Data Analytics for the Cryptocurrencies Behavior

Eduardo Sánchez✉, Jose A. Olivas and Francisco P. Romero

Dept. of Information Technologies and Systems
University of Castilla-La Mancha, Ciudad Real, Spain

Eduardo.Sanchez00@outlook.es✉
{JoseAngel.Olivas, Franciscop.Romero}@uclm.es

Abstract. The cryptocurrencies are a new paradigm of transferring money between users. Their anonymous and non-centralized is a subject of debate around the globe that paired with the massive spikes and declines in value that are inherit to an unregistered asset. These facts make difficult for the common daily use of the cryptocurrencies as an exchange currency as instead they are being used as a new way to invest. What we propose in this article is a system for the better understanding of the cryptocurrencies economical behavior against the global market. For that we are using Data Analytics techniques to build a predictor that uses as inputs said external financial variable. These forecasts would help determine if a coin is safe to trade with, if those forecasts can be precise by only using this external data. The results obtained indicates us that there is a certain degree of influence of the global market to the cryptocurrencies, but that is it not enough to correctly predict the fluctuations in price of the coins and that they care more about others factors and that they have their own bubbles, like the crypto collapse in late 2017.

Keywords: Cryptocurrencies, Data Analytics, Blockchain

1 Introduction

The cryptocurrency world is one of the most fascinating and unique paradigms that modern society is facing. It aims to change how money is generated and exchanged between people through what is called a blockchain. The main difference with classic currencies is the decentralization, most of the crypto coins are not associated to a single entity, corporation or country, although there are examples of the contrary. Said decentralization carries a degree of uncertainty, there is nothing behind the coin and the value of it depends entirely on the uses of the holders of the currency. It is also in a grey legal area, with more and more countries making steps for a more regulated crypto market.

So even though the original idea looks good in paper, there are many caveats left in the air that makes the trading with cryptocurrencies unsafe and unattractive for the

The final authenticated version is available online at http://doi.org/10.1007/978-3-030-27713-0

general public because sometimes it feels that nobody has control on the value of them and overnight you could potentially lose a great part of your savings.

For this reason, we, the authors of this document, are going to develop a model to obtain knowledge and conclusions on the fluctuation on the price of the coins according to external economic circumstances. This way we can verify if the cryptocurrency behaviors go along the global economy or if they are independent, making them harder to predict than normal classic markets.

1.1 Data Analytics and Big Data

Data Analytics is the process of studying raw data sets for the extraction of knowledge [4][6][8]. There are many directions and techniques in which we can approach a data analytics project and one of them is what is called Big Data.

Big Data is none less than the treatment of large groups of data that are usually obtained in real time and need quick treatment so they can be useful. Economical models, such as this one, take advantage of the Big Data approach since using a predictor in real time adds a lot of value to the decision making that the end users have to affront. Making long-term determinations in investments is not as important as making short-term ones, and for that we need quick and fluent data processing.

1.2 Cryptocurrencies

One of the most interesting and controversial finance paradigms in modern society is the cryptocurrency boom. Between its novelty using current technology and exceptional situation and uses, the crypto phenomenon is a trending topic in both general public opinion and in the scientific community [2] [5] [10].

A cryptocurrency is a digital asset that uses cryptography and other encryption techniques to regulate the generation of those said assets and the verification of the transactions. (Mining) Said coins are stored, sent and received through a Cryptocurrency Wallet.

The initial idea for this new type of currency was so it was decentralised from any government, company or entity, for that it is the user computers that make the verification of the transactions.

This is done by the blockchain technology, developed by an anonymous person or group that went by the alias of Satoshi Nakamoto [9] in 2009 for the developing of Bitcoin, the first cryptocurrency. This technology consists on the idea that everybody has a complete database of the blocks of the said blockchain. That way if everybody has the same data, said data must be true. This makes that all the transactions are public, but both the sender and the receptor are anonymous.

The workflow of the blockchain consists of that once a computer (node) receives a transaction it tries to solve a computationally difficult puzzle that once is done, if it is the first that solved it, it places the next block in the block chain and claims the rewards.

