다음 블록에 포함되는 비트코인 트랜잭션 예측 연구

고경찬 O*, 이채현 *, 우종수**, 홍원기*

*포항공과대학교 컴퓨터공학과

**포항공과대학교 정보통신대학원

{kkc90, chlee0211, woojs, jwkhong}@postech.ac.kr

A Study on Predicting Bitcoin Transaction Included in the Next Block

Kyungchan Ko O*, ChaeHyeon Lee*, Jongsu Woo**, James Won-Ki Hong*

*Department of Computer Science and Engineering, POSTECH

**Graduate School of Information Technology, POSTECH

요 약

데이터의 양이 많아지고 분석 기술이 발달하면서 과거의 축적된 데이터를 분석하여 미래에 일어날 일들을 예측하려는 시도들이 생겨나고 있다. 비트코인은 블록체인 기술을 기반으로 최초로 구현된 암호화폐이다. 이 암호화폐는 투명성이 적용된 공개 분산 원장이기 때문에, 누구나 비트코인 데이터에 접근하여 활용할 수 있다. 따라서 비트코인에서 생성된 과거 데이터를 분석하여 비트코인 플랫폼에서 향후에 발생할 이벤트 예측을 시도할 수 있다. 비트코인에서 트랜잭션이 블록에 포함된다는 것은 블록체인 원장에 저장된다는 중요한 의미를 가지고 있지만, 일반 사용자들은 어떤 트랜잭션이 다음 블록에 포함될 것인지 알 수 없다. 본 연구에서는 블록에 포함되는 비트코인 트랜잭션들의 특징을 분석하여 다음에 생성될 블록에 포함될 트랜잭션을 예측하는 실험의 결과를 보여준다. 더 나아가 포함될 만한 트랜잭션이 블록에 포함되지 않는 경우를 조사한다.

I. 서 론

비트코인 [1]은 P2P(peer-to-peer) 네트워크 구조로 구축되어 탈중앙화방식으로 운영되는 전자결제 시스템이다. 비트코인 시스템의 사용자들은 비트코인(화폐)을 다른 사용자에게 전달하기 위해서 이를 위한 트랜잭션(Transaction)을 만들어서, 자신과 연결된 피어(Peer)들에게 브로드캐스트(Broadcast) 한다. 이 트랜잭션을 수신한 피어는 트랜잭션 검증을 수행한 후 그와 연결된 다른 피어들에게 트랜잭션을 전파한다. 이렇게 P2P 네트워크에서 메시지를 전파하는 방식으로 트랜잭션은 비트코인 네트워크 전역으로 전달되게 된다. 하지만 P2P 방식으로 데이터가 전파되면 전파 지연(Propagation delay) 때문에 각 노드(Node)들이 트랜잭션을 수신하는 시간이 달라지며 심지어 중간에 손실되어 트랜잭션을 수신하지 못할 수도 있다. [2]

비트코인 네트워크의 각 노드들이 보유하는 원장은 모두 동일 해야하며, 이는 원장에 저장될 트랜잭션 순서의 획일화를 의미한다. 비트코인 시스템은 동일한 원장을 유지하기 위해서 작업증명(PoW, Proof of Work) [3]라는 합의 알고리즘을 사용한다. 작업증명은 계산 집약적인 연산을 통해서 가장 빠르게 문제의 정답을 찾은 노드에게 블록을 생성할 권한을 부여하는 것이다. 즉, 노드들이 각자 다른 트랜잭션 셋을 블록에 포함시키고 블록헤더 안에 있는 Nonce 값을 조정하면서 연속적으로 블록의 해시값을 계산하는데, Target value 보다 낮은 해시값을 가장 먼저 찾은 노드의 블록을 새로운 블록으로 인정하는 것이다. 작업증명을 수행하는 것을 마이닝(Mining)이라고 부르며 계산 집약적인 연산을 수행하기 위해서 컴퓨팅 리소스가 필요하고 많은 전력이 소모된다. 때문에 해당 작업의 동기를 부여하기 위해서 블록을 생성할 때 인센티브로서 비트코인(블록보상, 트랜잭션 수수료)이 제공된다. 현재는 보상을 얻기 위한 경쟁이 치열해져서 난이도가 많이 증가된 상태이기 때문에, 혼자서는 블록 마이닝에 성공하는 것이 불가능하다. 그래서 여러 마이너(Miner)들이 참여하는 마이닝 풀(Mining pool) [4]을 구성하여 함께 마이닝을 수행한다.

