**Data Analytics and Data Driven Decision**

Academic year 2017-2018

**University of L’Aquila, Italy**

**Cryptocurrencies**

Ebrima M. Bah - 254960

Fabio Giancristofaro - 255146

Hilmi nur Ardian - 255449

Stefan Ilic - 252212

Table of Contents

[1. Introduction 4](#_Toc516237607)

[1.1 What is a cryptocurrency? 4](#_Toc516237608)

[1.2 Why this data? 4](#_Toc516237609)

[2. Description of dataset 5](#_Toc516237611)

[3. Exploratory analysis 7](#_Toc516237612)

[3.1 Correlations Analysis 7](#_Toc516237613)

[3.2 NVT Bubble prediction 9](#_Toc516237614)

[4. Unsupervised learning 10](#_Toc516237619)

[4.1 Silhouette analysis 10](#_Toc516237620)

[5. Supervised learning 12](#_Toc516237621)

[5.1 Metcalfe’s law 12](#_Toc516237622)

[5.2 LSTM Neural network 13](#_Toc516237623)

[Conclusions 15](#_Toc516237625)

# Introduction

## What is a cryptocurrency?

Cryptocurrency is a kind of [digital currency](https://en.wikipedia.org/wiki/Digital_currency), [virtual currency](https://en.wikipedia.org/wiki/Virtual_currency) or [alternative currency](https://en.wikipedia.org/wiki/Alternative_currency). Cryptocurrencies use decentralized controlas opposed to centralized [electronic money](https://en.wikipedia.org/wiki/Electronic_money) and [central banking](https://en.wikipedia.org/wiki/Central_bank) systems. The decentralized control of each cryptocurrency works through [distributed ledger](https://en.wikipedia.org/wiki/Distributed_ledger) technology, typically a [blockchain](https://en.wikipedia.org/wiki/Blockchain), that serves as a public financial transaction database.

[Bitcoin](https://en.wikipedia.org/wiki/Bitcoin), first released as open-source software in 2009, is generally considered the first [decentralized](https://en.wikipedia.org/wiki/Decentralization) cryptocurrency. Since then, over 4,000 *altcoin* (*alternative coin*) variants of bitcoin have been created.

Cryptocurrencies are used primarily outside existing banking and governmental institutions and are exchanged over the Internet. While these alternative, decentralized modes of exchange are in the early stages of development, they have the unique potential to challenge existing systems of currency and payments. As of April 2018, total market capitalization of cryptocurrencies passes 400 billion USD.

The validity of each cryptocurrency's coins is provided by a [blockchain](https://en.wikipedia.org/wiki/Blockchain). A blockchain is a continuously growing list of [records](https://en.wikipedia.org/wiki/Record_(computer_science)), called *blocks*, which are linked and secured using [cryptography](https://en.wikipedia.org/wiki/Cryptography). Each block typically contains a [hash](https://en.wikipedia.org/wiki/Cryptographic_hash_function) pointer as a link to a previous block, a [timestamp](https://en.wikipedia.org/wiki/Trusted_timestamping) and transaction data.

## 1.2 Why this data?

We choose this dataset because is one of the most discussed topic in the last period. Lot of people is trying to study the trade of these coins, others are mining them. In the month of December 2017, the price of Bitcoin reached 20000 USD, an incredible value especially if we consider that 2 years before the price was near 300 USD.

Cryptocurrency is beginning to indirectly impact traditional exchanges, too. Interest in cryptocurrency and the Blockchain is high enough that any mention of these technologies can [cause a surge in share price](http://www.telegraph.co.uk/technology/2018/01/22/telecoms-minnows-shares-rise-130pc-adding-blockchain-name/).

The aim of this project is to show the trend of cryptocurrencies in the last 3 years, trying to answer to the most asked question: how the value of a cryptocurrency changes?

The most significant information is the value of Bitcoin, the father of all the other cryptocurrencies. We will use Bitcoin as reference for all other cryptocurrencies.

# Actually, there aren’t relevant economic data directly connected with the cryptocurrencies trades but we will try to use different way in order to show some good results.

