POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Engineering and Management

Tesi di Laurea Magistrale

Open-High-Low-Close-Volume, Blockchain & Social Cryptocurrency Data Analysis

Relatore prof. Luca CAGLIERO ***Candidato Antonio DELL'ANNA***

A.A 2019-20

RINGRAZIAMENTI

"Puoi chiamarti dottore, puoi chiamarti scienziato,
Puoi chiamarti ufficiale, puoi chiamarti soldato
Puoi persino morire
Comunque l 'amore è la dove sei pronto a soffrire
Lasciando ogni cosa al suo posto e partire "

Cesare Cremonini

Vorrei in primis ringraziare infinatemente il Professore Luca Cagliero che ha accettato di seguire un "gestionale" su una tematica nuova e lontana dal suo percorso di studi, accompagnandolo in tutti gli step di questo lavoro.

Vorrei inoltre dire grazie al Dottorando Giuseppe Attanasio, che con una cordiale chiaccherata mi ha illuminato sull' implemetazione e funzionamento di SVC and MLP.

Probabilmente sembrerà una frase acchiappa likes, ma questo traguardo è interamente frutto dei miei genitori (V&G), che mi hanno indirizzato su binari ambiziosi, nonostante non goda di grandi talenti. Ringrazio vivamente mio fratello (S) per tutta la vicinanza in questi anni e anche per avermi fatto crescrere un pochino spalle e petto, garzie alle sue schede di allenamneto.

Una ragazza poi (F), ha deciso di credere in me da un bel po di anni. Credeva in me anche quando ero nel buio piu totale. È attraente e anche carismatica, ma piu di tutto riesce a capirmi, nonosstante le grandi differenze caratteriali.

Due sconosciuti invece(S&C) mi hanno aiutato costantemente in questi ultimi anni. Abbiamo vissuto grandi esperienze con grande intensità e feeling. Ora sconosciuti non lo sono piu'.

Grazie a tutta la mia famiglia, zii, cugini e amici storici. Mi avete aiutato a crescere e stare al mondo.

Un ringraziamento va ovviamente ai nonni che non mi hanno insegnato sicuramente a programmare in Python o fare un NPV, ma mi hanno insegnato a vivere ed andare avanti sempre. In particolare volevo dedicare parte di questo lavoro al mio nonnino, che probabilmete risponderà durante la proclamazione, convinto che quel Dell'Anna Antonio sia lui. Probabilmente lo farà, anche da lassu'.

Grazie

OHLCV, Blockchain & Social Cryptocurrency Data Analysis
INDEX

Figure List

CHAPTER 1:

CHAPTER 2:

CHAPTER 3:

CHAPTER 4:

CHAPTER 5:

Table List

Abstract

Cryptocurrency Market today counts a market capitalization of $207 Billion, with more than 3000 coins and where main dominant Cryptos, as BTC, ETH, LTC have reached a clear popularity on Social Networks such Twitter, Facebook, Reddit and GitHub. Hence, it is today an important financial reality that attracts a lot of risk lovers and digital coin users. Nevertheless, the ambiguity of Market Nature and the huge volatility makes this market complex and approach to Cryptocurrency analysis a stiff process.

It looks distant from exchange market, which appears stable and with low volatility level, and appears more similar with stock. Both in fact, present high degree of risk, but Crypto market results more fragile. All this makes price forecasting an interesting and complex game.

Looking at the actual State of Art, the most interesting trend is the application of several machine learning algorithms, such as simple and multiple Linear Regression, Support Vector machine (SVM), Multilayer Perceptron (MLP) to OHCLV financial data.

But the lack of seasonality and the continuous volatility drastically afflict models accuracy. Throughout the recent years, Sentiment Analysis has been involved into the Cryptocurrency price forecasting. It is a tool, based on Opinion Mining and Natural Processing Language that allows extracting polarity from Social Posts and Text, a good proxy of investor Sate of Confidence about Market. Most of works consider just Twitter sentiment and Google Trend with daily data sampling frequency.

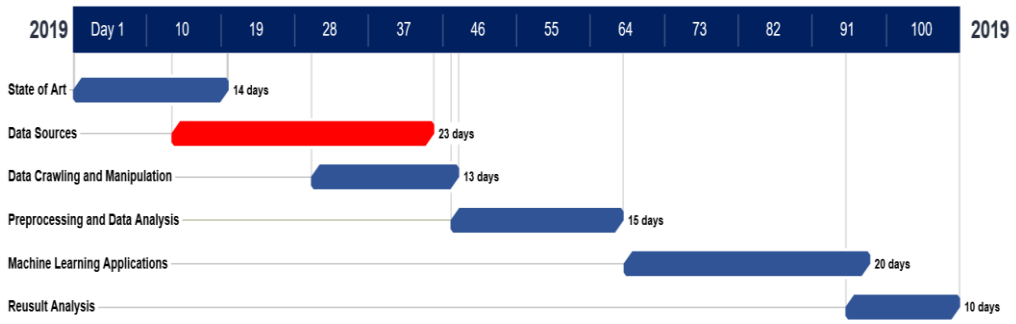
Today, few papers have inferred on Blockchain quantitative features as possible Price spread explanatory variables. Blockchain is the most underlying cryptocurrency technology and it is definable as a distributed, immutable and transparent ledger that allows emitting transactions stored by blocks. This innovation paradigm is impacting on several business areas, as Financial Transactions, Supply Chain and Politic, with a hype expectation that is touching the stars.

The scope of this work, is to explore the main Cryptocurrency Sources, and evaluating which kind of data is offered, with which granularity and time horizon and in which ways (REST APIs, Web Socket APIs, csv, excel adds-on).

Under this perspective, three kinds of data are stored: the OHLCV (Open, High, Low, Close, Volume) financial data, Social Data, including Facebook likes, Reddit posts, comments, GitHub activity and Blockchian data, as Block size in Byte, the number of Transactions, the Difficulty to add a new Block, the Miners Remuneration in USD.

Once Data Crawling is reached, the thesis proceeds inferring on the existence of possible correlation between financial data and Social and Blockchain data.
Finally, in order to evaluate the validity of the work done so far, two supervised classification algorithms (Multilayer Perceptron and Support Vector Machine) are trained on part of the datasets. The target class as defined as a sort of trading signals based on the Price Changing Value (Upper, Lower, No Signal), based on specific threshold (+/- 5%).
The work counts 5 chapters, that deeper explain the above steps and with takeaways, highlighting the fundamental concepts and results, reached in each chapter. In particular the thesis is scheduled as follows:

- *Chapter 1: it is a Cryptocurrency Market overview, where the main features are analyzed and evaluated. The chapter concludes with the Actual State of Art description;*
- *Chapter 2: It is an overview of the Blockchain technology, approaching to it main features and business areas involved in the revolution that it is pursuing;*
- *Chapter 3: It presents the Data Science steps, involved in this thesis work and then describes the Sources Data available for Cryptocurrency Market, which have selected for the next chapters empirical analysis;*
- *Chapter 4: This chapter starts the empirical phase of the thesis. It analyzes the correlations between Closing Price for five Coins and Social and Blockchain variables.*
- *Chapter 5: Multilayer Perceptron and Support Vector Machine applied to Blockchian daily data, and evaluation of Accuracy as forecasting performance;*



Abstract Fig: Thesis Gant Chart Representation

1. Cryptocurrency Market and Price Forecasting

Cryptocurrency Market is a youth Financial Market, with features and behaviors distant from traditional markets, as stock and exchange. Before evaluating them from quantitative perspective, it is fundamental to understand their meanings, so that it is possible to interpret Data Analysis result and extract knowledge from them. The first part of the chapter focuses on Financial Market features and theories and Cryptocurrency features are illustrated.

The last paragraph exposes the actual State of Art about Cryptocurrency Price Forecasting, evaluating which are the most diffused techniques and tools.

1.1. Financial Markets and Underlying Theories

Financial Market is the marketplace where financial operators can trade securities, such as stock, bond, forex, derivatives. Initially, it was a physical place, where different actors interest to trade met, but with the introduction of modern ICT it is now a virtual platform, where intermediaries, called brokers ensures the exchange of financial assets at the best market price. The price of securities traded in the markets, in fact, does not necessarily reflect their intrinsic value and this represents an important driver of trading strategy that can beat the market itself.

The stock market is probably the most known market and it allows investors to allocate their savings, buying or selling shares of publicly traded companies. Under this view, it is important to distinguish between primary market and secondary market. The former is the market place, where the company decides to sell shares for the first time through an Initial Public Offering (IPO) in order to achieve new external capital. The latter, enables the trade of each stock subsequent to IPO, involving intermediaries as brokers.

Another important market is the bond market, where an investor loans a fixed amount of money for a defined time interval and with a fixed interest rate. This kind of security is called Bond and it can be issued by corporations, states, sovereign governments and municipalities. Generally, Bond market is with low risk, and it is also called fixed-income market.

There exist then, a market where financial instruments, such as options and futures are traded. They are secondary securities, since their value depends on underling assets as stocks or bonds.

Derivatives play today an important role in hedging and speculative strategy and it largely involved in corporate financial strategy.

The forex market is instead, the marketplace where people can exchange, buy, sell and speculate on currencies. It is the most liquid market and it handles amount as 5 trillion of dollars per day, more than equity and derivatives markets together.

It is also crucial to consider the presence of market not regulated, the over the counter (OTC) market.

It is a decentralized, without a physical location in which investors trade securities without the presence of broker. Generally, companies traded on OTC present reduced sizes due the lest costs and regulation required into this kind of market.

There are two significant theories that allows analyzing the markets, Efficient Market Hypothesis (EMH) and Random Walk Theory (RWT).

The former asses that price of security immediately reflects the complete market information and when news come out, the market instantly corrects the security price, making arbitrage opportunities impossible. Although its lack of testability, today EMH represents the basic theory for Asset Pricing, as defined in the paper "Evaluating Sentiment in News Article" [12].

Random Walk Theory is similar to the previous one, so that all information is involved in the current price, but adds that short time price movement follows a completely random activity, es explained by Nadin Gullodin and Rand Shott in [13]. Hence, past movements or trends are not useful to predict the next ones.

Contrary to these theories in June 2002, Blake LeBaron, a researcher associate at the National Bureau of Economic Research and member at Santa Fe Institute, test the impacts of new information on financial market. He built an artificial stock market of simulated traders where trading decision could be manipulated. This led to discover that a lag period between the introduced information and when the market adjusts price exists. This delay supports the idea through which the market can be foresee following the introduction of new information and contrasts the instantaneously price adjustments theory.

On the basis of LeBaron studies, Gidofalvi arrived to demonstrate that exists twenty-minutes time opportunity before that market correct the security price with news.

Trading techniques, such as fundamental and technical analysis base their principles on the assumption of existing of this lag, so that it is possible to beat the market.

1.2. Fundamental and Technical Analysis

Fundamental and Technical analysis represent today the most known and spread trading techniques, but the principles behind them are completely different.

FA infer on the intrinsic values of an asset, looking at the fundamentals of an asset, defined as every kind of aspects that afflict its overall value. The pragmatic concept linked to FA is that, if it is possible that an asset has intrinsic value not reflected from its current market price, it is convenient to invest on it. In order to pursue this, FA analyzes public available information about securities, taking into account three main sources: overall Economy conditions, Industry scenario and Company fundamentals. The former suggests economic indicators as inflation, GDP, growth rate and interest rate, whereas the second takes into account factors such as government attitude, foreign entrants, cost structure. The last source is the company analysis where company growth rate, financials, competitive advantage, management quality, market shares are evaluated. The Financials are generally available on the Balance Sheet an Income Statements documents.

This step, evaluates also several financial ratios such as:

- Earnings per Share (EPS): It is defined as the ratio between Earnings and the Number of Outstanding Shares and allows to evaluate the evolution of the share prices among time;
- Price on Earnings (P/E): it is computed as the ration between Price of single share and EPS and it permit to evaluate how the price is close to the earnings per share. If P/E ratio is high, means that marketers expect the price continues to growth. Contrary, implies they expect to see a fall;
- Return on Asset (ROA): This ratio defined as (Net Income + Interest expense)/ Total Assets, allows evaluating how a company is managing its tangible assets;
- Return on Equity (ROE): it is the ratio between (Earnings/ Shareholder Equity) and describes how the shareholders' money are used;
- Market Capitalization: it reflects the global market value, due the formula (price* number of outstanding shares);

Looking at Crypto Market, it is possible to observe how most of Digital Coins projects do not follow traditional stocks companies, hence in this case FA seeks about factors as Target market, Competitors, Partnership and Community. But, being FA interested to infer on intrinsic asset value by evaluating its financial statements, it appears not the ideal solution for Cryptocurrency Market.

Differently form FA, Technical Analysis is not interested to trace the underlying asset value, but it approaches historical market prices and volumes in order to foresee future movements and trends.

Therefore, TA, concerns with capability to track past asset data and through statistical and pictorial tools to evaluate future patterns and trading signals.

The underlying principle of this approach, how explained by Jhoin J. Murfy in [14], is that price of security already contains all available information and it is based on three assumptions:

I. Price movements follow trends;

II. Market discounts everything;

III. Past patterns tend to repeat itself.

TA approach data through three different tools, the charting lines that illustrates the points of price reactions, Patterns, which shows the presence of framed movements and Indicators.

Indicators are statistical tools that allow to point out two kind of signals.

The overbought is the signal that an asset is traded above its intrinsic value and it defines the best time to sell it since it results expensive and a pullback is forecasted.

The oversold is instead the situation where the asset is traded below its value and it is the best time to buy the asset due the possibility to gain a good return.

In the below BTC/USD chart (Fig 1.1), Simple Moving Average, Relative Strength Index and MACD indicators are pictured.

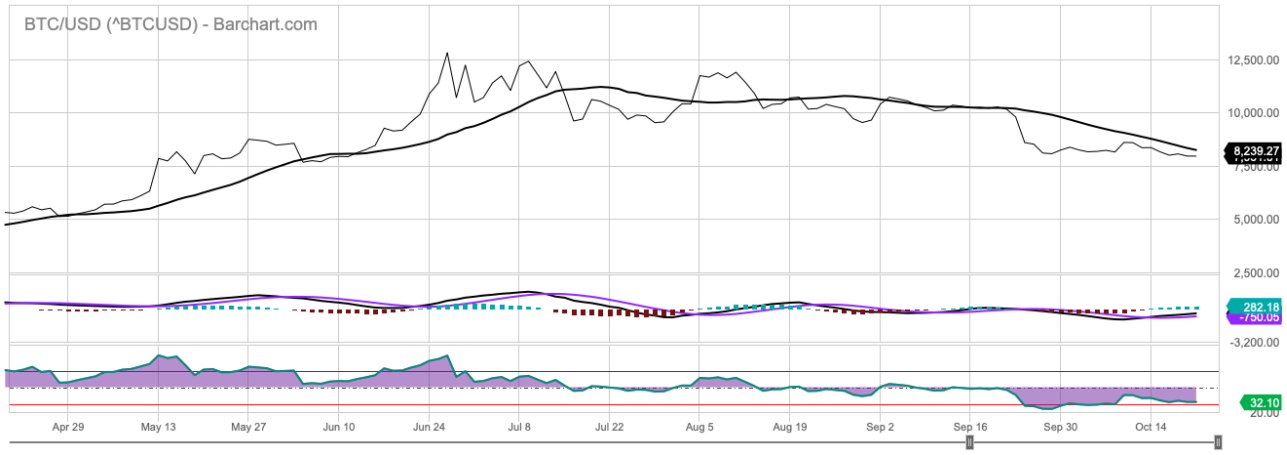


Fig 1.1: Chart of BTC price series exchanged in USD and representation of Technical Indicators as Moving Average, RSI and MACD from Barchart.co

The simple moving average is probably the most known indicator. It is the sum of the Closing price in a time interval [T_1, T_n] divided for the number of times taken into account (T_n –T_1). The direction of SMA line is a proxy of market trend. There would be a positive trend, if SMA is above price line, whereas it would be negative in the opposite case, how it is possible to observe in the first portion of

chart. In the second part, it is instead pictured the Moving Average Convergence/ Divergence (MACD), which is a momentum indicator. Final Formula requires three Exponential Moving Average, one in 12 days, EMA_12, one in 26 (EMA_26) and the last in 9 days, EMA_9 and two lines to evaluated the trend. The former is given by the difference of first two Moving Average (EMA_12 – EMA_26) and is called MACD line, whereas the latter coincides with EMA_9, called signal line. When the MACD is above signal line, upward trend is likelihood, when below, a downward trend would happen. The last portion of figure, contains the Relative Strength Index, that is in the range between 0 and 100. High value of RSI, implies that security is overbought, otherwise it is oversold.

1.3. Cryptocurrency Market Overview

With more than 3000 coins and a global market capitalization close to $207 Billion, as observable in the Fig 1.2, cryptocurrency market being an important financial reality.

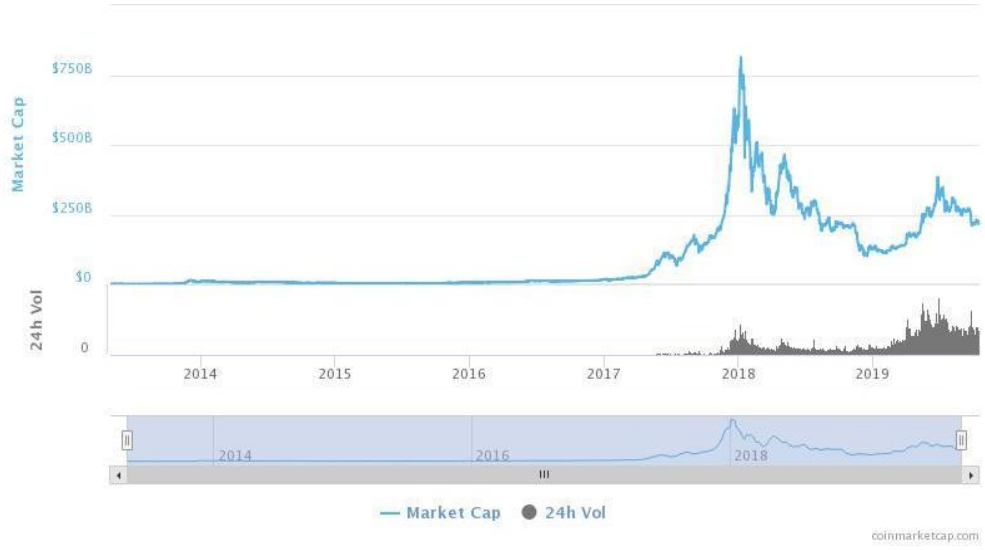


Fig 1.2: Chart of Total Cryptocurrency Market Capitalization.

Source: CoinMarket.com

Although decentralized coins were introduced before, the creation of Bitcoin in October 31st 2008 by a group of individuals, operating with the pseudonymous of 'Satoshi Nakamoto', is de-facto considered as the birth of revolutionary financial transaction methods. Nakamoto describes it in the Bitcoin Whitepaper as 'a purely peer to peer version of electronic cash, which would allow online payments to be sent directly from one party to another without going through a financial institution'.

Due to a consensus network and a cryptography protocol, Bitcoin is not managed by any governments or bank and the main purpose is to make easy, secure and transparent transactions of goods or services, attracting huge amounts of users and media attention. While exchange coins result primarily based on the issuing Governments, Cryptocurrency is totally based on the demand-supply game. The emission of new digital coin depends on specific process called Mining, where transaction are verified and added to the distributed network. People that pursues this are called Miners and generate a hash, a solution for complex mathematical problem required to validate the transactions and create new block. They need PCs with high computational power and high energy consumes to achieve Hashing solution. For this effort, they are rewarded with new Crypto coins. All this process contributes to makes the Cryptocurrency market stiff and ambiguous.

Forums, tweets, Reddit posts, blogs, Telegram has in fact contributed to boost the diffusion of Bitcoin and of cryptocurrency as whole. Social Media represents today an important proxy of people state of confidence about digital coins and could produce important correlation with price movements.

BTC result the most popular coin on Twitter and Reddit, with almost one million of users and shows also the most intensive project activities on GitHub (20 thousands) as illustrated in the below charts (Fig 1.3a and Fig1.3b).

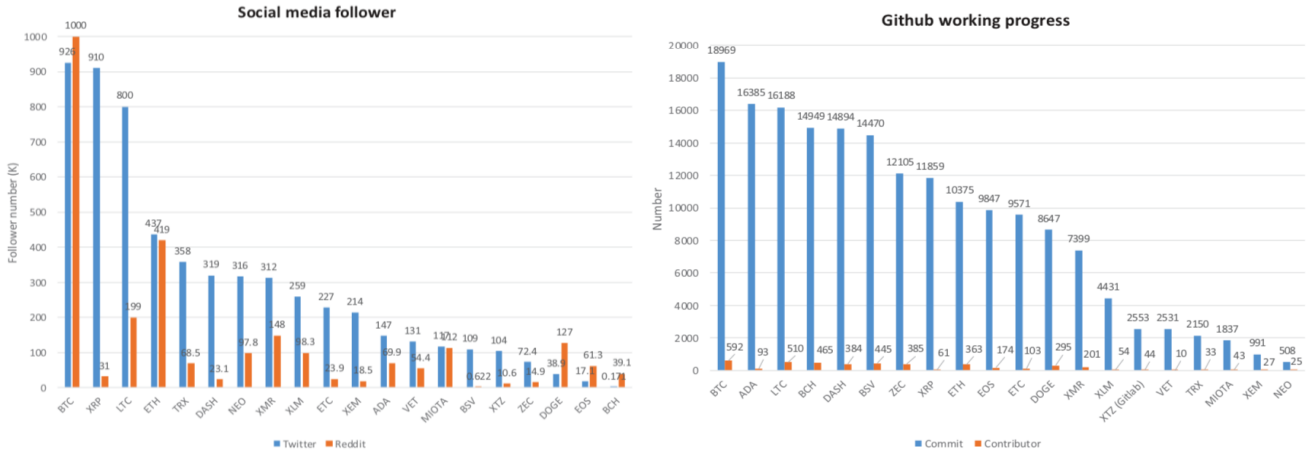


Fig 1.3a and Fig 1.3b: Chart of BTC price series exchanged in USD and representation of Technical Indicators as Moving Average, RSI and MACD from Barchart.com

Other popular coins are:

- Ethereum (ETH): Introduced in July 30th, 2015 it is the currency for Ethereum platform, a smart contract platform that allows developers to program the so called 'dapps', a sort of decentralized applications, idealized by Vitalik Buterin in 2013. Smart contracts, run on Blockchain and allows

executing automatically a deal, evaluating that the conditions are meet. ETH is today one of the most popular coins, following BTC and with a market capitalization of roughly $17 billions.

- Ripple (XRP) is a 'Real Time Gross Settlements System', a currency exchange system that must be validated from independent servers. The currency traded on this network is known as XRP, can be traded in different fiat currencies and transaction time is close to zero. It owns huge popularity on twitter, but very few on Reddit and GitHub compared with other currencies.

The actual XRP market capitalization is close to $12 billion.

- Litecoin (LTC): it is a peer to peer cryptocurrency network, created on the basis of Bitcoin protocol, but involving a different hashing algorithm. The main porpouse of Litecoin network is to reduce the block confirmation time from10 minute to 2,5 minutes, guaranteeing faster processing. LTC has today a market capitalization $3.19 Billions.
- Bitcoin Cash (BCH): It was born from a fork of Bitcoin in order to enable biggest block size dimensions so that the potential transactions volume on the network result improved.

  On august 2017, the first Bitcoin Cash software implementation was realized and miners validated new kind of transaction on new network. From that date, the transactions of Bitcoin and Bitcoin Cash were splitted, while those before it, are in common.

Bitcoin is today the most dominant cryptocurrency, with a market capitalization of $137.71 Billion, and Market Dominance close to 70%, how showed in below chart (Fig 1.4)

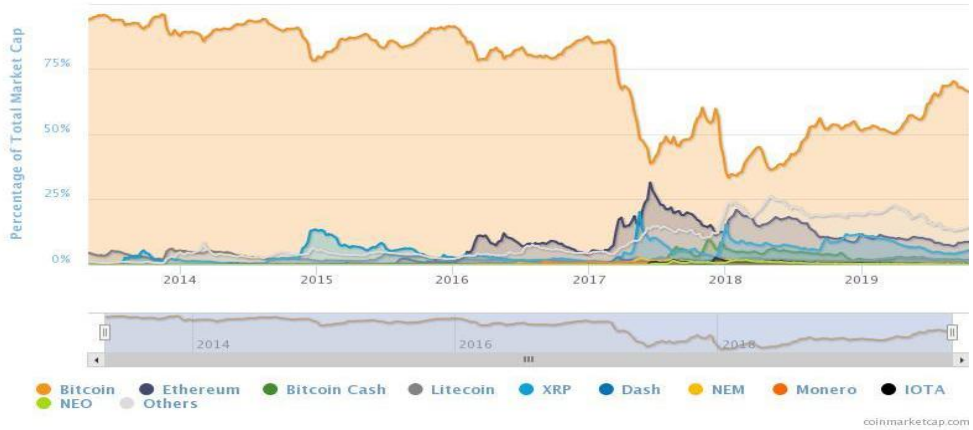


Fig 1.4 : Chart of BTC price series exchanged in USD and representation of Technical Indicators as Moving Average, RSI and MACD from Barchart.com

Since 2018, Litecoin has resulted the cryptocurrency with higher Market Dominance, excluding Bitcoin and followed from Ethereum and Ripple coin (XRP). It is also important highlight how the popularity of

specific coin shows similar attitude with the price movements and Market Capitalization and Market Dominance.

1.4. Cryptocurrency Market Features

Throughout 10 years, cryptocurrency market has become an important financial reality in the international scenarios. Although it was born as new form of currency based on distributed network and blockchain technology, its market dynamic makes its behavior distant from exchange market and the high volatility allows describing it as more similar to stock market.

The authors of 'Towards an Understanding of Cryptocurrency', inferred on the main cryptocurrency market, comparing it with Foreign Exchange and Stock market.

They considered daily close price for 50 cryptos that build the 90% of whole market capitalization, with time span from January 1, 2015 to November 30, 2018 and price considered in US dollar.

They took also into account 50 traditional currencies, including Gold, Silver and Platinum and 102 stocks, including S&P 100 Index.

