International Institute of
Information Technology, Bangalore

Business Understanding Document
DS 707 Data Analytics

Blockchain understanding and Cryptocurrency Analysis

Akanksha Dwivedi - MT2016006
Hitesha Mukherjee - MS2016007
Nayna Jain - MS2017003
Tarini Chandrashekhar - MT2016144

Instructors :
Prof. Ramanathan Chandrashekhar
Prof. Uttam Kumar

October 15, 2017

Contents

1 Determining Business Objectives

1.1 Background

Cryptocurrency (built over Blockchains), a mysterious new technology emerged seemingly out of nowhere, at its most fundamental level is a breakthrough in computer science – one that builds on 20 years of research into cryptographic currency, and 40 years of research in cryptography, by thousands of researchers around the world. It gives a way for one Internet user to transfer a unique piece of digital property to another Internet user, such that the transfer is guaranteed to be safe and secure, everyone knows that the transfer has taken place, and nobody can challenge the legitimacy of the transfer.
Blockchains (and the consensus protocols that support them) were invented as a result of developers trying to solve this bold problem of how to create digital, untraceable money. By combining cryptography, game theory, economics, and computer science, they managed to create an entirely new set of tools for building decentralized systems.

1.2 Business Goals

The business objectives of this project undertaking are:

- To understand/describe the sudden surge in interest in cryptocurrencies recently.
- To explore the volatile/unstable nature of the cryptocurrencies and co-relation between price fluctuations among them.
- To be able to predict the future prices of the cryptocurrencies.
- To identify and understand factors contributing to the overall behaviour of the cryptocurrencies, so that the prediction becomes easier.
- To identify fake or dangerous users, thereby preventing fraud.
- To grant cryptocurrency more legitimacy and thereby, greater adoption by performing in-depth analysis and pattern recognition across thousands of transactions, ensuring that users are protected.

1.3 Business Success Criteria

The success of our analytics endeavour depends on the value addition provided in terms of new information which the potential crypotocurrency adopters could benefit from and make use of, in their investment decisions.

- Correct prediction of future prices of blockchain tokens.
- Highlight fraudulent use and/or theft of cryptocurrency.

- Identification of factors leading to fluctuations in the currency evaluation of tokens.
- Conclusion of the better currency from the predicted trends and volatility.

1.4 Business benefits

- **Exploratory and Descriptive Analytics:** Based on analyzing the historical prices of different cryptocurrencies, we can predict the trends for the same, which will help potential investors make informed decisions.
- **Classification:** Identifying fraudulent transactions will help distinguish between legitimate and illegitimate transactions, preventing the case of a dishonest network and avoiding usage of the network for running scams.
- **Clustering:**
- **Association rules:** There can be established definite correlation between various factors and prices of the cryptocurrencies. This means the data can be analysed for frequent if-then relationships using the criteria of **support** and **confidence** to identify the most important relationships. This eventually leads to predict blockchain behaviour.

1.5 Target Users

Our analytics on cryptocurrencies will not only benefit the miners, who are actively engaged in the network but many other stakeholders who have actively been interested on the use of cryptocurrencies ever since its advent but have been holding back because of the lack of predictive information on its trends and patterns.

- The miners who validate transactions could benefit from the future prediction of a token price to know whether it's worth validating or not.
- Individual investors participating in a token sale would know the criteria before hand, to evaluate token sales.
- Consumers who pay with cryptocurrency which is evidently more stable.
- Merchants who accept cryptocurency will come to know about the fraudulent transactions, which will help them avoid and report those users.
- Entrepreneurs who are building new applications on top of blockchain technology can choose the more versatile blockchain technology based on the tokens used on top of it.

2 Assessing the situation

2.1 Inventory Of Resources

In the dataset we have the historical price information of some of the top cryptocurrencies by market capitalization. The currencies included are Bitcoin, Ethereum, Ripple, Bitcoin cash, Bitconnect, Dash, Ethereum Classic, Iota, Litecoin, Monero, Nem, Neo, Numeraire, Stratis, Waves.

Each currency has one csv file with the following attributes extracted from **coinmarketcap**. Price history is available on a daily basis from April 28, 2013. The columns in each of the csv file are:

- Date : Date of observation
- Open : Opening price on the given day
- High : Highest price on the given day
- Low : Lowest price on the given day
- Close : Closing price on the given day
- Volume : Volume of transactions on the given day
- Market Cap : Market capitalization in USD

25 Manually hand crafted features are available for Bitcoin and Ethereum currencies.

2.2 Requirements, Assumptions and Constraints

- The correlations calculated between non-stationary timeseries data are often spurious and are not representative of any actual correlation inherent between the data sets. So, we make the data stationary.
- We can ignore the cryptocurrencies which have comparatively less number of data points because of being relatively young.
- Our data has a constraint that it has more tokens than it has blockchains technologies, so at best, we can gauge more information about the tokens than about blockchains.
- We don't have individual block information in terms of network addresses, so performing anomaly detection and tagging fraudulent users/transactions will be a challenge.