# Description of dataset

Our dataset is composed by different cryptocurrencies information. The main coin, used as reference, is the Bitcoin. We tried to select as other coins the oldest (each coin has data at least from 2015), since they contain complete and accurate data.

Each coin has an abbreviation used in the notebook.

* Bitcoin (BTC)
* Decred (DCR)
* Dogecoin (DOGE)
* **Ethereum Classic** (ETC)
* Ethereum (ETH)
* Litecoin (LTC)
* Monero (XMR)
* Ripple (XRP)
* Dash (DASH)
* Vertcoin (VTC)

Since we are working with differente coins, we pull the data from one source. We’ll use [coinmarketcap.com](https://coinmarketcap.com).

All coins have the same 7 columns:

* date
* txVolume
* txCount
* marketcap
* **exchangevolume**
* generatedcoins
* fees

First, we have the **date**. We use the gregorian calendar with the format dd/mm/yyyy. Daily closes for our price quotes occur at 00:00 UTC.

Next, we have **txVolume(USD).**That’s what we’re talking about when we say “on-chain transaction volume.” Simply put, it’s a broad and largely unadjusted measure of the total value of outputson the blockchain, on a given day. This is an answer to the question “approximately how much value, denominated in USD, circulates on the Bitcoin blockchain a day?”.

The third column is **txCount.**That refers to the number of transactions happening on the public blockchain a day. For low-fee blockchains, it’s really easy to fabricate a whole bunch of transactions. Additionally, UTXO networks like Bitcoin can batch a whole bunch of transactions into one, so **txCount**underestimates those ones.

Next, **marketcap(USD).**This is of course the unit price multiplied by the number of units in circulation. Marketcap or network value is definitely flawed. It becomes less tethered to reality the smaller the float is. Float means the ratio of actual circulating units to the total number of units. Ripple, for instance, has a fairly small float, so one should probably be skeptical of its “market cap.”

**Price** is the opening price. We get it from CoinMarketCap.

**exchangevolume(USD)**is the dollar value of the volume at exchanges like GDAX and Bitfinex. It doesn’t include data on OTC exchanges, which is a meaningful portion of all global exchange.

Next, **generatedCoins.**This refers to the number of new coins that have been brought into existence on that day. We count up the actual number of newly-minted coins, rather than using the stated inflation figures (i.e. for bitcoin you should expect 12.5 per block, every ten minutes, giving you 12.5\*6\*24 = 1800 coins per day).

Lastly, **Fees**. Fees in our data are based on the native currency, not USD. Transaction fees for cryptocurrency depend mainly on the [supply](https://en.wikipedia.org/wiki/Supply_and_demand) of network capacity at the time, versus the [demand](https://en.wikipedia.org/wiki/Supply_and_demand) from the currency holder for a faster transaction. The currency holder can choose a specific transaction fee, while network entities process transactions in order of highest offered fee to lowest.

# Exploratory analysis

## Correlations Analysis

When the price of one cryptocurrency rises, another falls almost in lockstep.

This is highly apparent in the cryptocurrency world, as many of the currencies on the market are completely coupled with the movement of Bitcoin. Certain coins will move incredibly close to Bitcoin, or act as a hedge against falling prices. Ripple, instead, is the one with less correlations with all the other coins, this because it works in a different way and it’s not minable.

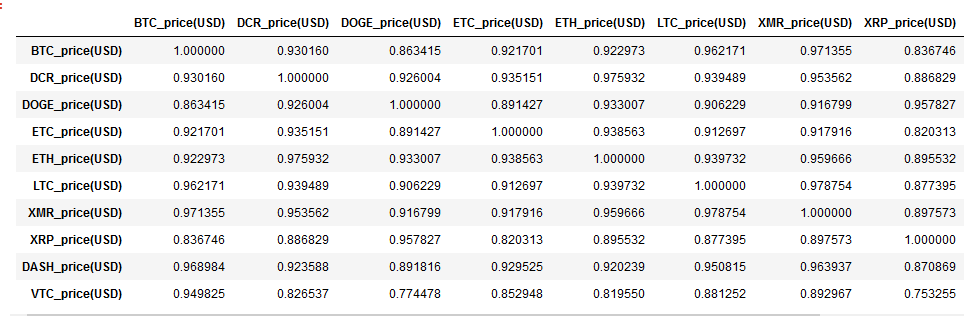


Figure Correlation matrix

In the financial world, correlation is a statistical measure of **how two securities move in relation to each other.**

Currency correlation, then, tells us whether two currency pairs move in the same, opposite, or completely random direction, over a designated period of time.