Basing on these datasets, authors used two main tools, correlation matrix and asset tree, in order to compare the three different markets, in term of clustering, volatility and stability.

A tree is a particular kind of graph and it is used in this case to evaluate which asset has the most influence and impact on global market. In the case of cryptocurrency market, Bitcoin has been the central node for a good time interval, but its role has become progressively weak and central node has diversified.

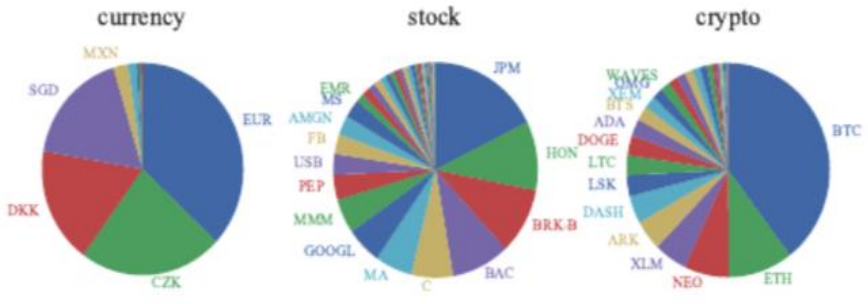


Fig 1.5: Distribution of central nodes in currency, stock, and cryptocurrency market.

In the Fig 1.5 the central node distribution for the three markets are defined. Looking at the pie charts, it is possible to affirm that cryptocurrency behavior is more similar to the stock, due the high diversification of central node, highlighting how the market is in not stable.

At the same time they analyzed the Robustness and Clustering Structure of the markets, concluding that traditional currency market is stable, stock market is less stable and cryptocurrency result absolutely fragile. The latter, however, shows no clustering structure and is subjected to frequent changes.
All this contributes to define crypto market as volatile market, that loads high risk factors and probably more attractive for speculators than digital coin users.

1.5. State of Art

As explained in paragraph 1.2, Fundamental Analysis is not the best approach for Cryptocurrency Market. Crypto companies in fact, are different from traditional stocks companies. They do not provide dividends and the access to traditional financials are more complicated.
Technical Analysis is instead takes into account, but just as preliminary analysis. TA is not objective and depends on the analysts 'interpretations.
Today, cryptocurrency price forecasting represents a new field of interest and researches on this focus are in the early phase. Most of the studies tries to foresee next price at different granularity with machine learning algorithms, trained on different kind of datasets.
The following Table sum-ups the main state of arte results, evaluating the involved algorithms, the Cryptocurrencies taken into account, the frequency of sampling (granularity), the experiments 'time horizon, the kind of attributes and data used to explain the crypto spread and the Sources where data are collected from.

Table 1.1: Cryptocurrency Analysis and Forecasting State of Art

	Publication Date	Algorithms and Tools	Cryptocurrency	Granularity	Data (Financial/ Social/ Blockchian)	Timespan	Sources
[1]	August 2018	Average one-dependence estimators (AODE)	Bitcoin (BTC) Ripple (XRP) Ethereum (ETH)	Day	• Historical price • Number of transactions • Number of view on communities • Number of replies of each comment • Number of Comments	July 9. 2015 – January 21, 2016	1.BitcoinTalk.org 2. CoinDesk.com 3.Forum.Ethereum.org 4. EtherScan.io 5.CoinMarketCap.com 6. XrpChart.com 7.ripplecharts.com

[2]	August 2019	Random Forests, involving alpha 101 factors	Bitcoin (BTC) Ethereum (ETH)	5 minutes	• Historical OHLCV • Alpha 101 Factors	February 6, 2018- August 6, 2018	1.Bincentive.com 2.BitcoinCharts. Com 3.Kaiko.Com 4.Binance.com
[3]	August 2018	Multiple Linear Regression, Sentiment Analysis based on TexBlob.polarity library	Bitcoin (BTC) Litecoin (LTC)	2 hours	• Historical Price • Social Sentiment from Twitter	January 1, 2018 – November 1, 2018	1.CoinDesk.com 2.Twitter.com
[4]	June 2015	SentiStrenght tool for Sentiment Analysis, trained on not standard trading vocabulary	Bitcoin (BTC)	Day	• Historical Price • Twitter Sentiment • Google Trend	January 1, 2015 – March 2015	1. Bitcointalk.org 2. Twitter. Com 3. GoogleTrend.com 4. SentiStrenght
[5]	April 2019	Neural Networks (NN), Support Vector Machine (SVM), Random Forests (RF)	Bitcoin (BTC) Ethereum (ETH) Bitcoin Cash (BCH)	Day	• Historical OHLCV • Twitter Sentiment	3 months	1. BitcoinCharts.com 2. Twitter.com 3. Bittrex.com
[6]	July 2019	Long Short Term Memory Networks, Sentiment Analysis	Bitcoin (BTC) Litecoin (LTC) Ethereum (ETH)	Day	• Open/ high/ low • Volatility • Sentiment Analysis	January 2016 - July 2018	1. CryptoCompare.com 2. Twitter. Com
[7]	January 2019	Long Short Term Memory, Sentiment Analysis through VADER python library	Bitcoin (BTC) Ethereum (ETH)	Hourly	• Historical Prices and Volume • Google Trend Volumes • Twitter sentiments data	January 1, 2017 – November 30, 2017	1. Kaggle.com 2. Cryptocoinsnews.com
[8]	May 2019	Correlation Matrix, Asset Tree	Top market capitalization cryptocurrencies	Day	• Historical Price	January, 2015- November 2018	1.Coinmarketcap.com 2.finance.yahoo.com
[9]	June 2015	Random Forests, Support Vector Machine and Generalized Linear Models (Binomial Logistic Regression)	Bitcoin (BTC)	Day	• Historical Price • Block Size • Cost per transaction • Difficulty • Transaction volume • Hash rate • Market Capitalization • Miners Revenue • Number of unique address • Number of transaction per block	January 1, 2009- January 1, 2015	1.Blockchian.info 2.Coinbase.com 3. OkCoin.com

[10]	June 2018	Auto Regressive, Auto regressive Integrated Moving Average model (ARIMA), Simple, Double, Exponential Smoothing and Holt-Winter's models (EXPSMOOTH) , Linear regression (LINREG) MultiLayer Perception (MLP), Support Vector Machine (SVC), Mulinomial Naïve Bayes (MNB), Random Forest Classifier (RFC)	ADA, BNB, BTC, BCH, XRP, BTG, DASH, DOGE EOS, ETH, IOT, LINK, LTC, NEO, QTUM, TEX, USDT, VEN, WAVES, XMR, XEM, ZEC, ZRX	Day	• Historical Prices	January 2011 – December 2018	
[11]	January 2018	Vector Auto Regression Model (VAR), Vector Error Correction Model (VECM) Sentiment Analysis based on Financial Dictionary	Bitcoin (BTC)	Day	• Historical Price • Social Data due polarity extracted from Twitter and BitcoinTalk.com.	January 1, 2012 – December 31, 2014	1. Bitcointalk.org 2. BitStamp Ltd 3. BitcoinCharts.com 4. Twitter.com

In [1], for instance, Time Series forecasting as Auto Regressive, Auto Regressive Integrated Moving Average, Linear Regression and Classification Algorithm as Support Vector Classifier (SVC), Multilayer Perception (MLP), Random Forests Classifier (RFC), tested on 8 years' horizon, are involved to predict the intraday prices for different coins and settle a Trading Strategy.

The authors of [2] instead considered Random Forests to build prediction on Bitcoin and Ethereum market, involving OHLCV data with alpha 101 factors. These are parameters that, combined with financial time series, allows predicting next financial instruments movements.

In this case, each created random forest did not implement all factors, but randomly selected some of them and finally integrated all the trees to get the classification results.

But the extreme market volatility and the lack of clear pattern in price movements, makes only financial data approach not exhaustive and able to explain good part of price spread. It is in fact, reasonable to

take into account factors such as Cryptocurrency Community activity, Social Media Discussion on the status of the Cryptocurrency itself.

This has probably lead authors of [3, 4, 6 and 11] to consider sentiment analysis. SA is text and opining mining technique that involves machine learning algorithm to process Natural Language.

It receives as input the text containing subjective ideas or opinion and outcome a numerical polarity reflecting the judgment. This tool is already largely diffused on fields as marketing, politics and financial stock market, where it allows processing the huge amount of data that customers of specific brand, electors and investors realized on Web Platform as Twitter, Facebook, and Telegram.

In paper [11], SA based on financial sentiment dictionary is used to extract polarity from social media sources as Twitter and Bitcoin forum (BitcoinTalk.com).

Vector Auto Regression Model (VAR) and Vector Error Correction Model (VECM) are used to empirically test the relationship between Bitcoin Price and Social Media Variables.

Error Measures as Forecast-Error variance decomposition (FEVD), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), are used to evaluate the model performance. They highlight that forum sentiment provides more powerful explanation of Bitcoin Market than Twitter Posts. The latter result too short and provide figurate words that financial dictionary is not able to understand, so that a clear polarity is complex to be assigned.

In [3], Sentiment Analysis is instead performed, extracting tweets related to Bitcoin and Litecoin in JSON format and labelled as positive, negative or neutral in the range of [-1,1] through TexBlob.sentiment.polarity Python library. Multiple Linear Regression Model is then used to perform the relationship between the crypto price and social factors and based on the following assumptions: linearity, equality of variance, normality and independence of error. Adjusted R^2, result for Bitcoin and Litecoin, respectively R^2(Bitcoin)=44% and R^2(Litecoin)=59%, highlighting how social factors play a crucial role on predicting Cryptocurrency Market Spread.

Authors in [4] investigated instead on the correlation between Bitcoin and Tweets sentiment and Google Trend. SA is, in this work, performed through SentiStyrenght, in order to estimate the polarity in short posts. It is a tool, based on sentiment dictionary and trained on adding rules for nonstandard text syntax and grammar. The Bitcoin historical price from January to March 2015 is compared with the number of tweets with positive polarity, the whole tweets volume and Google Trend, which is a tool that allows tracking the popularity for specific topic. The correlation analysis reflected these outcomes:

- Corr (Bitcoin, Tweets_Volume) =0.15, delay= 1 day

- Corr (Bitcoin, Tweets_Positive) =-0.35, delay= 3-4 days
- Corr (Bitcoin, Google_Trend) =0.64, delay= 0 day

In April 2019, authors of [5], evaluated the possibility to use Bitcoin sentiment movements as explanation features for Altcoins, as Ethereum (ETH) and Bitcoin Cash (BCH). They applied machine learning algorithms as Neural Network, Support Vector Machine, Random Forests and Naïve Bayes to daily OHLCV dataset of the three different coins and trained on a time horizon of three months.
The experiments say that Ethereum prediction has the highest accuracy (Ethereum_accuracy= 93.3%), which can be the result of including the Bitcoin sentiment as correlated factor.
Bitcoin predictive models perform better in time periods of less volatility (Bitcoin_accuracy= 85%), whereas Bitcoin Cash displays the lowest accuracy (Bitcoin_Cash_accuracy= 70%).

TAKEAWAYS CHAPTER I

I. With a Market Capitalization of $207 Billions, Cryptocurrency represents a consolidated financial reality.

II. Cryptocurrency Market appears fragile and volatile, and more similar to stock market then exchange market. Hence, it attracts at most risk lover speculators rather than revolutionary digital coin users.

III. FA is preferable for traditional stock companies, whereas TA is today a pictorial method, depending on technical analysts' interpretations. Machine Learning, based on financial time series, result today the main trend for cryptocurrency price prediction and automatic trading systems. But, the lack of seasonality and volatility, makes prediction quite complex.

IV. Sentiment Analysis is recently introduced into crypto analysis and allows taking into account important social factors into Price Forecasting process. Most of the works are based on tweets polarity and Google Trend data.

2. The Blockchain storm in the economic areas

Underlying technology for most cryptocurrencies is Blockchain. This chapter illustrate the main features of this disruptive innovation, so that it is possible to understand quantitative Blockchain data and the way in which they can afflict cryptocurrency market. Throughout recent years, Blockchain, is making resounding headlines in newspaper, covering several business spheres; Finance, IT, politics, supply-chain are the most vibrant. Conferences, Scientific papers and researches and a lot of solemn phrases about Blockchain are presented in each field.

2.1. Blockchain pillars and structure

Blockchain technology provides new method for represent assets, for value exchange and implement trust mechanism and appears like a storm impacting on the modern business. According to Gartner, the hype created on this innovation has reached the stars among the different actors involved in. Although the implementation of these solutions show potential fruitful results in different economic areas, the new paradigm represents an enigma, due its business uncertainty and technology risk. In order to progressively understand the possible related fields of applications, it's crucial to metabolize what Blockchain really is and how it works.

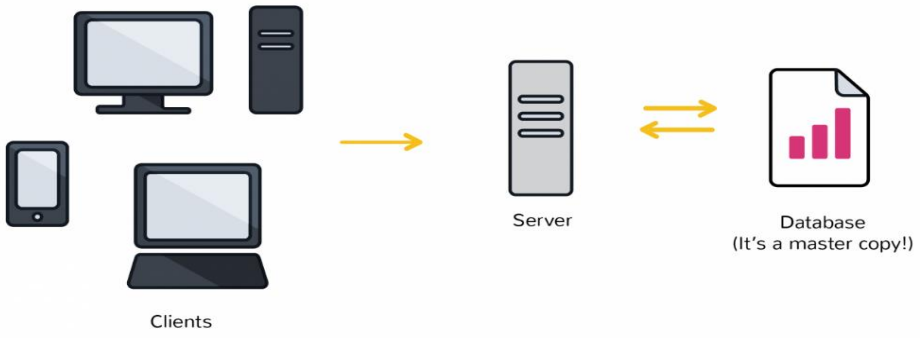
The most popular definition of Blockchain is developed by Don and Alex Tapscott :

"The Blockchain is an incorruptible digital ledger of economic transaction that can be programmed to record no just financial transactions but virtually everything of value "

In simplest terms, Blockchain is dispersed and decentralized database which consisting of a time-stamped series of immutable record of data that is managed by a group of computers, not owned by a central single authority. Chunks of record, called transaction, build the Blocks, which are secured and linked using cryptography. The main pillars that make unique and interesting the Blockchain are: Decentralization, Transparency and Immutability.

2.2. Decentralization and Distributed Database

At first glance, a Blockchain could not appear different from old data distribution network. With Blockchain, different people can write entries into a record, and a community of users can control the uploading and downloading of record information. Wikipedia entries, for instance, are not the product of a single publisher, so it could not be showing impressive differences from the new paradigm.
Conversely, going deeper, the features that make Blockchain unique, start to be clear.
Nevertheless, both run on distributed networks, Wikipedia works on World Wide Web, through client-server model. In this model a client, associated with its account, is able to change Wikipedia records stored on a centralized server (Fig 2.1). When the user access to the Wikipedia page, they will generate a 'mastery copy' of the Wikipedia entry, powering the Double Spending issues. De facto, the control and validation of the database remains with the Wikipedia administrators, that work as central entity.



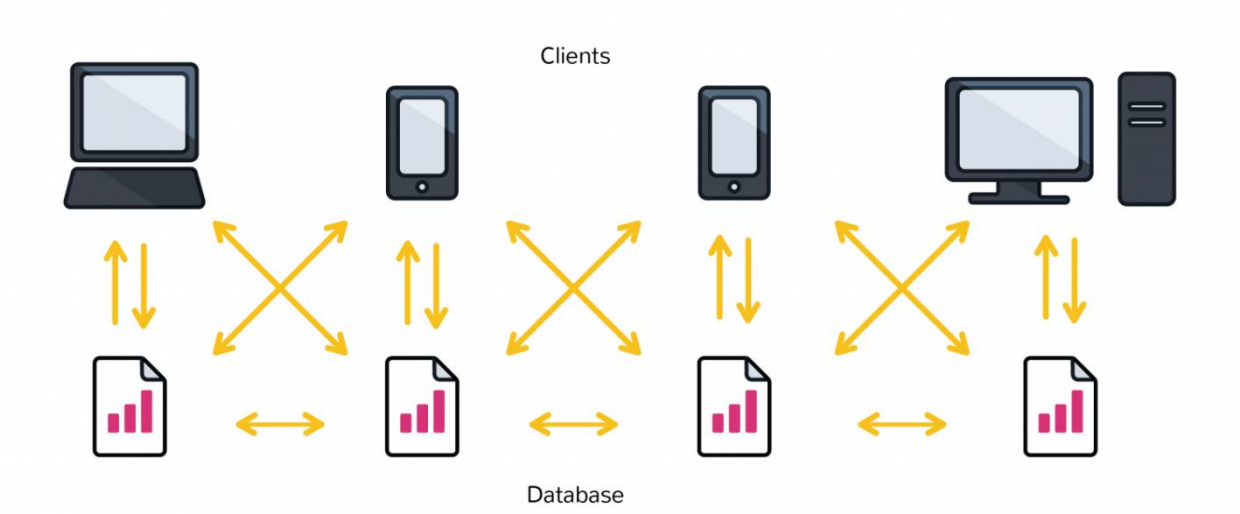
2.1. Client-server model designed by CoinDesk.com

The Wikipedia architecture is similar to the centralized database of governments, insurances companies and banks, in which a centralized entity stored all the data, and clients might interact just with this to get each kind of information, of the approving of each form of transactions with other clients.
Nevertheless, centralized architecture guarantees scalability and stability, the vulnerabilities, that such model presents are not trivial:

- Being centralized, all data are stored in one spot, making easier the hacking process;
- During a software upgrade, the central server represents a bottleneck, reducing the throughput of the system;
- If the centralized authority shut down, no client is able to access the information that is stored;

- Assuming a world without ethics, the central authority could become malicious and corrupted, impacting severely on client's data security;

The distributed database created by Blockchain has a completely different backbone; the information is not stored by one single entity and each client in the network owns the information and they can interact each other without the presence of omniscient authority, how showed in the below figure (Fig 2.2). However, every node in the network can update the record independently and no 'mastery-copy is realized', conferring more flexibility, dynamicity and more data integrity to the whole architecture.



2.2: Clients interaction model without centralized authority designed by CoinDesk.com

Therefore, the main advantages that this model confer are:

- Decentralized system allows to keep control of all owns information and transactions;
- Being without central point, the network is more likely to survive to malicious attacks, due the difficulty to intact each nodes of the network;
- It enables users to exchange something without the intermediation of a third party entity, reducing risk and enhancing privacy;
- Transaction times are drastically reducing to minutes;
- Without third party intermediaries involved into the system, the overhead costs and the transaction cost drop down;

- The decentralization implies that Blockchain data is consistent, accurate and available, guarantying transparency within the network;

2.3. Transparency as second pillar

The decentralization defined in the first pillar, entails privacy and data integrity within the system, and this generally leads to some confusion about how privacy and transparency can coexist. Blockchain balances this paradox through a clear and interesting concept of users' identity. The latter one is hidden behind powerful cryptography and represented only by its public address.
Cryptography is framework that used advanced mathematical transformation to storing and transmitting data in particular form, such that, just selected people can access to it.
In simplest terms, it is the process of encrypting and decrypting information. It's possible to identify two different directions of cryptography:

- Symmetric Cryptography: form of cryptography in which a unique key is used to encrypt and decrypt the information. An historical example of Symmetric Cryptography is Julius Caesar's encrypted military messages, in which exactly the same unique key, is used to both encrypt and decrypt them.
- Asymmetric Cryptography: form of cryptography in which two keys are required to unlock the information stored. The first, the Public Key, is used to encrypt the message and information, the second, the Private Key, is instead used to decrypt them.

Blockchain adopts the latter process to ensure transparency and security. Public and Private keys appear as digital assets, that combined, form a digital signature and ensure the correct transferring of assets, data or transactions.
The Public Key is generated from the Private one, trough hash function, that is easy to compute in one direction, but computationally complicate or impossible in the opposite way.
Several Cryptocurrencies adopt elliptic curve multiplication as the basis for their cryptography; it allows to produce a one-way function, easy to generate just in one direction. For this reason, owners of private key can supply their public key, without worrying about someone can reversely derive the private one.

Once the Public key is generated, it's possible to design the Chain Address. It is a simply string of alphanumeric characters, through which its possible to send or receive data, asset, or transactions. The address is derived via hash function, starting from Public Key input.

The Cryptography mechanism represent the Blockchain backbone for ensure transparency.

The combination of the user's Public and Private Keys creates the so called Digital Signature, which confirms that only Keys 'owners have accessed to that information and that requires consensus from the remaining chain's participants.

Formally speaking, the Digital Signature is mathematical scheme that allows to verify the authenticity of a digital asset or transaction. in particular, it plays three major roles:

I. It serves as proof that the Private Key' owner has authorized that transaction;
II. It allows to prove that the transaction is undeniable;
III. It serve as proof that the transaction authorized by the sign has not altered by the remaining chain's member;

Therefore, the transparency of the network, is achieved through the registration of every single transaction, making available its information for all user of the chain, at any time. This level of transparency has never existed before, especially in financial system. In fact, it confers an higher degree of fairness to accountability mechanism. Let's imagine a Blockchain, in which several companies exchange Cryptocurrencies; through the Public Key it's possible to look at all the transactions that they have engaged in, forcing them to be clear and transparent.

Another resonant example of Blockchain transparency is within the supply chain management. The technology can allow the tracking of each assets across the distribution network as structured in Fig 2.3. This entails that, the consumers can access to every kind of information, related to its good; if it is fair trade, what raw materials are used, or if it respects the workers' rights.

This form of transparency, impacts in positive way also on the consumers and company reputation, improving integrity and ethic.

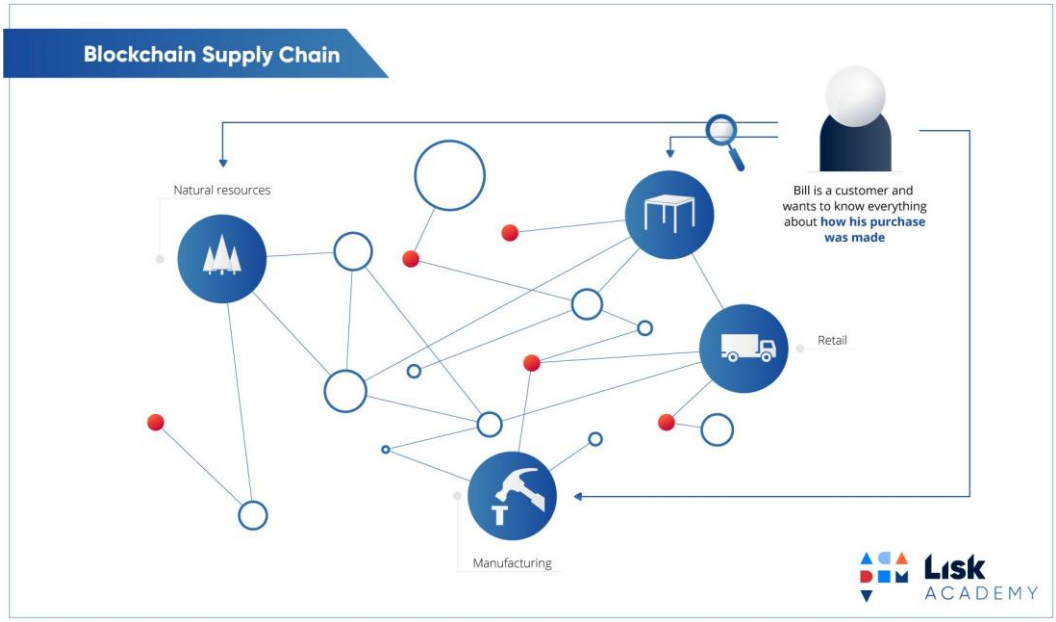


Fig 2.3: shows the way in which Blockchain technology could impact on supply chain network and and how it could result useful to enhance transparency from Upstream to Downstream.

This has further implications in critical market, such as luxury goods or drugs, where the transparent tracking mechanism ensures consumers to receive the correct and real product, treating the huge counterfeit market.

2.4. Immutability as third pillar

In the Blockchain context, immutability implies that, once an input enters into the chain, it cannot be manumitted. Also, this feature is given via cryptographic hash function. In simple terms, the hash process consists in giving an output of fixed dimension of an input string of any length.

For instance, the Bitcoin Blockchain uses Secure Hashing Algorithm (SHA-256) to hash transactions, displaying an output with fixed length of 256 bits.

This property results crucial, when the system dealing with a huge amount of transactions, that instead of remembering the huge input data, can just remember the hash. The Cryptographic hash function owns different properties, but one of the most impressive is the so called 'Avalanche Effect'[3], through which a small change in input, impacts in great way on the output, as showed in below figure (Fig 2.4)

INPUT	HASH
This is a test	C7BE1ED902FB8DD4D48997C6452F5D7E509FBCDBE2808B16BCF4EDCE4C07D14E
this is a test	2E99758548972A8E8822AD47FA1017FF72F06F3FF6A016851F45C398732BC50C

Fig 2.4: Representation of 'Avalanche Effect', through which small change in the input produce huge effect on the output.

Under this viewpoint, it's possible to define the Blockchain as a chain of Blocks, containing the transactions stored and a hash pointer. This is a pointer storing the address and the hash of the transactions o the previous Block. This architecture guarantees the immutability within the chain. In fact, if a hacker attacks a block, a little changing in its data, alters the hash drastically. This entails a chain reaction in changing of hashing of the other blocks. In order to reach this, all the blocks should be attacked or all the network users should approve the modifications. Both scenarios are complex to pursue, so this ensure high immutability to the system.

2.5. Blockchain structure

The Blockchain Blocks are linked in linear way and they are involved into the chain with regular intervals. The basic components constructing a blockchain are:

I. Node: Blockchain user and are physically defined by computers connected to the network;
II. Transaction: it involves information between the interacting user, the exchanging values, the Public Address and the Digital Signature;
III. Block: Collection of transactions, verified by blockchain participants;
IV. Ledger: It is the Public Ledger, through which the Block are linked together with a cryptography hashing function, forming a linear and chronological chain;
V. Hash: is the elementary operation, that maps a string of variable length into an unique and fixed dimension string. Each block is identified by a hash code and contains the hashing code of the previous block;

Each Block contains different fields of information, that depending on the behaviour and the taxonomy of the network. One possibility is shown in the Table 2.1.