데이터 양이 점점 방대해지고 분석 기술이 발달함에 따라서 과거 데이터를 분석하여 향후에 발생할 이벤트를 예측하려는 시도가 많아지고 있다 [5, 6]. 비트코인 트랜잭션 데이터도 누구에게나 공개되어 있고, 과거의 트랜잭션 데이터를 분석함으로써 비트코인 네트워크에서 향후에 일어날 일을 예측할 수 있다. 현재 새 블록의 생성은 거의 마이닝 풀들

에 의해서 이루어지기 때문에 블록에 포함될 트랜잭션들은 마이닝 풀 오퍼레이터가 결정하고 있다고 볼 수 있다. 그래서 일반 노드들은 다음에 어떠한 트랜잭션이 블록에 포함되는지 정확하게 알 수 없다.

다음 블록에 포함될 트랜잭션을 알 수 있으면 다음과 같은 이점들이 있다. 첫째, 고액의 비트코인이 거래소 주소로 이동하는 것은 많은 양의 비트코인이 명목화폐(USD, KRW)로 교환 될 것이라고 생각할 수 있다. 이것은 비트코인 가격에 영향을 미칠 것이고, 투자자들은 이러한 트랜잭션이 다음 블록에 포함될 것인지 파악함으로 트레이딩을 좀 더 수월하게 할 수 있다. 둘째, 블록에 포함될 것이라고 생각한 트랜잭션이지만 마이닝 풀에서 포함을 시키지 않은 트랜잭션들을 조사할 수 있다. 이것은 마이닝 풀이 올바르게 트랜잭션들을 블록에 포함시키고 있는지 검사하는데 사용할 수 있다. 본 연구는 과거의 멤풀 데이터를 분석해서 다음 블록에 포함되는 트랜잭션들을 예측할 수 있다는 것을 실험을 통해서 보여준다.

II. 관련 연구

Abdullah Al-Shehabi [7]는 원하는 시간에 트랜잭션이 컨펌(Confirm)되기 위해서 필요한 feerate (fee/size)을 예측하기 위해서 weight vector 와 함께 멤풀 상태정보를 사용한 기존과는 다른 접근법을 도입했다. 이때 저자는 트랜잭션이 포함되어야 할 여러 개의 블록 범위들에 대한 weight vector 들을 생성하기 위해서 perceptron machine learning 알고리즘을 사용했다.

Beltran Fiz [8]는 마이닝 풀의 트랜잭션 선택 정책에 변화를 탐지하는 방법을 제안했다. 저자들은 트랜잭션 선택 정책을 일종의 분류 문제로 간주했다. 이 연구에서는 그들의 시나리오를 기반으로 마이닝 풀에서 정책이 변한 것이 어떻게 탐지될 수 있는지 보여줬다.

본 논문의 저자 [9]는 비트코인 멤풀에 있는 트랜잭션들 중에서 다음 블록에 어떤 트랜잭션들이 포함될지 예측하는 연구를 수행했다. 본 논문은 이를 확장하는 연구로서 다양한 분류 알고리즘을 사용해서 더 높은 정확도를 보였고, 추가적으로 블록에 포함될 것이라고 예측했지만 포함되지 않았던 트랜잭션들을 분석했다.

III. 비트코인 모니터링 및 분석 시스템

본 연구는 데이터 수집, 데이터 전처리, 데이터 분석 세 단계로 진행되며, 이를 수행하기 위해서 비트코인 모니터링 및 분석 시스템을 구현했다. 아래의 그림 1 은 비트코인 모니터링 및 분석 시스템의 아키텍처를 보여준다.

비트코인 원장에 저장되는 데이터는 히스토리컬(Historical) 트랜잭션들 이다. 즉, 블록에 포함된 트랜잭션들만 저장되게 된다. 하지만 해당 연구를 수행하기 위해서는 과거에 멤풀이 어떤 상태였는데 어떤 트랜잭션이 블록에 포함되었다는 정보가 필요하기 때문에, 실시간으로 멤풀 상태 데이터를 수집하는 것이 중요하다.