Unless you plan on trading just one pair at a time, it’s crucial that you **understand how different currency pairs move in relation to each other.**

**Correlation Coefficient**

Correlation is computed into what is known as the correlation coefficient, **which ranges between -1 and +1.**

* Perfect positive correlation (a correlation coefficient of +1) implies that the two currency pairs will move in the same direction 100% of the time.
* Perfect negative correlation (a correlation coefficient of -1) means that the two currency pairs will move in the opposite direction 100% of the time.

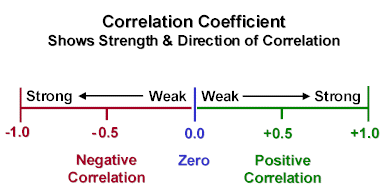


Figure Correlation Coefficient

We can see from the Scatter plot and the table, Correlation coefficients are on a high interval, which means that BTC and other coins have strong correlation and We can conclude that the prices behave similar**.**

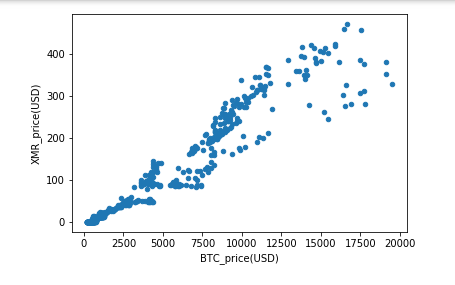
****

Figure 3Price correlation between XMR and BTC

## NVT Bubble prediction

# Over the course of the last year, a new study of cryptoeconomic ratio analysis emerged. The main idea behind this new field is to study the relationship between price of a cryptoasset and its fundamentals. One of the most widely known ratios is **Network Value to Transactions** or **NVT**. Some studies said that NVT can be successfully used to detect bitcoin price bubbles when valuation is not supported by fundamentals and differentiate them from consolidations.

https://cdn-images-1.medium.com/max/800/1*lIkoMsyDVeE5_krTNlCFCg.gif

# Trying to improve this ratio, we can say that Ratio has been smoothed using moving averages, 14 day forward and 14 day backward facing, that means:

https://cdn-images-1.medium.com/max/800/1*xfrGbkxX9uu8s-N-nJjcxw.gif

* 28 MA is “28-day Moving Average”
* NV is “Network Value in USD” (BTC Marketcap)
* TV is “Transaction Volume in USD”

# Also experiments have been done with different Moving Average periods and they came to an empiric conclusion that the optimal solution is to **divide daily Network Value by 90 days Moving Average of Transaction Volume**. So here’s a definition of new NVT ratio:

https://cdn-images-1.medium.com/max/800/1*T80ZvYvbrDveZiJMmpUQQQ.gif

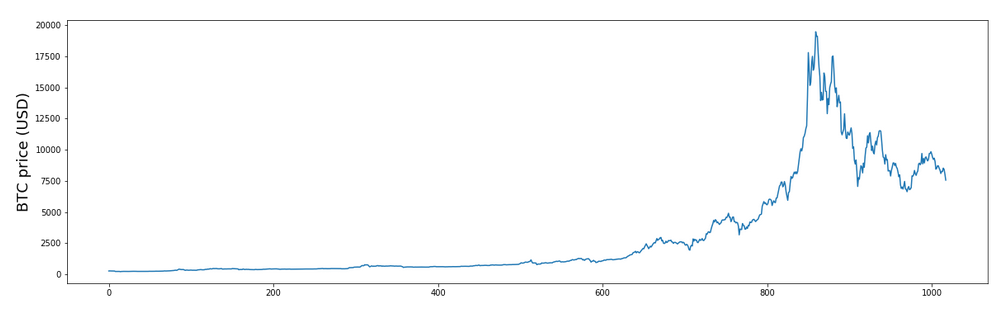


Figure BTC price trend

# 

Figure NVT\_28\_avg Ratio and NVT\_new Ratio

# Unsupervised learning

## 4.1 Silhouette analysis

This can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients as these values are referred to, near +1 indicate that the sample is far away from the neighbouring clusters.