Table 2.1 : Basic component of elementary blockchain block. It's important to notice how these elements are not the same for every blockchain network, but depends on its taxonomy and behavior.
Size Source: Bitcoin.com

Name of the field	Definition	SIze
BLOCK_ID	The unique number of the Block	4 bytes
TIME	The time of Block creation	4 bytes
USER_ID	The unique number of the user, who created the Block	5 bytes
SIGN	The sign includes PREVIOUS_BLOCK_HASH, HASH, TIME, LEVEL, MARLL_ROOT)	128-512 bytes
LEVEL	The level, at which the miner was at the time of the block creation	2 bytes
TRANSACTION	PUBLIC_ADDRESS, DIGITAL_SIGNATURE, information and features of each single transaction	up to 3 Mb

As defined in 2.3, each Block stores the cryptographic hash of the previous one, enhancing the immutability and the resilience. The 'Merkle Root', instead, contains all previous transactions and the associated hashes that following a tree structure as pictured in Fig 2.5

The Markle tree [17] encodes the blockchain data in a secure way, so that it pursues the quick verification of blockchain data and quick movements of huge amounts of data from one computer node on the peer to peer blockchain network.

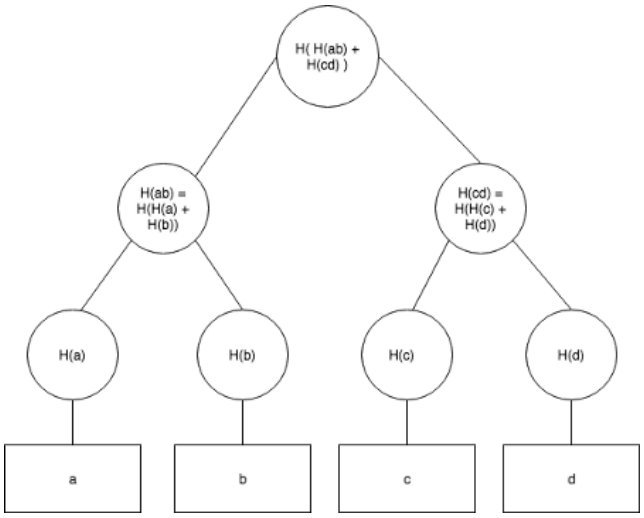


Fig 2.5: representation of a basic Marle Tree, in which a,b,c,d are the basic input transactions and H is the hash function. The intermediate nodes are abbreviated as H(cd) and H(ab), but formally speaking, the mathematical expression should be: H(H(H(a) + H(b)) + H(H(c) + H(d))))

Source: Investopedia.com

Another crucial feature, is the Timestamp, that indicates the creation time of the Block.

The possibility to access to the registration time for each transaction for chain's users, confers, as explained in 2.2 an important rate of transparency.

The transaction, represents the backbone of the block, and contains the Public address of interested nodes, features and related information to it and the Digital Signature, that authenticates and approves the same.

Under blocking viewpoint, a blockchain appears, as a ramification of block, linked through an hashing pointer and each block stores transaction via Markle Root or Tree. **(**Fig 2.6**)**

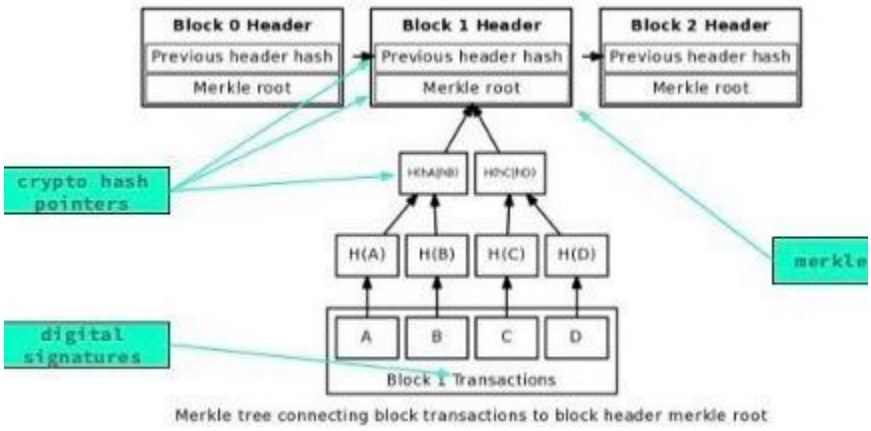


Fig 2.6: Top-down representation of blockchain architecture, where transaction hash function are embedded through merkle root structure and each block is linked through a crypto hash pointer.

Source: hackeroon.com

A new block of transaction requires to be verified, checked and cryptographied before it can be attached to the chain. Only this complex process, called Mining, can activate the new block and it is based on the so-called Proof of Work (PoW), which is an algorithm of security that network relies on. To ensure the accuracy of the new block, it's required to solve a complex mathematical problem. The optimal solution of the problem is reached via physically resource, such as specialized hardware for the computations and electricity power, spent to maintain the hardware itself.

2.6. Taxonomies of Blockchain

The Blockchain structure and tipology is continuously evolving among time, and there exists different classification. Basing on the way of access to the blockchain data, it's possible to distinguish:

I. Public Blockchain: it does not have any restrictions on reading of the blocks and on submitting of the transactions;
II. Private Blockchain: it restricts the reading blocks and the submitting of transaction to just selected branch of users;
III. Permissionless Blockchain: it does not have any restrictions for the users which are permitted to

create blocks of transactions;

IV. Permissioned Blockchain: It defines the list of the predefined users that engaged to process the transactions;

In reality these categories do not present in this scissed way, but most of the time, they exist as combination or as hybrid form. The most recurrent form are synthetized in Table 2.2

Table 2.2: Description of the main possible forms of Blockchain architecture

	Access	**Typology**
Control Typology	Close	Open
Concentrated	Permissioned or Private Blockchain: the access and the transactions validation are restricted to a list of users. This structure is shaped for private company applications and organizations, where the authentication of users and blocks are crucial	
Distributed	Hybride Blockchain: The data network access is limited to a branch of users, whereas the validation is required by all chain nodes	Permissionless Blockchain: The network access is open and the validation and the checking process depends on all the nodes

The Permissioned form is in general created with the scope to maintain the compatibility with existing applications. it can be fully private or consortium and generally the transactions do not regard on-chain assets, but off-chain assets.

The advantage of of the Permissioned blockchain is the scalability, due the smaller pre-selected participants involved, they can scale computing power if the number of transactions are increasing.
The Permissionless Ledger are open and does not present a leading property or central actor. The main objective of this category is to enable everyone to join at the upgrading and the evolution of the Ledger and to keep all the past data and transactions.
The Permissionless can be applied as form of 'Global Database' for that documents that need to be immutable in the time. The most important example of this form of network is given by Bitcoin Blockchain,

2.7. Blockchain as a paradigm shift

The Blockchain technology, described so far appears as an important form of innovation in the field of data accountability and distribution, based on a peer to peer network, and powerful cryptography mechanism. Some analysts define it as the new Internet of
Transactions, whether other define it as the way to create and trak digital assets.
In the realty, the blockchain cannot be defined as simple technology, but it should introduced as new technology paradigm.
A technology paradigm is a mixture of supply side and demand side components that combine and give the birth to a technology trajectory, called S-curve, that is viable for both companies and the market.
The blockchain represents the emergence of new ledger paradigm, called distributed ledger. The first form of ledger was the physical ledger, made of paper and physically touched and invented several centuries ago. This form of ledger remained the only available form until the widespread adoption of computers at the end of 20th century.
This has entailed the birth of Digital Ledger, which is digital file, or a database, It can be manipulated only by computers software, due it's impossibility to be physically touched.

2.8. Ledger and S-curve evolution

From its birth to recent years the dominant design of ledger has been defined by the Centralized Ledger. It is based on centralized logic, where exist a one-to-many relationships and all the transactions are managed by a single central authority.

The trust of the system is released to this central point, that owns huge powers and duties. This kind of the system is steel used in environment, where a centralized power is required or could be useful and scalability and stability are needed. In the recent years, the new paradigm has represented by the Decentralized Network, in which the centralization logic is redefined in local level. In this case, there is a replication of more One to many relationship, in which more local central entity dialogue each other. Unique central authority disappears, but more central points take its place and the trust of the peripheral users is conferred to a Governance that take the place of the unique entity.

The last paradigm is a revolutionary system, where central authorities disappears and it is called Distributed Ledger. It is based on a real and complete distributed logic, where the governance is built around a new concept of trust among each local users that become peer.

The different architectures are showed in the Fig 2.7

The evolution of the industry among the time can be easily evaluated, identifying a relevant performance indicator, which do not follow a linear model, but generally traces a sequence of S-curves.

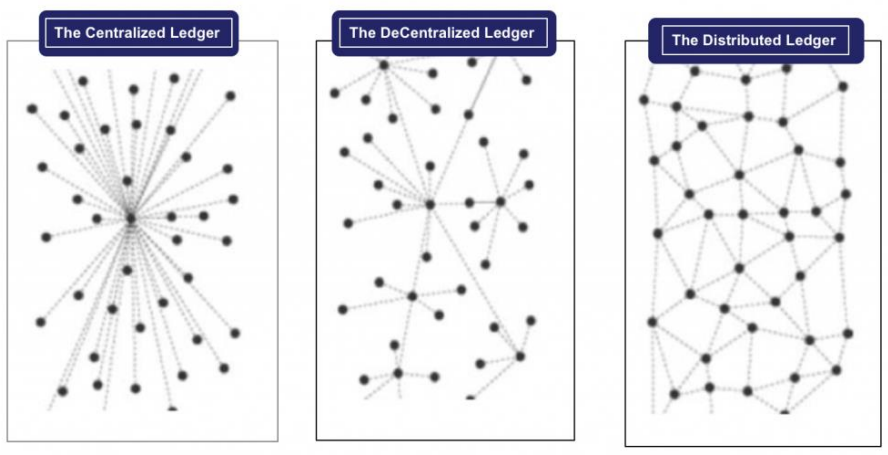


Fig 2.7: from left to right the evolution of Ledger paradigm are represented. The first show a central entity, charged to manage the system, whether the second replicate the concept of central authority but through a Governance. The last architecture is the Distributed Ledger, where each node own the same power and it is based on a shared consensus among all users.

S-curves show that, when a new technology emerges, the performance is low, until a sufficient rate of maturity is reached. Here, the performance grows rapidly until a technology limit is reached due its

intrinsic limitations in the technology.

When the technology limit comes, the firms interested to improve their products or services have to approach with a new technical solutions, choosing among a number of a new available candidates and decide the optimal time instant of entrance (t*).

t* must be defined, taking into account when the customers needs and beliefs could be matched by the new technology.

S-curves are in general, defined as the altering of evolutionary and revolution phases; the former occurs when moving along a given s-curve, whether the latter occurs when transitioning from an old s-curve to new one.

S-curve is a useful tool that permits to evaluate the lifecycle of technology and represent the diffusion or penetration of the innovation into the market. Moving along the s-curve, it is possible to distinguish three main stages, identified as Incubation, diffusion, maturity, as explained in [22].

During the incubation, when the performance and the penetration take off, it's crucial to fine tune the new technology that is giving the birth of the new trajectory. In this phases, the sales and cumulative sales are obviously null.

During the diffusion phase, adoption sales are peak and firms must convince the market of the utility of the new technology.

During maturity, Adoption sales are replaced by Additional Sales and Replacement Sales. At the end of maturity, the firms expect a period of revolutionary change that leads to the choice of the new paradigm.

2.9. Technology Forecasting Models

As explained in 1.2.1 the evolution of technology among the time defines the S-curve, which appears as a useful toolbox for scholars and innovators that are interested to define which is the actually technology phase and which is the next. Forecast the dynamic and volatile behaviour of innovation technology is not trivial, and in the most of the cases the foresing error result evident, but models represent the referment for management and for innovators that have to take decisions.

How defined, s-curve is the altering of evolutionary and revolutionary phases, that trace the evolutionary path of a certain industry [22].

Depending on which phase is moving, different approaches and models are required. In the case of revolutionary phase, statistical models cannot involved, due the absence of historical data. This leads to a approaches like the Dolphie method or some variants. Its basic version consists in interact through

multiple rounds of discussions with analysts and experts of the market, that provide quantitative forecasting values for the future phases, explaining the reasons and the assumptions that lead them to those outputs.

Moving among different meetings, different values are stored, and variance tend to decrease. A good result is reached when, the variance among experts is minimize, or when the values stored in the last discussion are similar to those of the previous.

Completely different approach is a possible with the evolutionary phase, well represented by an s-curve model. In this case statistical models, based to time series analysis are available and preferred. Three are the main models used, depending on the evolutionary phase of the curve(incubation, diffusion, maturity)

I. Linear model: when the data available are few, it is possible to use a linear model, where the performance value changing dv among the time dt is constant:

$$\frac{\mathrm{d}v}{\mathrm{d}t} = k$$

Integrating the differential equation, it outcomes:

$$v(t) = v_0 + kt$$

that boils down in this regression line:

$$v_t = v_0 + kt$$

This model is useful for short term forecasting and it produces good performance when it is in an incubation phase.

II. Exponential growth model: another linear model, that can be useful for short horizon forecasts, is the exponential growth model, where the time instant changing of the performance dv/dt is proportional to the instant value of performance v times a constant k.

The exponential growth model is mathematically represented from the differential equation:

$$\frac{dv}{dt} = kv$$

Integrating the equation, the outcome is:

$$v(t) = v_0 e^{k(t-t_0)}$$

that boils down in:

$$\ln(v_t) = \ln(v_0) - kt_0 + kt$$

which represent a new regression line in log form. This model allows to shape the s-curve until the inflection point is reached. Hence it results a good approximation of the incubation and diffusion phases as shown in Fig 2.8.

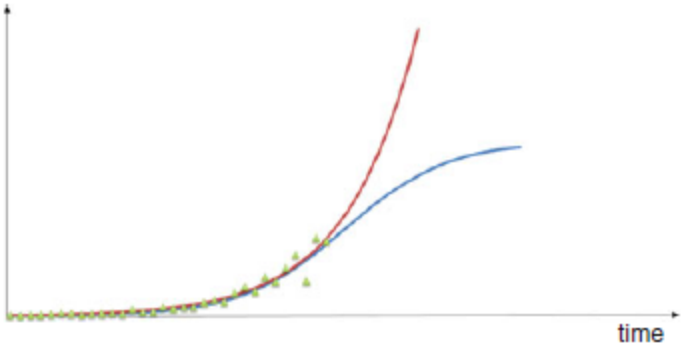


Fig 2.8 : The blue curve represents the s-curve, whereas the red one pictures the exponential growth model. What is really crucial is the fact that the model is good to shape the s-curves just in the firsts two phases: incubation and diffusion, whereas it is not adapt to shape and model the plateau in which maturity is reached. In order to model the last part of the curve, logistic model must be into account.
Source: Innovation Management and Product Development, Canatamessa , Montagna

III. Quadratic model: This model allows to represent the complete evolution of the parameter v among the independent variable t. This model is not linear anymore and the differential equation that represent this phenomenon is:

$$\frac{dv}{dt} = kv - bv^2$$

Integrating the differential equation, the following expression is reached:

$$v(t) = \frac{\frac{k}{b}}{1 + e^{-k(t-t_0)}}$$

The following equation represents the well-known logistic curve, where k/b define the plateau or asymptote for the s-curve. This equation entirely design the evolutionary curve, but suffer of several problems in inferring parameters.

2.10. Hype Effect and Fields of Application

According to Gartner, Blockchain is in incubation phase, where the new technology performance is still immature and a dominant design is not emerged yet. In this phase, most of technology suffers from Hyperinflated expectations, or simply Hype. Following the green line in Fig 2.9, when the technology appears, (technology trigger), markets fall in love with it and expectations peak, reaching the peak of inflated expectations. After this point, the market observes how this expectation is impossible to materialize and tend to classify it as a failure, reaching the trough of disillusionment. Finally the technology shows its realistic applications until it will affirm (plateau of enlightenment)

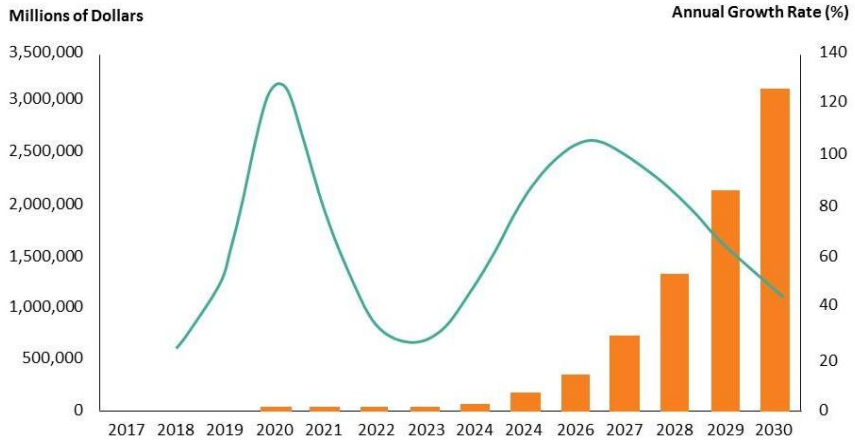


Fig 2.9: Hype Cycle representation for Blockchain technology. The green line define the expectation curve, whereas the orange chart pictures the market revenue growth from its introduction to 2030. The actual year shows how the innovation is still in incubation phase, with huge expectation and null revenues.

Source: Gartner

Following the Gartner research, the blockchain hype is touching the stars and the expectation for its application is overflowing. “(Blockchain) Is making us rethink the old ways of creating transactions, storing data, and moving assets, and that’s only the beginning. Blockchain cannot be described just as a revolution. It is a tsunami-like phenomenon, slowly advancing and gradually enveloping everything along its way by the force of its progression… Blockchains are enormous catalysts for change that affect governance, ways of life, traditional corporate models, society and global institutions”, said William Mougayar in his 2016 book: The business blockchain: promise, practice, and application of the next internet technology.

According to PWC survey the 46% of Blockchain applications **(**Figure 2.10**)** regard Financial Industry. Financial System today serves billion of customers, moving trillions of dollars per day. But, due its antiquated and centralized technology it suffers from several problems, such as adding fee costs, delays, and malicious attacks and fraud. According to Harvard Business Review, 45% of financial intermediaries, such as payment system, money transactions system and stock exchanges offer crime opportunity every year. The final result is the necessity of new regulations pushing and related adding costs that afflict the final consumers. However, the decentralization and the transparency of blockchain technology appears today as the perfect solution for these issues. It allows, for the first time in human history, exchanging money, equities, bonds, stocks, contracts without an intermediary entity, such as Bank or Government. In particular business actors, that not know each other, can sign an agreement through a peer to peer network, based on network consensus and cryptography. Another important problem, that banks take into account is the compliance costs. The Know Your Customers (KYC) practice requires huge budget and delay transactions, taking from 30 to 50 days to reaches a satisfactory level. Also in this case, a distributed ledger could automate the process, reducing compliance error. It not only remove the duplicated effort of KYC checks, but allows enabling updates of customers, providing historical record of all document.

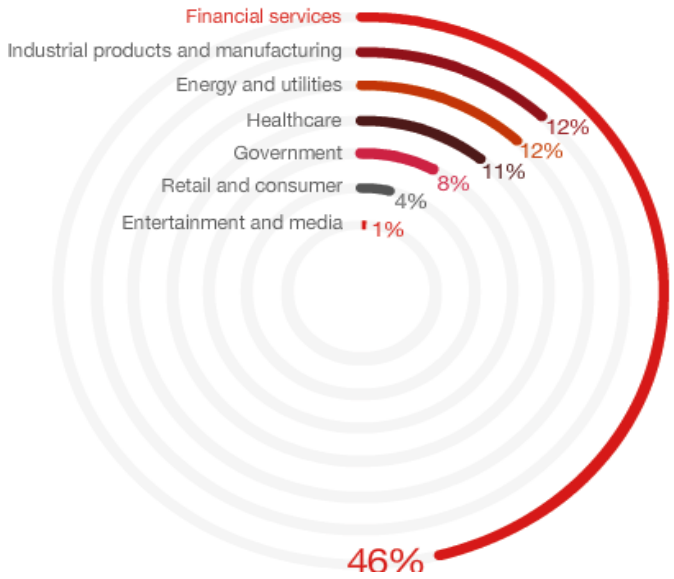


Fig 2.10: The most fields of application for Blockchain technology. The financial service shows its absolutely dominance across the other areas with a 46%. Nevertheless, Energy and Government are growing in term of popularity, due the several possibilities of implementing this technology in disruptive way.

Source: PWC Global Blockchain Survey, 2018

The result is a powerful impact on delaying transactions and Credit Risk issue. Several Financial Entities, that rely themselves on different financial intermediaries, are looking to this technology as an important disruptive and innovative paradigm for the future. Santander bank estimates potential savings of 20 $ billion per year. According to a recent article from Let's Talk Payments, 26 other banks are evaluating the use of this ledger for payment process. However, Capgemini foresees that the final consumers could save up to $ 16 billion in banking and insurances fee per year thanks this technology.

Another important aspect regards the classic debt/capital funding provided today by venture capital, private equity and banks and eventually culminating with an Initial Public Offering (IPO) on a stock exchange. This industry activates a lot of intermediaries such as investment bankers, auditors, crowdfunding platforms, like Kickstarter and lawyers. Also in this case, Blockchain appears as a possible solution, enabling companies of different sizes to raise money through a distributed share offerings. In this way, a company can sells a predefined number of digital token related to a specific ICO of investors, which hope that the token will perform well in the future, generating high return on the investment. The company holding ICO, are not just digital currency suppliers, but they can be interested to fulfill specific goals or launching new digital products. Actually, ICO could result an interest way to bypass the ultra-regulated capital raising process of venture capitalists or banks.

The transparency and immutability of Blockchain, allows to track every form of asset, and this has important impacts in Industrial Production and Manufacturing and Supply Chain. In these cases, the technology allows each member of the chain to monitor the products or services reaching important degree of ethics. The final customers can control which raw material are applied, which techniques are used to work and which other members are involved into the distribution and production network. A famous example of application in this field is Blockverify, an anti-counterfeit solution for supply chain in the directions of Pharmacy, diamonds, luxury items and electronics. This application allows tracking products across all the network guaranteeing the quality and products certifications that make sure that consumers receive the original equipment. Other examples are Bext360, that uses Blockchain technology to track the coffee trade and Maersk & IBM joint venture, which use Blockchain for more efficient and secure global supply chain. The automatized network allows each participant of the chain to follow the product transportations and the status of specific documentation.

Healthcare (11%) and Government (8%) are other important field of applications. For the former, most of the medical centers, use centralized electronic system that do not distribute easily the information. Hence, applications like MedRec, are developing solutions with Blockchain technology, providing secure transparent and scalable access to medical documentations and transactions. The Government seeks this technology as possible solution for electronic voting; projects such as Follow my Vote and E-Residency enable a platform allowing online voting. The easiness and fastness with which is possible to vote impact on the voter counters. Moreover, the same PWC report suggests how Blockchain centers of gravity are shifting. While US is today the most advanced country in developing this technology, in three to five years, China will be the leader, how showed in Fig 2.11.

All these applications highlight how blockchain expectation is overflowing, with at the maximum level as possible to see in the Gartner Graph in Fig 1.9. Today the technology is immature and a Dominant Design is not emerged yet. Blockchain has idealized a sort of miracle for every kind of business problem, but until now few success stories emerge. Not all economic areas will be revolutionized by Blockchain and according to Gartner 2020 will be the disillusionment year, where caution and skepticism will take the place of Hype. Most projects are falling and people will be unsure about the real power of this innovation, with the number of real application drastically reduced compared with the hype phase.

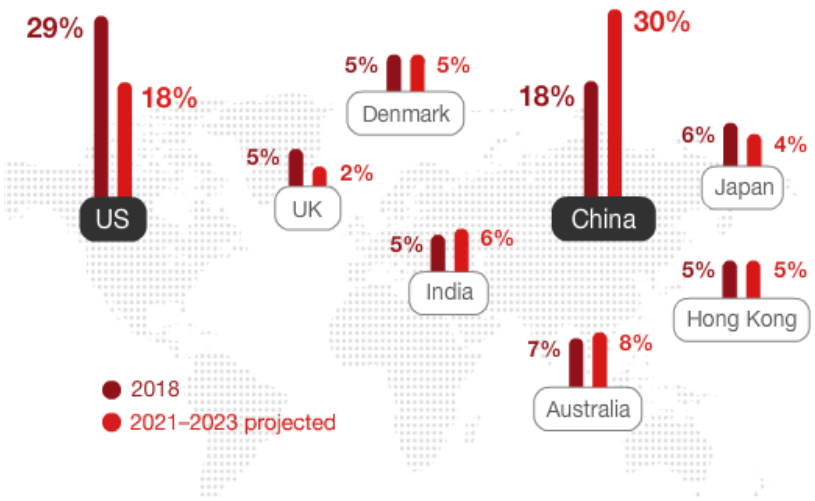


Fig 2.11: The most advanced territories in developing blochchain technology are represented. The dark red shows the actual condition, whereas the clearer pictures the future dynamics, highlighting how China will represent and important center of gravity, due huge capital and technological power.

Source: PWC Global Blockchain Survey, 2018

Hence, different barriers for Blockchian penetration appear. Following PWC survey the most braking aspect is the Regulatory Uncertainty; the most part of regulators is still coming for this innovation. Many territories are discussing and evaluating the terms, especially for financial services and healthcare, but the regulatory environment remains unsettled. Therefore, invest in a technology with no defined regulation and policy result absolutely difficult, boosting the market risk.

Another challenge for blockchain is the construction of trust in the network users. The traditional network presents a central authority that in most of cases has the skill and the attitude to manage problem arising among the chain. The lack of this authority could produce an empty that is difficult to manage and for single users could result difficult to establish trust with unknown people.

