Toward Characterizing Blockchain-Based Cryptocurrencies for Highly Accurate Predictions

Muhammad Saad, Jinchun Choi, DaeHun Nyang, Joongheon Kim, and Aziz Mohaisen, *Senior Member, IEEE*

***Abstract*—Recently, the Blockchain-based cryptocurrency market witnessed enormous growth. Bitcoin, the leading cryptocurrency, reached all-time highs many times over the year leading to speculations to explain the trend in its growth. In this article, we study Bitcoin and Ethereum and explore features in their network that explain their price hikes. We gather data and analyze user and network activity that highly impact the price of these cryptocurrencies. We monitor the change in the activities over time and relate them to economic theories. We identify key network features that help us to determine the demand and supply dynamics in a cryptocurrency. Finally, we use machine learning methods to construct models that predict Bitcoin price. Based on our experimental results using two large datasets for validation, we confirm that our approach provides an accuracy of up to 99% for Bitcoin and Ethereum price prediction in both instances.**

***Index Terms*—Bitcoin, Blockchain, Ethereum, prediction.**

I. Introduction

BLOCKCHAIN-BASED digital currencies have witnessed enormous change in value over the last few years [1]. Bitcoin, the most popular cryptocurrency, was launched in 2009, and stayed as the only Blockchain-based cryptocurrency for more than two years. However, today, the cryptocurrency world has more than 5000 cryptocurrencies [2] and more than 5.8 million active users [3]. Bitcoin leads the cryptocurrency market with 58% market share; corresponding to $4.9 Billion USD trade volume and over 12 000 transactions per h [4]. In December 2016, the price of 1 Bitcoin token was under $1000 USD, compared to about $19 000 USD in late 2017, and over $3600 USD in January 2019 [5]. These changes in the price led to a lot of interest in cryptocurrency and Bitcoin in particular. In this article, we carry out a study on Bitcoin and Ethereum to analyze their network features that capture the user behavior and in turn have an impact on their price.

Manuscript received September 14, 2018; revised January 15, 2019 and April 16, 2019; accepted June 19, 2019. Date of publication September 17, 2019; date of current version March 2, 2020. This work was supported in part by National Research Foundation of Korea under Grants NRF-2016K1A1A2912757 and 2017R1A4A1015675 and in part by a Chung-Ang University Research Grant (2019). This article was presented in part at the Hot Topics in Pervasive Mobile and Online Social Networking (HotPOST 2018), Honolulu, HI, USA, April 2018. *(Corresponding authors: Joongheon Kim; Aziz Mohaisen.)*

M. Saad and A. Mohaisen are with the University of Central Florida, Orlando, FL 32816 USA (e-mail: saad.ucf@knights.ucf.edu; mohaisen@cs.ucf.edu).

J. Choi is with the University of Central Florida, Orlando, FL 32816 USA, and also with the Inha University, Incheon 22212, South Korea (e-mail: jc.choi@knights.ucf.edu).

D. Nyang is with the Inha University, Incheon 22212, South Korea (e-mail: nyang@inha.ac.kr).

J. Kim is with the Chung-Ang University, Seoul 156-756, South Korea (e-mail: joongheon@cau.ac.kr).

Digital Object Identifier 10.1109/JSYST.2019.2927707

The underlying technology of every cryptocurrency is the Blockchain. Blockchain acts as a decentralized public database that preserves anonymity and augments trust between the users [6]. Trust in an anonymous peer-to-peer model is achieved by consensus protocols such as proof-of-work, proof-of-stake, proof-of-knowledge, and distributed consensus [7]. The decentralized environment and the append-only model prevent Blockchains failure and data tampering, and such features lay ideal foundations for cryptocurrency applications to be built on top of Blockchain.

Cryptocurrencies involve the exchange of digital assets (tokens) and have evolved from virtual currency to smart contracts and applications beyond currency. This transformation of cryptocurrencies is categorized as Blockchain 1.0, 2.0, and 3.0 [8]. Blockchain 1.0 solely involves transfer of digital currency between parties. Bitcoin is an example of Blockchain 1.0, since it only allows transfer of digital tokens (bitcoins). Blockchain 2.0 is an extension of Blockchain 1.0 that allows transfer of many other assets, offering more flexible protocols for the users to design their transactions, such as smart contracts [9] and decentralized autonomous organizations [10], which are among many useful applications of Blockchain 2.0 [11]. Blockchain 3.0 is yet another extension of this technology that envisions the use of Blockchain beyond digital currencies, with applications for distributed censorship resistant organization models, digital identity verification and decentralized domain name system [12].

New cryptocurrencies address shortcomings of older ones, with better throughput, scalability, and programmability. Although this gives a general idea why cryptocurrency markets have grown, many factors contributing to the rise in cryptocurrency prices are not well understood. In this article, we look at the dynamics of various variables in a cryptocurrency, namely Bitcoin and Ethereum, which can shed light on their price trends. We use the network features of these cryptocurrencies as an example and perform an in-depth analysis using the data obtained from their Blockchain and peer-to-peer network.

The key factor that influences the growth of the cryptocurrency market is the interest shown by the users toward the trade of the digital tokens. As more users engage in the market activity, the demand for the digital tokens increases, leading to a higher price. However, unlike fiat currency systems which are centralized and traceable, cryptocurrencies are (theoretically) decentralized and pseudoanonymous, lacking tangible digital footprints. Therefore, with insufficient knowledge it becomes challenging to measure the interest factor of the users and perform a user-based study aimed toward the understanding of changing price and market trends.

We address this challenge by arguing that despite anonymity and decentralization of cryptocurrencies; there are several network indicators that might be useful in demonstrating the interest of users and the overall market behavior. We show that these network indicators have a high correlation with the price of a cryptocurrency and can be used to accurately predict its price. Furthermore, these features can also be used to provide a rationale behind the network activity driven by the user behavior. To validate our reasoning with experiments, we construct a machine learning model that learns from the highly correlated network indicators and predicts the price of cryptocurrencies with high accuracy.

Prior efforts on prediction for cryptocurrency price used the past price indexes to forecast the future price [13]. This approach is inspired by a large body of work on stock market prediction [14], and has been tried in the cryptocurrency market. However, this method does not partake the volatile behaviors of the network entities that may indirectly influence the price, independent of the previous price indexes. For instance, a sudden decrease in the network hash rate can prolong the block publishing time and reduce the network throughput as well as the number of newly generated coins. Such a decline in hash rate is independent of the past price index, and therefore cannot be used to accurately model the future price. Lacking the ability to capture this behavior has led to low accuracy of prediction models in the prior art ($\approx$52%). Specific to our work, we take a tangential approach toward modeling the price by using network indicators that are strongly correlated with the cryptocurrency price and lead to better and more accurate prediction models.

Novelty: The novelty of our approach lies in: 1) the identification of the key network features that capture the changing price models, and can therefore be used for feature engineering as given in Section III-A1; 2) the distinguishing methodology from the prior work [13], in which only the past price indexes were used to predict the future price given in Section III-A; and 3) the methodical reasoning about correlation of identified features with cryptocurrency price, to enhance the understanding of user behavior and the cryptocurrency network provided in Section IV. After feature selection, we use standard machine learning techniques, including regression, long short-term memory (LSTM) networks, and conjugate gradient algorithm. As a result, our prediction models achieve a high accuracy of 99% for Bitcoin and Ethereum, outperforming the state-of-the-art [15] (52%), and validating the novelty and significance of our approach.

Contributions: In summary, we make the following contributions. 1) We study Bitcoin and Ethereum network and identify the key network indicators that affect their price. 2) We show how these features are driven by user and network activity, and provide a rationale behind their influence on price. 3) We adopt machine learning approach using regression and LSTM analyses to construct price prediction models for Bitcoin and Ethereum. 4) Our prediction models estimate the price of cryptocurrencies with high accuracy (99%), and outperform the state-of-the-art.

Organization: The rest of the article is organized as follows. In Section II, we review the related work. In Section III, we provide preliminaries of this work and outline our methodology and dataset attributes. In Section IV we perform data analysis to extract the most significant features that impact the price. In Section V we carry out our experiments and report the results. Concluding remarks are made in Section VI.

II. Related Work

In this section, we review the notable related work. We focus on analyses dedicated to understanding how cryptocurrencies influence the financial and other systems, general analysis of Bitcoin and Ethereum, and their price prediction.

Vigna *et al.* [16] analyzed how Blockchain based applications are challenging the global economic order by exploring the impact of Blockchain-based applications on the future of the financial system. Swan [8] proposed a possibility of cheaper, efficient, and secure economical models based on Blockchain. The use of Blockchain 3.0 is estimated to create new possibilities in Internet of Things (IoT), privacy management, and voting systems [17].

Blockchain 2.0 transformed cryptocurrency from mere exchange of tokens to smart contracts. Rose [18] analyzed the evolution of digital currencies and Omohundro [19] explored recent developments in cryptocurrency and smart contracts. Kosba *et al.* [9] explored different dimensions of smart contracts, including criminal smart contracts. Peters *et al.* [20] analyzed the future of banking system ledgers with Blockchain technology, transaction processing, and smart contracts.