2.3 Risks and Contingencies

Even though we conduct analytics to identify trends and predict prices, there are certain unforeseen risks and contingencies which we are completely unaware of:

- Even if we might have tagged a fraudulent transaction or user, it might be totally a false positive due to some other inexplicable error.
- The prices we predict might be subject to sudden unforeseen fluctuations due to a new world event, something our analysis might not account for.

2.4 Terminology

What is a 'Blockchain'? A blockchain is a digitized, decentralized, public ledger of all cryptocurrency transactions. Constantly growing as 'completed' blocks (the most recent transactions) are recorded and added to it in chronological order, it allows market participants to keep track of digital currency transactions without central record keeping. Each node (a computer connected to the network) gets a copy of the blockchain, which is downloaded automatically.

Originally, it was developed as the accounting method for the virtual currency Bitcoin, blockchains – which uses what's known as **distributed ledger technology (DLT)**. Currently, the technology is primarily used to verify transactions, within digital currencies though it is possible to digitize, code and insert practically any document into the blockchain. Doing so creates an indelible record that cannot be changed; furthermore, the record's authenticity can be verified by the entire community using the blockchain instead of a single centralized authority.

The blockchain technology and respective cryptocurrencies has gained popularity due to the following advantages:

- Efficiencies resulting from DLT can add up to some serious cost savings. DLT systems make it possible for businesses and banks to streamline internal operations, dramatically reducing the expense, mistakes, and delays caused by traditional methods for reconciliation of records. Electronic ledgers are much cheaper to maintain than traditional accounting systems; the employee headcount in back offices can be greatly reduced. Nearly fully automated DLT systems result in far fewer errors and the elimination of repetitive confirmation steps. Minimizing the processing delay also means less capital being held against the risks of pending transactions.
- Cryptocurrency uses a "push" mechanism that allows the cryptocurrency holder to send exactly what he or she wants to the merchant or recipient with no further information, thus preventing identity theft.
- Bitcoin contracts can be designed and enforced to eliminate or add third party approvals, reference external facts, or be completed at a future date

or time for a fraction of the expense and time required to complete traditional asset transfers, resulting in immediate settlement.

3 Determining Data mining goals

3.1 Data Mining Goals

- How did the historical prices / market capitalization of various currencies change over time?
- Predicting the future price of the currencies.
- Which currencies are more volatile and which ones are more stable?
- How does the price fluctuations of currencies correlate with each other?
- Seasonal trend in the price fluctuations.
- Market Trends visualization by constructing cross correlation maps between different cryptographic currencies.
- Blockchain Statistical Analysis of various fundamental factors affecting the network to draw basic inferences from the Bitcoin blockchain.

3.2 Data Mining Success Criteria

The process of data mining would be judged successfully from:

- The accuracy of different regression models (say specificity and precision) which predicts the sign of future change in price of cryptocurrencies at varying levels of granularity(eg 10 minute or 10 second interval time points). This evaluation metric measures the success percentage of our exploratory and predictive analytics on the historical data of cryptocurrencies.
- The R-square error of the predictive models estimating the change in prices or the future prices of the crypotcurrencies.
- Visualizations of the descriptive analytics done on various crytocurrencies and comparing individual cryptographic currencies with the overall market.
- Dual Evaluation metric (based on a combination of outliers in a user based graph and a transaction based graph) can be used to evaluate our identification of suspicious transactions and dubious users in the network.

4 Producing a Project Plan

4.1 Project Plan

Data mining can be defined as the extraction of implicit, previously unknown and potentially useful information from data. Machine learning provides the technical basis for data mining. In this project, we attempt to apply machine learning algorithms to predict Bitcoin price. Our data set consists of many features relating to the Bitcoin and payment network recorded over a period of time. Using this information we can predict the sign of the daily price change. Following steps can be taken during the course of the project:

- Firstly, cleaning and pre-processing of the data will be done along with interpolation of missing data.
- Regression and predictive models can be used to predict price change as well as future prices of the currencies.
- Unsupervised learning methods can be used for pattern recognition, which will lead to identify trends, as well as for anomaly detection,which will identify dubious transactions and users.
- Classification algorithms can be used to distinguish between legitimate and illegitimate transactions.

4.2 Initial Assessment of Tools and Techniques

We will use a combination of Python and R libraries to supplement our needs for machine learning in our analytics project. Various unsupervised and supervised algorithms used are available in the form of libraries. We can also use Tableau Tool for effective visualization.