Finally, the lack of standards that regulate the interaction between different blockchains, result a heavy factor. Standard can be defined as a 'set of specifications that provide value to the product because of its conformity to the standard'. In easy terms, standard allows providing value for the product due the possibility for consumers and producers to interact in easy way and without new training process. Standards for blockchain are the key that could trigger the penetration into the market, allowing the interaction of different network in the best way. In 2020 technology is entering in the through of disillusionment phase and it is difficult to foresee it will end. Gartner estimates that it ranges from two to five years, but it is not sure that it will be able to emerge from this stage as occurred for several

innovations, such as RSS Enterprise and Ultra-wide Band. Though technology and market risks are high, analysts appear confindent about blockchain diffusion, obviously not in all the market areas defined in the hype phase, but in financial services and supply chain, due the wide impacts and added value that it could be able to pursue.

TAKEAWAYS CHAPTER II

I. Blockchian is a distributed ledger, based on decentralized peer-to-peer network, that allow to track each form of asset in the network through a consensus mechanism;

II. Decentralization, Transparency and Immutability represents the main pillars that make it innovative and adapt for several business applications;

III. Block size (byte), hashing rate, hashing point, timespan represents the main architecture variables that afflict the way in which system works;

IV. Financial Services, Healthcare and Supply Chain result as the most shaped and adapt fields of application, due the great value that this innovation can pursue;

V. Blockchain appears as a new ledger paradigm, in incubation phase. 2020 will be the year of transaction from Hype to Disillusionment phase;

3. Data Source and Crawling Process

This chapter opens the empirical phase of this work. Looking at the complexity of the Crypto Market, and the actual State of Art, it is interesting to evaluate which variable of different nature can afflict price spread and movements. The first, and complex step is to evaluate from which sources and in which way it is possible to achieve Crypto data. The infancy of this market makes stiff this process. In particular, three different kind of historical attributes are evaluated: OHLCV financial data, Blockchain data and Social Metrics.

3.1 Data Mining Introduction

Data Mining represents the merger of different disciplines such as statistics and machine learning, applied to large databases, in order to insight possible information, useful for decision making process. Widely used in scientific applications, as bioinformatics and physic, the main driver that boosts the development of this discipline has been business potential. Data Mining, in fact appears as the response to the need of extract value from the huge business available data. Today, the big data available from different business units and modern data warehouse are crucial resources for company, but needs to techniques and tools to be valorized.

According to www.kdnuggets.com portal the main fields of Data Mining applications are Customer Relationship Management (CRM) (12%), Banking and Finance (14%), Direct Marketing (8%), Fraud Detection (5%), Insurance (6%), Retail (6%), Telecommunications (5%), Scientific Research (4%) and Health (4%).

Data Mining is today largely applied in Marketing and Product Development, where survey data are analyzed in order to discover possible patterns and customer's primary needs, crucial to address market pull product development.

Financial Market as well, is an important field of interest for Data Mining. The huge moles of data that users, traders and operators realized for dummy financial activities, has improved the centrality of Data processing and Analysis.

Data Mining, in reality, belongs to a wider process, known as Knowledge Discovery from Data, which is represented in the below figure (Fig 3.1) and culminate with the extraction of Knowledge from data.

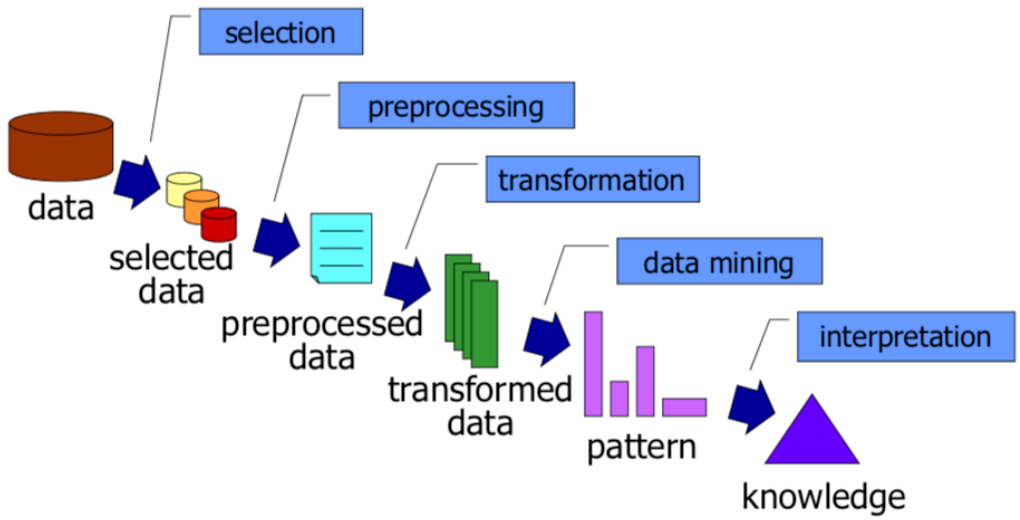


Fig 3.1: Illustration of the main phases of Knowledge Discovery from Data process. It starts with the Data Selection and culminate with the interpretation of the discovered pattern. Source: Data Science Course Slide at Politecnico di Torino.

The first phase of the process is the Collection of Data, which consists of collecting data from different sources and to crawl them into one database.

In the Selection phase, data is then selected on the basis of specific criteria, in order to consider just interesting data for future steps. This data defines the selected or target data.

The next step is the data preprocessing, where outliers and null values are detected and replaced with specific techniques and data is standardized in order to compare data of different scales.

The crucial phase is the Data Mining process, where specific tools and approaches allows identifying meaningful patterns that explains significant data attitude.

Finally, patterns are supervised by domain experts, so that they can extract knowledge from the data and can use it for decision making process.

3.2 Crypto Data Sources

Today, Cryptocurrency Market represents an important financial reality. The huge market volatility is a crucial feature that attracts a lot of risk lovers. Most of the Trading Strategy requires to collect data and factors that insight out some possible correlation with the crypto financial time series. Blockchain statistics, for instance could represent an important explanatory variable, that lead part of the price spread, or Social Data, just crucial for Stock Market Forecasting, could play an important role on the Crypto Market Analysis.

Although, for Stock and Exchange Markets, today it is possible to access to several metrics and information, the youth of Cryptocurrency, makes this process complicate.

Speculators or people interested to crypto as new payment currency, can take into account Exchanges Websites or Market Analysis Page. With a powerful and complete Trade View interface BitStamp Ltd, for instance, represents one of the top Bitcoin Exchange by volume. It provides premium access to Cryptocurrency trading for individuals and institutional clients. It shows the main financial information about Bitcoin, Ethereum, Bitcoin Cash, XRP and Litecoin in USD, EUR or BTC. Through REST API and WebSocket API it is possible to access to data screenshots and real time streaming data. Similarly, CoinDesk allows accessing to price and volume charts of several Cryptos. It is, today, one of the most leading digital media, events and information services company for the cryptocurrency asset and Blockchain Community and ensures to inform it through the daily news dedicated to these fields. However, it gives access to the 'Crypto-Economic Explorer', a tool that help to visualize and analyze cryptocurrency network, including Social, Developing, Blockchain activities and Exchange metrics.

The main drawback related to these sources is the difficult to access to historical time series, that is fundamental for empirical and research activities.

One of the most complete source, in this optic, is NeuroSantiment, a community website which permits the access to several Financial, Social and Blockchain metrics, through REST and Websocket API via GraphQl or Sanpy Python package. Chain data is collected directly from running nodes, making it faster and reliable, by eliminating potential point of failure. Blockchain metrics are available for BTC, EOS and ETH and include:

- **Daily Active Address**: It includes the number of unique addresses that participated in the transfers of given token during the day;

- **Network Growth**: It shows the number of new addresses being created on the network each day;
- **Token Age Consumed**: It shows the number of tokens changing addresses on a certain date multiplied the number of blocks created on the Blockchain since they last moved;
- **Average Token Age Consumed in Days**: it represents the average number of days that the tokens were idle before being moved on a certain date;
- **Exchange Flow Balance**: it shows the combined values of tokens moving in and out of exchange wallets on a certain date;
- **Exchange Flow**: This graph shows the amount of inflow and outflow plotted separately, based on the previous metric;
- **Transaction Volume**: it shows the aggregate number of tokens across all transactions that happened on the network on a certain date;
- **Token Circulation**: Spread of idle over time;
- **Velocity of Tokens**: Average number of times that a token changes wallets each day;
- **Gas used**: return gas used by a Blockchain for the token supply computations. It is just available for ETH;
- **Miners balance**: it returns the Miners Balances over time. It is just available for ETH;
- **Top 100 Transactions**: List of the top 100 transactions by volume in a given time frame;
- **MVRV Ratio**: it shows the whole market value divided by the realized value of the network;

Financial metrics includes Historical Price, volume, market capitalization and OHLC data for more than 2000 cryptocurrencies with daily granularity, whereas Social data offers granularity ranging from minute to week and different metrics are available:

- **Social Dominance**: it returns the dominance of a certain asset has over a social channel in percentage, compared to all projects mentioned in that channel;
- **Social Volume**: it returns a list of mentions count for a given project and time interval;
- **Topic Search**: it returns the lists with the mentions of the search phrase from the selected source (Telegram, Professional Traders Chat, Reddit, Discord)
- **History Twitter data**: it counts the number of followers of the official Crypto profile;
- **Github Activity**: it returns the GitHub activities for a given project and time interval;
- **News Collection**: It returns the News for a given digital asset in a specific timespan;

NeuroSantiment allows unlimited access to historical advanced metrics through Premium and Corporate versions, while the free plan guarantees accessing to last three-month data, excluding last 24h, with 20 API calls/minute constraint. This limited timeframe appears inadequate for empirical researches, especially for machine learning applications, where Data Dump is needed for training and testing phases. A concrete alternative to NeuroSentiment is CryptoCompare [38], which is a Global cryptocurrency data provider, producing at granular level, trade, order book, Blockchain, social and historical price data. Deeper, financial data are available for 2000 coins and USD and BTC exchange, including the following data:

- **Historical Daily OHLCV**
- **Historical Hourly OHLCV**
- **Historical Minute OHLCV**
- **Historical Daily OHLCV All Markets**
- **Historical Day OHLCV for a timestamp**
- **Historical Day Average Price**
- **Historical Daily Exchange Volume**
- **Historical Hourly Exchange Volume**
- **Toplist by 24H Volume Full Data**
- **Toplist by 24H Top Tier Volume Full Data**
- **Toplist by Market Cap Full Data**
- **Top Exchanges Volume Data by Pair**
- **Top Exchanges Full data by Pair**
- **Toplist by Pair Volume**
- **Toplist of Trading Pairs**

Social data is available with hourly and daily granularity and allows accessing to Social Statistics from several sources, such as Twitter, Facebook, Reddit and accessing to news related to a specific coin. In particular, includes the following data:

- **Latest Coin Social Stats Data**
- **Historical Day Social Stats Data**
- **Historical Hour Social Stats Data**

- **Latest News Articles**
- **List News Feeds**
- **News Article Categories**
- **List News Feeds and Categories**

The free plan permits 100,000 calls/month, more than 35 Market Data endpoints, Full access to Daily and Hourly historical data, 7 days' access to minute granularity data, Order Book snapshot from Binance, FAQ support and Personal use only as license. At a glance, CryptoCompare appears as a good data provider platform for research activities.
For Bitcoin Blockchain, Blocckchain.com represent the top reference due the wide information provided, easy to download in csv format and with daily granularity. It includes these metrics:

- **Total Transaction Fees:** it shows the total transaction fees paid to the miners;
- **Confirmed Transaction per Day**: the number of confirmed Bitcoin transactions,
- **Output Value**: the total value of all transactions per day, including coins returned to the sender as change;
- **Estimated Transaction value**: the total estimated value of transactions on th4 Bitcoin Blockchain, not including the coins returned to the sender as change;
- **Estimated USD Transaction Value**: the estimated value in USD;
- **Miners Revenue**: total value of coin Block rewards and transaction fees paid to miners;
- **Cost % of Transaction Volume**: it shows miners revenue as percentage of the transaction volume;
- **Cost per Transaction**: it shows the revenue divided by the number of transactions;
- **Hashing Difficulty**: A relative measure of how difficult it is to find a new block.

Similarly, Craft Source Maker [39], a machine learning powered data and analytics platform, extend Blockchain Insights to several chains. Data presents granularity of 1 day and it is possible to achieve data back to first year emission of the asset. Data is updated every day and CSV download is possible. For financial data, instead, CryptoDataDownload appears as a valid solution; it provides OHLCV data organized by exchange and the possibility to access free to all historical daily, hourly and sometimes minute data.
It includes the exchange of different countries:

- US & UK (Gemini, GDAX, Coinbase, Kraken, Bitstamp, Bitfinex, Cexio, QuadrigaCx, Coinfloor, Luno, Tidex, itBit);
- EU & Russia (HitBTC, Exmo, BitBay, Bitmarket, Tobit, Liquui)
- Asia Pacific (Binance, Bithumb, OkCoin, Quiniex, Bitflyer, BTCMarkets, Cryptodia, Zaif)
- Other international (Bitso, BTCXIndia, Unocoin, Bit2C)

For Social data, instead, one of the most powerful platform is BittsAnalytics [40], which provides advanced metrics through a premium account (99 $/month) via REST and WebSocket APIs for more than 200 digital coins. The historical data insights these metrics:

- **Daily Social Sentiment**: this API Endpoint provides access to historical daily sentiment data for over 200 cryptocurrencies. Historical data starts on 14 August 2017. Daily social sentiment is the average sentiment of cryptocurrency on twitter in last closed day in UTC time. Sentiment is determined by applying advanced machine learning models on texts of social media posts on twitter that mention individual coins;
- **Daily Social Mentions**: this API Endpoint provides access to historical daily mentions on social media for over 200 cryptocurrencies. Historical data starts on 14 August 2017;
- **Hourly Social Sentiment**: this API Endpoint provides access to historical hourly social media sentiment data for over 200 cryptocurrencies;
- **Hourly Social Mentions**: this API Endpoint provides access to historical hourly social media mentions data for over 200 cryptocurrencies;

It also permits accessing to real time data, including the following daily, hourly, minutely features:

- **ticker**: symbol of cryptocurrency,
- **price_usd**: price of cryptocurrency in USD,
- **market_cap**: market capitalization of cryptocurrency in USD,
- **trading_volume**: trading volume of cryptocurrency in USD (last 24 hours),
- **momentum score**: proprietary indicator calculated as equal-weighted z-scores (from population of 200+ cryptocurrencies) for the following categories: 24-hour change in price, change in daily sentiment, change in daily number of mentions, change in daily trading volume. Momentum score

indicates the composite momentum of a cryptocurrency in price, trading volume, social media sentiment and social media buzz.

- **hourly_social_sentiment**: we calculate social media sentiment of coins by using advanced machine learning models from texts of social media posts on twitter that mention individual coins. Hourly social sentiment is the average sentiment of a coin in the last closed hour in UTC time.
- **hourly_social_mentions**: we calculate number of mentions of coins in texts of social media posts on twitter. Hourly social mentions is the number of mentions of a given coin in the last closed hour in UTC time.
- **hourly_social_buzz**: is the number of mentions of coins on twitter in last 3 hours normalized by historical, similar 3 hour windows. Hourly social buzz indicates how much is the coin mentioned on social media recently as compared to historical averages for this coin.
- **hourly_social_sentiment_momentum**: is the average sentiment of coins on twitter in last 3 hours normalized by historical, similar 3 hour windows. Hourly social sentiment momentum indicates how positive or negative is the sentiment of coin as compared to historical averages for this coin.
- **daily_social_sentiment**: is the average sentiment of coin on twitter in last closed day in UTC time.
- **daily_social_mentions**: is the number of mentions of a given coin in last closed day in UTC time.

Finally, it is possible to include as Social Data, tool which allows to measure the popularity of specific topic or keyword, such as Google Trend and Wikipedia Statistics. Similarly, Reddit Statistics provides measures as Subscribers Number per day, Comments per day and Posts per day for a specific Subreddit. These kind of sources, highlight the frequency of interest of specific fields and represents a proxy of people sentiment about it.

The complete report of evaluated sources is available in Appendix II.

3.3. Data Crawling Process and Crypto Metrics

The Sources used for empirical application in this thesis work are CryptoCompare.com and Craft Source Maker. The former is used to collect OHLCV data with daily and hourly granularity, and hourly Social metrics, whereas the latter is used to extract Daily Blockchian data.

CryptoCompare allows accessing REST APIs through an API Key, which is possible to do for free on the site, requesting and querying data by different ways, such as Python programming, Javasvript or Google Sheet Adds-on.

In this case, Python is used to pursue the data crawling process. It is a high-level and general purpose programming language introduced in 1991 as successor of ABC language by Guido van Rossum. It is today one of the most popular language due its wide and powerful applications. Python is, in fact, used by many biggest companies, such as PayPal, Google, Facebook, Instagram, Netflix, Dropbox, Uber, Reddit. Advanced libraries, as Pandas, Matplotlib, NumPy and SciPy make it a perfect language for Data Analysis and machine learning applications.

In order to call the API is required to define the coin Symbol for OHLCV data and coin Id for Social metrics. Symbols and Id of the main famous cryptos are reported in the Table 3.1.

Table 3.1: It shows the Coin Symbols and Coin Id for several cryptocurrencies, which are required to access to Social and Financial data via CryptoCompare APIs.

Cryptocurrency name	Symbol	Cod ID
Eos	EOS	166503
Bitcoin	BTC	1182
Ripple	XRP	5031
Litecoin	LTC	3808
Dash	DASH	3807
Ethereum	ETH	7605
Groestcoin	GRS	13070
Bitcoin Sv	BSV	926591

Bitcoin Cash	BCH	202330
Dogecoin	DOGE	4432
Monero	XMR	5038

However, there are several Python libraries to do an http request, in this application requests library is involved. The following lines code show how to produce an API call, once the API key is obtained, for DASH coin with hourly granularity.

For each call, different parameters must be settled. Api_key requires the key needed to call the API, whereas fsym and tsym defines the symbols of interested coin and conversion coin.

ToTs returns historical data before that specific timestamp and limit allows setting the number of retuned data, with a maximum of 2000 for call. Hence, in order to produce a full historical data, different calls with different timestamps are needed.

```
#import libraries
import requests
import pandas as pd

apiKey = "d30ec53fdcaf63a963b3ea7287f087f8d015ff19a15aeecf56b590a13c4edf1f"

#import OHLCV data with hourly granularity
url = "https://min-api.cryptocompare.com/data/histohour"

payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "limit": 2000
}

result_1 = requests.get(url, params=payload).json()

df1 = pd.DataFrame(result_1['Data'])

print(df1.head())

payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1563440400,
```

Similarly, it is possible to generate call for hourly Social Data, changing thefsym with the CoinId, how showed in the below lines:

```
#import social data
url = "https://min-api.cryptocompare.com/data/social/coin/histo/hour"

payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "limit": 2000
}

social_1 = requests.get(url, params=payload).json()

db1 = pd.DataFrame(social_1['Data'])

print(db1.head())

payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1563440400,
    "limit": 2000
}

social_2 = requests.get(url, params=payload).json()

db2 = pd.DataFrame(social_2['Data'])

print(db2.head())

payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1556240400,
    "limit": 2000
}
```

Both Social and OHLCV data are framed into Pandas Data frames and are aligned and concatenated into one single dataset, named 'Financial&Social', using pd.concat() pandas function and then exported as csv evoking .to_csv function.

```
#Scial database given by merging all social dataframes

social_dataframe=pd.concat([db12,db11,db10,db9,db8,db7,db6,db5,db4,db3,db2,db1
],sort=True)

xport_csv = social_dataframe.to_csv (r'DASH_SOCIAL.csv',
 index = None, header=True)

print(social_dataframe.info)

#OHLCV database given by merging all financial dataframes

final=pd.concat([df12,df11,df10,df9,
df8,df7,df6,df5,df4,df3,df2,df1])

xport_csv = final.to_csv (r'DASH_OHLCV.csv', index =
None, header=True)
print(final.info)


financial_and_social =
pd.concat([final,social_dataframe], axis=1, sort=False)

xport_csv =
financial_and_social.to_csv(r'Financial&Social.csv'
, index = None, header=True)

```

Atthe same way, OHLCV with daily granularity are collected, whereas Blockchain data are extracted from Craft Source Maker. It allows downloading on-chain daily metrics for different coins with extended time horizon in csv format.

The complete code for Crawling process is available in the Appendix III.

The crawling process output is the following. OHCLV data is collected both with hourly and daily granularity. The former is aligned with Social Data that have constant timeframe for ach crypto. The latter is aligned with Blockchain Data that presents different timespan dimension. For most of them it has been possible to store data back to its quasi market emission.
The crypto taken into account and respective timespan are presented in the table below.

Table 3.2: It shows the Time Horizon taken into account for each coin and for both Social and Blockchain data

Cryptocurrency (SYMBOL)	Social data (Hourly)	Blockchain Data (Daily)
Bitcoin (BTC)	1/1/19, 0.00 12/10/19, 11.00	17/07/10 12/10/19
Dash (DASH)	1/1/19 0.00 12/10/19 11.00	08/02/14 12/10/19
Ethereum (ETH)	1/1/19 0.00 12/10/19 11.00	07/08/15 12/10/19
Litecoin (LTC)	1/1/19 0.00 12/10/19 11.00	24/10/13 12/10/19
Monero (XMR)	1/1/19 0.00 12/10/19 11.00	2015-01-29 12/10/19

Calling dataframe.info () function it is possible to obtain information about the number of entries and attributes of specific dataframe.

Applying it to hourly Social and daily Blockchain dataframes, for DASH coin, the result is in the Fig 3.2 and Fig 3.3

```
Python 3.7.4 (default, Jul  9 2019, 00:06:43)
[GCC 6.3.0 20170516] on linux
<bound method DataFrame.info of             time       date  close  ... TxTfrValM
edUSD   TxTfrValNtv   TxTfrValUSD
0     1.391818e+09    2/8/2014   0.07  ...      1.491126  1.915961e+06  2.286838e+0
5
1     1.391904e+09    2/9/2014    0.1  ...      1.307980  4.759990e+05  9.813357e+0
4
2     1.391990e+09   2/10/2014    0.1  ...      0.804860  8.942498e+04  2.230517e+0
4
3     1.392077e+09   2/11/2014    0.1  ...      1.564131  4.493588e+05  1.309713e+0
5
4     1.392163e+09   2/12/2014    0.1  ...      2.057161  5.159810e+05  1.592489e+0
5
...            ...         ...    ...  ...           ...           ...           ..
.
2069  1.570579e+09   10/9/2019  74.27  ...      0.742515  1.928737e+05  1.432102e+0
7
2070  1.570666e+09  10/10/2019  72.74  ...      0.727062  2.713168e+05  1.972621e+0
7
2071  1.570752e+09  10/11/2019  69.68  ...      0.697484  1.930038e+05  1.346062e+0
7
2072  1.570838e+09  10/12/2019  72.42  ...      1.316559  1.565513e+05  1.108123e+0
7
2073           NaN         NaN    NaN  ...           NaN           NaN           Na
N

[2074 rows x 27 columns]>
```

Fig 3.2: It shows the dimension of daily Blockchain and financial dataframe for DASH coin. It counts 2074 daily entries and 27 attributes. The time horizon depend on specific coin, hence the entries varies for each coin.

```
You should consider upgrading via the 'pip install --upgrade pi
p' command.
<bound method DataFrame.info of       close   high  ...  total_
page_views  trades_page_views
0     80.24  80.37  ...           1834013              26658
1     79.75  80.37  ...           1834029              26658
2     79.63  80.18  ...           1834039              26658
3     79.89  80.11  ...           1834043              26658
4     80.35  80.47  ...           1834046              26658
...     ...    ...  ...               ...                ...
6759  70.49  70.87  ...           1914818              27679
6760  70.37  70.65  ...           1914822              27679
6761  70.44  70.60  ...           1914828              27679
6762  70.20  70.52  ...           1914831              27679
6763  70.29  70.46  ...           1914835              27679

[6764 rows x 28 columns]>
>
```

Fig 3.3: It shows the dimension of hourly Social and financial dataframe for DASH coin. It counts 6764 hourly entries and 28 attributes. The time horizon in this case, is constant for each coin.

The hourly dataframe contains OHLCV data and the following Social metrics:

- **analysis_page_views**: It counts the number of views for Analysis CryptoCompare.com page;
- **charts_page_views**: It provides statistics about charts page view from CryptoCompare.com;
- **code_repo_closed_issues**: It counts the times that a repository related to specific coin is closed on GitHub community. Generally, it comes through the syntax Close or Fix followed by the number of Issue. For example, to close the issue number 200, just needs the phrase "Closes#200" in the pull request description;
- **code_repo_forks**: It takes into account the time where a new copy of a repository for specific coin project is produced. By the way, a fork is a copy of a repository, that allows to experiment with changes without afflicting the original project;
- **code_repo_stars**: It counts the number of new repository for specific coin on GitHub;
- code_repo_subscribers: It takes into account the number of subscribers for specific coin developing activity project on GitHub;
- **fb_comments**: It insights the number of Facebook Comment under Coin Posts;
- **fb_likes**: It reports the number of Facebook likes on Coin Posts;
- **followers**: It defines the number of followers of Facebook coin page;
- **forum_page_views**: It counts the number of views for Coin forum page on CryptoCompare.com;
- **influence_page_views**: It counts the number of views for most influencing Coin news;
- **markets_page_views**: It counts the number of views for Coin market page on CryptoCompare.com;
- **overview_page_views**: It counts the number of views for Coin overview page on CryptoCompare.com;
- **forum_posts**: It counts the number of posts for Coin forum page on CryptoCompare.com;
- **reddit_active_users**: It insights the number of active users for coin subreddit on Reddit community;
- **reddit_comments_per_hour**: It insights the number of hourly comments for coin subreddit on Reddit community;

- **reddit_posts_per_hour**: It defines the number of hourly posts for coin subreddit on Reddit community;
- **reddit_subscribers**: It counts the number of hourly subscribers for coin subreddit on Reddit community;
- **total_page_views**: It returns the sum of views from all coin pages taken into account;
- **trades_page_views**: It defines the number of views of Coin trading pages on CryptoCompare.com;

The daily dataframe contains OHLCV data and the following Blockchain metrics:

- **AdrActCnt**: It shows the sum count of unique addresses that were active in the that day. Individual addresses are not double-counted if previously active;
- **BlkCnt**: It defines the sum count of blocks created that day that were included in the chain;
- **BlkSizeMeanByte:** It gives mean size in bytes of all blocks created that day.
- **DiffMean** The mean difficulty of finding a hash that meets the protocol-designated requirement (i.e., the difficulty of finding a new block) that day. This is a proxy of how difficult is to find a new block that day;
- **FeeMeanUSD**: It gives the USD value of the mean fee per transaction that day;
- **FeeTotUSD**: It defines the sum USD value of all fees paid to miners that day;
- **IssTotUSD:** The sum USD value of all new native units issued that day;
- **NVTAdj** The ratio of the network value (or market capitalization, current supply) divided by the adjusted transfer value. Also referred to as NVT;
- **SplyCur**: It defines the sum of all native units ever created and visible on the ledger of that day. For account-based protocols, only accounts with positive balances are counted.
- **TxCnt** It insights the sum count of transactions that day. Transactions represent a bundle of intended actions to alter the ledger initiated by a user (human or machine). Transactions are counted whether they execute or not and whether they result in the transfer of native units or not (a transaction can result in no, one, or many transfers);
- **TxTfrCnt** The sum count of transfers that day. Transfers represent movements of native units from one ledger entity to another distinct ledger entity. Only transfers that are the result of a transaction and that have a positive (non-zero) value are counted.

