ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 25, No. 7, pp. 329-334, 2019. 7
https://doi.org/10.5626/KTCP.2019.25.7.329

머신러닝 기술을 사용한 비트코인 특성들의 비트코인 가격에 미친 영향 분석

(Analyzing Impact of Bitcoin Features to Bitcoin Price via Machine Learning Techniques)

윤 성 욱 [†]

(Seongwook Youn)

요 약 2009년에 처음 소개된 비트코인은 전 세계적으로 출시되어있는 암호 화폐들 중 가장 대중적이다. 출시 될 당시에 가치는 아주 낮았고, 다른 암호 화폐처럼 대중적이지도 않았다. 비트코인에 오늘날 많은 사람들이 관심을 가지고 있다. 비트코인은 일반 화폐와 교환될 수도 있고 지불 용도로 사용할 수도 있다. 비트코인은 많은 온라인 상점들과 온라인 서비스에서 통용되고 있다. 암호 화폐는 특정 당국에 의해 규정되지 않고 일반적인 수요와 공급에 따라 규정되지도 않는다. 비트코인은 최근에 상당한 성공을 이루었다. 비트코인 가격이 어느 주요한 요소들에 의해 결정되는지에 대한 호기심이 본 연구를 진행하게 되었다. 가장 대중적인 비트코인 디지털 월렛(www.blockchain.com) 으로부터 Blockchain Wallet API에 기반해서 데이터셋으로부터 특징들을 뽑았다. 머신러닝 기법으로 polynomial regression을 이용하여 특징들의 영향도를 계산하였다. 결과적으로 어떤 특징들이 비트코인 가격 변동과 연관이 깊은지 살펴봤다.

키워드: 비트코인, 블록체인, 머신러닝, 예측분석, 다항 회귀

Abstract Bitcoin (BTC ticker) is the most popular crypto-currency in the world, the first release of which took place in 2009. In this respect, in a release date for this currency the price was equal basically to none and was not considered as popular as other known crypto-currency available at the time. Today bitcoin is in high demand from users around the globe. It can be exchanged through special exchanges for ordinary money, or used directly as a means of payment for anything of choice by the users. Bitcoin is accepted by many of the largest online stores and online services for products, goods and services worldwide. Quotes of crypto currency are not regulated by any legislative or legal authorities, and therefore are considered fluid, whereby the value depends totally on the current natural demand and supply. Recently, bitcoin has achieved a great success in use within open global markets. By being motivated from a review of these factors, we decided to take a deep look into the main factors and features characteristic of bitcoin. In this project, we tried to take a vision regarding the use and impact of bitcoin features from a dataset based on Blockchain Wallet API, which is derived from one of the most popular bitcoin digital wallets – Blockchain Wallet API [1]. As a target, we set different phases for the analysis and gathering of relevant data from API. In these terms, for calculating the impact of feature we have used machine learning techniques; which included a pure linear regression and expended version of a linear regression-polynomial regression.

Keywords: bitcoin, blockchain, machine learning, analysis of prediction, polynomial regression

· 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2017R1A2B1009139)

[†] 정 회 원 : 한국교통대학교 소프트웨어 교수
(Korea Nat'l Univ. of Transportation)
youn@ut.ac.kr

논문접수 : 2019년 3월 5일
(Received 5 March 2019)
논문수정 : 2019년 4월 10일
(Revised 10 April 2019)
심사완료 : 2019년 4월 23일
(Accepted 23 April 2019)

정보과학회 컴퓨팅의 실제 논문지 제25권 제7호(2019. 7)

1. Introduction

Bitcoin is innovational digital currency system based on P2P architecture. Opportunities such as no control organ and centralized bank system give bitcoin whole advantages to exist as a universal and perfect payment system. Bitcoin is an open-source, so anyone who wants can look inside to bitcoin and ones more can participate and contribute bitcoin's net. Due to its unique properties, bitcoin opens up new horizons of possibilities, which have not been provided before by any payment system. As bitcoin is purely peer-to-peer version of electronic cash, it would allow online payments to be sent directly from one party to another without going through a financial institution. Digital signatures provide part of the solution, but the main benefits are lost if a trusted third party is still required to prevent double-spending[2].