Those rewards are in fact a certain number of cryptocurrencies. The number of digital coins that provides as a reward depends on how many other nodes are mining that certain cryptocurrency and other factors that depends of the cryptocurrency.

There are a lot of different cryptocurrencies ready to be used. Alongside Bitcoin, the other most important and relevant cryptocurrency is Ethereum, but there is a plethora of smaller and less used coins that can be traded as well such as Monero, Ripple, Aion, Waves or Litecoin.

There have been some successes in the matter of correctly analysing and study the economical behaviour of the cryptocurrencies, such as an accurate forecasting on the price trend by the use of Gradient Boost Decision Trees [11] or a dissection of the bubbles in the Bitcoin history [12].

1.3 Workflow

This study explanation is divided by the following parts, starting in section 2. In 2.1 we are going to define what is this project and what is not. For 2.2 we are explaining the caveats on the transformation of the cryptocurrencies timeseries to a characterization that allows us to work with it easily. 2.3 corresponds to the methodology for choosing which coins that represent the total the best are we going to study. Feature selection of the external economical variables used in the next stage for prediction is done in 2.4, and the prediction in 2.5. Evaluation of this prediction is detailed in 2.6.

Conclusions of the study are in section 3 in which we talk about the discoveries of our investigation. Section 4 is Acknowledgements and section 5 References.

2 Analysis on the cryptocurrencies behaviour

The first step to take was the selection of what cryptocurrencies we were going to use to perform the study. This is a very important step since the goal is to obtain generalizations of the coins, not a case by case study. We knew that every coin was going to have different needs and different results, but if we could select a few ones that were representative of the conglomerate we can make conclusions of the general behavior of the cryptocurrencies. The selection was performed over 73 different coins which data was obtained from Coinmetrics.io, a webpage with the sole intent of providing raw data on cryptocurrencies and its related characteristics. Once we had the data, the selection was made as follows.

2.1 Scope of the model

There are a lot of details to consider, an almost impossible task to cover in a single project, that is why in this document we are focusing on the relation between global economic variables and the cryptocurrencies price fluctuation. Using Artificial Intelligence, Data Analytics and Machine Learning the goal is to obtain knowledge about how related the fluctuations of the digital currencies with the global economy are, helping the interested user to forecast if it is safe to hold onto these new types of coins, and if so which of the coins are the safest. For that we are going to use supervised and unsupervised learning in conjunction. With unsupervised learning we are segmenting the coins to select the most relevant ones and generalizations. Supervised learning completes this process by taking those coins and using Machine Learning algorithms construct a predictor for each of them to forecast the value of them in the near future. This model development is done by two Knowledge Discovery in Databases [7] processes that are done in order, not in parallel. The first one uses unsupervised learning to obtain this more relevant coins and the second one with supervised learning forecasts the value of the coins by only using external economic factors.

This knowledge is not aimed at investors, although it could be used by then, the goal is not to turn on a profit with this model, but to help final users make better decisions if they want to start exchanging money through this method.

2.2 Cryptocurrency characterization

A cryptocurrency database is made by several characteristics, called features, such as the price or market cap, that had a value registered for each day from the moment the currency was coined to the last day that we stipulated to stop retrieving data in order to have the same end for every coin. This form to express data is called time series, and it is just a linear way to store information. It is widely used in economics where time and seasonality matters. For this project we used days as our grain, but we could had use months, semesters or even years, the timeseries model is not only thought to work with days [1]. Also, not all the coins had the same structure, some had more technical data like the block size, while others lacked it. We had to drop features and in some cases databases in order to have the same structure, so we could start working. Nonetheless, this structure was not useful to select the most representative coins, so we had to do a transformation and convert the data into one that can help us on the task.