For better applications, the security attack surface of Blockchain is also explored, including the 51% attack, selfish mining, double-spending, block withholding, block forks, and distributed denial-of-service attacks [21]; arguably the most prevalent attack [22].

Limited research is done on the feature-based price analysis. Indera *et al.* [13] developed a nonlinear autoregressive Bitcoin price prediction model using the opening and closing past prices to predict future price. McNally [15] explored various machine learning approaches to predict Bitcoin price using Bitcoin price index, achieving a maximum accuracy of 52% with LSTM networks. This article is an extension of our previously published work in [23]; concurrent to that work, Jang and Lee [24] performed a time series analysis of Bitcoin to improve predictive performance. They use Bayesian neural network with other linear and nonlinear benchmark models to explain volatility in Bitcoin price.

In this article, we explore other features, besides past prices, to establish patterns in price. We investigate various network features and identify the highly correlated ones that determine the price. Using those features, we train and test our models, which achieves a near-perfect prediction accuracy.

III. Preliminaries

The main goal of this article is broad and aims to provide the initial step toward characterizing Blockchain-based cryptocurrencies for predictions. Toward that, we perform a detailed analysis for the top two cryptocurrencies in the market, namely Bitcoin and Ethereum. We select them due to their widespread popularity, extensive user-base, and high market cap. Our approach toward price characterization based on network features can be extended to other cryptocurrencies.

In Fig. 1, we plot the price change trend of five major cryptocurrencies over the last one year. The difference in the actual price value of each currency is high, and cannot be plotted in one graph. We use the min–max normalization to scale the data in the range [0,1] and plot the normalized price. The min–max scaling is conducted as $z = \frac{x_i - \min(x)}{\max(x) - \min(x)}$.

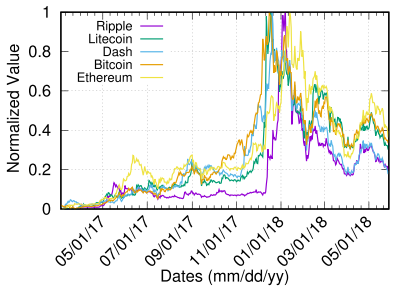

Fig. 1. Cryptocurrency price change in 2017–2018. Notice that there is a high correlation in the price fluctuation in all cryptocurrencies. Furthermore, it can be observed that toward the end of 2017, each cryptocurrency reached its highest price index. In 2018, the price index decreased.

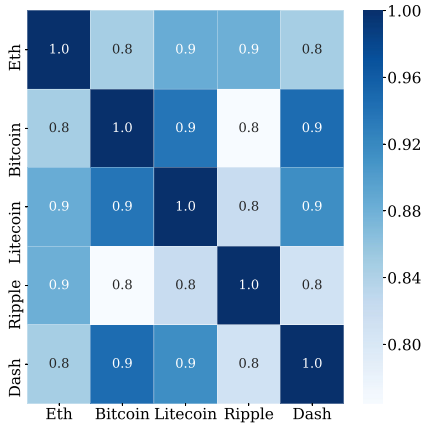

Fig. 2. Correlation of major cryptocurrencies exemplified through a heatmap.

In Fig. 1, we observe an increase in the price of every cryptocurrency over the year 2017, and particularly toward the end of the year. The growing trend started around April 2017, and kept on increasing. Toward the end of 2017, the rise in the price has been very steep. It is commonly conceived that these cryptocurrencies are competitors in the market and price hikes in one leads to a price fall in another. However, from the plots we observed that there is an almost monotonic change in the price of all the currencies simultaneously. They all followed similar trends of rise and fall over time. It can be further observed in Fig. 1 that the price of each cryptocurrency decreased sharply at the start of the year 2018. Although, the price has been fluctuating over the year, it is noteworthy that there has been a monotonic change in the price across all cryptocurrencies, indicating the presence of a correlating factor among all.

To further analyze the similarity in their trends, we use the Pearson correlation coefficient between the price in all currencies over time, defined as $\rho(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}}$. We report our results in Fig. 2. While the pairwise correlation is high across all currencies, supporting the initial premise of this work, we observe significant correlation between Bitcoin, Dash, and Litecoin price growth. Furthermore, we found a significant correlation between the price trend of Ethereum and Ripple. As such, the growth in one major cryptocurrency, derives the growth in another similar currency, highlighting the speculative nature of the interdependent interactions between the currencies' prices, and hinting on the potential generality of findings to other systems.

A. Methodology

In this section, we outline our methodology for characterizing price of Bitcoin and Ethereum, spanning data collection, data characteristics, and our approach. We outline the relevance of key indicators in our dataset toward the broader goal of making a price prediction model.

1) Data Collection: For this article, we crawled data related to the network features of Bitcoin and Ethereum using online resources. For Bitcoin, we used the public Blockchain and application programming interface (API) provided by the exchange company "Blockchain" [5], that maintains data related to Bitcoin network. One of the features used in our prediction model is the "number of wallets." These are the wallets created solely on the exchange of "Blockchain" and are not related to other exchanges such as "Coinbase." The memory pool (mempool) data shown in our article is also related to the information maintained by the mempool of "Blockchain" full node. It is worth mentioning that mempool of nodes in the peer-to-peer settings of Bitcoin may vary due to peer positioning and nature of transaction relay. However, all other features such as hash rate, price, number of bitcoins, number of transactions are consistent across all exchanges and nodes in the network. From "Blockchain" API, we collected data from April 2016 to May 2018. The dataset consists of features including the number of wallets, unspent transaction outputs (UTXO's), mempool size, block size, mean confirmation time, miner's income, transactions per day, transactions per block, unique Bitcoin addresses, cumulative network's hash rate, network's difficulty, fee, fee per transaction, system-wide total bitcoins, trade volume, and the market price of Bitcoin.

For Ethereum, we followed the same procedure and collected data using the information provided by an Ethereum exchange "Etherscan" [25]. We collected data from April 2016 to May 2018 including features such as transaction growth, address count, ether supply, market cap, transaction fee, hashing power, difficulty, block time, gas limit, and gas used.

The price of a cryptocurrency can be influenced by internal features, external features, or both. Internal features include indicators that represent the network behavior such as mempool size, hash rate [23], etc. On the other hand, external features include crude oil price, government policies toward cryptocurrency exchanges, electricity charges, public sentiment [26], etc. In this article, we focus on collecting internal features and determine their effect on price. Our rationale for this approach is driven by the fact that the internal features eventually accommodate for the impact of external policies. For instance, if electricity cost is increased, some mining pools shut down [27]. As a result, the internal features including the hash rate and the block publishing time change. Since the external factors influencing the cryptocurrency market are eventually manifested in the internal network behavior, we primarily focus on collecting and analyzing internal features. External features, however useful, fall outside the scope of this article.

2) Data Characteristics: Bitcoin and Ethereum involve the exchange of digital tokens, and their operations may vary at the application level. As mentioned earlier, Bitcoin belongs to the first generation of Blockchain (Blockchain 1.0) that only involves exchange digital coins. Ethereum on the other hand, belongs to Blockchain 2.0, that offers development of smart contracts atop Blockchains. Smart contracts enable the users to make conditional changes in the exchange of coins by offering greater programmability with a broader use-case. Due to that, the dataset includes some