Jain et al. forecasted bitcon pricie using web search and social media data. One of the unique features of bitcoin is that its price fluctuation depends mostly on people's opinions instead of institutionalized money regulation. As an example, at November 2017, Warren Buffet tweeted optimistic opinion on cryptocurrency and offered 1 $BTC to everyone who retweeted his post if bitcoin hits $12,500 by the next day[3].

2. Bitcoin, Blockchain and Features

A distributed database called blockchain arose with the advent of bitcoin and the first distributed blockchain was conceptualized by an anonymous person or group known as Satoshi Nakamoto, developer of bitcoin. Blockchain is inextricably linked with bitcoin, because they emerged as an inseparable part of each other. Blockchain is a continuously growing list of records, called blocks, which are linked and secured using cryptography. Each block typically contains a hash pointer as a link to a previous block, a timestamp and transaction data. By occurring transactions some open digital wallets provide API with real-time data about bitcoin and blockchain variables. This variables can vary by numbers but value is identical. Bitcoin price is one of that variables. Each new block includes a cryptographic signature, formed on the basis of the previous one. So the blocks are linked together, forming a "chain of blocks", "blockchain". The chain of blocks can branch out, but in the end, the branch of the block that most of the miners work on is acknowledged. So self-regulation of the network is carried out.

3. Preliminary Work

As an API provider we chose one the most popular and reliable digital wallet for bitcoin – Blockchain Wallet API. Blockchain Wallet API affords values to about 25 features in a JSON format. In the beginning we gathered data with all features via python script to csv file. Data gathering flow is shown in Fig. 1.

From our view, we are going to find some relations between bitcoin price and specific feature or group of them. In this analysis we tried to use different machine learning techniques, traditionally beginning from linear regression. To support prediction rate we also used special case of linear regression – polynomial regression. As bitcoin has specific character, using vector of features is more effective than simple feature variable. In the case of Blockchain Wallet API we can easily assemble data without any api keys. Screenshot of data gathering is shown in Fig. 2.

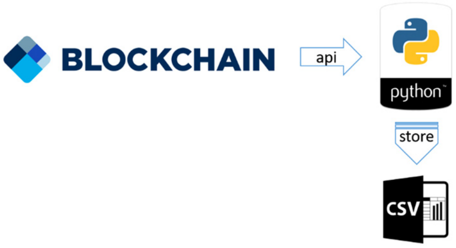
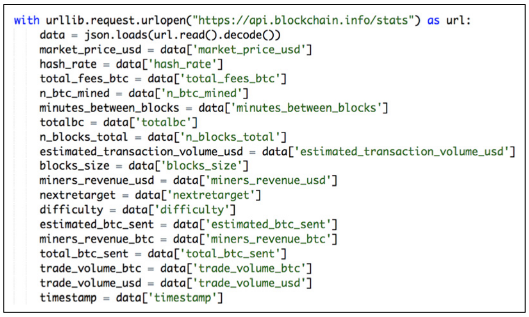


Fig. 1 Main data gathering flow



```python
with urllib.request.urlopen("https://api.blockchain.info/stats") as url:
    data = json.loads(url.read().decode())
    market_price_usd = data['market_price_usd']
    hash_rate = data['hash_rate']
    total_fees_btc = data['total_fees_btc']
    n_btc_mined = data['n_btc_mined']
    minutes_between_blocks = data['minutes_between_blocks']
    totalbc = data['totalbc']
    n_blocks_total = data['n_blocks_total']
    estimated_transaction_volume_usd = data['estimated_transaction_volume_usd']
    blocks_size = data['blocks_size']
    miners_revenue_usd = data['miners_revenue_usd']
    nextretarget = data['nextretarget']
    difficulty = data['difficulty']
    estimated_btc_sent = data['estimated_btc_sent']
    miners_revenue_btc = data['miners_revenue_btc']
    total_btc_sent = data['total_btc_sent']
    trade_volume_btc = data['trade_volume_btc']
    trade_volume_usd = data['trade_volume_usd']
    timestamp = data['timestamp']
```