In our case, for this data we used Silhouette to find the optimal value between an interval [2, 20], out of this we obtained an optimal value (ie Silhouette coefficient) of 0.510570786696885, which corresponds to k = 2. This value represents the number of clusters.

Grouping set of objects in a way that objects in a way that objects in a group are similar than objects in another groups.

We use the k-mean algorithm clustering, where the number of clustering K is a parameter of the the algorithm. Assume the spherical shapes(ball-like) clusters, and we computed the Euclidean distance, with an initial centroid from the clusters.

)2

Where: ci is the collections of the centroids in the set C, and and x being the data points.

We standardized the data and ploted on the xy-plane the BTC\_Price(USD) and LTC\_Price(USD) clustering results.

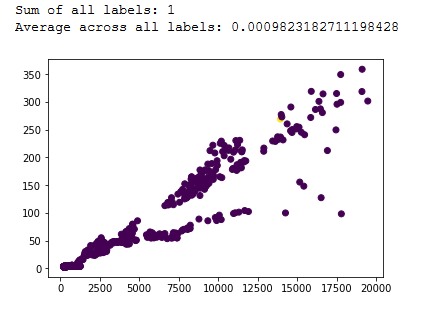


Figure Result of clustering

# Supervised learning

## 5.1 Metcalfe’s law

Metcalfe's Law states that the value of a network is proportional to the square of the number of users on the network. The classic example is a fax machine: a fax machine is useless by itself but is very useful if a few of your friends have one. If the number of fax machine user’s doubles, the value of the network increases exponentially.

[Fundstrat](http://www.businessinsider.com/bitcoin-price-movement-explained-by-one-equation-fundstrat-tom-lee-metcalf-law-network-effect-2017-10#ampshare=http://www.businessinsider.com/bitcoin-price-movement-explained-by-one-equation-fundstrat-tom-lee-metcalf-law-network-effect-2017-10) had the idea to try and apply this formula to the price of bitcoin. Unique bitcoin addresses are used as a proxy for number of network users. He also added to this formula the number of transactions per user. Fundstrat found a formula by regressing the price of bitcoin against both unique addresses squared and transaction volume per user. This model explained 94% of the variation in the cryptocurrency price since 2013.

*BTC* = *x \* n* 2 *\***t/n*

Where X is a constant, n is number of addresses, and t is transactions

In this notebook, we will try to replicate these results, trying to modify the main formula. We will use the transaction volume instead the t/n value.

*BTC* = *x* \* *n* 2 \* *v*

Price prediction is done in different way: in the first method the prediction is done from the entire dataset and then it is compared with the actual price. In the second, the dataset is splitted in two parts and the prediction is completed starting from the first part.