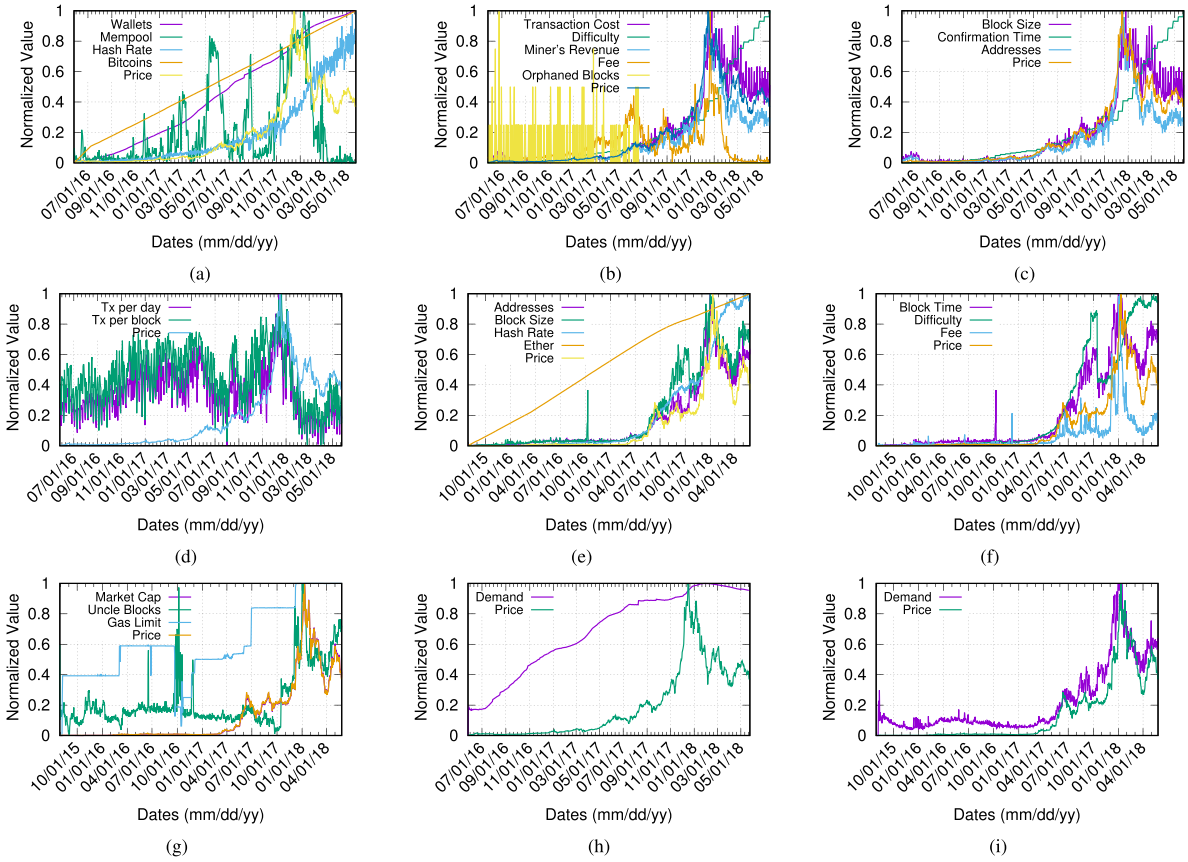

Fig. 3. Trends in the features captured from our dataset. Notice that the hash rate, the difficulty, and the transaction cost are highly correlated with the price. The increase in demand (Total Wallets / Total Bitcoins) has led to an increase in the price. The features in Ethereum dataset are more correlated with price than Bitcoin dataset. Furthermore, Ethereum also captures the demand and supply trend more accurately than Bitcoin. (a) Normalized Features (Bitcoin). (b) Normalized Features (Bitcoin). (c) Normalized Features (Bitcoin). (d) Normalized Features (Bitcoin). (e) Normalized Features (Ethereum). (f) Normalized Features (Ethereum). (g) Normalized Features (Ethereum). (h) Demand and Price (Bitcoin). (i) Demand and Price (Ethereum).

common features among both cryptocurrencies such as hash rate, block size, etc., and some unique features such as gas limit, gas price, etc.

The number of wallets gives an estimate of how many new users join the platform everyday. Although this measure is specific to the exchange, the other parameter known as "unique addresses" captures the growth of users in the overall cryptocurrency. For Bitcoin, we collected a total of 24 867 899 wallets and 464 173 unique addresses, while for Ethereum we collected a total of 812 183 addresses. In cryptocurrencies, mempool is a repository for unconfirmed transactions prior to the mining process. The size of mempool varies depending on the rate of the incoming transactions, the transaction backlog, and the rate of transaction mining.

In Bitcoin, the size of blocks is fixed at 1 MB and the average block computation time is 10 min. In Ethereum, the block size is adjustable depending upon the transaction backlog and mean confirmation time. The average block computation time in Ethereum is between 10 and 20 s. We observed in our dataset that the maximum hash rate of Bitcoin was equal to 11 941 671 Terahashes per second (TH/s) with a difficulty parameter of 1 590 896 927 258, and the maximum hash rate of Ethereum was equal to 268 134 Gigahashes per second (GH/s) with a difficulty parameter of 3218.953. The total coins in Bitcoin and Ethereum, at the time of our data collection were 17 055 012 and 99 687 139, respectively.

3) Analysis Metrics and Approach: In this article, we analyze the attributes of the cryptocurrency system, exemplified by Bitcoin and Ethereum, that are influential on their price. To determine the contributing features toward price, we found the most highly correlated features in the dataset to explore general trends and insights about the two cryptocurrencies.

Next, we estimated the change in user behavior (characterized by various attributes associated with users) that led to increase or decrease in the price. For example, if the number of wallets is increasing, then there is a likelihood that more users are joining the network, which leads to a higher demand for the fixed number of coins in the system. With the limited coin supply and high collective purchase power, the price (naturally) goes up. Using those highly correlated features, we train machine learning models to predict the price of Bitcoin and Ethereum over time. Toward that, we divided our data into a training dataset and a test dataset, and cross validated the predicted outcome. With good accuracy, we were able to construct models for the top two cryptocurrencies that help in explaining their price trends.

IV. Data Analysis and Trends

1) General Trends: We analyze the trends in features of dataset for each cryptocurrency. In order to do that, we normalize the data using the min–max normalization and plot various normalized features over time in Fig. 3. In Fig. 3(a) and (b) we observe

that the number of wallets, the hash rate, the number of bitcoins, the cost per transaction, the difficulty, and the miner's revenue change monotonically with the price. In Bitcoin, the mempool size and and the fee varied over time, although had an identical trend to one another; the correlation between the fee and the mempool size was 0.82. When the mempool size grows, for sudden high demands, or while the Bitcoin network is under flood attacks [28], users naturally pay more to prioritize their transactions, which explains the high correlation between the mempool size and the transaction fee.

We also observed that in the Ethereum dataset, the features including addresses, hash rate, block time, and gas limit closely followed the changing trends in price. In Blockchain applications, it is possible that two miners come up with a valid block and only one of them gets accepted into the main chain. In Bitcoin, those rejected blocks are known as the "Orphaned Blocks" and in Ethereum they are called "Uncle Blocks."

From Fig. 3(b), it can be observed that in Bitcoin there is no link between the rate of orphaned blocks and price, but in Ethereum, from Fig. 3(g), there is a high correlation between the rate of uncle blocks and the price. One possible explanation to that is block time in each cryptocurrency. In Bitcoin, the average block time is 10 min and it is less likely that two miners can come up with same block within that time period. However, in Ethereum, the block time is very short and when the price is increasing more miners attempt to mine blocks which increases the possibility of uncle blocks.

2) Supply-and-Demand Trends: In cryptocurrencies, new coins are generated in the system as when a block is published. Since the average block time is constant, therefore, the supply of new currency in the system is deterministic and linear. When new users join the cryptocurrency, new wallets and addresses are created. In Fig. 3(a), we observed that the number of wallets and addresses have increased nonuniformly in Bitcoin and Ethereum, raising the demand for the limited number of coins. Since the number of wallets grew at a higher rate than new coins, we can formulate this as a demand and supply model: A growing rate of wallets denotes that more users are joining Bitcoin, which leads to an increase in demand for the coins. Since the increase rate of coins is a small constant, the new coin supply to system is less than the demand, which explains the primary cause of price rise with growing wallets number.

We plot the min–max normalized number of wallets per available coins for Bitcoin and Ethereum in Fig. 3(h) and (i). We first calculated the number of wallets per bitcoin, and then normalized the number using the min–max normalization. We observed that there is an increase in the demand, which contributes to the price hike. We also noticed that correlation between demand and price in Ethereum was higher (0.96) than Bitcoin (0.74).

3) Examining External Features: It has been postulated in the literature [29], that the crude oil price may influence trends in the cryptocurrency market. The crude oil price affects the electricity tariffs worldwide, which in turn affect the operations of the mining pools. High electricity price can force mining pools to shut down, and as a result, the hash power and the throughput of a cryptocurrency might decrease.

To examine that, we collected the price indexes of crude oil in the international market and observed its correlation with the price of Bitcoin and Ethereum. In Fig. 4, we plot the normalized price indexes of cryptocurrencies with crude oil. Notice that the overall trend in oil prices differs in Bitcoin compared to Ethereum. Especially, since the start of 2018, and while the price of cryptocurrencies decreased, the oil prices have increased considerably.

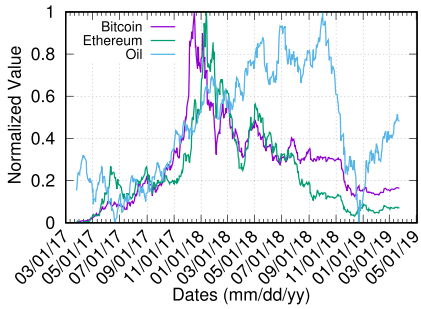

Fig. 4. Price trends observed in Bitcoin, Ethereum, crude oil. Notice that over time, while Bitcoin and Ethereum show similar trends in price changes, the oil prices have been distinctly different.

TABLE I
REGRESSION ANALYSIS RESULTS, HIGHLIGHTING A MODERATE POSITIVE CORRELATION BETWEEN CRUDE OIL PRICES, BITCOIN (0.55), AND ETHEREUM (0.44)

	Slope	Y-Intercept	Correlation Coefficient	Standard Error
Bitcoin	0.41	0.09	0.55	0.02
Ethereum	0.35	0.08	0.44	0.03

However, the crude oil price is not used as a prediction feature in this article because it is below the correlation threshold (0.6), used for feature selection.

To further observe the patterns of similarity, we performed linear regression (LR) analysis to model the relationship between the independent variable (crude oil price) and the dependent variables (Bitcoin and Ethereum prices). We report our results in Table I. Overall, the results show a positive correlation between the crude oil price and the price of cryptocurrencies. In particular, Bitcoin has a comparatively high correlation coefficient (0.55) compared to Ethereum (0.44). However, and as we show later in the subsequent paragraph, for our prediction models, we only select features that have a minimum correlation coefficient of 0.6. Since the correlation coefficient of crude oil is below our baseline criteria, we do not include it among the selected features for the prediction task.