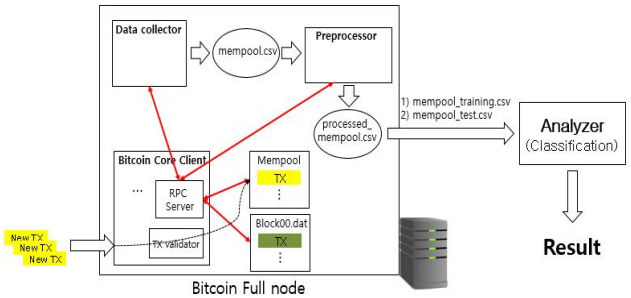


그림 1. 비트코인 모니터링 및 분석 시스템

- **Data collector:** Data collector 는 RPC(Remote Procedure Call) [10] 기반으로 비트코인 네트워크를 모니터링하여 블록(Block), 트랜잭션, 멤풀 상태 데이터를 수집한다. 풀 노드에서 동작하는 Bitcoin core 로 RPC 을 보내서 수신한 응답을 정리하여 csv 파일로 저장한다. 이때 저장하는 멤풀 상태 데이터를 다음에 생성되는 블록의 크기를 임계치(Threshold)로 설정해야 하는데, 비트코인 플랫폼에서 가끔 트랜잭션이 없는 Empty block [11]과 블록의 최대 크기에 훨씬 못 미치는 Abnormal block 이 있기 때문이다. 이러한 일반적이지 않은 데이터를 배제하여 데이터를 수집한다.

- **Data preprocessor:** Data preprocessor 는 수집한 멤풀 상태 데이터에 담겨있는 실시간 트랜잭션들과 생성된 다음 블록에 포함된 트랜잭션들을 비교하여 어떠한 트랜잭션이 블록에 포함 되었는지에 대한 레이블(Label)과 멤풀과 관련된 추가정보를 구한다. Data collector 에서 데이터를 수집할 때는 해당 트랜잭션이 다음 블록에 포함되어 있는 여부를 알 수 없기 때문에 전처리과정에서 이를 분류해주어야 한다. 또한, 비트코인은 UTXO [12]기반의 시스템으로서 트랜잭션의 입력 정보는 원장을 검색해서 얻어야 하며, 이 때문에 Request 들이 추가적으로 발생하게 된다. 그래서 블록 안에 수많은 트랜잭션이 포함되어 있으면 데이터를 수집하는 도중에 실시간으로 해당 트랜잭션들의 디테일 정보를 구하기 어렵다. 아래 표 1 은 전처리 된 데이터에 포함되는 요소들을 설명한다.

Data	Description
txid	transaction id
fee	transaction fee in BTC
size	virtual transaction size as defined

	in BIP 141
feerate	fee / size
txchainCnt	number of related (connected) transactions in mempool
txchainFee	sum of fees of related (connected) transactions in mempool
txchainSize	sum of sizes of related (connected) transactions in mempool
txchainFeerate	txchainFee / txchainSize
timepast	past time after a transaction enters in mempool
num-inputs	number of inputs in a transaction
num-outputs	number of outputs in a transaction
sum-inputs	sum of input values in a transaction
sum-outputs	sum of output values in a transaction
tps	transaction per second
memsize	mempool size
prebs	size of previous block
avg-feerate	feerate on average in mempool
pos-feerate	position of feerate in mempool (in descending order)
c-avg-feerate	sum of feerates of related transactions on an average in the mempool
c-pos-feerate	position of the sum of feerates of related transactions in mempool (in descending order)
next-height	height of the next block
label	included in the next block(1) or not(0)

표 1. 전처리 된 데이터 셋

- **Analyzer:** Analyzer 는 Data preprocessor 를 통해서 전처리 된 Training dataset 과 Testing dataset 을 입력으로 받는다. 실시간 트랜잭션이 다음 블록에 포함될 것인가 여부는 포함(1)되거나 미포함(0)되는 Binary classification 문제로 해석될 수 있기 때문에, Analyzer 는 분류 문제를 해결할 수 있는 머신러닝 모델들(Distributed Random Forest [13], Extremely Randomized Trees [14], Gradient Boosting Machine [15])을 사용하여 결과값을 도출한다. 입력 받은 Training dataset 을 이용해서 위에서 언급한 각 머신러닝 모델들을 학습하고, Test dataset 으로 학습된 모델의 성능을 평가한다.