- **TxTfrValAdjNtv** The sum of native units transferred that day removing noise and certain artifacts.
- **TxTfrValAdjUSD** The USD value of the sum of native units transferred that day removing noise and certain artifacts.
- **TxTfrValMeanNtv** The mean count of native units transferred per transaction (i.e., the mean "size" of a transaction) that day.
- **TxTfrValMeanUSD:** The sum USD value of native units transferred divided by the count of transfers (i.e., the mean "size" in USD of a transfer) that day.
- **TxTfrValMedNtv:** The median count of native units transferred per transfer (i.e., the median "size" of a transfer) that day.
- **TxTfrValMedUSD**: It is the median count of USD value of native transfer that day
- **TxTfrValNtv**: The sum of native units transferred (i.e., the aggregate "size" of all transfers) that day. Hence it is a proxy of the aggregate size of all transfers that day;

TAKEAWAYS CHAPTER III

I. The youth of Cryptocurrency Market makes stiff the Data Selection process.

II. Most of free data sources, allows accessing to instantaneous information or limited time horizon. Platforms that gives access to complete historical data require payments.

III. CryproCompare allows accessing REST APIs through API key, quering data by different ways, such as Python programming, Javascript, Google Sheet adds-on. From this source, hourly and daily Financial and hourly Social data are extracted.

IV. Blockchian data is collected from Craft Source Maker, a machine learning powered data and analytics platform with granularity of one day.

V. For Each considered coin, two dataframes are created. The former includes hourly OHLCV and Social Data. The latter, includes daily OHLCV and Blockcchian data.

4. Preprocessing and Data Analysis

Once the data is stored, preprocessing step is required so that bias in Data interpretation and Modelling are avoided. In this Chapter, data is cleaned from outliers and null values, replaced with the median. Concluded this phase, HeatMap correlation Matrix is generated for Social and Blockchain dataframes, in order to insight on the main correlated variables with the Closing Price of that coin in the next time instant (t+1). These steps are crucial for the final experiment.

4.1. Null Values Detection

Null value represents the first noisy element within a dataset. Hence, it is important to detect it and replace through a specific procedure. The null value detection process has been applied to Social and Blockchain Datasets for BTC, LTC, ETH, XMR, XRP.

The percentage of Null value in BTC social dataframe is:

- close_h-1 0.000148

The percentage of Null value in BTC blockchain is:

- volumefrom 0.000592
- volumeto 0.000592
- AdrActCnt 0.000592
- BlkCnt 0.009476
- BlkSizeByte 0.000592
- BlkSizeMeanByte 0.000592
- DiffMean 0.000296
- FeeMeanUSD 0.000592
- FeeTotUSD 0.000888
- IssTotUSD 0.000592
- NVTAdj 0.000592
- TxTfrValAdjNtv 0.000592

- TxTfrValAdjUSD 0.000888
- TxTfrValMeanNtv 0.000592

The percentage of Null value in LTC social dataframe is:

- close_h-1 0.000148

The percentage of Null value in LTC blockchain is:

- close_d_1 0.000482

The percentage of Null values in XMR social dataframe is:

- code_repo_contributors 0.002957
- close_h-1 0.000148

In XMR blockchain is:

- close_d-1 0.000582

The percentage of Null value in ETH social is:

- close_h_1 0.000148

The percentage of Null value in ETH blockchain is:

- close_d_1 0.000654

The percentage of Null value in DASH social is:

- close_h_1 0.000148

The percentage of Null value in DASH blockchain is:

- close_d_1 0.000482

These statistics shows a low level of null value in analyzed datasets, giving important insights about the accuracy of the source taken into account. Null values detecting and replacing are important steps for Data Analytics, so that the further analysis are not biased and Knowledge extracted from data can result likelihood.

4.2. Outliers Detection and Replacing

The second bias element within a dataset is Outliers, which are data values that greatly differ from the major part of dataset values. Interquartile Range is a measure of Statistical dispersion, it is defined as the difference between the first (0.25), and the third quartiles (0.75) .In order to detect the outliers, the interquartile range method (IQR) is applied. It is proxy of dataset variability, assuming it dived in quartiles. The Outliers detection process based on IQR measure is defined from the following algorithm:

START

1. *Compute the 0.25 percentile: Q1 = quantile (0.25)*
2. *Compute the 0.75 percentile: Q3 = quantile (0.75)*
3. *Compute the Interquartile Range as: IQR = Q3 - Q1*
4. *Add 1.5 * IQR to third quartile and subtract 1.5* IQR from the first quartile*
5. *Check for any greater or lower number. These are Outliers:*

*Outliers = (db.data < (Q1 - 1.5 * IQR)) |(db.data> (Q3+ 1.5 * IQR))*

END

Once Outliers and Null values are detected, they are replaced with median values, as showed in the following code lines.

```
#replacing otliers and null values
median_1=db_1.median()
Q1_1 = db_1.quantile(0.25)
Q3_1 = db_1.quantile(0.75)
IQR_1 = Q3_1 - Q1_1
outliers=(db_1 < (Q1_1 - 1.5 * IQR_1)) |(db_1> (Q3_1 + 1.5 * IQR_1))

db_1[outliers] = np.nan

db_1.fillna(median_1, inplace=True )
```

The percentage of outliers is reported in the below tables for each coins and for Blockchain dataframe.

Table 4.1: Outliers % statistics for Blockchain dataset

Blockchin attributes	BTC	ETH	DASH	LTC	XMR (Monero)
AdrActCnt	0	0.9156	1.5429	3.0247	-
BlkCnt	4.353	3.9895	5.0145	3.8038	24.7818
BlkSizeByte	0	0	2.8447	2.8414	0.8144
BlkSizeMeanByte	0	0	2.7483	2.8414	0.8144
DiffMean	19.6624	0	0.3857	2.1082	3.025
FeeMeanUSD	14.6876	8.2407	8.1967	9.3492	9.7731
FeeTotUSD	15.3983	10.5952	11.8611	10.7699	14.1943
IssTotUSD	13.9177	7.1288	9.4021	7.3327	-
NVTAdj	4.8564	8.8947	0	6.4161	-
SplyCur	0	0	3.4716	0	0
TxCnt	0	0	1.6393	3.0247	1.4543
TxTfrCnt	0.2369	1.5043	6.702	3.0706	3.8394
TxTfrValAdjNtv	1.5102	4.3819	8.92	4.4913	-
TxTfrValAdjUSD	9.1797	6.7364	11.4272	9.8533	-
TxTfrValMeanNtv	12.9109	8.3061	12.8255	6.187	-
TxTfrValMeanUSD	8.7948	18.8358	14.1755	11.1824	-
TxTfrValMedNtv	8.7949	0	11.2825	3.8497	-
TxTfrValMedUSD	8.795	11.6416	9.3057	8.5243	0
TxTfrValNtv	8.7951	13.4729	10.9932	10.1283	0
close	8.7953	10.4644	17.4542	1.1916	13.4962
close_d-1	8.7954	10.4644	17.4542	1.1916	13.4962
high	8.7955	10.2027	17.9364	1.0541	12.5073
low	8.7956	10.399	18.756	1.3291	11.5183
open	8.7957	10.5298	17.8881	1.1916	13.4962
time	0	0	0	0	0
volumefrom	5.6263	5.6246	0	0	0
volumeto	0	0	0	0	0

Monero (XMR) lacks of such attributes as IssTotUSd, NVTAdj, so '-' symbol replaces the outlier percentage referred to that attribute. However, the table offers the possibility to say that the number of outliers in Blockchain datasets are limited; hence, no complex replacing approaches are required.

The percentage of outliers is reported in the below tables for each coins and for Social dataframe

Table 4.2: Outliers % statistics for Social dataset

Social attributes	BTC	ETH	DASH	LTC	XMR (Monero)
close	0	0.1183	0	0	0
analysis_page_views	0	0	0	0	0
charts_page_views	0	0	0	0	0
close_h-1	0	0.1183	0	0	0
code_repo_closed_issues	2.7942	0.0739	9.5801	24.8669	0.207
code_repo_forks	1.2714	1.2123	0	0	0
code_repo_stars	2.4985	1.2271	0	0	0
code_repo_subscribers	2.8681	1.4045	0	0	0.0739
comments	0	0	0	0	0
fb_likes	0	0	0	0	0
followers	0	0	0	0	0
forum_page_views	0	0	0	0	0
high	0	0.1331	0	0	0
influence_page_views	0	0	0	0	0
low	0	0.1183	0	0	0
markets_page_views	0	0	0	0	0
open	0	0.1183	0	0	0
overview_page_views	0	0	0	0	0
posts	0	0	0	0	2.0698
reddit_active_users	2.5872	0.3992	0.1183	0.547	0.1478
reddit_comments_per_hour	3.8143	5.7658	6.1798	5.618	6.2093
reddit_posts_per_hour	5.5884	1.9959	1.9219	1.0645	2.5872
reddit_subscribers	0	0	0	0	10.0828
total_page_views	0	0	0	0	0
trades_page_views	0	0	0	0	0
volumefrom	8.6044	8.427	0	9.1218	0
volumeto	9.5653	8.6931	10.1419	0	9.7428

How highlighted from above table, also Social datasets present low percentage of outliers, detected through Interquartile method. Hence, in order to make fast and easy the Cleaning process, Outliers and Null values are replaced with Median.

In such case Median replacing is not suggested, due to the strong bias that data could continue to show.

4.3. HeatMap tool and Data Visualization

Once cleaned dataset, it is important to evaluate the possible correlation across different variables. In particular, the scope of this work is to evaluate possible relationship between the Closing Price at the next time instant t+1 and the different attributes at the time instant t.

In simple terms, what it has been verified is the possibility of existing relationships between financial time series with Social and Blockchain variables.

In order to do this, HeatMap correlation matrices are generated for each coin, so that it is possible to infer on the existence of correlations. The following lines code display how HeatMap Pearson Correlation Matrices are built for Ethereum coin, through MatPlotlib Python library.

```
# heat correlAtion matrix for social dataframe
def heatMap(df):
#Create Correlation df
corr = df.corr()
#Plot figsize
fig, ax = plt.subplots(figsize=(10, 10))
#Generate Color Map
colormap = sns.diverging_palette(220, 10, as_cmap=True)
#Generate Heat Map, allow annotations and place floats in map
sns.heatmap(corr, cmap=colormap, annot=True, fmt=".1f")
#Apply xticks
plt.xticks(range(len(corr.columns)), corr.columns);
#Apply yticks
plt.yticks(range(len(corr.columns)), corr.columns)
#show plot
plt.show()
plt.savefig('Socialmap.png')


image=heatMap(db_1)
```

ETH Blockchain and Social HeatMaps, generated with MatPlotlib, are reported below:

Fig 4.1: Blockchain ETH MatplotLib HeatMap

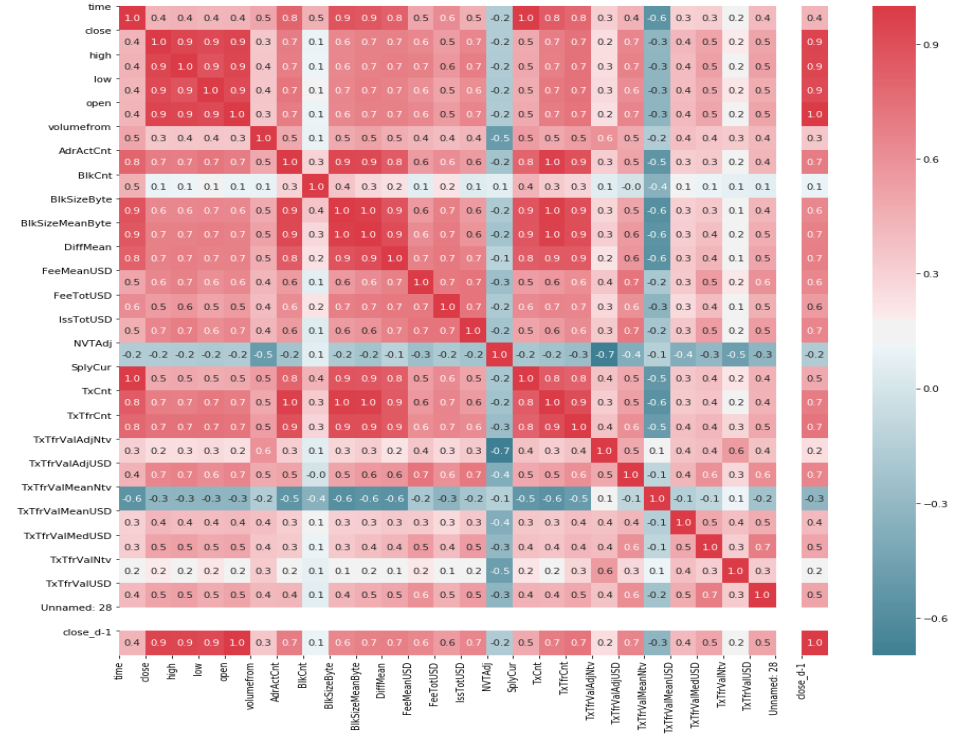
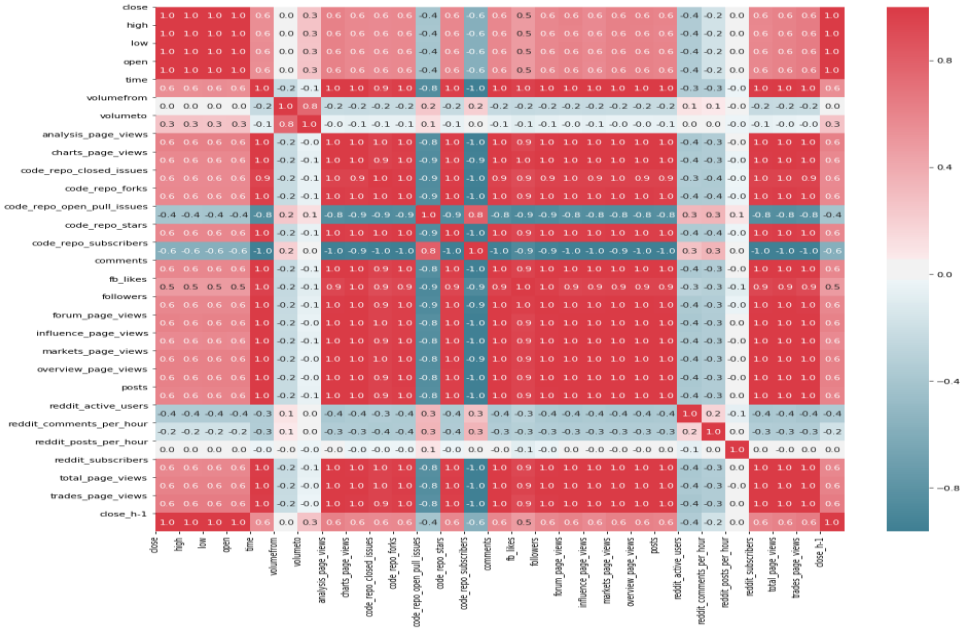


Fig 4.2: Social ETH MatplotLib HeatMap



The Pearson correlation presents value in the range of [-1 , +1], where -1 indicates the presence of maximum negative correlation, and +1 the maximum positive one. Positive means that an increase of

one variable implies an increase in the correlated one. Contrary, an increase in one variable implies a decrease in the other, if negative correlated.

Zero shows the absence of correlation between the variables, while intermediate positive values between 0 and 1 shows a positive gradient of correlation, negative in the case of intermediates values between]-1,0[.

The python function HetaMap.Corr(db) uses this pictorial representation that allows identifying the most correlated variable, through Colors. The intense red insights the maximum correlation, whereas, most strength grey implies strength negative correlation.

Both Social and Blockchain HeatMaps shows correlation between Closing price at t+1 and High, Low, Close and Open at the time instant t. It is not a surprise, Financial Data Presents an obvious correlation, and what we expected is a top linkage between closing price at t+1 and closing at t, because of they are similar value. Closing at t+1 is a new column added to dataset, simply shifting of one entry the Closing price. It is important to notice that, due to the implementation of different granularity in the two dataframes, it will be used Closing d+1, when referring on Blockchian dataset and Closing h+1, when referring to Social data, reflecting the daily and hourly sampling frequency.

There is not a clear relationship with volume; in the time series with hour granularity they show a Pearson correlation of 0.3, whereas in 0.2 in the daily dataset.

An interesting result is the strong pattern defined frim close d+1 and Block metrics. The correlations with DiffMean, FeeMeanUSD and TxTfrValMedUSD are 0.7, a good level of positive correlation.

In order to better understand, what positive relationship means, it is possible to plot correlated time series. Data shows different dimension and unit measure; hence data normalization is required, assuming the normality distribution of the data, under Central Limit Theorem hypothesis. Doing this, StandardScaler () function is used, once preprocessing module is imported from sklearn library.

```
1. from sklearn.preprocessing import StandardScaler
2. s = StandardScaler()
3. scaled=s.fit_transform(close_1)
```

The following MatplotLib plots (Fig 4.1 and 4.2) compare the Ethereum Closing price at t+1 with FeeMeanUSD and TxTfrValMedUsd. The former defines the value of the mean fee per transaction that day, whereas the latter the mean size of a transfer that day.

Both variables provide a good degree of correlation and how pictured in the plots below, they follow the price movements with limited lag. However, also in the timespan with high volatility, these features appear able to react and to strike a pose, comparable with the price oscillations.

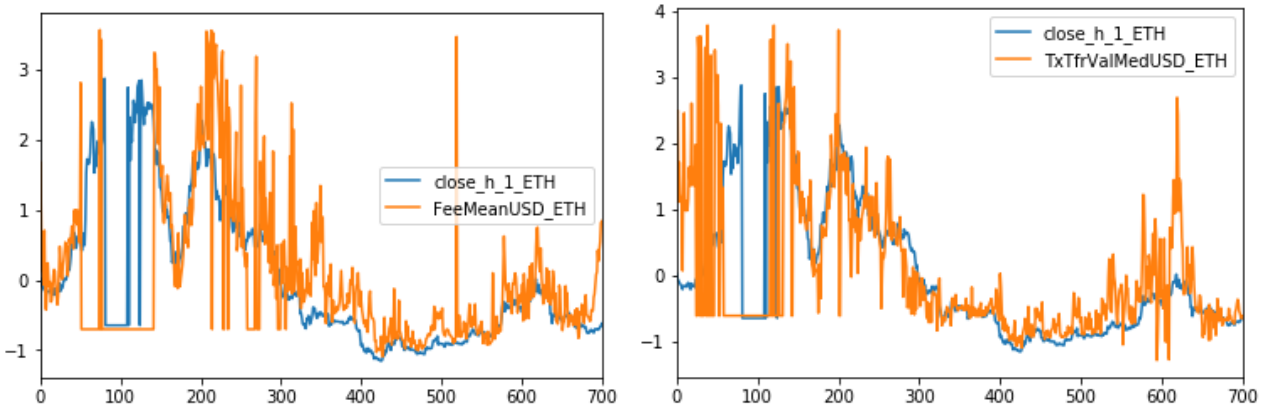


Fig 4.2 and Fig 4.3: ETH MatplotLib plots, where Block variables are compared with Closing Eth at t+1 in the last 700 days

Different is the relationship between Social Variables and Closing ETH. In the first plot, price is compared with hourly Reddit posts about Ethereum, whereas the second reports the relationship between price and Official Ethereum posts' likes.

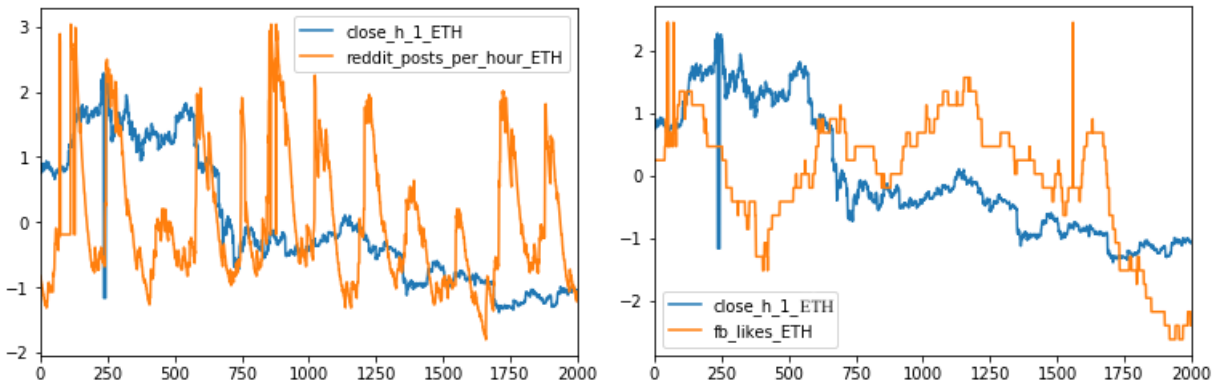


Fig 4.4 and Fig 4.5: ETH MatplotLib plots, where Social variables are compared with Closing Eth at t+1, in the last 2000 hours

The hourly Reddit posts movement follows a completely Random Walk and does not show meaningful signal about price. Facebook likes, instead, show a more similar movement, but a price oscillation is

overrated. When the price goes up, the sentiment appears over expected, whereas, when it goes down, sentiment drastically falls.

BTC Close (t+1) dilspay a different behaviour, how observable in the below charts. Social data, in this case strongly afflicts BTC Financial Time Series, most of the attributes show a Person correlation value of 0.9. Code_repo_subscribers showes in the selected time horizon of Fig 4.5 a clear movments commonality, and just in few time instants there is a reaction lag.

Facebook likes (fb_likes_BTC), presents good level of positive correlation as well; but in how illustrated in Fig 4.6, Social Features do not proportionally reflect the Price peaks.

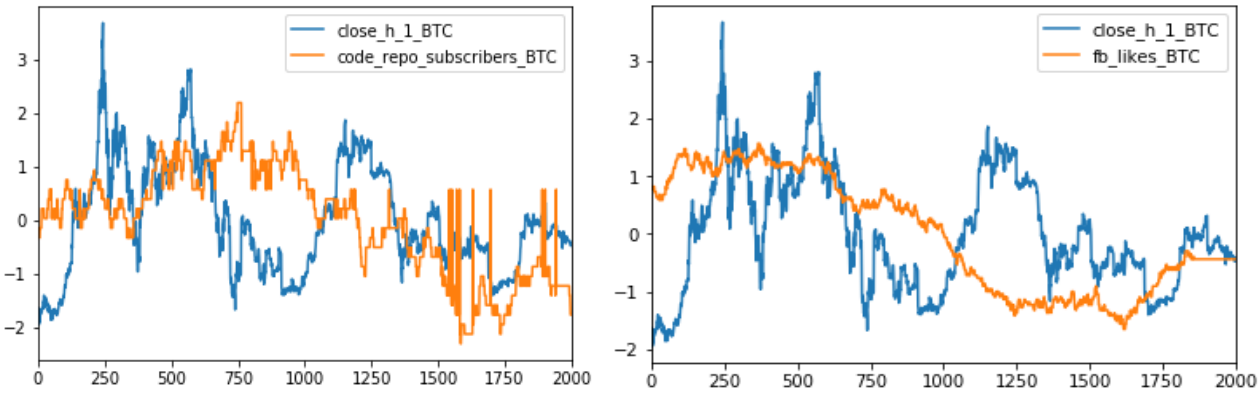


Fig 4.6 and Fig 4.7: BTC MatplotLib plots, where Social variables are compared with Closing Bth at t+1 from 1/7/19 2.00 to 12/10/19 11.00

Looking at the blockchain BTC HeatMap instead, TxTfrValMedUSD_BTC shows an interesting correlated oscillation with Close (d+1) (roughly 0.7), also in huge volatility periods.

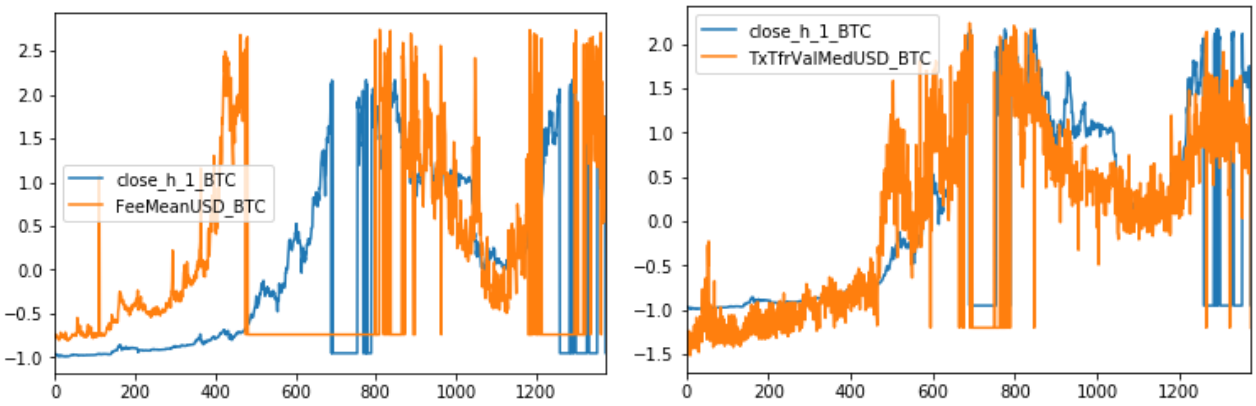


Fig 4.8 and Fig 4.9: BTC MatplotLib plots, where Block variables are compared withClosing Bth at t+1 from 18/09/19 to 12/10/19

Daily Miners Fee Mean (FeeMeanUSD_BTC), has a Pearson value of 0.4, and how represented in Fig 4.7, the plotted time series altering phase of good match and phase with clear reaction lag.