4) Features for Price Prediction: To determine the most useful features in our dataset for price estimation, we calculated the correlation matrix of all data attributes. We report a subset of correlation matrix in Figs. 5 and 6. It can be observed from the figures that the features in Ethereum dataset are more highly correlated with the price than the features in Bitcoin dataset. In Ethereum, the minimum and maximum correlation factor of the features with price is 0.7 and 0.9, respectively, while in Bitcoin, the minimum and maximum correlation of the features with the price is 0.4 and 1.0, respectively. For our regression model and prediction, we selected features with correlation coefficient greater than 0.6.

A. *Effects of User Activity on Price*

In this section, we try to explain the user activity, determined by highly correlated features, affects the price. Among them the features such as the number of wallets, the hash rate, and the UTXO's, determine the number of new users coming into the

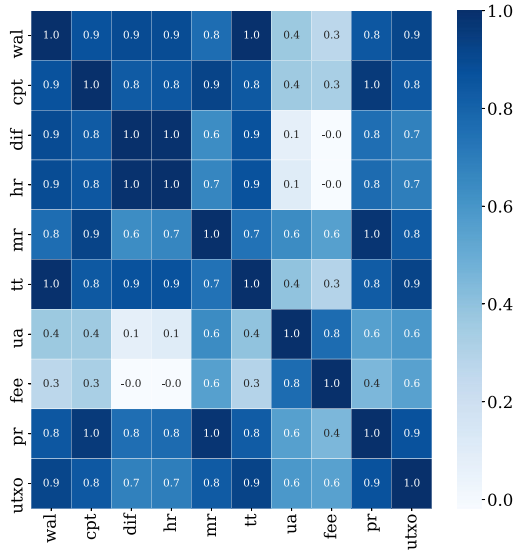

Fig. 5. Correlation matrix of Bitcoin. Here, wal, cpt, dif, hr, mr, tt, ua, fee, pr, and utxo denote number of wallets, cost per transaction, difficulty, hash rate, mining revenue, total transactions, unique addresses, fee, price, and UTXO's respectively.

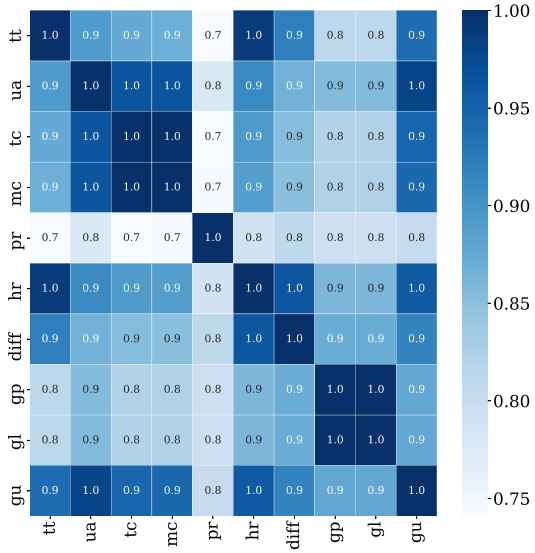

Fig. 6. Correlation matrix of Ethereum. Here tt, ua, tc, mc, pr, hr, diff, gp, gl, and gu denote total transactions, unique addresses, market cap, price, hash rate, difficulty, gas price, gas limit, and gas used, respectively.

network, new miners joining the mining pools, and the aggregate spendable balance of all the users.

1) Wallets and Unique Addresses: As mentioned earlier, the increase in the number of wallets corresponds to greater demand of the limited coins in the system, which results in a price hike. This reasoning can also be extended to the number of unique addresses and the number of transactions per day. The growth in these two features indicates more users coming into the system and making more transactions. As such, the increase in the number of users and user activity (transactions) corresponds to (possibly) more cash is flowing into the system. Since cash flow in Bitcoin increases, the (collective) purchase power of users also increases. This implies that for fixed assets (bitcoins) owned by a user A in the system, there is some user B in the system who is willing to pay more for the same set of assets. In economics, the trend above is captured by a theory known as the "greater fool theory" [30], which states that the price of a commodity is determined by the expectations of users rather than by the commodity's intrinsic value.

2) Difficulty and Hash Rate: Computing a block generates new coins in the system, which are given to the miner as a Coinbase reward. Miners earn coins from the Coinbase rewards and fee paid by the users for transaction processing. As the price grows, the corresponding value of miner's income (in USD) also grows. In Fig. 7(a) and (d), we plot the miner's income from our two datasets. We observed that the Coinbase rewards and fee have increased over time. With the growing incentive of income, more miners are joining the mining pools hoping to capitalize on the increasing monetary reward, which explains why the hash rate grows with the price.

3) Bitcoin: In Bitcoin, the difficulty is a measure of how long it takes to compute a block, which is defined by a target value set by the network [31]. Based on the hashing power, the *target* is adjusted every two weeks to keep block mining time within 10 min. The difficulty is recomputed based on the hashing power: if hashing power increases, the probability of finding a block within under 10 min increases. To adjust the probability, the difficulty is raised by increasing the target. In Fig. 7(b) and (e), we plot the difficulty along with the network's hashing rate for Bitcoin and Ethereum. In (1), we show how the block computation time, $T(B)$, is affected by the hashing rate, H_r, the *target*, Target, the probability of finding a block, $P_r(B)$, and the average number of hashes required to solve the target, H

$$P_r(B) = \frac{\text{Target}}{2^{256}} \quad H = \frac{1}{P_r(B)} \quad T(B) = \frac{H}{H_r}. \tag{1}$$

Since the difficulty remains constant for 2016 blocks, we analyze how the mining pool size affects the price and the average block computation time. From our dataset, we found a window of time where the difficulty was constant and the hashing rate was reduced. At the same time interval, we found the mean confirmation time for transactions and the price. From (1) we inferred that, with constant $P_r(B)$, the block time $T(B)$ increases if H_r is reduced, leading to a higher confirmation time for transactions and less Coinbase rewards per time unit, therefore leading to a fall in the price. In Fig. 7(c), we plot one case that happened in October 2017.

4) Ethereum: In Ethereum, the difficulty is adjusted after every block using Homstead method described in [32]. Since Ethereum follows a different set of protocols than Bitcoin, its difficulty measure does not remain constant for a deterministic period of time (2016 blocks in Bitcoin). Due to that, there is no price fluctuation observed in Ethereum related to the change in the hash rate and constant difficulty. However, within our dataset, we noticed that on October 15, 2017, the difficulty measure of Ethereum decreased by 52% over night while the hash rate was constant. Although, it did not affect the price, but it decreased the block computation time by 52%. We plot this observation in Fig. 7(f).

5) UTXO's: In Bitcoin, another important feature that contributes toward the price is the set of UTXO's. UTXO's are the spendable transactions in wallets that are confirmed in Blockchain. UTXO's determine the number of sellers in Bitcoin. Just as the increase in the number of wallets indicates

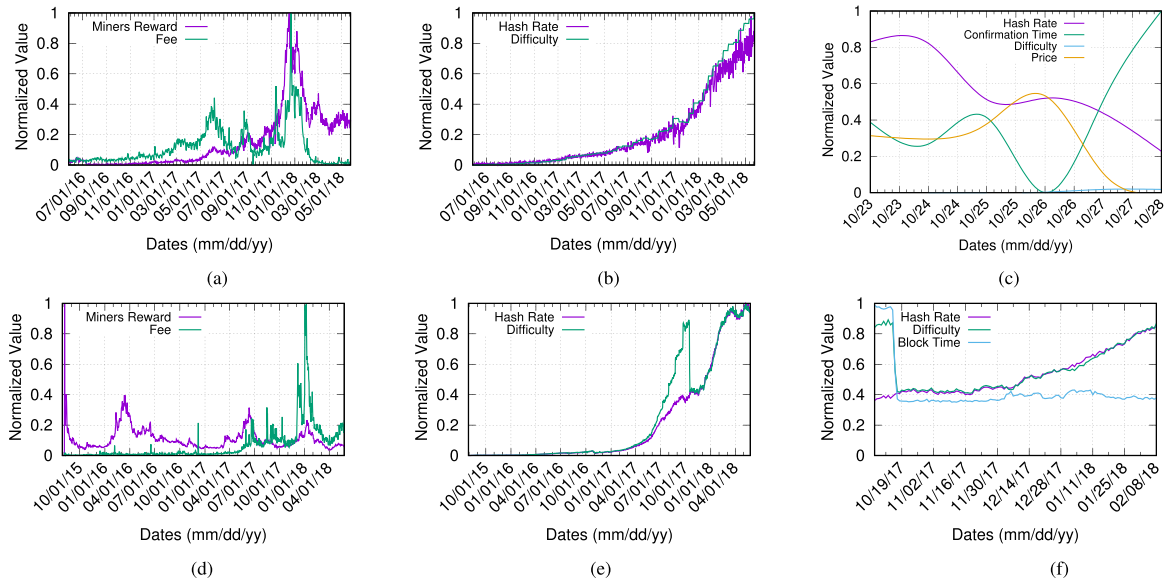

Fig. 7. In (a) miner's revenue is indicated by the Coinbase reward. (b) shows the increasing hash rate and the network's difficulty. Notice in (c) when the network's difficulty is constant and the hash rate decreases, the price also decreases. (a) Miner's revenue and fee paid (Bitcoin). (b) Trend of hash rate and difficulty (Bitcoin). (c) Price with difficulty and hash rate (Bitcoin). (d) Miner's revenue and fee paid (Ethereum). (e) Trend of hash rate and difficulty (Ethereum). (f) Ethereum change in difficulty.

more buyers in the system, more UTXO's indicate more sellers. UTXO's depend on the number of coins produced and the nature of the ongoing transaction. In our dataset, we observed that there is a high correlation between the price and the UTXO set.