IV. 실험

비트코인 모니터링 및 분석 시스템으로 데이터를 수집하고 분석을 진행했다. 실험 환경은 Dell PowerEdge R610 서버(Intel(R) Xeon(R) CPU X5650 @ 2.67GHz, 48G RAM)이다. 서버에는 Bitcoin Core 클라이언트(Bitcoin Core 0.17.0 version) [16]를 구동시켜 하나의 노드로서 비트코인 네트워크에 참여했다. Training dataset 을 만들기 위해서 블록 번호 607,702~608,304 에 걸쳐서 데이터를 수집했고, 포함된 트랜잭션이 총 1,927,843 개 이다. Test dataset 은 블록 번호 608,327 ~ 608,345 에 걸쳐서 수집된 데이터로 구성되어있고, 포함된 트랜잭션이 총 140,584 개 이다.

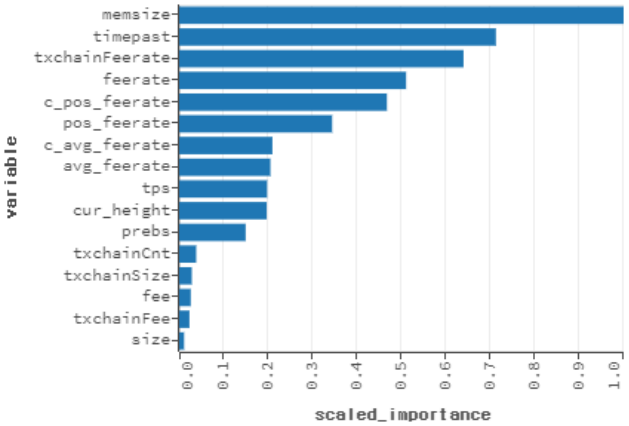
1. 트랜잭션 포함 예측 결과

위에서 설명한 데이터를 기반으로 Analyzer 에서 Distributed Random Forest (DRF), Extremely Randomized Trees (XRT), Gradient Boosting Machine (GBM) 모델을 학습시키고 예측의 성능을 평가했다. 아래의 표 2 는 실시간으로 생성되어 멤풀에 보관되고 있던 트랜잭션들 중에서 다음 블록에 포함될 트랜잭션들을 예측한 실험의 결과를 보여준다.

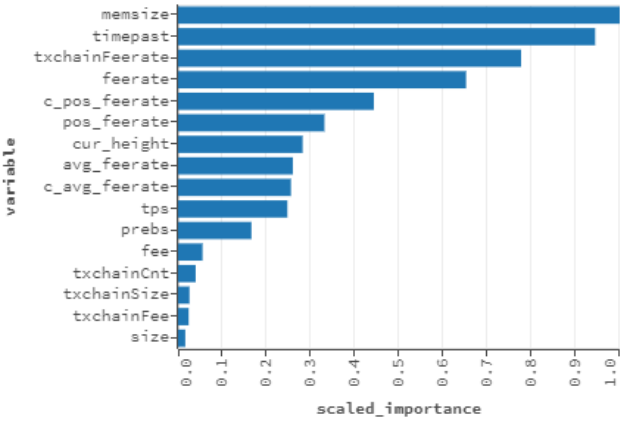
	Precision	Recall	Accuracy	F1-score
DRX	0.8204	0.8960	0.9187	0.8581
XRT	0.8350	0.8688	0.9179	0.8547
GBM	0.8325	0.8697	0.9173	0.8495

표 2. 각 모델에 대한 예측 결과

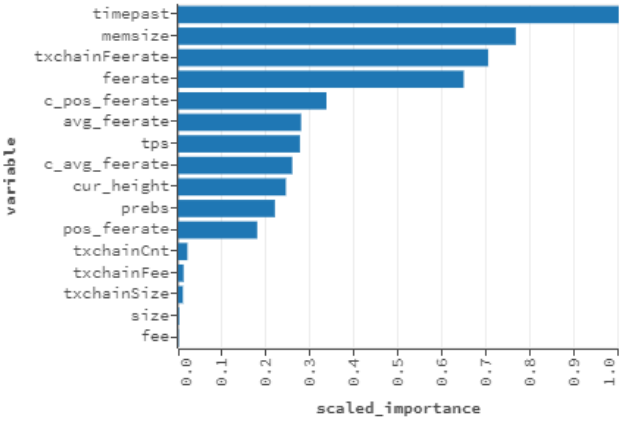
표 2 의 결과를 보면 전처리 된 데이터를 이용해서 학습한 DRX, XRT, GBM 모델 모두 높은 정확도를 보인다는 것을 알 수 있다. 이것은 수집한 데이터를 이용해서 다음 블록에 포함될 트랜잭션을 어느정도 예측할 수 있다는 것을 의미한다. 다음 그림 2 은 각 분류 모델의 Feature Importance [17]를 보여준다. 각 그래프를 보면 해당 모델에서 분류 결과를 도출하는데 있어서 학습한 데이터에서 어떠한 특징이 주요한 영향을 미쳤는지 확인할 수 있다.



