	0.029710	0.002371	0.997629
Boosted Decision Tree Regression (BDTR)	High	6.335036	11.777895	0.044749	0.004293	0.995707
	Low	7.112280	12.821600	0.053311	0.005927	0.994073
	Open	7.232510	11.964275	0.052683	0.004821	0.995179
	Close	6.052190	11.077272	0.043789	0.004055	0.995945

The results seem to provide a high coefficient of determination with the minimum value being a significantly large 0.989459. The results suggest that the regression-based analysis algorithms could potentially yield more usable results for the Bitcoin price prediction. In predicting the Bitcoin price movement, some other potential features that are not directly tied to price data such as transaction count, on-chain transaction volume, value of mined coins, market cap, and exchange volume of the Bitcoin blockchain can be considered.

4. Conclusions

The cryptocurrency market is a rapidly expanding canvas of trade and investment that has garnered the attention of traders, investors, entrepreneurs on a worldwide scale that is unprecedented in this century. By providing comparative studies and findings from the price data of cryptocurrency markets, it will further help document the behavior and habits of such a lucratively challenging and rapidly expanding market.

In conclusion, this research deals with regression-based analysis algorithms in order to predict the price direction of Bitcoin. The Microsoft Azure Machine Learning Studio were used in conducting the experiments with the Bitcoin dataset. Future work may explore other datasets with more features that could contribute in predicting a more reliable and accurate bitcoin price rate.

Acknowledgement

This project is supported by the Ministry of Higher Education Fundamental Research Grant Scheme,Vot 1609.

References

[1] Nakamoto S (2008), Bitcoin: A peer-to-peer electronic cash system.

[2] McNally S (2018), Predicting the price of Bitcoin using Machine Learning. In: *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing* (PDP), 339-343. IEEE.

[3] Georgoula I, Pournarakis D, Bilanakos C, Sotiropoulos D, Giaglis GM (2015), Using time-series and sentiment analysis to detect the determinants of bitcoin prices.

[4] Shah D, Zhang K (2014), Bayesian regression and Bitcoin. In: 2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton), 409-414. IEEE.

[5] Matta M, Lunesu I, Marchesi M (2015), Bitcoin spread prediction using social and web search media. In: *Proceedings of DeCAT*, 2015.

[6] L. Kristoufek (2015), What are the main drivers of the bitcoin price? Evidence from wavelet coherence analysis, PloS one, 10(4), p. e0123923.

[7] Vidal RD (2014), The fractal nature of bitcoin: Evidence from wavelet power spectra.

[8] Matta M, Lunesu I, Marchesi M (2015), The predictor impact of web search media on bitcoin trading volumes. In: 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), vol. 1, 620-626.

[9] Greaves A, Benjamin A (2015), Using the bitcoin transaction graph to predict the price of bitcoin.

[10] Madan SS, Zhao A (2015), Automated bitcoin trading via machine learning algorithms.

[11] Hitam NA, Ismail AR (2018), Comparative Performance of Machine Learning Algorithms for Cryptocurrency Forecasting. *Indonesian Journal of Electrical Engineering and Computer Science*, 11(3).

[12] Guo T, Antulov-Fantulin N (2018), Predicting short-term bitcoin price fluctuations from buy and sell orders. arXiv preprint arXiv:1802.04065.

[13] Jang, H., & Lee, J. (2018). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *IEEE Access*, 6, 5427-5437.

[14] Kim YB, Jun GK, Wook K, Jae HI, Tae HK, Shin JK, Chang HK (2016), Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one*, 11, no. 8, e0161197.

[15] Brown MS (2015), What IT Needs to Know About the Data Mining Process, available from https://www.forbes.com/sites/metabrown/2015/07/29/what-it-needs-to-know-about-the-data-mining-process/#4ade2420515f

[16] Wirth R, Hipp J (2000), CRISP-DM: Towards a standard process model for data mining.

[17] Yan, X., & Su, X. (2009). Linear regression analysis. Hackensack, N.J.: World Scientific.

[18] Specht, D. F. (1991). A general regression neural network. IEEE transactions on neural networks, 2(6), 568-576.

[19] Minka, T.P. (2009). Bayesian linear regression.

[20] Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802-813.