As the UTXO set increases, there is more spendable balance in the system which leads to an increase in the exchange of transactions (trade), which in turn increases the price of Bitcoin. The fall in Bitcoin price in 2018 can also be attributed to the fall in the UTXO set, indicating less spendable balance in the system and limited trade avenues for the users. This can be further linked to the decreasing interest of people in Bitcoin which explains the decrease in price.

6) Gas: In Ethereum, gas is the "fuel" unit used in the execution of smart contracts. Each operation code instruction in a smart code consumes different units of gas which is summed up toward the end of smart contract execution to compute the total units of gas used. The transaction fee is calculated using gas price and the units of gas used during the process. In our dataset, we observed that the amount of gas used in Ethereum had a high correlation with the price, indicating a user behavior related to the interest in smart contracts. A high use of gas can (possibly) mean that more smart contracts are being run on Ethereum virtual machine, or more computation intensive operations are being performed while running smart contracts. In each case, it is indicative of a high user interest in Ethereum and smart contracts which explains the price hike.

V. Predictions: Experiment and Results

In this section we build price prediction models for Bitcoin and Ethereum using features in our dataset. For prediction models we take supervised learning approach using regression, LSTM networks, and conjugate gradient algorithm. Our results validate that network features can be used to accurately predict the price of a cryptocurrency.

A. Regression Approach

We consider three popular approaches: the LR, regression with gradient boosting (GB), and regression with random forest (RF). We test our datasets with each method to find the optimum technique useful toward the price prediction of Bitcoin and Ethereum. In the following, we review the conceptual primitives required for understanding each of those algorithms.

1) Linear Regression: LR is a method of predicting the future value of an unknown dependent variable by learning the values of known independent variable [33]. Provided data in the format $x = x_1, x_2, \ldots x_n, y = y_1, y_2, \ldots y_n$, where x is the independent variable and y is the dependent variable to be predicted, LR finds a line of best fit, $y = mx + b$, where m is the coefficient of regression of y on x, and b is the y-intercept. For example, if the regression coefficient m, of y on x is 0.45 units, that will imply that y will increase by 0.45 if x increases by 1 unit. The accuracy of LR is determined by calculating the coefficient of determination R^2, also known as the least square fit. Least square fit calculates the minimum (min) between the predicted value and the real value as mentioned below

$$R^2 = \sum_{i=1}^{n} (\Delta y_i)^2 = \sum_{i=1}^{n} [(mx_i + b) - y_i]^2 = min. \tag{2}$$

The value of regression coefficient m, and y-intercept b is computed by taking partial derivative of R^2 and setting to 0

$$m = n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i / n \sum_{i=1}^{n} ({x_i}^2) - \left(\sum_{i=1}^{n} x_i \right)^2 \tag{3}$$

$$b = \sum_{i=1}^{n} {x_i}^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i / n \sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i \right)^2. \tag{4}$$

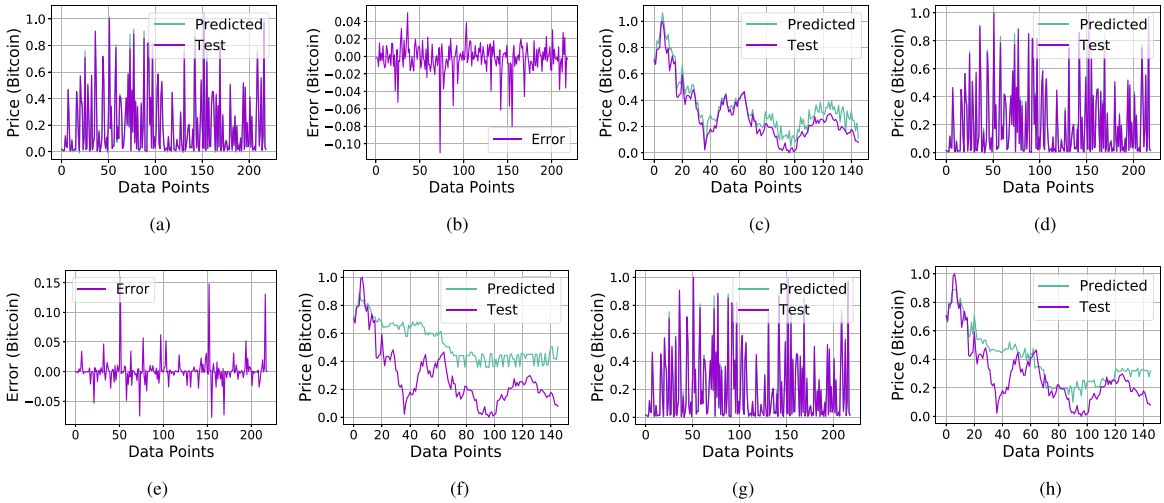

Fig. 8. Results obtained from regression model applied on Bitcoin dataset with 30% test data. Notice in (a) high similarity in prediction and test values indicate high accuracy. Also notice that random sampling always achieved a higher accuracy than design-based sampling. Due to low accuracy in design based sampling as shown in (c), (f), and (h), there is a significant difference in the predicted and test price. (a) Predicted and test values (LR). (b) Error in predicted/test val (LR). (c) Design-based sampling (LR). (d) Predicted and test values (RF). (e) Error in predicted/test val (RF). (f) Design-based sampling (RF). (g) Predicted and test values (GB). (h) Design-based sampling (GB).

The approach of using LR for modeling has been widely adopted in many applications. De Cock *et al.* [34] introduced a protocol for performing LR over a dataset in multiple parties. Roy *et al.* [35] propose predict financial market behavior based on a LR model.

2) Gradient Boosting: GB uses residual fitting to minimize the loss function and improve the accuracy. The loss functions, root-mean-square error (RMSE) and mean absolute error (MAE) are defined as

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} (y_i - y_{i_p})^2}$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - y_{i_p}|}{n}. \tag{5}$$

where y_i is ith target value, y_{i_p} is ith prediction value. To minimize loss function value, gradient descent approach is used to update predictions based on a learning rate, α

$$y_i^p = y_{i_p} - \alpha * 2 * \sum_{i=1}^{n} (y_i - y_{i_p}). \tag{6}$$

GB allows updating the prediction values so that the sum of the remainders is minimum and the predicted values are close to the actual values. This approach is used for many applications as well. For example, Alonso *et al.* [36] research the wind energy prediction problem using Gradient Boosted Regression. Zhang *et al.* [37] propose GB regression tree method to improve travel time prediction.

3) Random Forest: RF is one of supervised learning algorithms that builds multiple decision trees and to make precise predictions [38]. RF creates random subsets of the features by drawing bootstrap sample Z^* of size N from training data and growing a RF tree T_b using these subsets recursively. It outputs the ensemble of trees $\{T_b\}_1^B$ and makes prediction over a new point x with regression using $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$. RF is robust against outliers and avoids overfitting. Various structures are predicted in the literature using this approach. Lin *et al.* [39] show that the prediction of wind speed and direction using RFs. Sadeghi-Mobarakeh *et al.* [40] use RF model to predict the values in the electricity market.

For our first experiment, we formulated our problem as a multiple regression model based on highly correlated features in the dataset. We applied the random sampling method for data division and trained the model on LR, RF regression, and GB. We changed the percentage of training and test data in for each regression model and evaluated the performance using regression coefficient R^2, RMSE, and MAE, defined in (2), and (5). We applied both random sampling and design-based sampling [41] for each regression model. In design-based sampling, we split the dataset into 80% training data and 20% test data based on the time series. In other words, we trained data from April 2016 to January 2018, and predicted results from January 2018 to May 2018. We report our results in Figs. 8 and 9. We found higher accuracy in Ethereum price prediction compared to Bitcoin. We also noticed that compared to random sampling, the design-based sampling achieved lower accuracy. This can be observed in Fig. 8(c), (f), and (h), as well as in Fig. 9(c), (f), and (h), which show a big difference between the predicted curve and the test curve. Unlike random sampling, the design-based sampling does not capture characteristics of data that lies in the unknown regions. It is similar to using past indexes to predict the future. Therefore, it does not lead to high accuracy. For more details about random and design-based sampling, we refer the reader to [41]. To further investigate the accuracy of prediction we varied the percentage of test data from 5% to 50% and noticed the change in the accuracy and error. We report our results in Table II. From our experiment, we made the following observations. 1) LR achieved highest accuracy and low error in Bitcoin dataset, followed by the RF and GB, respectively. 2) GB achieved highest accuracy in the Ethereum dataset, followed by the RF and LR. 3) As the percentage of the training data decreased, the accuracy decreased and the error increased. 4) The maximum accuracy achieved in Bitcoin dataset was 0.9957 with 10% test data. 5) The maximum accuracy achieved in the Ethereum dataset was 0.9999 with 10% test data.