Said transformation consisted into making a characterization of each coin considering the overtime price increases and decreases. The new database with all the characterization of the coins had the following features or variable

name: Name of the coin.
mean_daily_price_variation: Mean of the daily price variation in %.

cv_daily_price_variation: Coefficient of variation (relative standard deviation) of the daily price variation in %.

ratio_of_price_increases: From -1 to 1 it indicates the number of days that were positive of negative for the coin, meaning that -1 indicates that all days were negative and 1 that every day was positive and 0 that there was an equal number of price increases and decreases.

linear_correlation_coefficient_price (Pearson's Correlation Coefficient): Measure of the strength of the linear relationship between number of days passed and price.

max_price_increase: Biggest increase for a day in %.

max_price_decrease: Biggest decrease for a day in %.

mean_active_addresses: Mean of the active addresses.

cv_active_addresses: Coefficient of variation (relative standard deviation) of the daily active addresses.

linear_correlation_coefficient_active_addresses (Pearson's Correlation Coefficient): Measure of the strength of the linear relationship between number of days passed and active addresses.

2.3 Representative coins

Now that we had all the coins defined by one line in a database, we applied algorithms to make segments out of the entire database, and from there take the most representative coins. In this case we had chosen Principal Component Analysis, a statistical procedure to project a set of points into a reduced number of new linear uncorrelated values, to transform all these features into 2 data projections that we can show in a graph and K-means clustering [3] to obtain a determined number of clusters out of them. K-means clustering is a method of vector quantization that aims to divide several samples into a specified number, k, of clusters. Each sample belongs to the cluster with the nearest mean. With this algorithm we settled the number of clusters that we considered to be the best and obtained the most representative coins out of them. The k number of clusters at the start were 5 but got reduced to 4 since we got better results and the selected coins were Bitcoin, Ethereum, 0x, Icon and Waves (figure 1). The first two were selected since they are the most important cryptocurrencies at the time of writing this document, while the other three were the centroids of the resulting clusters. We did not select the 4th centroid of the last cluster since it was a very small cluster with just 2 coins and we considered that it was not going to contribute anything useful.

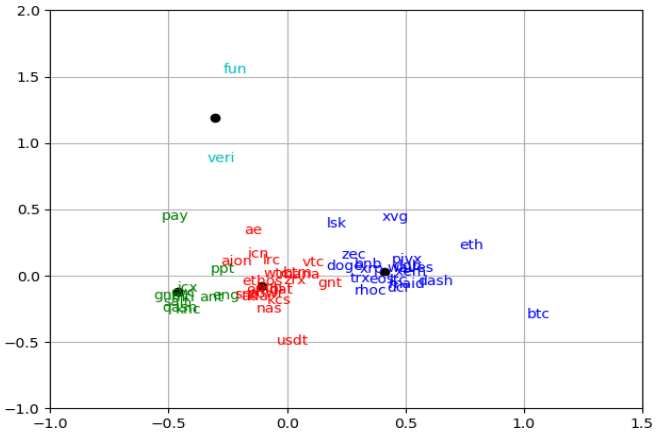

Fig. 1. KMeans clusters result with k = 4

2.4 Feature selection

Once we had the coins selected, we had to choose the external economic factors from which we are going to base our model. After some investigation we concluded that the factors that were best suited for the problem were the price of Oil, Gold, the valuation of the market indexes Nikkei225, Financial Times Stock Exchange 100 and Dow Jones Industrial Average and the exchange in dollars of the Euro and Japanese Yen. The structure of the databases that contain this data is very similar to the cryptocurrencies databases, a timeseries with the grain put in days but instead of multiple features there was only one, the value in dollars.

What we did next was to finally calculate the importance that each of the selected coins gave to the external economic features. For that we used a correlation test to prune the economic features that were very similar considering the linear correlation with the market cap of the cryptocurrency. If two external variables acted identically, that could cause overfitting in out model, so we had to delete one of them. The one that was deleted was always whichever had the least linear correlation with the market cap of the coin at study.