(가) DRF 의 Feature Importance



(나) XRT 의 Feature Importance



(다) GBM 의 Feature Importance

그림 3. 각 분류 모델의 Feature Importance 결과

각 모델의 Feature Importance 결과를 보면 상위 5 개의 특징이 동일하다. 즉, 다음 블록에 트랜잭션이 포함되는지 여부는 memsize, timepast, txchainFeerate, feerate, c_pos_feerate 특징들에 영향을 가장 많이 받는다는 것을 확인했다. 블록의 최대 크기가 제한되어 있기 때문에 블록이 생성될 당시 멤풀에 저장되어 있는 트랜잭션의 수가 많을 수록 더 다음 블록에 포함되기 어려운데, memsize 는 이러한 특징이 많이 영향을 끼치는 것을 나타내고 있다. Fee 는 마이너가 수익을 얻는 수단중에 하나이기 때문에 feerate 이 높은 트랜잭션을 우선으로 블록에 포함시킨다는 것을 나타낸다. 이것은 txchainFeerate, feerate, c_pos_feerate 의 높은 Feature Importance 가 설명해주고 있다. 멤풀에 먼저 들어온 트랜잭션일수록 마이너가 블록에 포함될 트랜잭션을 선정하는데 우위를 차지하는데, timepast 의 높은 feature Importance 가 이를 나타내고 있다.

2. 트랜잭션 배제 분석

트랜잭션이 다음 블록에 포함될지 예측한 실험에서 timepast 와 txchainFeerate 는 높은 Feature Importance 를 보였다. 이 중요한 특징들을 이용하여 역으로 다음 블록에 포함될 것으로 예측되었지만 포함이 안된 트랜잭션에 대해서 분석을 하였다. Filtering 의 첫째 조건은 블록에 포함될 것이라고 예측된 트랜잭션들의 txchainFeerate 의 중간값보다 높은 값을 갖는 트랜잭션이 다음 블록에 포함되지 않은 경우 이다. 둘째 조건은 멤풀에 들어온 이후로 경과된 시간이다. 본 실험에서 기준 시간 제약은 임의로 20 분으로 설정했다. Filtering 의 목적은 충분히 높은 txchainFeerate 을 가지고 있고 멤풀에 들어 온지 충분히 긴 시간인 20 분이 지났지만 아직 다음 블록을 통해서 생성되지 않은 트랜잭션을 추출하는 것이다. 아래 그림 4 는 해당 분석 내용을 그래프로 그린 것이다. 그림 4 의 빨간색 점은 앞에서 언급한 filtering 조건 두개를 모두 만족하는 트랜잭션 들이다. (1), (2) 부분을 보면 다음 블록에 포함될 가능성이 높은 많은 수의 트랜잭션들이 실제로는 마이닝 풀에 의해서 블록에 포함되지 않았다는 것을 생각해 볼 수 있다. (1)은 각각 607798, 607798 번째블록이다. (2)는 각각 607810, 6078111, 707812 번째 블록을 의미한다. 공교롭게도 (1), (2)의 블록들은 모두 Poolin 이라는 마이닝 풀에서 채굴하였다.