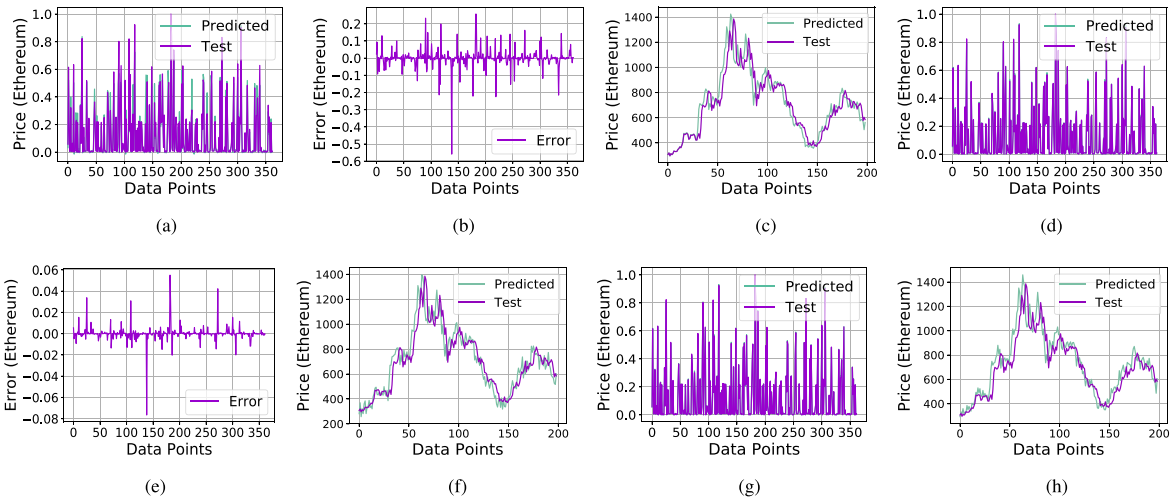

Fig. 9. Results obtained from Ethereum dataset. Notice that unlike Bitcoin, GB and RF achieve higher accuracy of prediction with low error compared to LR. Design-based sampling achieves lower accuracy than random sampling, however, it is higher compared to Bitcoin. (a) Predicted and test values (LR). (b) Error in predicted/test val (LR). (c) Design-based sampling (LR). (d) Predicted and test values (RF). (e) Error in predicted/test val (RF). (f) Design-based sampling (RF). (g) Predicted and test values (GB). (h) Design-based sampling (GB).

TABLE II
RESULTS OBTAINED FROM REGRESSION MODELS APPLIED ON BITCOIN AND ETHEREUM DATASETS WITH VARYING TEST DATA PERCENTAGE

	Test Data (%)	Linear Regression			Random Forest			Gradient Boosting		
		R^2	RMSE	MAE	*R^2*	RMSE	MAE	*R^2*	RMSE	MAE
Bitcoin	5	0.9937	0.0207	0.0143	0.9970	0.0141	0.0072	0.9968	0.0146	0.0076
	15	0.9956	0.0175	0.0121	0.9933	0.0215	0.0108	0.9924	0.0228	0.0108
	25	0.9949	0.0179	0.0117	0.9914	0.0234	0.0105	0.9924	0.0221	0.0105
	35	0.9951	0.0170	0.0112	0.9893	0.0251	0.0109	0.9927	0.0206	0.0100
	50	0.9952	0.0162	0.0106	0.9899	0.0235	0.0105	0.9933	0.0191	0.0094
Ethereum	5	0.9559	0.0486	0.0289	0.9999	0.0028	0.0014	0.9999	0.0022	0.0012
	15	0.8897	0.0718	0.0316	0.9984	0.0088	0.0026	0.9996	0.0041	0.0016
	25	0.8964	0.0651	0.0298	0.9981	0.0087	0.0027	0.9992	0.0056	0.0017
	35	0.9113	0.0593	0.0267	0.9978	0.0093	0.0028	0.9994	0.0050	0.0017
	50	0.9277	0.0563	0.0262	0.9972	0.0110	0.0031	0.9995	0.0049	0.0016

LR performs best with 10% test data while gradient decent and RF perform best with 5% test data.

6) Design-based sampling always achieved lower accuracy (a maximum of 0.901) compared to the random sampling. 7) In design-based sampling, LR outperformed gradient boost and the RF. 8) There is a more linear relationship among the Bitcoin features with its price, compared to Ethereum.

B. LSTM Approach

LSTM units are units of RNNs that can be used for prediction by keeping a continuous set of data for a long time [42]. RNNs constructed from LSTM units are also called an LSTM networks, and are popular for making predictions based on time series data. For prediction purposes, three types of deep learning approaches are typically used, recurrent neural networks (RNNs), convolutional neural networks (CNNs), and autoregressive integrated moving average (ARIMA). Our choice of RNNs over other alternatives is driven by the results obtained in the literature. In one such work [15], it has been shown comparatively how RNNs and ARIMA perform in predicting Bitcoin price. In particular, it is noted that RNNs significantly outperformed ARIMA and achieved higher accuracy. Moreover, another work has shown that CNNs are well suited to perform predictive analysis on image or text-based samples [43]. To this end, and because RNNs (particularly, LSTM-based RNNs) are more suitable to our goal of prediction, we use them in this article.

Technically, an LSTM consists of memory cells where each cell is composed of three gates, a forget gate, an input gate, and an output gate. The gates are responsible for managing the information of each cell. The forget gate layer determines the information transfer based on the results of the sigmoid layer, f_t, where W is weight, b is bias, and $\circledcirc$ is element-wise vector product as defined below

$$f_t = \sigma(W_f \circledcirc [h_{t-1}, x_t] + b_f). \tag{7}$$

Input gate layer and tanh layer decide the nature of the information to be stored in the cell. The sigmoid layer determines the value to update (i_t) and the tanh layer creates a new candidate, $\widetilde{C}_t$, that is the state value of the cell

$$i_t = \sigma(W_i \circledcirc [h_{t-1}, x_t] + b_i) \tag{8}$$

$$\widetilde{C}_t = \tanh(W_C \circledcirc [h_{t-1}, x_t] + b_C). \tag{9}$$

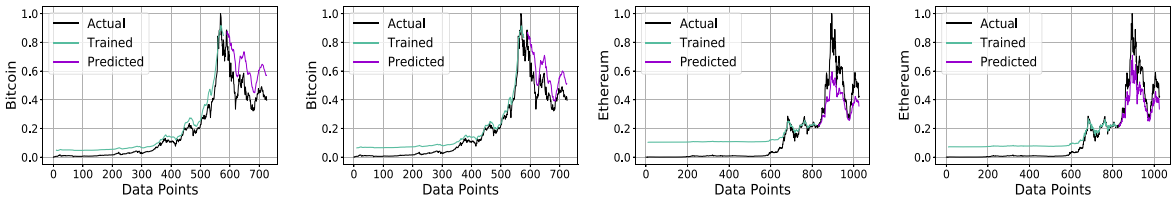

Fig. 10. LSTM cell overview. C_t, h_t, x_t, f_t, i_t, and o_t denote state of the cell, hidden state, input, output of forget gate, output of the input gate, and output of the output gate. C_{t-1} and h_{t-1} denote previous cell's state. (a) Epoch = 10 (Bitcoin). (b) Epoch = 30 (Bitcoin). (c) Epoch = 10 (Ethereum). (d) Epoch = 30 (Ethereum).

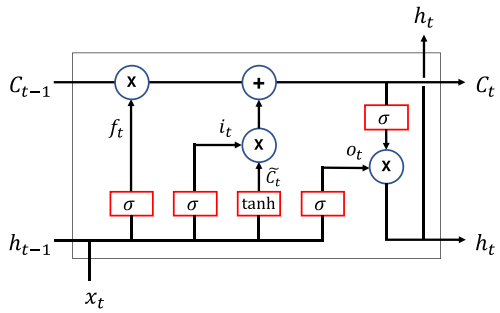

Fig. 11. Results obtained by applying LSTM networks prediction over the dataset. Note that prediction over Bitcoin is more accurate than Ethereum. (a) Epoch = 10 (Bitcoin). (b) Epoch = 30 (Bitcoin). (c) Epoch = 10 (Ethereum). (d) Epoch = 30 (Ethereum).