Once we did this initial feature selection, we used a Random Forest predictor to calculate the importance. Random forests or random decision forests are an ensemble learning method for both regression and classification which build the model by using a N number of decisions trees for the training and making the model by selecting a portion of them. A decision tree is a tree-like flowchart used for decision support where each of the branches is a decision and the nodes are states. It can store a large amount of information such as chances that certain decisions are taken, cost of them

and etcetera. Decision Tree learning implements that decision tree model to obtain conclusions about an item using the branches as observations. It is very useful specially to represent the decision-making process in data mining projects but can be also used as a predictor. Note that we are not using the Random Forest predictor to predict anything yet, we are only extracting the importance that it gives to each of the features while fitting the given data.

2.5 Behavior prediction with external data

To better validate that the obtained relevancies were on point, we developed a K-nearest neighbours model that considered the importance of each feature. For that we used a scaler to transform the numeric values of the global economic features into values from 0 to 1 and then multiplied then by the relevancy, which goes from 0 to 1 as well. After that transformation we operated like a normal predictor model, we parametrized the model by using a Cross-Validation test. Cross-validation (CV) is a statistical technique that help us determine the usefulness of a machine learning model. More in concrete, is a resampling procedure used to evaluate these models with preestablished data samples.

Once we parametrized the algorithm correctly, we divided the data into two sets, the training and test datasets. The training dataset is used to fit the model, to train it and adjust the predictor, and the testing dataset is used to validate and compare the predictions that the model can do with real data. Once the model is trained, we proceeded to forecast the price of the coins in 6 months for now, and that forecast was then compared with the testing dataset, obtaining the results showed in figures 2, 3, 4, 5 and 6.

Fig. 2. Bitcoin six months prediction against real data

In the Bitcoin prediction we can observe that it gets accurate the first weeks, but as time passes, the normal behavior of the coin would be to go up in value. What the predictor cannot take into account are things like the collapse of the cryptocurrencies in the fall of 2017 and early 2018, which is an event completely unrelated with global economics.

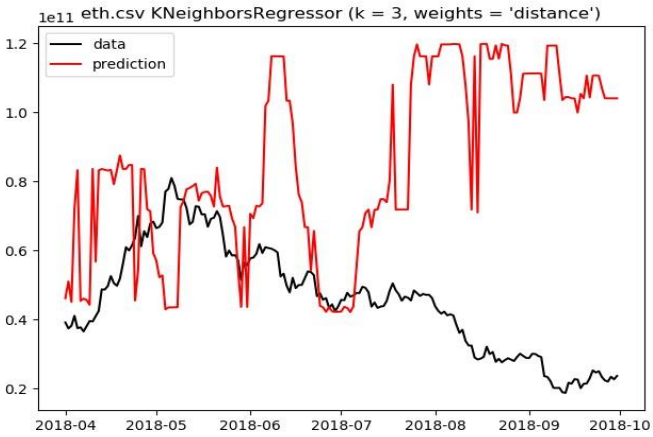

Fig. 3. Ethereum six months prediction against real data

Ethereum had the same type of result as Bitcoin, an overoptimistic prediction since the economical behavior of the coin did not follow the global market.

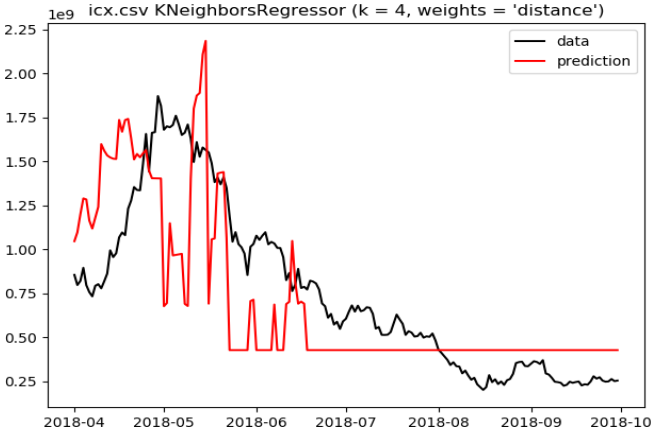

Fig. 4. Icon six months prediction against real data

Icon had an interesting prediction, it was not affected by the 2017 collapse like Bitcoin and Ethereum, this coin follows the global economy behavior much closer.

The prediction falls down to a flat line due to the lack of samples that have that low of a value. This means the predictor could not tell that a new low was coming for this coin.