Therefore, Blockchain variables entails important correlation with daily ETH price, and with shallow intensity with daily BTC. Contrary, Social data affects drastically on financial BTC time series, whereas inconsistent relationship with hourly ETH price is emerged from the analysis.

At the same time, it is important to notice that correlation plots focus on a limited time horizon, in order to frame with more detailed the time series oscillation. Extending the plot to whole dataset, Social Data reflects the observation exposed above, whereas Block attributes displays a strange pattern, observable on the below figures. (Fig 4.9 and 4.10)

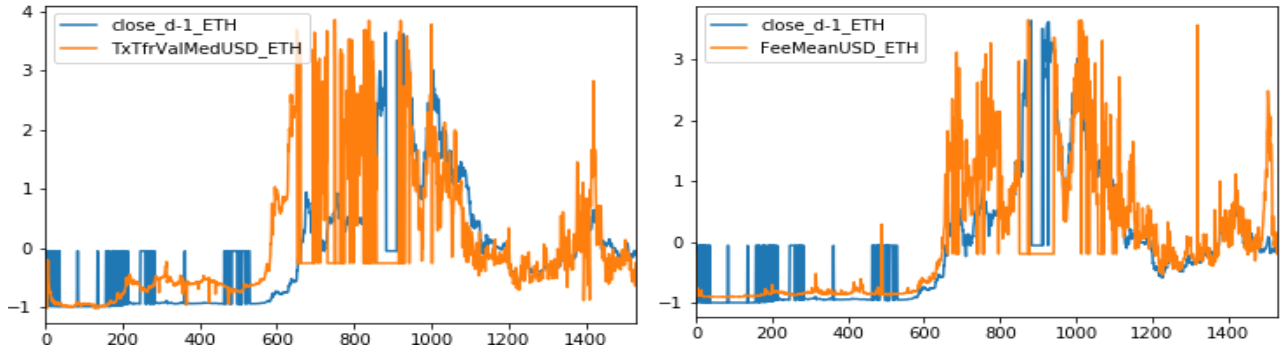


Fig 4.10 and Fig 4.11: ETH MatplotLib plots, where Block variables are compared with Closing Eth at t+1 for whole time horizon

Blockchain data show a common attitude, they are able to strike a pose similar with financial movements, but the there is a clear transitory phase close to the coin emission, where the correlation is completely out of the picture.

The HeatMap analysis is extended to XMR, DASH and LTC coins, as well.

Due to its extreme infancy, XMR HeatMap does not provides interesting relationship with the Social data. Launched on April 18th , 2014, it is the most unpopular coin on Social and Developing platforms from analyzed coin, explaining the lack of meaningful correlation. Difficult to explain, is instead the lack of deep connection with Block features. Just Trade and Market Pages Views show a timid affinity, with Pearson values of 0.4.

LTC Social date displays a good feeling with Closing trends, and appears as important candidate to explaining part of price spread. Trade Page Views, how observable in Fig 4.13, reflects at most Closing oscillations, evidencing a discrete dose of kinship through a Pearson Correlation of 06.

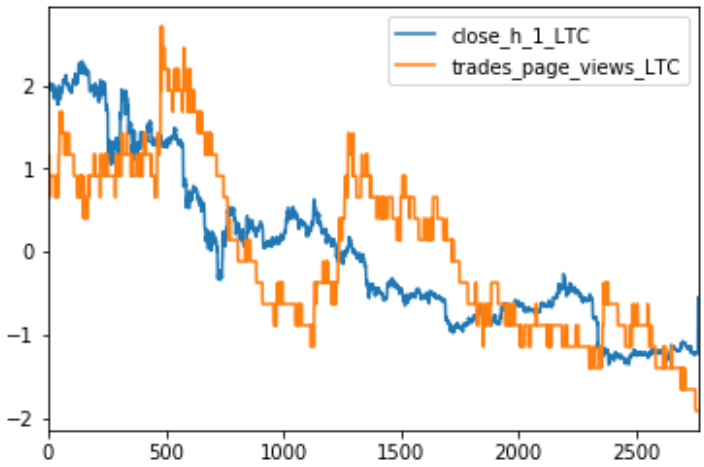


Fig 4.12: LTC MatplotLib, where Social Variables are compared with Closing LTC at t+1 from 1/7/19 2.00 to 12/10/19 11.00

Contrary Blockchain data, looks few interests to track Daily Closing Price at the next time instant t+1, and most of chain attributes appears strongly negative correlated, as illustrated in Fig 4.14

The lack of such Block attributes, instead affects the HeatMap DASH analysis. The matrix is undersized compared with the others. Nevertheless, variables as Block Size (Byte), Difficulty to Find new block and Currency Supply, display a discrete level of feeling (0.6), whereas features as FeeMeanUSD and FeeTotUSD demonstrates a negative correlation. This entails the presence of an ambiguous connection between DASH financial data and Blockchian attributes.
Social factors instead show interesting correlations, such as:

- Corr (Close_DASH (d+1), Total_page_view) = o.6
- Corr (Close_DASH (d+1), Market_page_view) = o.7
- Corr (Close_DASH (d+1), Fb_likes) = o.5
- Corr (Close_DASH (d+1), Posts) = o.6

The correlation analysis, combined with data visualization is a powerful tool that allows inferring on data insights and deciding which data can empirically implemented into Deep Learning phase.

For instance, can be interesting to design a machine algorithm experiment, trained on hourly OHLCV and Blockchain data for ETH, BTC.

TAKEAWAYS CHAPTER IV

I. *Outliers and Null Value within Blockchain and Social Dataset are limited. They have been detected with IQR method and replaced with median value.*

II. *Financial Social Attributes show interesting correlation with BTC, ETH and DASH coins;*

III. *Blockchain attributes displayed good feelings with ETH.*

IV. *XMR does not strike similar pose nor with Siocial, nor Blockchain attributes;*

5. MLP and SVC Simulations and Result Analysis

This chapter formalizes the applications of Multilayer Perceptron, a form of Neural Network algorithm, and Support Vector Machine to the data analyzed so far. The first part is dedicated to the main properties of these machine learning algorithms, whereas the rest of the chapter focuses on the experiments design and the running outcomes.

5.1. Artificial Neural Network and Multilayer Perceptron

Artificial Neural Network is one the most recognized example of mimesis, through which an interconnection of nodes replies in simpler way Biological Brain Neurons. In practice, it is a computing system that tries to learn specific task basing on provided examples.

ANN is defined from a collection of specific nodes, called artificial neurons, linked through specific connections similar to biological synapses, that allow transmitting signals from input to output neurons.

For this specific application, Multilayer Perceptron, a form of Artificial Neural Network, has been applied. MLP is a supervised machine learning algorithm, that elaborates input data through the presence of one or more hidden layers and transmit a final output.

Formally it can be represented as oriented graph (Fig 5.1) that counts the following elements:

1. **Input nodes**: They are the features of input that model must process and once define relationship, understand how can be transformed through an activation function f(z) that allows reaching specific target value;
2. **Hidden Layers**: These are intermediates nodes with the main goal to understand which are the relationship between input and output through training examples. The number of hidden layers are proportional to the complexity of relationship between input-outputs. Overestimation of that number can lead model to enrich inexistent relationships.

3. **Output nodes**: They are the nodes that reproduce the target values, once all the input data are processed and involved into activation function

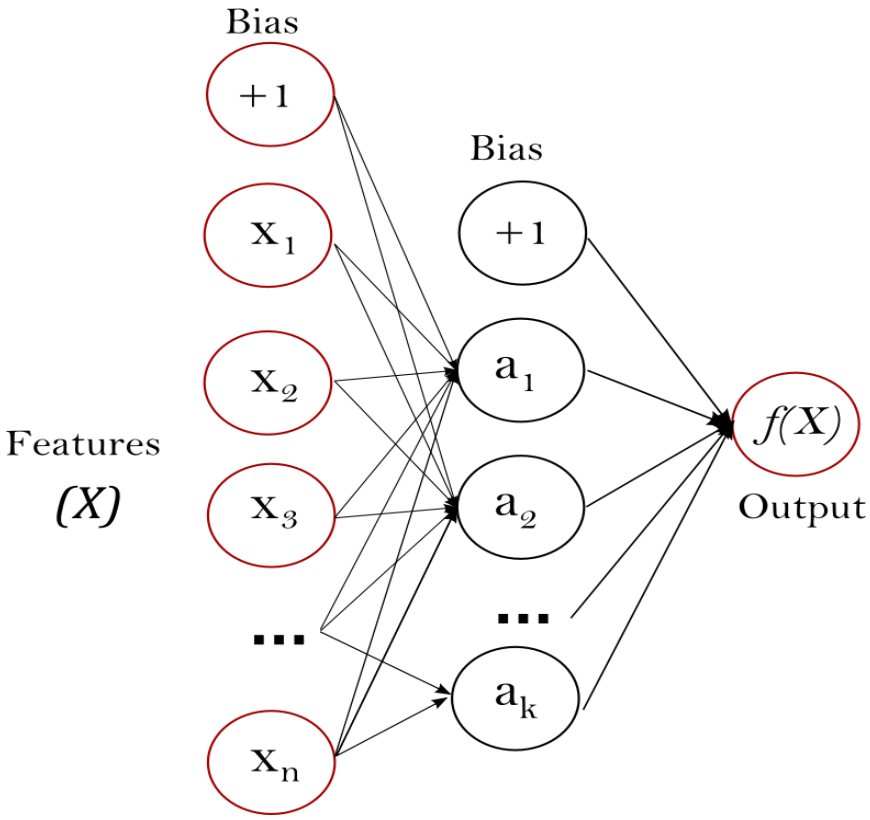


Fig 5.1: Oriented Graph showing how ANN works

In mathematical terms the model tries to train a learning function $f(z)$ on the basis of m input variables or features $\{ Xi \mid X1, X2, X3 \ldots Xm \}$ that suppose have a specific influence on the outcome. Supposing to have one hidden layer h with n nodes, each neuron inside it transform the input values with a weighted linear summation of different features and bias:

$$Z^i = W^i{}_1 * X1 + W^i{}_2 * X2 + \ldots + W^i{}_m * Xm + Bias \ , \ \forall i \in \text{hidden layer}$$

followed by a nonlinear activation function, so that it possible to get the input for the next layer. Supposing to have five features and two nodes in the hidden layer, the matrix formula is:

$$\begin{bmatrix} W1,1 & W2,1 & W3,1 & W4,1 & W5,1 \\ W1,2 & W2,2 & W3,2 & W4,2 & W5,2 \end{bmatrix} \times \begin{bmatrix} X1 \\ X2 \\ X3 \\ X4 \\ X5 \end{bmatrix} + \begin{bmatrix} Bias1 \\ Bias2 \end{bmatrix} = \begin{bmatrix} Z1 \\ Z2 \end{bmatrix}$$

This summation is replaced for each hidden node on that layers, until output node is reached.

$$Y = f\left(\sum_{i=1} Z\,i + Bias\right)$$

The training process pursues to adjust weight for each layer connection through different approaches. One of the most common is the Backpropagation that combined with descent gradient computation tries to optimize W_i.

At glance, what this method does, is to consider a cost function, like MSE and minimize it, computing the function gradient respect weight W_i and bias.

5.2. Support Vector Classifier

Support Vector Machine is another kind of supervised machine learning algorithm that trained on a set of examples, allows assigning new inputs to specific class. A SVC model is a representation of training datasets as points in geometrical space (Fig 5.2), so that it is possible to distinguish examples in net categories, divided by a clear gap that must be as wide as possible.

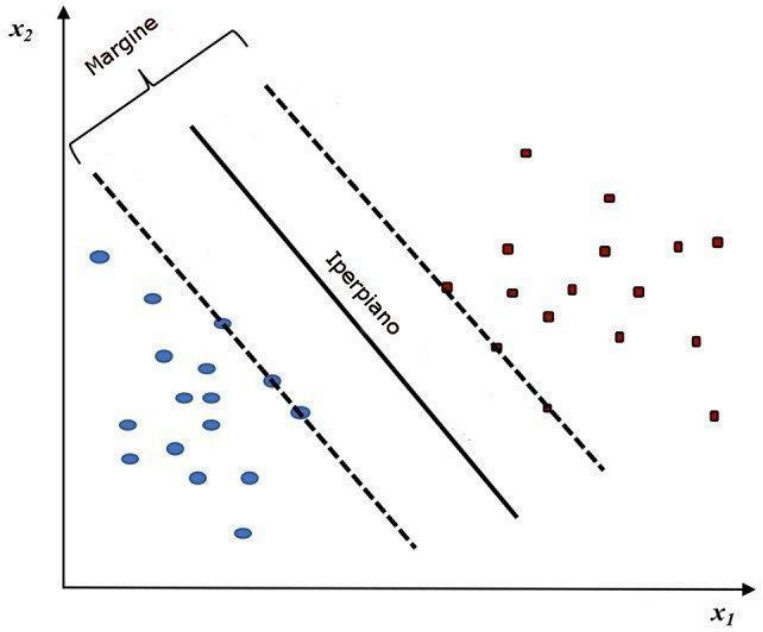


Fig 5.2: SVM hyperplane and margin representation in geometrical space

Basing on this step, new inputs are analyzed and a belonging class is defined. From mathematical point of view, the underlying concept of this algorithm is the idea of searching a hyperplane able to divide the classes and that maximizes the distance between both first point classes, called Margin.

Given a training dataset of n points $\{(Xi,Yi) \mid (X1,Y1), (X2,Y2)...(Xn,Yn)\}$ where Yi can be -1 or +1 and indicates the Xi belonging class. Hence Hyperplane can be expressed by the following linear combination:

$$W_1 * X1 + W_2 * X2 + \ldots + W_n * Xn + Bias = 0$$

and the margin equation allows identifying belonging class Yi

$$W_1 * X1 + W_2 * X2 + \ldots + W_n * Xn + Bias \geq 1 \quad if\ Yi = +1$$

$$W_1 * X1 + W_2 * X2 + \ldots + W_n * Xn + Bias \leq -1 \quad if\ Yi = -1$$

These constraints can be put together and in order to achieve the optimal Hyperplane it is required to solve the following quadratic model:

$$\textbf{o.f. : } \min \quad \| W \|$$

$$\textbf{st: } \quad Y_i \left(W_1 * X1 + W_2 * X2 + \ldots + W_n * Xn + Bias \right) \geq 1$$

The lack of linearity of this model requires the application of Lagrange multipliers and Karrush-Kun-Taker conditions. This mathematical formulation represents the easiest case, due to the assumption of classification linearity. But most of applications requires not linear classification, based on more complex mathematical concepts, that are out of the thesis proposals. However, as explained in the next paragraph, the machine learning underlying optimization tools are treated as black-box, so that it is possible to stay focus on simulation results.

5.3. Experimental design

Sklearn Python library allows approaching Multilayer Perceptron and Support Vector Classifier, without coding algorithms from scratch, and considering behind optimization mathematics as a black-box. This has the great advantage of spending more time on simulations and result analysis rather than coding. Both method can be classification and regression algorithm, that in simple terms means that produce as output respectively discrete and continuous values. This work considers the former approach, so it is important to define a discrete targeting value within experimental datasets. In order to do this, a new attribute has added to dataset, called price changing computed as:

$$Price\ changing\ (t) = \frac{Closing\ Price\ (t+1) - ClosingPrice(t)}{ClosingPrice(t)}$$

and then target class is qualified as:

- ***Class (t) = 'Upper', if Price Changing (t) $\geq$ + 0.05***
- ***Class (t) = ' No Signal', if -0.05 $\prec$ Price Changing (t) $\prec$ 0.05***
- ***Class (t) = ' Lower', if Price Changing $\preccurlyeq$ - 0.05***

Usually, Trading System involves value threshold of +/- 0.1%. Following experiments involves a value of +/- 0.5%, in order to give more feasibility and dynamicity to these simulations.

This has been involved in python, through the use of lambda syntax, as shown in the following code lines:

```
def target (r):
   if r>= 0.005:
      class_value= 'Upper'
   else: class_value = 'No signal'
   return(class_value)

db_2['classe']=db_2.close_change.apply(
      lambda x: target(x) if x>-0.005


      else 'Lower')
```

For sake of simplicity, the implementation of two algorithms are settled through two phases: trainingand testing, avoiding the validation phase. Training percentage is settled to standard value of 67 %, so that the remaining dataset (1 -0.67) is used for testing evaluation. The training settings fix a training matrix of attributes and a training target vector as graphically displayed in the below figures (Fig 5.3 and 5.4).

Record	open	volumeto	.....	BlkSizeByte
1	0.6747	371.79	.....	3282693
2	3.0	1438.16	.....	3508878
3	1.2	0.0	.....	3167541
4	1.2	0.0	.....	3316883
5	1.2	7419.73	.....	3653834
	.....	.....	.....	.....
	.....	.....	.....	.....
	.....	.....	.....	.....
	.....	.....	.....	.....
1877	0.998	7729.17	.....	5053
1878	0.9	11933.21	.....	5028
1879	0.75	2346.43	.....	5079
1880	0.88	1866.15	.....	5119

class
Lower
Lower
No Signal
No Signal
Lower
.....
.....
.....
.....
Upper
Upper
Lower
Upper

Fig 5.3: Blockchain and OHLCV matrix and class training vector

record	open	Facebook_likes	.....	comments
1	133.49		.....	79895
2	134.07		.....	79895
3	133.49		.....	79895
4	132.89		.....	79895
5	133.65		.....	79895
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....
4025	163.42	163.47	.....	2065.04
4026	162.47	163.67	.....	8143.63

class
Lower
Lower
No Signal
No Signal
Lower
.....
.....
.....
.....
Upper
Upper
Lower
Upper

Fig 5.4: Social and OHLCV matrix and class training vector

This phase is implemented in Python through the following commands:

```
train_perc = 0.66
y = db_2['classe'].values
print(db_2.head())
db_2 = db_2.drop(columns=['classe','time'])

X = db_2.values
print(type(X))

clf_mlp = MLPClassifier()
clf_svc= SVC()

rows, cols = X.shape # (365, 20)
train_size = int(rows*train_perc)

X_train = X[:train_size, :]
y_train = y[:train_size]
X_test = X[train_size:, :]
y_test = y[train_size:]
print (y_test)
print(X_test)

clf_mlp.fit(X_train, y_train)
y_pred_mlp= clf_mlp.predict(X_test)
```

Classification Accuracy has also evaluated as performance metrics, and has been computed for both MLP and SCV algorithms outcomes for each coin. It is one of the most diffused way to measure Machine Learning outputs and it is defined from the following formula:

$$Accuracy = \frac{Correct\ predictions\ number}{Total\ prediction\ number}$$

It is the ratio of number of correct predictions to the total predictions' number. It is a sort of naive and intuitive approach that offers global insight about classification performances.

5.4. Results Analysis

Table 5.1 frames the performances achieved for each coin. The yellow side displays accuracies of Multilayer Perceptron and Support Vector Classifier trained on Datasets with daily sampling frequency. The blue side proposes coins 'accuracy of MLP and SVC, settled on hourly granularity. Granularity is not the only difference between the sides, while the former defines algorithms performances based on Blockchain and OHLCV data, the latter refers to Social and OHLCV data.

Table 5.1: Accuracy values each coins

	MLP_DAY	SVC_DAY	MLP_HOUR	SVC_HOUR
BTC	0.283972	0.39459	0.6743475	0.730869
ETH	0.447867	0.483478	0.5678897	0.687688
XMR	0.451282	0.442735	0.5656521	0.578695
LTC	0.45148	0.47708	0.54608	0.59304
DASH	0.448226	0.523445	0.5147826	0.599565

The first insight is the evident difference of performances between the two sides. The average MPL_DAY accuracy computed across the five coins is 0.4077, whereas SVC_DAY one is 0.4672. The average values for MLP_HOUR and SVC_HOUR respectively are 0.590326 and 0.6573. This is not a surprise due to the wider amount of data that hourly dataset offers compared with daily one. The first contains 6764 entries, whereas the second depends on coins emission date, but with an average entries counting of 2745. This difference affects the algorithm training process and as consequence the accuracy. This leads to an important consideration about algorithms' evaluations. While for the hourly datasets, targets

forecasting yields are directly comparable, daily ones are function of the amount of data collected. For instance, it is likelihood that ancient coins, as Bitcoin and Litecoin get superior accuracy.

What is probably more interesting is the difference of performance between the two machine learning algorithms. Excluding Monero (XMR) coin, where MLP and SVC own aligned efficiency, the Support Vector Classifier emerges as the most performant.

Hourly Bitcoin Price Forecasting have produced the highest accuracy, SVC_Accuracy (BTC) = 0.73089 and MLP_Accuracy (BTC)=0.6748 are satisfactory values and confirm what explained in the Data Analysis chapter: Social Variables, combined with Financial time series result interesting Bitcoin spread explanation factors. On the other side, daily BTC price forecastings result inaccurate for both algorithms (0.2839 and 0.3945), but also this can be easily explained. The low accuracies are fruit of lack of greatest correlations between coin closing (t+1) movements and Blockchain variables and the limited data through algorithm are trained on. Support Vector Classifier has guaranteed higher performance due to its higher capability to explain output, without requiring extremely huge amount of training data.

Ethereum coin partially reflects the consideration of chapter 4. Due to the interesting relationship with Blockchain variables, daily accuracy improves (0.44768 and 0.4838), but does not justify at all the expectations that movements correlation created. In this case, it results stiff to infer on what percentage of fault is allocable to the limited training datasets.

Difficult to explain is also the optimistic result, given from hourly accuracy (0.56789 and 0.6878). Social data in fact, just report a timid affinity, making more than suspicious the presence of good randomality dose within algorithms execution.

Although the lack of deep connection between XMR closing (t+1) and Blockchain attributes, daily target forecastings present discrete values (0.451282 and 0.4427) compared with the mean values (0.4077 and 0.4672). The blue side do not provide particular insights, hourly XMR accuracies (0.56565 and 0.578695) are probably attributable at most to wider available data dump.

Strange is instead the case of DASH simulation outcomes; although the limited number of Blockchain attributes, daily SVC forecasting achieves the best result (SVC_DAY (DASH) $\simeq$ 0.523445), whereas hourly forecasting is comparable with the mean (0.51478 and 0.59952).

5.5. Conclusion and Future Perspective

This thesis work is just a preliminar and shallow approach to the complex and immature world of Cryptocurrency Analysis and forecastings. Due to its Market Capitalization of $207 Billion, it is today an important financial reality that throughout ten years have reached great popularity. The lack of

seasonality and the huge volatility makes Cryptocurrency Price Movements similar to Random Walk and beat on it appears absolutely risky. The first step of this work has been the State of Art analysis that allowed taking confidence with this complex and stiff world. The central and expensive phase has defined from the exploration of the main available data sources, that have represented the real work bottleneck, as represented in the below Gantt chart (Fig 5.3).

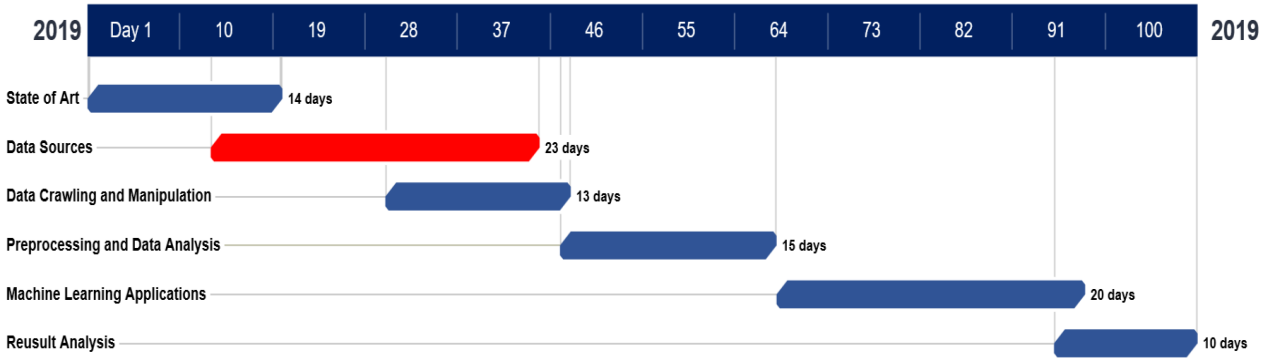


Fig 5.5: Gantt chart showing the work thesis scheduling

Although for Stock and Exchange Markets, financial and not, quantitative data with flexible sampling frequency is available, for digital coin it is today complicate to crawl it. The , main goal of this thesis has been inferring on the main cryptocurrency features and exploring available quantitative variable that can lead to a sort of correlation with price spread.5

After focussed analysis of collected sources, two of them have taken into account. CryptoCompare.com and Craft Market maker. The former offers historical OHLCV and Social Data with Hourly and Daily granularity through Rest APIs, whereas the latter allows downloading interesting Blockchain metrics as Block Size in Byte, Difficulty to find new Block, Miner remuneration in csv format. By the way, two datasets are built for each coin, one with daily granularity that combines OHLCV and Chain metrics and a hourly one, involving OHLCV and Social data as Reddit, Facebook likes, posts, comments or GitHub activities. This phase has been defined in the Gantt chart as Data Crawling and Manipulation.