The current state of the cell can be calculated by multiplying the old state of the cell, C_{t-1} by the result of the forget gate, f_t, and adding the result of $i_t * \widetilde{C}_t$, which is the product of the output of the input gate and new candidate values

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t. \tag{10}$$

The hidden state, h_t, based on the state of the cell that is revised, is obtained by selecting the parts of the cell state to be output at the output gate, o_t, and putting the current cell state into the tanh layer and multiplying by the result of the sigmoid layer. The formula for computing o_t and h_t is defined below in

$$o_t = \sigma(W_o \circledcirc [h_{t-1}, x_t] + b_o) \tag{11}$$

$$h_t = o_t * \tanh(C_t) \tag{12}$$

while the model of this LSTM networks is illustrated in Fig. 10. We used LSTM approach on our datasets of Bitcoin and Ethereum to build a price prediction model. Similar to our methods in LR, we used min–max normalization on dataset features and selected the features with a correlation factor greater than 0.6. We split the dataset into 80% training and 20% test subsets, and varied the number of epochs to observe the change in the prediction model. We set the batch size (subset size of training sample) to 1 and the look back value to 1. The look back value is the number of previous time steps to be utilized as input variables for prediction of the next time period. We tested various look back values (1–5 and 10–50), and chose 10 for our experiments based on the performance. We report our results in Fig. 11. Our results indicate that the error values, captured by RMSE and MAE in test data for Bitcoin were low at 50 epochs (0.11 and 0.095), while the error values in test data for Ethereum were low at 30 epochs (0.13 and 0.1091). For the train data, the error values decreased as the number of epochs increased for both Bitcoin and Ethereum. In Table III and IV, we enlist the values of RMSE and MAE for training and test data obtained from our experiments. The results show that with LSTM, Bitcoin achieves higher accuracy with minimum error on each epoch. This also validates our results obtained in regression analysis.

TABLE III
RESULTS OBTAINED FROM LSTM MODEL USED ON BITCOIN DATASET

Epochs	Train Data		Test Data	
	RMSE	MAE	RMSE	MAE
10	0.05	0.042485	0.17	0.163407
20	0.05	0.045599	0.14	0.135538
30	0.05	0.046075	0.13	0.118998
40	0.05	0.044236	0.12	0.106008
50	0.04	0.040621	0.11	0.094958

The results show that with 50 epochs RMSE and MAE for test data was minimum.

TABLE IV
RESULTS OBTAINED FROM LSTM MODEL USED ON ETHEREUM DATASET

Epochs	Train Data		Test Data	
	RMSE	MAE	RMSE	MAE
10	0.09	0.08287	0.15	0.123918
20	0.07	0.06937	0.14	0.114183
30	0.06	0.057834	0.13	0.109132
40	0.05	0.050322	0.14	0.113484
50	0.05	0.043871	0.15	0.125366

The results show that with 30 epochs RMSE and MAE for test data was minimum.

C. Conjugate Gradient Approach

We also built a neural network and used conjugate gradient algorithm with linear search for price prediction. We normalize and split the data into 20% test and 80% training subsets. We train our network on 100 epochs and compute the training and validation errors. For this model evaluation, if the training and validation errors are high, the model is considered to be underfitting, and overfitting otherwise. In our experiment, we found the training error for Bitcoin was 0.00013, where the corresponding validation error was 0.00089. For Ethereum, the training and validation errors were found to be 0.00026 and 0.00095, respectively. From this experiment, we notice that the error, while small, is slightly higher than the training error. Such a model is considered to be a good fit and we report our results in Fig. 12. For comparison, we also used the Hessian gradient

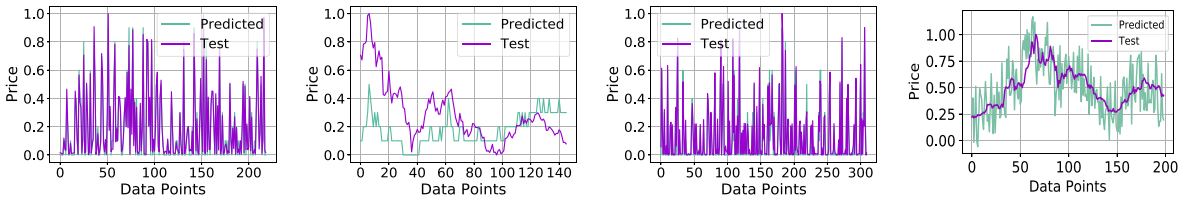

Fig. 12. Results obtained from neural network for Bitcoin and Ethereum. With random sampling, the accuracy was as higher than design-based sampling with low training and validation error. Prediction accuracy was higher for Bitcoin dataset. (a) Predicted/test values (Bitcoin). (b) Design-based sampling (Bitcoin). (c) Predicted/test values (Ethereum). (d) Design-based (Ethereum).

decent optimization which reduces training and validation error at a faster rate by choosing second derivative information for better gradient direction. However, the overall margin of error with Hessian algorithm was more than the conjugate gradient's.

VI. Conclusion and Future Work

In this article, we look into analyzing cryptocurrency market price through a correlation analysis with various cryptocurrency attributes, exemplified by Bitcoin and Ethereum. We collect data spanning more than 20 months and estimate the most significant features that influence the price. We computed the correlation between features such as hash rate, number of users, transaction rate, total bitcoins, and price. We map the change in features on users and network activities to understand the dynamics of the cryptocurrencies. We used our findings to construct a machine learning model that accurately predicts Bitcoin and Ethereum prices with the minimum error rate, based on other attributes than past price. Compared to the previous work that predicts Bitcoin price based on previous price observations, our approach is highly accurate.

References

[1] P. M. Krafft, N. D. Penna, and A. S. Pentland, "An experimental study of cryptocurrency market dynamics," in *Proc. Conf. Human Factors Comput. Syst.*, Montreal, QC, Canada, Apr. 2018, p. 605. [Online]. Available: http://doi.acm.org/10.1145/3173574.3174179

[2] A. Sonewane, "Top 10 cryptocurrency 2017 | best cryptocurrency to invest," 2017. [Online]. Available: https://goo.gl/D1cafv

[3] G. Hileman and M. Rauchs, "Global cryptocurrency benchmarking study," Cambridge Centre Alternative Finance, Cambridge, U.K., 2017.

[4] K. Sedgwick, "Statistics that reveal growing demand for the cryptocurrency," 2017. [Online]. Available: https://goo.gl/yK5dyh

[5] B. Community, "Bitcoin block explorer - blockchain," 2018. [Online]. Available: https://blockchain.info/

[6] M. Saad, L. Njilla, C. A. Kamhoua, and A. Mohaisen, "Countering selfish mining in blockchains," in *Proc. 2019 Int. Conf. Computing, Networking and Communications (ICNC)*, Feb. 18-21, 2019, doi: 10.1109/ICCNC.2019.8685577.

[7] A. Ahmad, M. Saad, M. Bassiouni, and A. Mohaisen, "Towards blockchain-driven, secure and transparent audit logs," in *Proc. Int. Conf. Mobile Ubiquitous Syst.: Comput., Netw. Serv., MobiQuitous*, New York City, NY, USA, Nov. 2018, pp. 443–448. [Online]. Available: https://doi.org/10.1145/3286978.3286985

[8] M. Swan, *Blockchain: Blueprint for a New Economy*, 1st ed. Sebastopol, CA, USA: WO'Reilly Media, Inc., 2015. [Online]. Available: https://goo.gl/7wck2T

[9] A. E. Kosba, A. Miller, E. Shi, Z. Wen, and C. Papamanthou, "Hawk: The blockchain model of cryptography and privacy-preserving smart contracts," in *Proc. 37th IEEE Symp. Secur. Privacy*, San Jose, CA, USA, May 2016, pp. 839–858. [Online]. Available: https://doi.org/10.1109/SP.2016.55

[10] A. Norta, "Creation of smart-contracting collaborations for decentralized autonomous organizations," in *Proc. Int. Conf. Bus. Inform. Res.*. New York, NY, USA: Springer, 2015, pp. 3–17.

[11] B. Community, "Crypto 2.0 comparison spreadsheet," 2017. [Online]. Available: https://www.reddit.com/r/CryptoCurrency/comments/2921em/created_a_crypt% o_20_comparison_spreadsheet_and/

[12] V. Buterin *et al.*, "A next-generation smart contract and decentralized application platform," White Paper, vol. 3, p. 37, 2014.

[13] N. Indera, I. Yassin, A. Zabidi, and Z. Rizman, "Non-linear autoregressive with exogeneous input (NARX) bitcoin price prediction model using PSO-optimized parameters and moving average technical indicators," *J. Fundam. Appl. Sci.*, vol. 9, no. 3S, pp. 791–808, 2017. [Online]. Available: https://goo.gl/9pojUX

[14] K. Kohara, T. Ishikawa, Y. Fukuhara, and Y. Nakamura, "Stock price prediction using prior knowledge and neural networks," *Int. Syst. Accounting, Finance Manage.*, vol. 6, no. 1, pp. 11–22, 1997.

[15] S. McNally, "Predicting the price of bitcoin using machine learning," Ph.D. dissertation, School Comput., Nat. Col. Irl., Dublin, Ireland, 2016.

[16] P. Vigna and M. J. Casey, *The Age of Cryptocurrency: How Bitcoin and the Blockchain Are Challenging the Global Economic Order*. New York, NY, USA: Macmillan, 2016. [Online]. Available: https://goo.gl/tTJN2j

[17] P. Marc *et al.*, "Blockchain technology: Principles and applications," 2016.

[18] C. Rose, "The evolution of digital currencies: Bitcoin, a cryptocurrency causing a monetary revolution," *Int. Bus. Econ. Res. J. (Online)*, vol. 14, no. 4, pp. 617–622, 2015.