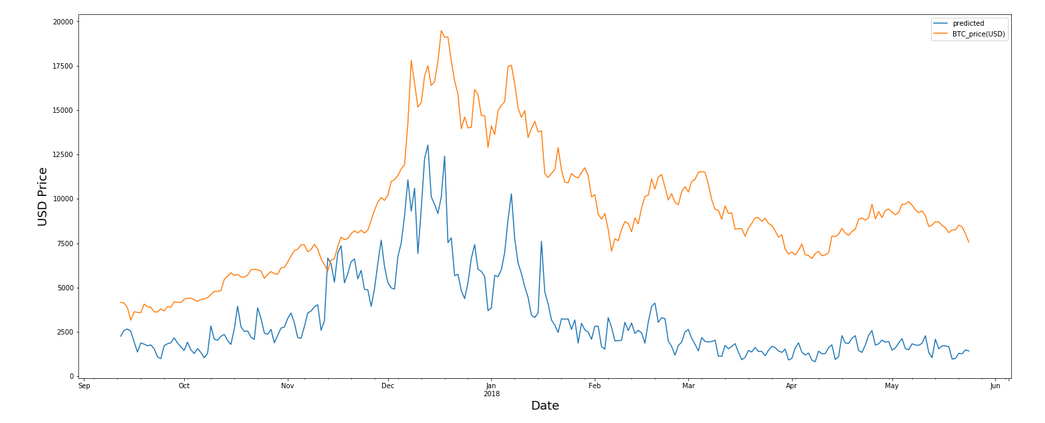


Figure 7 Metcalfe's law Using one coin and splitting

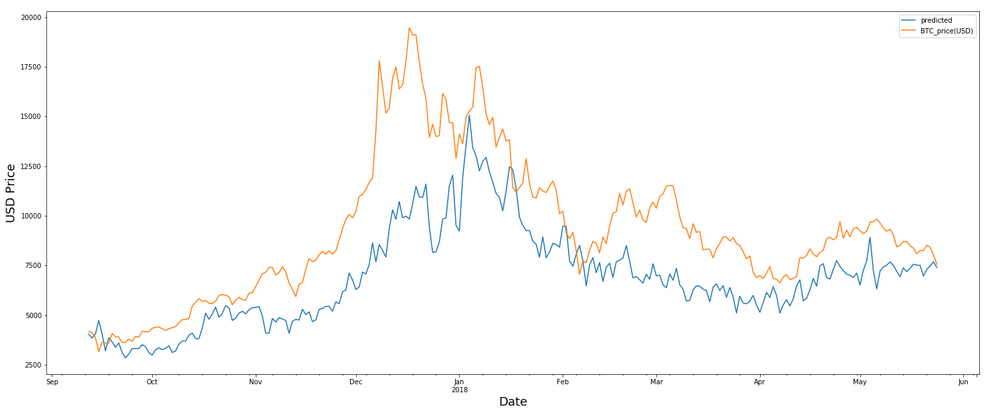
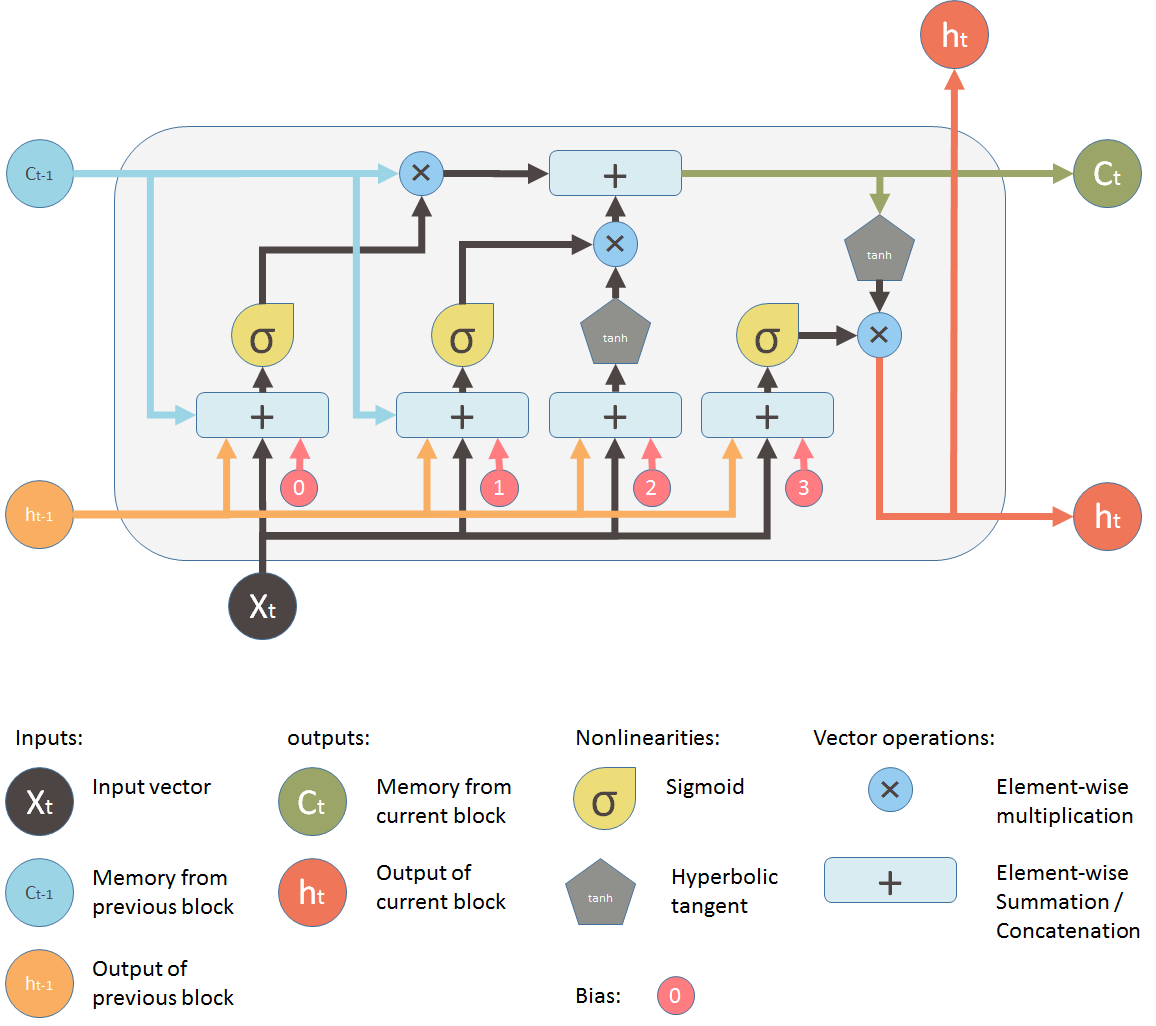