Fig. 2 Main data gathering source code from a python script

4. Related Work

Quite recently, Shah and Zhang [2] described their application of Bayesian regression to bitcoin price prediction, which achieved high profitability. This work, however, does not explore or disclose the relationship between bitcoin price and other features in the space, such as market capitalization. However, we used short-time dataset. Of course, more long- time dataset can perform more accurately. Modeling the price prediction problem as a binomial classification task, experimenting with a custom algorithm that leverages both random forests and generalized linear models. These results had 50-55% accuracy in predicting the sign of future price change using 10 minute time intervals. This can be likely attributed to the long time interval between data points leading to dampening of the price fluctuation within the actual bitcoin market. Additionally, the higher percentage of true positives compared to true negatives suggests to us that over the longer term bitcoin prices are generally rising [4]. In another work [5, 6], there are analyses of whether social media activity or information extracted by web search media could be helpful and used by investment professionals. There are several works that present predictive relationships between social media and bitcoin price where the relative effects of different social media platforms (Internet forum vs. microblogging) and the dynamics of the resulting relationships, are analyzed using cross-correlation such as [7] or linear regression analysis such as [8] or [9].

5. Data and Methods

5.1 Features

Main aspect in this work remains in features. Quantity of them is about 25, but we will take a look to the main of them. Basically, features from Blockchain Wallet represent data about block summary, transaction summary, mining cost, hash rate and electricity consumption [7]. These features available in a real-time plaintext query api and main of them listed in Table 1.

Features for analysis were selected manually on the basis of our research and views. Data set format for test case and training same as well. Training set and test set rate was captured as 75% to 25%.

5.2 Dataset

As mentioned, dataset is represented as csv local file. In this state handling data can be more elastic and can increase reusability. So JSON format data after editing with python script is stored to csv file as this state (example of one row):

[01/10/2017,15:03:55,4311.62,6307791157.03,9843199108,143750000000,11.72,1659703750000000,487763,651020666.64,105995088,6622364.3,487871,1103400932964,15099191503272,1535,160116699338071,17015.21,73363234.24,1506837518000]

Data from above index with labels from below:

[current_date, current_time, market_price_usd, hash_rate, total_fees_btc, n_btc_mined, minutes_between_blocks, totalbc, n_blocks_total, estimated_transaction_volume_usd, blocks_size, miners_revenue_usd, nextre-

Table 1 Main Features and their Definition

Feature	Definition
Market Price	Real-time price of one bitcoin in USD
Hash Rate	Estimated number of hashes per second
Total Transaction Fees (BTC)	Total value of all transaction fees paid to miners
Blocks Mined	Number of mined blocks
Time between Blocks	Average time between blocks in minutes
Total Bitcoin	Total Bitcoins in circulation (delayed by up to 1 hour)
Total Number of Blocks	Total number of mined blocks
Estimated transaction Volume	Total estimated value of transactions on the blockchain
Blocks size	The 24 hour average block size
Miners Revenue (USD)	Total value of coinbase block rewards and fees paid (USD)
Difficulty	Current difficulty target as a decimal number
Miners Revenue (BTC)	Total value of coinbase block rewards and fees paid (BTC)
Total Sent Bitcoins	Total number of sent Bitcoins over network performance
Trade Volume (BTC)	Total BTC value of trading volume on bitcoin exchanges
Trade Volume (USD)	Total USD value of trading volume on bitcoin exchanges

target, difficulty, estimated_btc_sent, miners_revenue_btc, total_btc_sent, trade_volume_btc, trade_volume_usd, timestamp]

For getting feel, we can take a look to dataset from time-series representation:

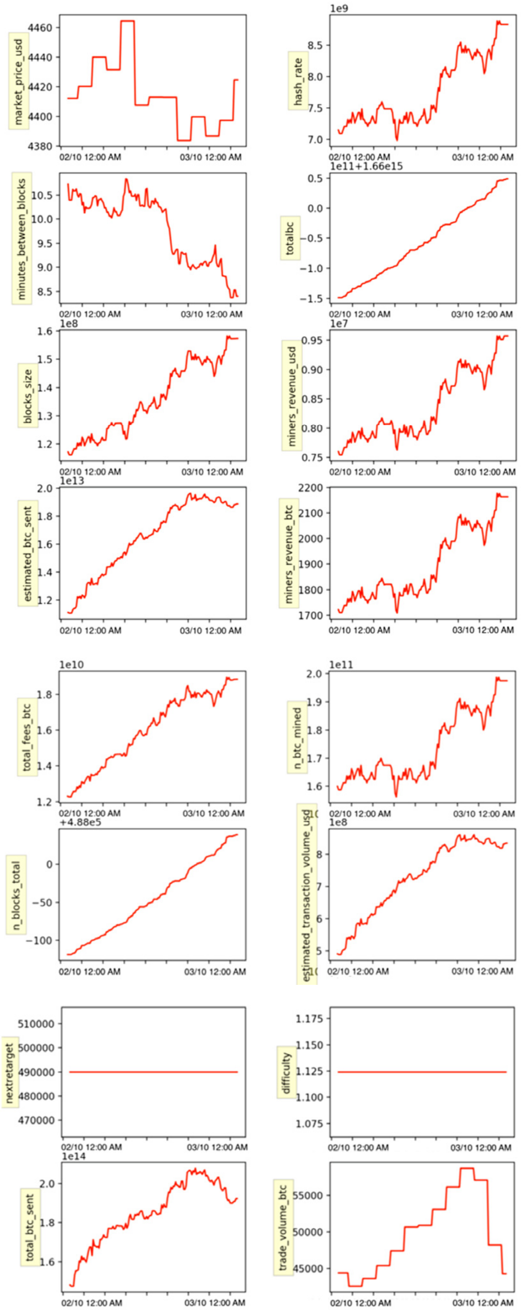


Fig. 3 One day time-series representation of a dataset

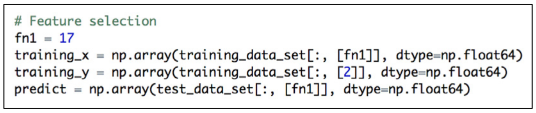
From Fig. 3, we can find out that some features take constant nature. Some features are increasing naturally by their aspect. As the saying goes, garbage in, garbage out. However, the dataset contains enough relevant features and some irrelevant ones. Before entering to the main part, we did feature engineering.

5.3 Experiment

As mentioned, for advantage results we used special case of linear regression, polynomial regression. Main idea is how you select your features. Looking at the multivariate regression with 2 variables (x1 and x2). Linear regression will look like below equation.

$$y = a1 * x1 + a2 + x2$$

In linear regression as one variable we are using time and as another one of our features. The main work is fitting training data. Some values of features are big. So while fitting we can specify the type. In the case of pure linear regression we can fit array of values from single feature as Fig. 4.



```
# Feature selection
fn1 = 17
training_x = np.array(training_data_set[:, [fn1]], dtype=np.float64)
training_y = np.array(training_data_set[:, [2]], dtype=np.float64)
predict = np.array(test_data_set[:, [fn1]], dtype=np.float64)
```

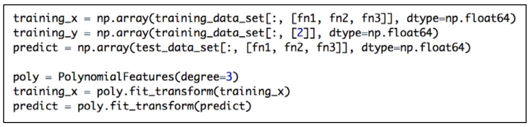
Fig. 4 Training data preparing by using array object for a single feature

As we found out from first results, with this expression we couldn't get any valuable prediction. So here polynomial regression gave breakthrough. We added additional features, so we got more complex "linear regression". That is the two-degree polynomial regression as below equation.

$$y = a1 * x1 + a2 * x2 + a3 * x1 * x2 + a4 * x1^2 + a5 * x2^2$$

This nicely shows an important concept curse of dimensionality, because the number of new features grows much faster than linearly with the growth of degree of polynomial. As we had more variables, we need feature engineering step for polynomial regression. We set degree value same as our variables quantity like Fig. 5.