Cleaned from null and outlier values datasets are analyzed, in order to insight on possible correlation between Closing coin price at the next time instant t+1 and metrics at time t. This analysis has confirmed

the ambiguity and diversification that afflict Cryptocurrency market. Social metrics, as Reddit Subscribers numbers or Facebook Likes have resulted able to strike a pose similar to price oscillations in case of Bitcoin and Dash. At the same time, Blockchain features have displayed evident price movement signal with Ethereum, although transitory phase appeared completely out of the picture. Such coins as XRP and Litecoin have not produced juicy insights. This work phases has been baptized as Preprocessing and Data Analysis. In order to explore the price volatility explanatory power of collected metrics, classification algorithms are applied. The main simulation purpose is to forecast a target class ('Upper', 'Lower', 'No Signal') related to Trading Signals, basing on the OHLCV, Blockchain and Social input features. The accuracies of daily datasets have resulted poor, and this highlight how Cryptocurrency daily data are insufficient to train machine learning algorithm, due to its young existence. Nevertheless, it is evident that coins showing good correlation with Blockchain features, improves its performances. It is the case of ETH and DASH. Hourly dataset offers more accurate classification outcomes, thanks to the extended training data, but also in this case correlation evidences impacts on accuracy measures, as demonstrated from BTC that accounts a value of 0.7308.

At glance, what this work says is that the correlation analysis between price time series and Global metrics allows inferring on which factors affect market volatility.

Social sentiment and Blockchain metrics, tested in this experiments show affinity with such coins, but represented just a drop in the ocean.

Cryptocurrency price today spread appears guided by multitude factors, and such of them are lead from irrationality and randomality.

This thesis is the basis for future works, that can start from some pillars:

- It is always need to take an eye to Data Source and explore new one, that offer quantitative data with fittest granularity about Cryptocurrency.
- Explore correlations between Cryptocurrency prices and news factors even if they do not provide rationale feeling at first.
- Collect Data Dump for training Machine Learning algorithm. Hence could be interesting to evaluate data with granularity of hour or minute extended on adequate time horizon (4-5 years).
- Involve both classification and regression machine learning algorithms and deep learnings algorithms.

TAKEAWAYS CHAPTER IV

I. Cryptocurrencies Daily datasets result innacurate for training machine learning algorithms, due the short amount of records offered;

II. *Coin showing interesting correlation with Social Attributes, present the highest Forecasting Accuracy;*

III. *SVC perform better rather than MLP.*

IV. *Future Work may consider datasets with fittest granularity and extended time horizon, combining Blockchain and Social Attributes as algorithm inputs.*

APPENDIX I

1. NeuroSantiment:

I. Source: https://neuro.santiment.net/

II. Data access: REST API and WebSocket API via GraphQl, or through sanpy Python package. On chian data is collected directly from running nodes, making it faster and reliable, by eliminating potential point of failure.

III. Accessibility: Free plan that allows accessing to:

- *last 3 months data, excluding the last 24 hours*
- *20 API calls / minute*
- *standard metrics*

Premium and Corporate versions allow unlimited historical access data advanced metrics.

IV. Granularity: Most metrics are daily, whereas social metrics can be setted on hourly.

V. Blockchain features: they are daily data available for BTC, ETH and EOS and include the following metrics:

- *Daily Active Address: It includes the number of unique addresses that participated in the transfers of given token during the day. It is available for BTC, ETH and EOS;*
- *Network Growth: It shows the number of new addresses being created on the network each day;*
- *Token Age Consumed: It shows the amount of tokens changing addresses on a certain date multiplied the number of blocks created on the blockchain since they last moved;*
- *Average Token Age Consumed in Days: it represents the average number of days that the tokens were idle before being moved on a certain date;*
- *Exchange Flow Balance: it show the combined values of tokens moving in and out of exchange wallets on a certain date;*
- *Exchange Flow: This graph shows the amount of inflow and outflow plotted separately, based on the previous metric;*
- *Transaction Volume: it shows the aggregate number of tokens across all transactions that happened on the network on a certain date;*
- *Token Circulation: Spread of idle over time;*
- *Velocity of Tokens: Average number of times that a token changes wallets each day;*
- *Gas used: return gas used by a blockchain for the token supply computations. It is just available for ETH;*
- *Miners balance: it returns the Miners Balances over time. It is just available for ETH;*
- *Top 100 Transactions: List of the top 100 transactions by volume in a given time frame;*
- *MVRV Ratio: it shows the whole market value divided by the realized value of the network,*

- *NVT Ratio: It returns the ratio of Market Value to Realized Value, calculated daily for Ethereum and available ERC-20 tokens;*

VI. *Financial Data: they include:*
- *History Price: It shows the daily price, volume and market capitalization for more each digital asset in USD or BTC;*
- *'OHLC': it allow accessing to daily Open/HIgh/Low/Close Price for each digital asset;*

VII. *Social Data: It includes differents sources and time range. In particular it offers granularity ranging from minute to week and different social Volume Type are available:*

- *- "PROFESSIONAL_TRADERS_CHAT_OVERVIEW" - shows how many times the given project has been mentioned in the professional traders chat*
  - "TELEGRAM_CHATS_OVERVIEW" - shows how many times the given project has been mentioned across all telegram chats, except the project's own community chat (if there is one)
  - "TELEGRAM_DISCUSSION_OVERVIEW" - the general volume of messages in the project's community chat (if there is one)
  - "DISCORD_DISCUSSION_OVERVIEW" - shows how many times the given project has been mentioned in the discord channels

  - *Social Dominance: it returns the dominance of a certain asset has over a social channel in percentage, compared to all projects mentioned in that channel;*
  - *Social Volume: it returns a list of mentions count for a given project and time interval;*
  - *Topic Search: it return the lists with the mentions of the search phrase from the selected source (Telegram, Professional Traders Chat, Reddit, Discord);*
  - *History Twitter data: it count the number of followers o the official Crypto profile;*
  - *Github Activity: it returns the GitHub activities for a given project and time interval;*
  - *News Collection: It returns the News for a given digital asset in a specific timespan;*

2. Alternative.me:

I. *Source: https://alternative.me/crypto/api/*
II. *Data access: API*
III. *Accessibility: The access is completely free, with the limit of 60 requests per minute*
IV. *Granularity: The data are updated each five minutes;*

V. *Financial Data: they are available for each cryptocurrency and are updated each 5 minute, but is not possible to collect the time series value of past timespan. it is possible to access to:*

- *Ticker: It provides Coin and Token prices updated every 5 minute and shows the following parameters:*
  - *name*
  - *id*
  - *symbol*
  - *website_slug*
  - *rank*
- *Global: It provides global market information at a glance. It involves the following parameters:*
  - *total_market_cap_usd*
  - *total_24h_volume_usd*
  - *bitcoin_percentage_of_market_cap*
  - *active_markets*

Ticker and Global includes the following exchanges conversion:

- *CRYPTO/USD*
- *CRYPTO/EUR*
- *CRYPTO/GBP*
- *CRYPTO/RUB*
- *CRYPTO/JPY*
- *CRYPTO/CAD*
- *CRYPTO/LTC*

VI. *Social data: Crypto Fear & Greed Index data are available with a granularity of 1 day and time span from Feb 01, 2018 until now. This index reflects the emotional behaviour of the crypto market over time and it ranges from 0 to 100. This index, viable only for Bitcoin takes into account different sources, with different weights:*

- *Bitcoin Volatility (25%)*
- *Market Momentum/Volume (25%)*
- *Twitter Sentiment Analysis (15%)*
- *Surveys (15%)*
- *Dominance over Crypto Market (10%)*
- *Trends (10%)*

3. CoinApi.io:

I. *Source: https://docs.coinapi.io*
II. *Data access: REST API and WebSocket API*

III. *Accessibility: Free plan that allows accessing to 100 daily requests, WebSocket access for Trades tests only, but no access to Streaming data. Howerver it allows accessing to Bitstamp symbols and quotes data only. Premium and Corporate versions allow 10k Daily requests Trades and Quotes streaming data;*

IV. *Granularity: It ranges from 1 minute do 5 years for the most of the available dat. The starting and the ending date for each crypto varies. The complete lists is define in the following link: https://www.coinapi.io/integration.*

V. *Financial Data: They provides financial information for every cryptocurrency and in particular it allows accessing to:*

- *OHLCV: It returns time series data in ascending order for the defined time span. it includes the following parameters:*
  - *time_period_start*
  - *time_period_end*
  - *time_open*
  - *time_close*
  - *price_open*
  - *price_close*
  - *volume_traded*
  - *trades_count*

- *Trades: it describes calls related to executed transactions data and include the following parameters:*
  - *Symbol_id*
  - *time_exchange*
  - *time_coinapi*
  - *uuid*
  - *price*
  - *size*
  - *taker_side*

- *Quotes: it describes calls related to quotes data, including the following variables:*
  - *symbol_id*
  - *time_exchange*
  - *CoinApi_time*
  - *ask_price*
  - *ask_size*
  - *bid_price*
  - *bid_size*
  - *last_trade*

4. CryptoCompare:

I. *source: https://www.cryptocompare.com*
II. *Data access: REST API with API key needed;*
III. *Accessibility: The free plan includes the following features:*
- *100,000 calls/month*
- *35+ Market Data endpoints*
- *Full access to Daily and Hourly historical data;*
- *7 days access to minute granularity data;*
- *Order Book snapshot from Binance*
- *FAQ support*
- *Personal use only as license;*

IV. *Granularity:Minute, Day, Hour;*
V. *Financial Data: they are available for each several cryptocurrencies with timespane from april,24, 2014 to current date and in USD or BTC conversion. In particular they include the following data:*
- *Historical Daily OHLCV*
- *Historical Hourly OHLCV*
- *Historical Minute OHLCV*
- *Historical Daily OHLCV All Markets*
- *Historical Daily OHLCV (Deprecated)*
- *Historical Hourly OHLCV (Deprecated)*
- *Historical Day OHLCV for a timestamp*
- *Historical Day Average Price*
- *Historical Daily Exchange Volume*
- *Historical Hourly Exchange Volume*
- *Top Exchanges Volume Data by Pair*
- *Top Exchanges Full data by Pair*
- *Toplist by Pair Volume*

VI. *Social Data: It allows accessing Social Statistics from sources, such as Twitter, Facebook and Reddit and accessing to news related to a specific coin/chain. In particular includes:*
- *Latest Coin Social Stats Data*
- *Historical Day Social Strats Data*
- *Historical Hour Social Stats Data*
- *Latest News Articles*
- *List News Feeds*
- *News Article Categories*
- *List News Feeds and Categories*

5. Blockchain.com:

I. *Source: https://www.blockchain.com/it/stats*
II. *Data access: API and direct csv download*
III. *Accessibility: The access is completely free*
IV. *Granularity: daily ;*
V. *Financial Data: they are available for Bitcoin and include:*
- *price with daily granularity and time horizon ranges from last 30 days to every historical data;*
- *Volume exchanged related to the current data;*

VI. *Blockchain data: It shows bitcoin blockchain data and includes the following parameters:*
- *Total Transaction Fees: it shows the total transaction fees paid to the miners;*
- *Confirmed Transaction per Day: the number of confirmed Bitcoin transactions,*
- *Output Value: the total value of all transactions per day, including coins returned to the sender as change;*
- *Estimated Transaction value: the total estimated value of transactions on th4 Bitcoin blockchain, not including the coins returned to the sender as change;*
- *Estimated USD Transaction Value: the estimated value in USD;*
- *Miners Revenue: total value of coinbase Block rewards and transaction fees paid to miners;*
- *Cost % of Transaction Volume:it shows miners revenue as percentage of the transaction volume;*
- *Cost per Transaction: it shows the revenue divided by the number of transactions;*
- *Hashing Difficulty: A relative measure of how difficult it is to find a new block. The difficulty is adjusted periodically as a function of how miSource: Data access: REST API and WebSocket API ;*
- *Accessibility: This platform makes available social data with premium account and different payment plans:*
  - *Standard: 29 US /month*
  - *Premium 49 $/ month*
  - *Advanced 99$/ month*

  The advanced plan is the only plan that authorize the access to API

- *Granularity: Minute, hour, day;*
- *Financial Data: They allow accessing to Real-time and historical endpoints for more then 200 cryptocurrencies. The former includes the following categories:*
- *ch hashing power has been deployed by the network of miners;*
- *Hash Rate: the estimated number of tera hashes per second the bitcoin is performing,*

6. CoinDance:

I. *Source: https://coin.dance*
II. *Data access: direct download in XLSX*

III. *Accessibility: It allows directly accessing to the following Blockchain statistics:*
- *Bitcoin (BTC)*
- *Bitcoin Cash (BCH)*
- *Bitcoin SV (BSV)*

IV. *Granularity: It ranges from 5 hours to 1 day, depending on the queried metics*

V. *Financial Data: they allows accessing to the Cryptocurrency Market Capitalization and to compare Bitcoin one with Altcoin one. The granularity is one day and timespan is in the range of 2016-02-19 to current date*

VI. *Blockchain Data: It allows the comparison of the following statistics for the defined Blockchains:*
- *Hash rate %;*
- *Network nodes;*
- *Transactions;*
- *Block Sizes;*
- *Daily Average Block Size;*
- *Daily Average Bitcoin Transactions per Block;*
- *Daily Average Bitcoin Fees by Network;*
- *Daily Accumulated Bitcoin Bitcoin Blockchain Growth by network;*

VII. *Social Data: They have daily granularity and timespan in the range of 2011_02-12 to three months before the current date and shows:*
- *Bitcoin Search Volume on Google Trends;*
- *Blockchain Search Volume on Google Trends;*
- *Bitcoin Community Demographics;*
- *Bitcoin Community Affinities;*

7. CryptoDataDownload.com:

I. *Source: https://www.cryptodatadownload.com/data/*

II. *Data access: direct download in CSV format;*

III. *Accessibility: CryptoDataDownload makes available free data;*

IV. *Granularity: Minute, hour, day;*

V. *Financial Data: they include OHLCV and volume data organized by exchange and with the possibility to access all existing data. It includes the exchange of different countries:*
- *US & UK (Gemini, GDAX, Coinbase, Kraken, Bitstamp, Bitfinex, Cexio, QuadrigaCx, Coinfloor, Luno, Tidex, itBit);*
- *EU & Russia (HitBTC, Exmo, BitBay, Bitmarket, Tobit, Liquui)*
- *Asia Pacific (Binance, Bithumb, OkCoin, Quiniex, Bitflyer, BTCMarkets, Cryptodia, Zaif)*
- *Other international (Bitso, BTCXIndia, Unocoin, Bit2C)*

8. DataLight:

I. *Source: https://datalight.me*

II. *Data access: REST API*

III. *Accessibility: API Market Data, API Blockchain and Social Media are available through Business Plan (195$ /mo);*

IV. *Granularity: It ranges from 5 minute to 1 day for financial data, whereas from 1 minute to 1 day for social and Blockchain Data;*

V. *Financial Data: they allows accessing to the most famous Cryptocurrencies amd to evaluate the following fields:*

- *Crypto Price in USD, ETH,XRP, EOS, BNB and since 2013-04-28*
- *all time low/high value of price in $;*
- *all time low/high value of market capitalization in $*
- *all time high value of volume in $*
- *Price BTC change in the last 1h/ 24 h/ 1 day/ 7day, since 2018-09-26 with minimum granularity of 5 minute*
- *Sharpe Ratio for 30 day since 2013-04-28 with minimum granularity of 1 hour;*
- *Cryptocurrency exchange volume in USD/ BTC in the 24 hour since 2013-04-28;*
- *Exchange liquidity of cryptocurrency each 5 minute in the 24 hour since 2013-04-28*
- *Mayer ratio with minimum granularity of 5 minute and since 2013-04-28;*

VI. *Blockchain Data: All Blockchain Data allows accessing to data with minimum granularity of 1 minute and accessing date since 2013-04-28 for the most fields, whersas all historical transactions are available for some metrics. They include:*

- *Available supply of Crypto Asset;*
- *Maximum Supply of Crypto Asset,*
- *24 Hour transaction sum, available for BTC, ETH; TRX, NEO;*
- *Transaction Count for the last 24 Hours, available for BTC, ETH, TRX, NEO;*
- *The number of address that sent a transaction in the last 24 hours and available for BTC, ETH, and tokens based on ETH, TRX, NEO ;*
- *The number of address that receive a transaction in the last 24 hours available for BTC, ETH, TRX, NEO ;*
- *Max value of transaction in USD in the last 24 hours , available for BTC, ETH, TRX, NEO ;*
- *Average value of transactions in USD in the last 24 hours for BTC, ETH, TRX, NEO ;*
- *Datalight_usage define the blockchain network usage as the Active Address times Transactions count;*
- *Nvt ratio defines the Cryptocurrency analogue of classic stock market P/E, defined as Market Capitalization/ Transaction Sum in 24 h in $;*
- *Daalight_Blockchain_Indicator is ratio /Market capitalization/ Usage Index) shows how overvalued/undervalued a cryptocurrency is relative to its blockchain activity,*
- *Metcalfe_usage in 24 h is the squared number of active users, following the Metcalfe's law;*

VII. *Social Data: The granularity of these data is 1 hour, and timespan since 2017-01-01 and include the following metrics:*

- *Telegram Hype Index that evaluates the number of unique users and messages on Telegram;*
- *Twitter Hype Index that evaluates the number of unique users and messages and reactions to them in Telegram;*
- *TG_mood_24 is the average value of talks tonality of one post for Telegram crypto groups and channels;*
- *TG_hype_change_24h is the Telegram Hype change in the last 24 hours ;*
- *TW_hype_change_24h is the Twitter Hype change in the last 24 hours ;*
- *Wikipedia page views count in the last 30 days with granularity of 1 day;*
- *Traffic visit is the Website visits per month ;*
- *MarketCap_per_TG-Hype_24h is the ratio that shows how big the value of the asset is relative to its Hype. The highest the indicator, more overestimated the asset is relative to its real popularity;*
- *arketcap_per_TW_hype_24h: is similar to before ratio, but the popularity is measured via Twitter Hype;*

9. BittsAnalytics:

I. *Source: http://bittsanalytics.com/*
II. *Data access: REST API and WebSocket API ;*
III. *Accessibility: This platform makes available social data with premium account and different payment plans:*
   - *Standard: 29 US /month*
   - *Premium 49 $/ month*
   - *Advanced 99$/ month*

   The advanced plan is the only plan that authorize the access to API

IV. *Granularity: Minute, hour, day;*
V. *Financial Data: They allow accessing to Real-time and historical endpoints for more then 200 cryptocurrencies. The former includes the following categories:*

- *ticker: symbol of cryptocurrency,*
- *price_usd: price of cryptocurrency in USD,*
- *market_cap: market capitalization of cryptocurrency in USD,*
- *trading_volume: trading volume of cryptocurrency in USD (last 24 hours),*
- *momentum score: proprietary indicator calculated as equal-weighted z-scores (from population of 200+ cryptocurrencies) for the following categories: 24 hour change in price, change in daily sentiment, change in daily number of mentions, change in daily trading volume. Momentum score indicates the composite momentum of a cryptocurrency in price, trading volume, social media sentiment and social media buzz.*
- *hourly_social_sentiment: we calculate social media sentiment of coins by using advanced machine learning models from texts of social media posts on twitter that mention individual*

coins. Hourly social sentiment is the average sentiment of a coin in the last closed hour in UTC time.

- *hourly_social_mentions: we calculate number of mentions of coins in texts of social media posts on twitter. Hourly social mentions is the number of mentions of a given coin in the last closed hour in UTC time.*
- *hourly_social_buzz: is the number of mentions of coins on twitter in last 3 hours normalized by historical, similar 3 hour windows. Hourly social buzz indicates how much is the coin mentioned on social media recently as compared to historical averages for this coin.*
- *hourly_social_sentiment_momentum: is the average sentiment of coins on twitter in last 3 hours normalized by historical, similar 3 hour windows. Hourly social sentiment momentum indicates how positive or negative is the sentiment of coin as compared to historical averages for this coin.*
- *daily_social_sentiment: is the average sentiment of coin on twitter in last closed day in UTC time.*
- *daily_social_mentions: is the number of mentions of a given coin in last closed day in UTC time.*

The historical data instead includes the following categories:

- *Daily Social Sentiment: this API Endpoint provides access to historical daily sentiment data for over 200 cryptocurrencies. Historical data starts on 14 August 2017. Daily social sentiment is the average sentiment of cryptocurrency on twitter in last closed day in UTC time. Sentiment is determined by applying advanced machine learning models on texts of social media posts on twitter that mention individual coins.*
- *Daily Social Mentions: this API Endpoint provides access to historical daily mentions on social media for over 200 cryptocurrencies. Historical data starts on 14 August 2017*
- *Hourly Social Sentiment: this API Endpoint provides access to historical hourly social media sentiment data for over 200 cryptocurrencies.*
- *Hourly Social Mentions: this API Endpoint provides access to historical hourly social media mentions data for over 200 cryptocurrencies*

10. CryprocurrencyChart.com:

I. *Source:* https://www.cryptocurrencychart.com
II. *Data access: REST API;*
III. *Accessibility: This platform makes available currency and historical data with different payment plans. The free plan includes*
   - *free API access;*
   - *2,000 requests per month;*
   - *3 download per month*

IV. *Granularity: Day;*

V. *Financial Data: They allow accessing to timestamped and historical endpoints for a full list of cryptocurrencies.*

- *View Coin History: it retrieves coin data of the specified type for the provided data range. Start and end date can be at most 2 years and include the following parameters:*
  - *coin id/ name/ symbol*
  - *start date*
  - *end date*
  - *dataType (market capitalization, volume, price ecc)*
  - *base Currency (USD/EUR/BTC)*
- *View Coin: it shows several data about specific coin for specific date and includes the following values:*
  - *coin id/ name/ symbol*
  - *start/ end date*
  - *price*
  - *open/close price*
  - *market capitalization*
  - *trade volume*
  - *fiat trade volume*
  - *rank*
  - *supply*
  - *trade Health*
  - *sentiment (bearish, bullish)*
  - *first data*
  - *most recent data*
  - *status*
- *get Coins: it provides a list of available crypto currencies and includes coin id / name/ symbol;*
- *get base Currencies: provides a list of base currencies used for the view COin and view Coin history calls;*

11. Craft Source Maker:

I. *Source: https://craft.co/*
II. *Data access: directly download in zip ;*
III. *Accessibility: It allows accessing to historical data with different metrics for supported blockchains.*
IV. *Granularity: Most of the data presents daily granularity ;*
V. *Financial Data: They allow accessing to historical data with access to all data for several Blockchains with different retrieved accessibility. For the most famous, full history is available and includes the following values:*

- *AdrActCnt: The sum count of unique addresses that were active in the network (either as a recipient or originator of a ledger change) that day. All parties in a ledger change action (recipients and originators) are counted. Individual addresses are not double-counted if previously active.*
- *BlkCnt: he sum count of blocks created that day that were included in the main (base) chain.*
- *BlkSizeMeanByte: The mean size (in bytes) of all blocks created that day.*
- *DiffMean: The mean difficulty of finding a hash that meets the protocol-designated requirement (i.e., the difficulty of finding a new block) that day. The requirement is unique to each applicable cryptocurrency protocol. Difficulty is adjusted periodically by the protocol as a function of how much hashing power is being deployed by miners.*
- *FeeMeanUSD: The USD value of the mean fee per transaction that day.*
- *FeeMedUSD: The USD value of the median fee per transactions*
- *FeeTotUSD: The sum USD value of all fees paid to miners that day. Fees do not include new issuance.*
- *NVTAdj: The ratio of the network value (or market capitalization, current supply) divided by the adjusted transfer value. Also referred to as NVT.*
- *NVTAdj90: The ratio of the network value (or market capitalization, current supply) to the 90-day moving average of the adjusted transfer value. Also referred to as NVT.*
- *PriceBTC: The fixed closing price of the asset as of 00:00 UTC the following day (i.e., midnight UTC of the current day) denominated in BTC.*
- *PrivceUSD: The fixed closing price of the asset as of 00:00 UTC the following day (i.e., midnight UTC of the current day) denominated in USD. This price is generated by Coin Metrics' fixing/reference rate service.*
- *SplyCur: The sum of all native units ever created and visible on the ledger (i.e., issued) as of that day. For account-based protocols, only accounts with positive balances are counted*
- *TxCnt: The sum count of transactions that day. Transactions represent a bundle of intended actions to alter the ledger initiated by a user (human or machine). Transactions are counted whether they execute or not and whether they result in the transfer of native units or not (a transaction can result in no, one, or many transfers). Changes to the ledger mandated by the protocol (and not by a user) or post-launch new issuance issued by a founder or controlling entity are not included here.*
- *VtyDayRet180d: The 180D volatility, measured as the deviation of log returns;*
- *VtyDayRet30d: The 30D volatility, measured as the deviation of log returns*
- *VtyDayRet60d: The 60D volatility, measured as the deviation of log returns*

12. ***CryptoFinance:***

I. *Source:* https://chrome.google.com/webstore/detail/cryptofinance/bhjnahcnhemcnnenhgbmmdapapblnlcn

II. *Data access: it is a Google Sheets Add-on;*

III. *Accessibility: it is essentially a formula that operates as an add-on that pulls the latest information from different sources , tracking 1900+ currencies with 30+ fiat currencies and include access to several exchanges (AltCoinTrader,Bilaxy, Binance, Binance Jersey, Bitbank, BitcoinTrade, Bitfinex, Bitflyer, Bitforex, BitHumb, Bitmax, Bitmex, Bitso, Bitstamp, Bittrex, Bitzeus, BTCMarkets, Coinbase, Coinbase Pro (GDAX),Coinbene, Coineal,Cryptopia, Digifinex, Gate.io, Gemini, HitBTC, Huobi, Huobi Pro, Ice3x, IDAX, IDEX, IndependentReserve, Kraken, Kucoin, LiveCoin, LocalBitcoins, Luno, Lykke, Mercado, Okex, Poloniex, QTrade, Trade Ogre, Uniswap, Upbit, VALR, Yobit) Some metrics are free but with a limitations of 25 calls/day and historical data not available. To access to all data a payment plan is required.*

IV. *Granularity: Most of the data presents daily granularity ;*

V. *Financial Data: The main financial and historical data are available through CryptoCompare APIs and are:*

- *Market Capitalization;*
- *24h volume ;*
- *Total Supply;*
- *Circulating Supply;*
- *Maximum Supply;*
- *Change Percentage;*
- *Currency Rank;*
- *Historical Exchange volume;*
- *Historical OHLCV data with full accessibility and daily granularity;*
- *All Time High Prices ;*
- *All Time HIgh Date;*
- *All Time High Volume;*
- *Bitcoin Dominance;*
- *Total Market Cap;*
- *Total 24h Volume;*
- *ROI returned per year, quarter, month;*
- *Market Order Book Liquidity;*
- *Exchanges Order Book Data*