[19] S. Omohundro, "Cryptocurrencies, smart contracts, and artificial intelligence," *AI Matters*, vol. 1, no. 2, pp. 19–21, 2014.

[20] G. W. Peters and E. Panayi, "Understanding modern banking ledgers through blockchain technologies: Future of transaction processing and smart contracts on the internet of money," in *Banking Beyond Banks and Money*. New York, NY, USA: Springer, 2016, pp. 239–278.

[21] A. Sapirshtein, Y. Sompolinsky, and A. Zohar, "Optimal selfish mining strategies in bitcoin," in *Proc. 20th Int. Conf. Financial Cryptography Data Secur.*, FC, Christ Church, Barbados, Feb. 2016, pp. 515–532. [Online]. Available: https://doi.org/10.1007/978-3-662-54970-4_30

[22] M. Vasek, M. Thornton, and T. Moore, "Empirical analysis of denial-of-service attacks in the bitcoin ecosystem," in *Proc. Financial Cryptography Data Secur. - FC Workshops*, Christ Church, Barbados, Mar. 2014, pp. 57–71. [Online]. Available: https://doi.org/10.1007/978-3-662-44774-1_5

[23] M. Saad and A. Mohaisen, "Towards characterizing blockchain-based cryptocurrencies for highly-accurate predictions," in *Proc. IEEE Int. Workshop Hot Topics Pervasive Mobile Online Social Netw.*, Honolulu, HI, USA, Apr. 2018, pp. 704–709. [Online]. Available: https://doi.org/10.1109/INFCOMW.2018.8406859

[24] H. Jang and J. Lee, "An empirical study on modeling and prediction of bitcoin prices with Bayesian neural networks based on blockchain information," *IEEE Access*, vol. 6, pp. 5427–5437, Dec. 2018.

[25] Etherscan, "The ethereum block explorer. ethereum charts and statistics," 2018. [Online]. Available: https://etherscan.io/charts

[26] R. Bohme, N. Christin, B. Edelman, and T. Moore, "Bitcoin: Economics, technology, and governance," *J. Econ. Perspectives*, vol. 29, no. 2, pp. 213–38, 2015.

[27] L. Wang and Y. Liu, "Exploring miner evolution in bitcoin network," in *Proc. Int. Conf. Passive Active Meas.*, New York, NY, USA, Mar. 2015, pp. 290–302. [Online]. Available: https://doi.org/10.1007/978-3-319-15509-8_22

[28] M. Saad, M. T. Thai, and A. Mohaisen, "POSTER: Deterring DDoS attacks on blockchain-based cryptocurrencies through mempool optimization," in *Proc. Asia Conf. Comput. Commun. Secur.*, Incheon, South Korea, Jun. 2018, pp. 809–811. [Online]. Available: https://goo.gl/4kgiCM
[29] M. Gronwald, "The economics of bitcoins–market characteristics and price jumps," *J. Econ. Perspectives*, 2014.
[30] M. Oberholzer, *Share Prices: Critical Perspective of the Greater Fool Theory*. Potchefstroom: Noordwes-Universiteit, Potchefstroomkampus (Suid-Afrika), 2010.
[31] B. Community, "Difficulty in Bitcoin," 2018. [Online]. Available: https://en.bitcoin.it/wiki/Difficulty
[32] P. Szilagyi, "Ethereum block validator," 2018. [Online]. Available: https://goo.gl/sBLoD4
[33] R. G. Ahangar, M. Yahyazadehfar, and H. Pournaghshband, "The comparison of methods artificial neural network with linear regression using specific variables for prediction stock price in Tehran stock exchange," 2010, *arXiv:1003.1457*. [Online]. Available: http://arxiv.org/abs/1003.1457
[34] M. D. Cock, R. Dowsley, A. C. A. Nascimento, and S. C. Newman, "Fast, privacy preserving linear regression over distributed datasets based on pre-distributed data," in *Proc. 8th Assoc. Comput. Machinery Workshop Artif. Intell. Secur.*, Denver, CO, USA, Oct. 2015, pp. 3–14. [Online]. Available: https://goo.gl/fr6Wti
[35] S. S. Roy, D. Mittal, A. Basu, and A. Abraham, "Stock market forecasting using LASSO linear regression model," in *Proc. Afro-Eur. Conf. Ind. Advancement*, Addis Ababa, Ethiopia, Nov. 2014, pp. 371–381. [Online]. Available: https://goo.gl/majFkf
[36] Á. Alonso, A. Torres, and J. R. Dorronsoro, "Random forests and gradient boosting for wind energy prediction," in *Proc. 10th Int. Conf. Hybrid Artif. Intell. Syst.*, Bilbao, Spain, Jun. 2015, pp. 26–37. [Online]. Available: https://goo.gl/RuDhbH
[37] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transp. Res. Part C: Emerg. Technol.*, vol. 58, pp. 308–324, 2015.
[38] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics. Berlin, Germany: Springer, 2009. [Online]. Available: https://doi.org/10.1007/978-0-387-84858-7
[39] Y. Lin, U. Krüger, J. Zhang, Q. Wang, L. A. Lamont, and L. E. Chaar, "Seasonal analysis and prediction of wind energy using random forests and ARX model structures," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 5, pp. 1994–2002, Sep. 2015. [Online]. Available: https://goo.gl/kjWEor
[40] A. Sadeghi-Mobarakeh, M. Kohansal, E. E. Papalexakis, and H. M. Rad, "Data mining based on random forest model to predict the california ISO day-ahead market prices," in *Proc. IEEE Power Energy Soc. Innovative Smart Grid Technol. Conf.*, Washington, DC, USA, Apr. 2017, pp. 1–5. [Online]. Available: https://goo.gl/pQHMHi
[41] A. Abadie, S. Athey, G. W. Imbens, and J. M. Wooldridge, "Sampling-based vs. design-based uncertainty in regression analysis," Preprint, 2017, *arXiv:1706.01778*.
[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[43] S. Mobin, B. Cheung, and B. A. Olshausen, "Convolutional vs. recurrent neural networks for audio source separation," 2018. [Online]. Available: http://arxiv.org/abs/1803.08629

Muhammad Saad is working toward the Ph.D. degree with the Department of Computer Science at the University of Central Florida (UCF), Orlando, FL, USA.

At UCF, he is a member of the Security and Analytics Lab (SEAL) advised by Prof. Aziz Mohaisen. His research spans blockchain with emphasis on their attack surface.

Mr. Saad work has appeared in reputable venues including Association for Computing Machinery (ACM) ASIA Conference on Computer and Communications Security (ASIACCS) 2018, Hot Topics in Pervasive Mobile and Online Social Networking (HotPOST2018), and Distributed Ledger of Things (DLoT 2018). He won the Best Paper Award at DloT 2018, for his work on blockchain-based audit logs.

Jinchun Choi received the B.Eng. and M.S. degrees from the Inha University, Incheon, South Korea, in 2011 and 2014, respectively. He is working toward the Ph.D. degree with the Department of Computer Science, the University of Central Florida, Orlando, FL, USA, and with the Department of Computer Information Science of Inha University, Incheon, South Korea (joint Ph.D. program).

He is a member of the Global Research Lab on Big Data Security and is conducting research in the field of information security. In particular, his research interests include biometrics, network security, user authentication and Internet of Things security.

DaeHun Nyang received the Ph.D. degree in computer science from Yonsei University, Korea, in 2000.

He is a Full Professor with the Department of Computer Science and Engineering, Inha University, Seoul, South Korea. His research interests include cryptography, privacy, usable security, network security, and system security.

Dr. Nyang is a member of the board of directors and the editorial board of the Korean Institute of Information Security and Cryptology, and a Section Editor for the ETRI-Journal.

Joongheon Kim (M'06–SM'18) received the B.S. and M.S. degrees from Korea University, Seoul, Korea, in 2004 and 2006 respectively, and the Ph.D. degree from the University of Southern California (USC), Los Angeles, CA, USA, in 2014.

He has been an Assistant Professor with Chung-Ang University, Seoul, Korea, since 2016. In industry, he was with LG Electronics, Seoul, Korea, 2006–2009; InterDigital, San Diego, CA, USA, 2012; and Intel Corporation, Santa Clara, CA, USA, 2013–2016.

Dr. Kim was a recipient of the Annenberg Graduate Fellowship with his Ph.D. admission from USC (2009) and the Haedong Young Scholar Award (2018).

Aziz Mohaisen (SM'12) received the M.Sc. and Ph.D. degrees from the University of Minnesota, Minneapolis, MN, USA, in 2012.

Before joining UCF in 2017, he was an Assistant Professor with SUNY Buffalo (2015–2017) and a Senior Research Scientist with Verisign Labs (2012–2015). He is currently an Associate Professor with the University of Central Florida (UCF), Orlando, FL, USA, where he only directs the Security and Analytics Lab (SEAL). His research interests include the areas of networked systems and their security, online privacy, and measurements.

Dr. Mohaisen is an Associate Editor for IEEE TRANSACTIONS ON MOBILE COMPUTING, and is a Senior Member of ACM (2018).