Figure 8 Metcalfe's law Using ten coins and splitting

## 5.2 LSTM Neural network

Long Short Term Memory networks are a special kind of Recurent Neural Networks (RNN), capable of learning long-term dependencies. They were introduced by Hochreiter & Schmidhuber (1997). LSTM are networks used in sequence prediction problems. They introduce a memory unit, called the cell into the network. This is the diagram of a LSTM building block.



In our case LSTM network is used to predict future bitcoin prices. To train this network we have split our dataset in training dataset consist of 75% of data and testing dataset consisting of rest of the data. We split our training data in small arrays (windows) of length 8 and we fed them to neural network to train giving the bitcoin price as a result. We trained our network 50 epochs with a batch size of 1. LSTM network trained in this way gave us following result on our testing data set:

# index

# Conclusions

# In exploratory analysis we have seen that coins are strongly correlated. The scattered plot and tables the correlations coefficient has a range [0.7, 1], which we can conclude that there is a strong correlation. We see that due strong correlation when one coin increases in price, the others also increase.

# The data set was analysed by different statistical too. The supervised and unsupervised learning mechanism, a general discussion is argued.

There is no single indicator that can accurately predict the price of BTC (or others) as there are too many variables to consider. However, if we accept the premise that blockchain networks that are predominantly in the speculative stages of adoption behave like online telecommunications networks, then Metcalfe may help us to better understand where usage and price intersect and when one has significantly outpaced the other.

The prediction has been done starting from one coin (BTC), then we tried to add other coins till 10. We can see that prediction with Metcalfe’s law improve comparing more than one coin but there isn’t a big improvement after adding more than 3 coins.

The unsupervised learning, in which we used the k-means clustering algorithm. We used silhouette score to find optimal number of cluster, which in our case was 2. But due strong correlation between coins we see that algorithms in unable to separate them, clustering all results in a single cluster.

In supervised learning, we also used LSTM neural network trying to predict bitcoin price trend. In results obtained we see that network gives decent prediction but that it suffers from lag. One of possible improvements would be to use stateful LSTM-s.

In conclusion we can say that, cryptocurrency market forecasting is a difficult topic and that it requires more complex prediction models.