VI. *Blockchain Data: they present data about Network Data and Balance & Transactions and are provided by CoinMetrics APIs. 74 assets are supported (ADA, AE, AION, ANT, BAT, BCH, BNB, BSV, BTC, BTG, BTM, CENNZ, CTXC, CVC, DAI, DASH, DCR, DGB, DOGE, DRGN, ELF, ENG, EOS, ETC, ETH, ETHOS, FUN, GAS, GNO, GNT, GUSD, ICN, ICX, KCS, KNC, LOOM, LRC, LSK, LTC, MAID, MANA, MTL, NAS, NEO, OMG, PAX, PAY, PIVX, POLY POWR, PPT, QASH, REP, RHOC, SALT, SNT, SRN, TRX, TUSD, USDC, USDT, VEN, VERI, VTC, WAVES, WTC, XEM, XLM, XMR, XRP, XVG, ZEC, ZIL, ZRX) and include the following parameters:*

- *tx_volume_usd;*

- *adjusted_tx_volume_usd;*
- *tx_count;*
- *marketcap_usd;*
- *price_usd ;*
- *exchange_volume_usd ;*
- *realized_cap_usd, generated_coins ;*
- *fees, active_addresses ;*
- *average_difficulty ;*
- *payment_count ;*
- *median_tx_value_usd ;*
- *median_fee, block_size, block_count ;*

VII. *Social Data: Provided by Solume.io it's possible to get over then 200 currencies mention count, change and sentiment analysis from the Web at large, Reddit and Twitter. In particular it provides 24h metrics such as:*

- *Cryptocurrency Sentiment;*
- *Cryptocurrency Mention on the Web;*
- *Cryptocurrency Mention on Twitter ;*
- *Cryptocurrency Mention on Reddit ;*

13. CryptoControl.io :

I. *Source: https://cryptocontrol.io/en/*

II. *Data access: it is a cryptocurrency news aggregator that allows accessing data through APIs;*

III. *Accessibility: CryptoControl offers access to news articles from over 1000+ sources, reddit feeds and a twitter feed, for all the different crypto coins.for free. it offers also Sentiment APIs that give the polarity of news articles for each cryptocurrencies for 9$/month;*

IV. *Social Data: it allows accessing several information such as:*

- *Top/ Latest News articles for category or specific coin ;*
- *Top/ Latest Reddit Post for category or specific coin;*
- *Top/ Latest Twitter Post for category or specific coin;*
- *Top Feed, combining Reddit, Twitter and Articles for a coin;*
- *Reddit/ Tweets/ Article for a coin sorted by time;*
- *Coin details for a specific coin;*
- *Historical Sentiment for a specific coin from specific source;*

14. Google Trends:

I. *Source: https://trends.google.it/trends/?geo=IT*
II. *Data access: it is a Google site that allows analyzing the popularity of search queries in Google Search across different languages and countries;*
III. *Accessibility: the download of the data is free, and it is available in csv format;*
IV. *Granularity: it ranges from 1 minute to 5 days and limited timeframe ;*
V. *Social Data: Counting of cryptocurrency researches is a proxy of social sentiment and trend. In particular it is possible to search this popularity setting:*
   - *country*
   - *country's region*
   - *timespan*
   - *category of interest*
   - *kind of search (Google News, Google Shopping, Google images, YouTube)*

15. Wikipedia Pageview Statistics:

I. *Source: https://tools.wmflabs.org/pageviews/?project=en.wikipedia.org&platform=all-access&agent=user&range=latest-20&pages=*
II. *Data access: It is a tool which allows to see the popularity of an article during a given period. It is possible to download the data in*
   - *CSV;*
   - *JSON*
   - *PNG*
III. *Accessibility: the download of the data is free and include data back to July 1 , 2015;*
IV. *Granularity: it could be settled on daily or monthly ;*
V. *Social data: the frequency of views for specific page could result a good proxy of social sentiment.*

16. Rddit Statistics:

I. *Source:* https://subredditstats.com/
II. *Data access: It allows accessing Reddit Statistics for free and exporting it in JSON format;*
III. *Accessibility: It is possible to access data back to 2013 for Subscribers Counting and November 2018 for Comments and Posts per day counting.*
IV. *Granularity: Daily;*
V. *Social Data: it offers three diffents kind of information:*

- *Subscribers: it counts the number of subscribers per day for a specific subreddit;*
- *Comments per Day: it counts the number of comments per day for a specific subreddit;*
- *Posts per Day: it counts the number of posts per day for a specific subreddit ;*

APPENDIX II: PYTHON CODING

This appendix displays the most important code lines implemented for experimental part of this thesis work. Due to its simplicity and high level property, it has been possible to compile it , with owning deep programming skills.

The main open source libraries imported in the environment are: Pandas for dataframe manipulation, NumPy for matrix, vector and numerical analysis implementation, MatPlotlib for mathematical tool as Correlation matrix and plotting functions, Seaborn, which is built on MatPlotlib and allows facilitating statistical plotting charts and Sklearn, a powerful framework for automatic learning, that includes classification and regression machine learning algorithms.

Reflecting the order analysis, three main code pieces are defined below: Data Crawling, Preprocessing, Data Analysis and Visualization and Machine Learning applications.

The code is proposed for one coin, but it is scalable and has been replicable to all coins' dataframe.

1. DATA CRAWLING

```
import requests
import pandas as pd

apiKey = "d30ec53fdcaf63a963b3ea7287f087f8d015ff19a15aeecf56b590a13c4edf1f"

#import OHLCV data
url = "https://min-api.cryptocompare.com/data/histohour"
payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "limit": 2000
}
result_1 = requests.get(url, params=payload).json()
df1 = pd.DataFrame(result_1['Data'])
print(df1.head())

payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1563440400,
    "limit": 2000
}


result_2 = requests.get(url, params=payload).json()
df2= pd.DataFrame(result_2['Data'])
```

```
print(df2.head())


payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1556240400,
    "limit": 2000
}


result_3 = requests.get(url, params=payload).json()

df3 = pd.DataFrame(result_3['Data'])

print(df3.head())


payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1549036800,
    "limit": 2000
}


result_4 = requests.get(url, params=payload).json()

df4 = pd.DataFrame(result_4['Data'])




payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1541836800,
    "limit": 2000
}
result_5 = requests.get(url, params=payload).json()

df5 = pd.DataFrame(result_5['Data'])


payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1534636800,
    "limit": 2000
}
```

```
result_6 = requests.get(url, params=payload).json()

df6 = pd.DataFrame(result_6['Data'])


payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1527436800,
    "limit": 2000
}


result_7 = requests.get(url, params=payload).json()

df7 = pd.DataFrame(result_7['Data'])

payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1520236800,
    "limit": 2000
}


result_8 = requests.get(url, params=payload).json()

df8 = pd.DataFrame(result_8['Data'])


payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1513036800,
    "limit": 2000
}


result_9 = requests.get(url, params=payload).json()

df9 = pd.DataFrame(result_9['Data'])


payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1505836800,
    "limit": 2000
}
```

```
result_10 = requests.get(url, params=payload).json()

df10 = pd.DataFrame(result_10['Data'])


payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1498636800,
    "limit": 2000
}


result_11 = requests.get(url, params=payload).json()

df11 = pd.DataFrame(result_11['Data'])



payload = {
    "api_key": apiKey,
    "fsym": "DASH",
    "tsym": "USD",
    "toTs" : 1491436800,
    "limit": 2000
}


result_12 = requests.get(url, params=payload).json()

df12 = pd.DataFrame(result_12['Data'])

#OHLCV database given by merging all financial dataframes

final=pd.concat([df12,df11,df10,df9,df8,df7,df6,df5,df4,df3,df2,df1])

xport_csv = final.to_csv (r'DASH_OHLCV.csv', index = None, header=True)
print(final.info)


#import social data
url = "https://min-api.cryptocompare.com/data/social/coin/histo/hour"


payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "limit": 2000
}


social_1 = requests.get(url, params=payload).json()

db1 = pd.DataFrame(social_1['Data'])
```

```
payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1563440400,
    "limit": 2000
}


social_2 = requests.get(url, params=payload).json()

db2 = pd.DataFrame(social_2['Data'])


payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1556240400,
    "limit": 2000
}


social_3 = requests.get(url, params=payload).json()

db3 = pd.DataFrame(social_3['Data'])


payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1549036800,
    "limit": 2000
}


social_4 = requests.get(url, params=payload).json()

db4 = pd.DataFrame(social_4['Data'])


payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1541836800,
    "limit": 2000
}

social_5 = requests.get(url, params=payload).json()

db5 = pd.DataFrame(social_5['Data'])


payload = {
```

```
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1534636800,
    "limit": 2000
}

social_6 = requests.get(url, params=payload).json()

db6 = pd.DataFrame(social_6['Data'])


payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1527436800,
    "limit": 2000
}

social_7 = requests.get(url, params=payload).json()

db7 = pd.DataFrame(social_7['Data'])


payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1520236800,
    "limit": 2000
}

social_8 = requests.get(url, params=payload).json()

db8 = pd.DataFrame(social_8['Data'])



payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1513036800,
    "limit": 2000
}

social_9 = requests.get(url, params=payload).json()

db9 = pd.DataFrame(social_9['Data'])



payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1505836800,
    "limit": 2000
}
```

```
social_10 = requests.get(url, params=payload).json()

db10 = pd.DataFrame(social_10['Data'])


payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1498636800,
    "limit": 2000
}

social_11 = requests.get(url, params=payload).json()

db11 = pd.DataFrame(social_11['Data'])



payload = {
    "api_key": apiKey,
    "coinId": 3807,
    "toTs" : 1491436800,
    "limit": 2000
}

social_12 = requests.get(url, params=payload).json()

db12 = pd.DataFrame(social_12['Data'])

#Scial database given by merging all social dataframes

social_dataframe=pd.concat([db12,db11,db10,db9,db8,db7,db6,db5,db4,db3,db2,db1],sort=True)

xport_csv = social_dataframe.to_csv (r'DASH_SOCIAL.csv', index = None, header=True)

print(social_dataframe.info)

financial_and_social = pd.concat([final,social_dataframe], axis=1, sort=False)

xport_csv = financial_and_social.to_csv (r'Financial&Social.csv', index = None, header=True)
```

2. DATA PREPROCESSING

```
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler



#read dataframe with hourly granularity
db_1=pd.read_csv('/Users/antonio/Desktop/tesi/crypto/DASH/DASH_hour_1.csv')
if 'date' in db_1.columns :
  del db_1['date']

#shift price series of 1 entry
db_1['close_h_1']=db_1['close'].shift(-1)

print(db_1.close_h_1)
Social_missing = db_1.isna()
Social_num_missing = Social_missing.sum()
Percentage_missing_social= Social_num_missing / len(db_1)

median_1=db_1.median()
Q1_1 = db_1.quantile(0.25)
Q3_1 = db_1.quantile(0.75)
IQR_1 = Q3_1 - Q1_1
print(IQR_1)
outliers=(db_1 < (Q1_1 - 1.5 * IQR_1)) |(db_1> (Q3_1 + 1.5 * IQR_1))
Percentage_otliers_social=outliers.sum()/len(db_1)


#read dataframe with daily granularity
db_2=pd.read_csv('/Users/antonio/Desktop/tesi/crypto/DASH/dash_day.csv')
if 'date' in db_2.columns :
  del db_2['date']

db_2.drop(db_2.columns[[-2]], axis=1, inplace=True)
db_2.drop(db_2.tail(1).index,inplace=True)
db_2['close'] = db_2['close'].str.replace(',', '').astype(float)
db_2['open'] = db_2['open'].str.replace(',', '').astype(float)
db_2['high'] = db_2['high'].str.replace(',', '').astype(float)
db_2['low'] = db_2['low'].str.replace(',', '').astype(float)

db_2['close_d_1']=db_2['close'].shift(-1)

print (db_2.close_d_1)



Block_missing = db_2.isna()
print(Block_missing)
Block_num_missing = Block_missing.sum()
Percentage_missing_block= Block_num_missing / len(db_2)

median_2=db_2.median()
```

```
Q1_2 = db_2.quantile(0.25)
Q3_2 = db_2.quantile(0.75)
IQR_2 = Q3_2 - Q1_2
print(IQR_2)
outliers_2=(db_2 < (Q1_2 - 1.5 * IQR_2)) |(db_2> (Q3_2 + 1.5 * IQR_2))
Percentage_otliers_block=outliers_2.sum()/len(db_2)


print(Percentage_missing_social)
print(Percentage_missing_block)
```

3. HAETMAP AND DATA VISUALIZATION

```
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.neural_network import MLPClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score



#read dataframe with hourly granularity
db_1=pd.read_csv( '/Users/antonio/Desktop/tesi/crypto/ETH/ETH_hour.csv')
if 'date' in db_1.columns :
  del db_1['date']

#shift price series of 1 entry
db_1['close_h_1']=db_1['close'].shift(-1)

print(db_1.close_h_1)
Social_missing = db_1.isna()
Social_num_missing = Social_missing.sum()
Percentage_missing_social= Social_num_missing / len(db_1)

median_1=db_1.median()
Q1_1 = db_1.quantile(0.25)
Q3_1 = db_1.quantile(0.75)
IQR_1 = Q3_1 - Q1_1
print(IQR_1)
outliers=(db_1 < (Q1_1 - 1.5 * IQR_1)) |(db_1> (Q3_1 + 1.5 * IQR_1))
db_1[outliers] = np.nan
```

```
db_1.fillna(median_1, inplace=True)



# heat correlation matrix for social dataframe
def heatMap(df):
    #Create Correlation df
    corr = df.corr()
    #Plot figsize
    fig, ax = plt.subplots(figsize=(16, 10))
    #Generate Color Map
    colormap = sns.diverging_palette(220, 10, as_cmap=True)
    #Generate Heat Map, allow annotations and place floats in map
    sns.heatmap(corr, cmap=colormap, annot=True, fmt=".1f")
    #Apply xticks
    plt.xticks(range(len(corr.columns)), corr.columns);
    #Apply yticks
    plt.yticks(range(len(corr.columns)), corr.columns)
    #show plot
    plt.show()
    plt.savefig('Socialmap.png')


image=heatMap(db_1)

# standardization of variable
s = StandardScaler()
close_1= db_1[['close_h_1','fb_likes']].loc[4000:6000]
scaled=s.fit_transform(close_1)
pd.DataFrame(scaled, columns=['close_h_1_ETH', 'fb_likes_ETH']).plot()
plt.savefig('Socilaplot.png')



#read dataframe with daily granularity
db_2=pd.read_csv('/Users/antonio/Desktop/tesi/crypto/LTC/ltc_day.csv')
if 'date' in db_2.columns :
  del db_2['date']

db_2.drop(db_2.columns[[-2]], axis=1, inplace=True)
db_2.drop(db_2.tail(1).index,inplace=True)
db_2['close'] = db_2['close'].str.replace(',', '').astype(float)
db_2['open'] = db_2['open'].str.replace(',', '').astype(float)
db_2['high'] = db_2['high'].str.replace(',', '').astype(float)
db_2['low'] = db_2['low'].str.replace(',', '').astype(float)

db_2['close_d_1']=db_2['close'].shift(-1)



median_2=db_2.median()
Q1_2 = db_2.quantile(0.25)
Q3_2 = db_2.quantile(0.75)
IQR_2 = Q3_2 - Q1_2
print(IQR_2)
```

```
outliers_2=(db_2 < (Q1_2 - 1.5 * IQR_2)) |(db_2> (Q3_2 + 1.5 * IQR_2))
db_2[outliers_2] = np.nan
db_2.fillna(median_2, inplace=True)

#heta correlation matrix for blockchain dataframe
def heatMap(df):
    #Create Correlation df
    corr = df.corr()
    #Plot figsize
    fig, ax = plt.subplots(figsize=(16 ,10))
    #Generate Color Map
    colormap = sns.diverging_palette(220, 10, as_cmap=True)
    #Generate Heat Map, allow annotations and place floats in map
    sns.heatmap(corr, cmap=colormap, annot=True, fmt=".1f")
    #Apply xticks
    plt.xticks(range(len(corr.columns)), corr.columns);
    #Apply yticks
    plt.yticks(range(len(corr.columns)), corr.columns)
    #show plot
    plt.show()
    plt.savefig('Blockcha')


image=heatMap(db_2)


# standardization of variable
s = StandardScaler()
close_2= db_2[['close_d_1','TxTfrValMedUSD']].loc[800:1500]
scaled=s.fit_transform(close_2)

pd.DataFrame(scaled, columns=['close_h_1_ETH', 'TxTfrValMedUSD_ETH']).plot()
plt.savefig('Blockchain.png')
```

4. MACHINE LEARNING ALGORITHMS APPLICATION (MLP AND SVC)

```
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib as mpl
```

```
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.neural_network import MLPClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score


#read dataframe with daily granularity
db_2=pd.read_csv( '/Users/antonio/Desktop/tesi/crypto/LTC/ltc_day.csv')
if 'date' in db_2.columns :
  del db_2['date']

#convert string value in float
db_2.drop(db_2.columns[[-1]], axis=1, inplace=True)


#generate the close(d+1)
db_2['close_d_1']=db_2['close'].shift(-1)


#outliers replacing process
median_2=db_2.median()
Q1_2 = db_2.quantile(0.25)
Q3_2 = db_2.quantile(0.75)
IQR_2 = Q3_2 - Q1_2
print(IQR_2)
outliers_2=(db_2 < (Q1_2 - 1.5 * IQR_2)) |(db_2> (Q3_2 + 1.5 * IQR_2))
db_2[outliers_2] = np.nan
db_2.fillna(median_2, inplace=True)



#add closing price variation as column
db_2['close_change']=(db_2.close_d_1-db_2.close)/db_2.close


#defimimg function for target evaluation
def target (r):
    if r>= 0.005:
        class_value= 'Upper'
    else: class_value = 'No signal'
    return(class_value)


db_2['classe']=db_2.close_change.apply(
        lambda x: target(x) if x>-0.005
        else 'Lower')


train_perc = 0.66
y = db_2['classe'].values
print(db_2.head())
```

```
db_2 = db_2.drop(columns=['classe','time'])


X = db_2.values
print(type(X))

clf_mlp = MLPClassifier()
clf_svc= SVC()

rows, cols = X.shape # (365, 20)
train_size = int(rows*train_perc)

X_train = X[:train_size, :]
y_train = y[:train_size]
X_test = X[train_size:, :]
y_test = y[train_size:]
print (y_test)
print(X_test)


clf_mlp.fit(X_train, y_train)
y_pred_mlp= clf_mlp.predict(X_test)

accuracy_mlp=accuracy_score(y_test, y_pred_mlp)

clf_svc.fit(X_train, y_train)
y_pred_svc = clf_svc.predict(X_test)

accuracy_svc=accuracy_score(y_test, y_pred_svc)

print('MLP Daily Accuracy: '+str(accuracy_mlp))
print('SVC Daily Accuracy '+ str(accuracy_svc))


#read hourly dataframe
db_1=pd.read_csv()
if 'date' in db_1.columns :
  del db_1['date']



db_1['close_h_1']=db_1['close'].shift(-1)

#outliers replacing process for hourly dataframe
median_1=db_1.median()
Q1_1 = db_1.quantile(0.25)
Q3_1 = db_1.quantile(0.75)
IQR_1 = Q3_2 - Q1_1
print(IQR_1)
outliers_1=(db_1 < (Q1_1 - 1.5 * IQR_1)) |(db_2> (Q3_1 + 1.5 * IQR_1))
db_1[outliers_1] = np.nan
db_1.fillna(median_1, inplace=True)
```

```
#add closing price variation as column
db_1['close_change']=(db_1.close_h_1-db_1.close)/db_1.close



db_1['classe']=db_1.close_change.apply(
        lambda x: target(x) if x>-0.005
        else 'Lower')


y_1= db_1['classe'].values


db_1 = db_1.drop(columns=['classe','time'])


X_1= db_1.values
print(type(X_1))
print(db_1.head())



clf_mlp_1 = MLPClassifier()


rows, cols = X_1.shape # (365, 20)
train_size_1 = int(rows*train_perc)

X_1_train = X_1[:train_size_1, :]
y_1_train = y_1[:train_size_1]
X_1_test = X_1[train_size_1:, :]
y_1_test = y_1[train_size_1:]
print (y_1_test)
print(X_1_test)


clf_mlp_1.fit(X_1_train, y_1_train)
y_pred_mlp_1= clf_mlp_1.predict(X_1_test)

accuracy_mlp_1=accuracy_score(y_1_test, y_pred_mlp_1)


clf_svc_1= SVC()

clf_svc_1.fit(X_1_train, y_1_train)
y_pred_svc_1 = clf_svc_1.predict(X_1_test)
```

```
accuracy_svc_1=accuracy_score(y_1_test, y_pred_svc_1)

print('MLP Hourly Accuracy: '+str(accuracy_mlp_1))
print('SVC Hourly Accuracy '+ str(accuracy_svc_1))
```

Bibliografia

[1] Quantitative cryptocurrency trading: exploring the use of machine learning techniques .*[https://smartdata.polito.it/publications/quantitative-cryptocurrency-trading-exploring-the-use-of-machine-learning-techniques-3/]*

[2] Using machine learning for cryptocurrency trading:
[*https://ieeexplore-ieee-org.ezproxy.biblio.polito.it/stamp/stamp.jsp?tp=&arnumber=8780358*]

[3] Forecasting Price of Cryptocurrencies using Tweets Sentiment Analysis:
[*https://ieeexplore-ieee-org.ezproxy.biblio.polito.it/stamp/stamp.jsp?tp=&arnumber=8530659*]

[4] Bitcoin Spread prediction using Social and Web Search Media:
[*https://www.researchgate.net/publication/279917417_Bitcoin_Spread_Prediction_Using_Social_And_Web_Search_Media*]

[5] Overview of the Blockchain Technology Cases: [*https://www.rese*
https://www.researchgate.net/publication/329396515_Overview_of_the_Blockchain_Technology_Case
sarchgate.net/publication/329396515_Overview_of_the_Blockchain_Technology_Cases]

[6] A Predictive Model for the Global Cryptocurrency Market: [*https://ieeexplore*
https://ieeexplore.ieee.org/document/8699292.ieee.org/document/8699292]

[7] A study of Opinion Mining and Data Mining Techniques to Analyze the Cryptocurrency Marke
[*https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8768762xplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8768762*]

[8] What drives Cryptocurrency prices? An investigation of Google Trends and Telegram Sentiment:
[*https://dl.acm.org/citation.cfm?id=3308955*]

[9] Cryptocurrency Price Prediction Using News and Social Media Sentiment:
[*https://pdfs.semanticscholar.org/c3b8/0de058596cee95beb20a2d087dbcf8be01ea.pdf*]

[10] Towards an Understanding of Cryptocurrency: A comparative Analysis of Cryptocurrency, Foreign Exchange and Stock:
[*https://www.researchgate.net/publication/332786433_Towards_an_Understanding_of_Cryptocurrency_A_Comparative_Analysis_of_Cryptocurrency_Foreign_Exchange_and_Stock*]

[11] How Does Social Media Impact Bitcoin Value? A Test of the Silent Majority Hypothesis:
[*https://www.researchgate.net/publication/324118265_How_Does_Social_Media_Impact_Bitcoin_Value_A_Test_of_the_Silent_Majority_Hypothesis*]

[12] Evaluating Sentiments in News Article. Source:
[*https://www.sciencedirect.com/science/article/pii/S0167923612000875*]

[13] Dynamic Random Walk, Elsevier Science, 2006

[14] John J Murphy Technical Analysis Of The Financial Market. [*https://www.academia.edu/4075580/John_J_Murphy_Technical_Analysis_Of_The_Financial_Markets*]

[15] Blockchain is here. [*https://www.pwc.com/gx/en/issues/blockchain/blockchain-in-business.html]*

[16] Blockchaoin: Il protocollo digitale del business peer to peer *[https://www.spindox.it/it/blog/blockchain-protocollo-digitale-business-peer-to-peer/]*

[17] What is a Bokckchain? A step by step guide for Beginners [https://blockgeeks.com/guides/what-is-blockchain-technology]

[18] Difference Between Client-Server and Peer-to-Peer network [https://techdifferences.com/difference-between-client-server-and-peer-to-peer-network.html]

[19] What is Blockchain? [https://lisk.io/academy/blockchain-basics/benefits-of-blockchain/blockchain-transparency-explained]

[20] Public and Private Key [*https://www.mycryptopedia.com/public-key-private-key-explained/*]

[21] Management of Innovation and Product Development, Cantamessa, Montagna. Springer, 2015

[22] Cryptocompare.com [*https://www.cryptocompare.com]*

[23] Craft Maker.com *[https://craft.co*]

[24] Data Science, Guida e principi alla scienza dei dati, Sinan Ozdemir, APOGEO

[25] Interquantile Range Method in Python [*https://medium.com/datadriveninvestor/finding-outliers-in-dataset-using-python-efc3fce6ce32*]

[26] Heatmap Correlation Matrix in Python [*https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec*]

[27] Machine Learning in Python [*https://scikit-learn.org/stable/]*

[28] Support Vector Machine [*https://it.wikipedia.org/wiki/Macchine_a_vettori_di_supporto*]

[29] A beginner Guide to Neural Network Applications [*https://skymind.ai/wiki/neural-network]*

[30] Data Science Slides, Elena Baralis, Politecnico Torino

[31] CoinMetrics Community Metrics [*https://coinmetrics.io/community-data-dictionary/*]

[32] CryptoFinance Data [*https://cryptofinance.ai/docs/bitcoin-crypto-blockchain-data/*]

[33] Medium Crypto Trading [*https://medium.com/treybrunson/crypto-trading-toolkit-534579f01863*]

[34] Sanpy Python Library [*https://pypi.org/project/sanpy/*]

[35] Blockchain.Chart [*https://www.blockchain.com/charts*]

[36] CoinDance APIs [*https://coin.dance/]*

[37]Public and Private Key *[https://www.mycryptopedia.com/public-key-private-key-explained/]*

[38] CryptoCompare.com [*https://www.cryptocompare.com/]*

[40] ByttsAnalytics [*http://www.bittsanalytics.com/*]