As features we tried many variants of features pairs by research and by their aspect. The most valuable result we got from "Miners' revenue" and "Block height of the next difficulty retarget".



```
training_x = np.array(training_data_set[:, [fn1, fn2, fn3]], dtype=np.float64)
training_y = np.array(training_data_set[:, [2]], dtype=np.float64)
predict = np.array(test_data_set[:, [fn1, fn2, fn3]], dtype=np.float64)

poly = PolynomialFeatures(degree=3)
training_x = poly.fit_transform(training_x)
predict = poly.fit_transform(predict)
```

Fig. 5 Training data preparing by using polynomial feature degree transforming

In summary, we extract about 25 features in a JSON format using Blockchain Wallet API. JSON data is transformed into csv file using python script. After some preprocessing, we can select reliable features from time-series representation of dataset. Finally, we can predict blockchain price by applying linear regression and polynomial regression model.

6. Results

We occurred some prediction results by fitting several features to our regression models to predict. In fact, as expected bitcoin prediction price results of simple linear regression were non-reliable as well. As the matter of fact, bitcoin price grows swiftly, prediction outcomes are in Fig. 6. Red colored line is real price, and blue colored line is predicted price.

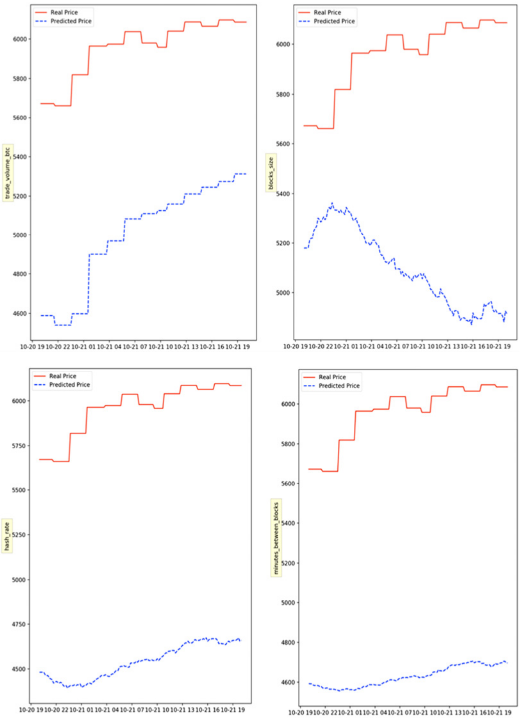


Fig. 6 Linear Regression with the most reliable features

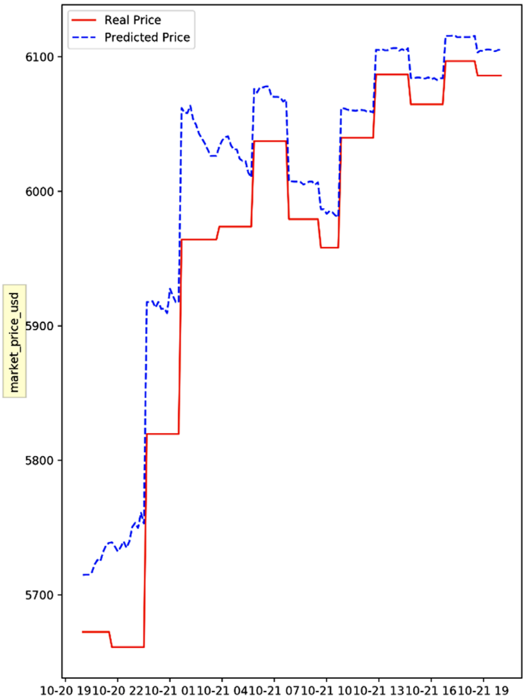


Fig. 7 Polynomial Regression result plot

A statistical measure of how close the data are to the fitted regression line, R-squared score is about 86%. As a first phase we can impact these two features. We can see that transaction factors didn't give better results. We can tell from the graph that the bitcoin's bubble burst was predictable. The result is shown in Fig. 7.