Fig. 5. 0x six months prediction against real data

0x does not follow the global economy as closer as the other coins, the prediction started to fail around summer, the slower months for economy. This did not hinder the movement of 0x.

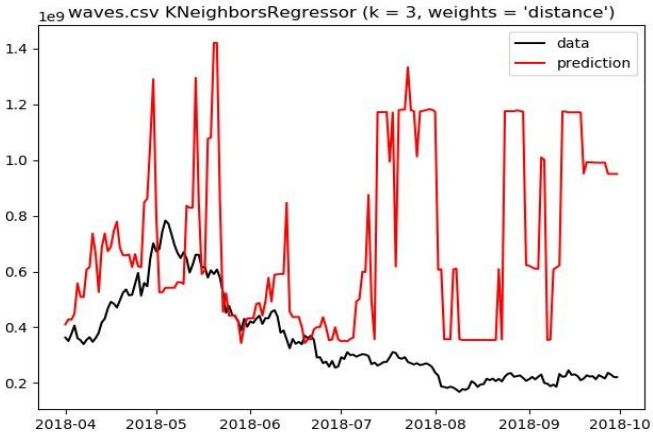

Fig. 6. Waves six months prediction against real data

Waves price does not follow the global market tendencies as we can see in the erroneous prediction.

2.6 Evaluation

The results obtained where then validated with another 3 different coins, Aion, Salt and Dash, each one from the 3 clusters resulting in the previous step. Said results were quite like their centroids counterparts so it proved that the predictor does not discriminate between the clusters of coins.

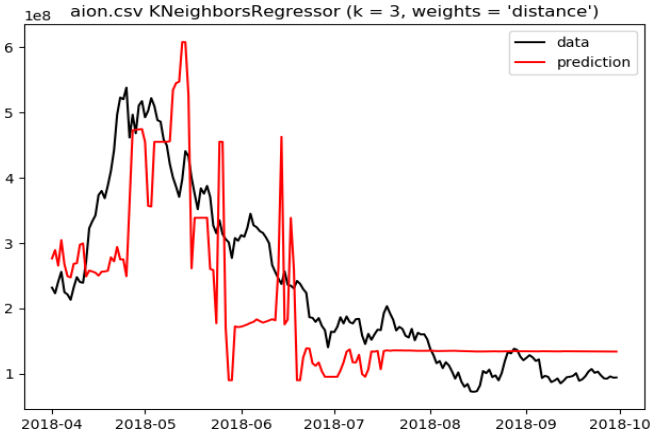

Fig. 7. Aion six months prediction against real data

Aion has similar behavior to Icon and 0x. The predictor can not estimate the new lows that a coin can have since it is out of the range of the training.

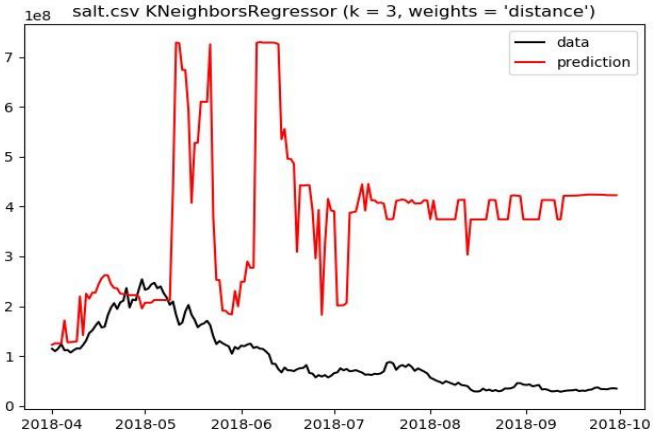

Fig. 8. Salt six months prediction against real data

As for Salt and Dash the predictor could not take into account the collapse of 2017, the same case as Bitcoin and Ethereum.

Fig. 9. Dash six months prediction against real data

Most of the predictions commence to err in the start of summer. This could mean that the predictor is not fit to forecast the price of a coin in a long-term situation of more than two or three months and that the economic slowness of these months do not affect the coins price.