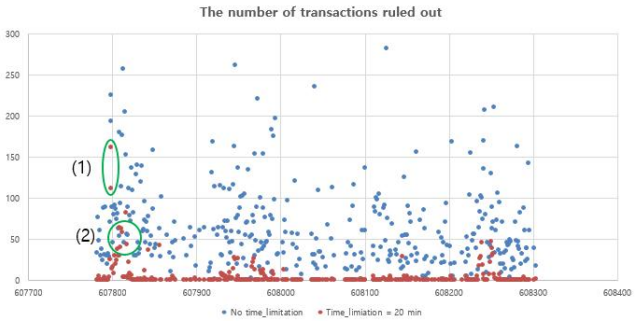


그림 4. 예측이 잘못된 트랜잭션 분석

분석된 내용만 봐서는 특정 마이닝 풀의 오퍼레이터가 의도적으로 filtering 된 트랜잭션들을 배제했는지 알 수 없다. 하지만 해당 분석을 통해서 정상적으로 블록에 포함되어야 하는 트랜잭션들이 얼만큼 배제되고 있는지 확인할 수 있다.

V. 결론 및 향후 연구

비트코인에서 마이너가 최대의 수익을 얻기 위해서 크기당 수수료(fee/size)가 높은 트랜잭션을 최대한 많이 블록에 포함시켜야 하기 때문에, 일반적으로 feerate 이 높은 트랜잭션들이 우선으로 블록에 포함된다고 알고있다. 본 연구에서는 비트코인 네트워크에서 수집한 과거 멤풀 상태 정보와 히스토리컬 트랜잭션 정보를 분석하여, 어떤 특징들이 트랜잭션이 다음 블록에 포함 될지를 결정하는데 많은 영향을 미치는지 확인했다. 분석 결과 memsize,

timepast, txchainFeerate, feerate, c_pos_feerate 특징들이 가장 많은 영향을 미치는 요소임을 알 수 있었다. 또한, 수집한 데이터를 학습한 세 종류의 분류 모델들 모두 90%이상의 높은 예측 정확도를 보여주었다. 향후 연구로는 여러 노드의 멤풀 정보를 수집하여 실험을 확장할 예정이며 Deep learning 알고리즘을 사용한 결과와 예측 정확도를 비교해볼 예정이다.

ACKNOWLEDGMENT

본 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구 임 (No.2018-0-00539)

참 고 문 헌

[1] Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system." (2008).
[2] 고경찬, 이채현, 홍원기, "비트코인 노드 메모리 풀 유사도 분석", KNOM Conference 2019, Daegu, Korea, May. 30, 2019, pp. 16-18.
[3] Zheng, Zibin, et al. "An overview of blockchain technology: Architecture, consensus, and future trends." 2017 IEEE international congress on big data (BigData congress). IEEE, 2017.
[4] “Mining pool”, https://en.wikipedia.org/wiki/Mining_pool, Accessed: April 1, 2020.
[5] Bunker, Rory P., and Fadi Thabtah. "A machine learning framework for sport result prediction." Applied computing and informatics 15.1 (2019): 27-33.
[6] Yan, Xiang-Bin, et al. "Research on event prediction in time-series data." Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826). Vol. 5. IEEE, 2004.
[7] Wang, Aiping, et al. "An incremental extremely random forest classifier for online learning and tracking." 2009 16th ieee international conference on image processing (icip). IEEE, 2009.
[7] Al-Shehabi, Abdullah. "Bitcoin Transaction Fee Estimation Using Mempool State and Linear Perceptron Machine Learning Algorithm." (2018).
[8] Pontiveros, Beltran Borja Fiz, Robert Norvill, and Radu State. "Monitoring the transaction selection policy of Bitcoin mining pools." NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium. IEEE, 2018.
[9] Ko, Kyungchan, et al. "Prediction of Bitcoin Transactions Included in the Next Block." International Conference on Blockchain and Trustworthy Systems. Springer, Singapore, 2019.
[10] Nelson, Bruce Jay. "Remote procedure call." (1981).
[11] “Empty block”, https://www.blockchain.com/ko/btc/block/625377, Accessed: April 11, 2020.
[12] “UTXO”, https://bitcoin.org/en/glossary/unspent-transaction-output, Accessed: April 11, 2020.
[13] “Distributed Random Forest”, http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html, Accessed: April 11, 2020.
[14] Wang, Aiping, et al. "An incremental extremely random forest classifier for online learning and tracking." 2009 16th ieee international conference on image processing (icip). IEEE, 2009.
[15] Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.
[16] “Bitcoin Core 0.17.0. released”, https://bitcoin.org/en/release/v0.17.0, Accessed: April 11, 2020.
[17] “Variable importance”, http://docs.h2o.ai/h2o/latest-stable/h2o-docs/variable-importance.html, Accessed: April 11, 2020.