7. Conclusion

In this paper, we studied feature impaction to bitcoin price. This impaction value we occurred through prediction of bitcoin price by using Machine Learning Techniques, such as linear regression and polynomial regression. We used regression to determine an equation that fits that data, and how well it fits. Obviously none of these things are surefire ways to make correct predictions, and maybe they wouldn't even be that useful. How under the circumstances, 2-degree polynomial regression performed better results. Variables that we user are "Miners' revenue" and "Block height of the next difficulty retarget". Surprisingly, these two variables fitted regression model more preferable. Main role was in side of "Miners' revenue". Due to the fact, that a single retarget never changes the target by more than a factor of 4 either way to prevent large changes in difficulty.

8. Future Work

To improve upon our results in the future, we can use more long-time dataset. However, there are no open dataset with the same feature collection. So we need gather data by ourselves. Then, we need more accurate feature engineering for bitcoin price prediction, where we can use leading Machine Learning techniques. Additionally, if we can consider other social media data like google trend to predict bitcoin price, we can get more accurate result which is more influencing to the bitcoin price. Dataset from Blcokchain Wallet API Advanced techniques require their own format to input dataset. By this way, we can additionally perform more breakthrough techniques, such as k-nearest clustering or even practice with recurrent neural network (RNN), a class of artificial neural network. Also, to predict bitcoin price well we need to know crypto-currency market, crypto-currency related technologies, etc. Bitcoin miners give us a preview of the future of planet-scale computing[10].

References

[1] Blockchain Wallet API [Online]. Available: http://www.blockchain.com/explorer

[2] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008.

[3] R. Jain, R. Nguyen, L. Tang, and T. Miller, "Bitcoin Price Forecasting using Web Search and Social Media Data," *Proc. of the SAS Global 2018 Conference, 2018*. [Online]. Available: https://www.sas.com/content/dam/SAS/support/en/sas-globalforum-proceedings/2018/3601-2018.pdf

[4] D. Shah, and K. Zhang, "Bayesian regression and Bitcoin" [Online]. Available: https://arxiv.org/pdf/1410.1231.pdf

[5] I. Madan, S. Saluja, and Aojia Zhao, "Automated Bitcoin Trading via Machine Learning Algorithms" [Online]. Available: http://cs229.stanford.edu/proj2014/Isaac%20Madan,%20Shaurya%20Saluja,%20Aojia%20Zhao,Automated%20Bitcoin%20Trading%20via%20Machine%20Learning%20Algorithms.pdf

[6] A. Mittal, and A. Goel, "Stock Prediction Using Twitter Sentiment Analysis," [Online]. Available at http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

[7] Blockchain Info Stats. [Online]. Available: https://www.blockchain.com/stats?

[8] Social media and markets: The New Frontier, [Online]. Available: https://www.hedgechatter.com/wp-content/uploads/2014/09/GNIP-SocialMedia-and-Markets-TheNewFrontier.pdf

[9] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, Vol. 2. No. 1, pp. 1-8, 2011.

[10] P. Treleaven, R. Brown, and D. Yang, "Blockchain Technology in Finance," *IEEE Computer*, Vol. 50, No. 9, pp. 14-17, 2017.

윤 성 욱

1997년 서강대학교 컴퓨터공학과(학사) 2002년 M.S. Electrical Engineering, University of Southern California. 2009년 Ph.D. Computer Science, University of Southern California. 2012년~2013년 LG전자 CTO연구원. 2015년 9월~현재 한국교통대학교 소프트웨어전공 교수. 관심분야는 데이터 분석 및 예측, 온톨로지, 센서 네트워크