3 Conclusions

What we learned is that both Bitcoin and Ethereum, the two largest and more known cryptocurrencies, follow the global market way closer than the rest of the samples. But at the same time, as we can see by looking at the predictions, these are also the two that are the hardest to predict with just these variables. This is not considered a contradiction, a coin can be very related to a certain feature, but without a global vision of the rest of the important features that affect it we cannot make a good estimation if the importance of the variables that we are studying is low.

What this really means is that even when those two cares a lot about the price of the oil, they are not conditioned by it and favor other features. Such features could be the number of miners currently mining the blockchain, the comments in social media and so on. It is as we explained earlier, from the 100% of the price, a Y percent is the economic external factors, from which that said percent, 80% is the value of the oil in dollars, but if the Y percent is really low, then it does not matter at all if the price of the crude has a lot of relevancy because in the grand scheme is not important at all.

The other three, smaller coins care a lot more about the economic features fluctuation since the prediction is more on point there, by a lot, and since they are varied in what features do, they give importance it indicates us that they are being traded by investors that also trade other assets. These cryptocurrencies could be used as a

scapegoat when one of the assets that those investors have capital in declines in value, while the blob of the investors of Ethereum and Bitcoin are enthusiasts or specialized capitalists that only care about those coins. This is not unusual, since they are the better-known ones that attracted a lot of attention from tech individuals and just common people that liked the idea of digital money, which never invested in the past. That also makes sense why the price of the crude is the most important feature for those two, since it is the better indicator of how the home economy go, the more money that the common

Acknowledgments

This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grant MERINET: TIN2016-76843-C4-2-R (AEI/FEDER, UE).

References

1. C. Chatfield. “Time-Series Forecasting”, Chapman & Hall/CRC, UK, 2001.
2. D. Wilson-Nunn and H. Zenil. “On the Complexity and Behaviour of Cryptocurrencies Compared to Other Markets”, arXiv:1411.1924 [q-fin.ST], 2014.
3. I. Sutskever, R. Jozefowicz, K. Gregor, D. Rezende, T. Lillicrap, O. Orio Vinyals. “Towards Principled Unsupervised Learning”, arXiv:1511.06440, November 2015
4. M. Rouse SearchEnterpriseAI. Retrieved from https://searchenterpriseai.techtarget.com/definition/AI-Artificial-Intelligence, November 2010.
5. M. Haferkorn and J. M. Quintana. “Seasonality and Interconnectivity Withing CryptoCurrencies – An Analysis on the Basis of Bitcoin, Litecoin and Namecoin”, Enterprise Applications and Services in the Finance Industry, 7th International Workshop, FinanceCom 2014, Sydney, Australia, 22 January 2016.
6. N. Elgendy and A. Elragal. “Big Data Analytics: A Literature Review Paper”, In: Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects. ICDM 2014. Lecture Notes in Computer Science, vol 8557. Springer, Cham, 2014
7. U. Fayyad, G. Piatetsky-Shapiro and P. Smyth. “The KDD process for extracting useful knowledge from volumes of data”, Communications of the ACM. 39 11, pages 27-34, November 1996.
8. P. Perner. “Advances in Data Mining. Applications and Theoretical Aspects: 14th Industrial Conference”, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014.
9. S. Nakamoto. “Bitcoin: A Peer-to-peer Electronic Cash System”, Retrieved from https://bitcoin.org/bitcoin.pdf, November 2008
10. Y. B. Kim, J.G. Kim, W. Kim, J.H. Im, T.H. Kim and S.J. Kang “Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies”, PLoS ONE 11(8): e0161197, 2016.
11. X. Sun, M. Liu, Z. Sima, A novel cryptocurrency price trend forecasting model based on LightGBM. Financ. Res. Lett. (2018), in press, available online 27 December 2018.

12. N. Aslanidis, A. F. Bariviera, O. Martinez-Ibañez, An analysis of cryptocurrencies conditional cross correlations. Financ. Res. Lett. 31, 130-137 (2019).