Empirical Analysis ot the Top 800 Cryptocurrencies using Machine Learning Techniques

Teresa Riedl

A dissertation submitted to the Faculty of Commerce, University of Cape Town, in partial fulfilment of the requirements for the degree of Master of Philosophy.

January 9, 2019

MPhil in Mathematical Finance,
University of Cape Town.

Declaration

I declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Master of Philosophy to the University of Cape Town. It has not before been submitted for any degree or examination.

Signed by candidate

Teresa Riedl

January 9, 2019

Abstract

The International Token Classification (ITC) Framework by the Blockchain Center in Frankfurt classifies 795 cryptocurrency tokens based on their economic, technological, legal and industry categorization. This work analyzes cryptocurrency data to evaluate the categorization with real-world market data. The feature space includes price, volume and market capitalization data. Additional metrics such as the moving average and the relative strengh index are added to get a more in-depth understanding of market movements. The data set is used to build supervised and unsupervised machine learning models. The prediction accuracies varied amongst labels and all remained below 90%. The technological label had the highest prediction accuracy at 88.9% using Random Forests. The economic label could be predicted with an accuracy of 81.7% using K-Nearest Neighbors. The classification using machine learning techniques is not yet accurate enough to automate the classification process. But it can be improved by adding additional features. The unsupervised clustering shows that there are more layers to the data that can be added to the ITC. The additional categories are built upon a combination of token mining, maximal supply, volume and market capitalization data. As a result we suggest that a data-driven extension of the categorization in to a token profile would allow investors and regulators to gain a deeper understanding of token performance, maturity and usage.

Acknowledgements

I would like to express my very great appreciation to Dr. Co-Pierre Georg for his guidance throughout the past year together with the AIFMRM team and especially Lameez and Lizzy. Thank you David for giving us exposure to your network and the supporters of AIFMRM. I would like to offer my special thanks to Luca Fragnani and Philipp Sandner at the Blockchain Center in Frankfurt for providing the Token Classification and support in drawing the initial hypotheses. Although the data could not be used for the analysis, I would like to thank the Santiment team for their openness and willingness to collaborate. I could not have succeeded throughout this year without Dustin being by my side. Thank you for the positivity through all the ups and downs, home-cooked meals, parkruns and morning swims. Pursuing this degree would not have been possible without the support of my family in Germany. Mum, dad thank you for always being there for me and letting me pursue my dreams and adventures. Nick Hops your help was greatly appreciated. And last but no least a great and most special thanks to my fellow students and especially Ashleigh, Jonjon, Bryony and Masego. You will all strive and change the world for the better. I cannot believe it is over!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Ten years after the invention of Bitcoin it is indisputable: Blockchain Technology will revolutionize the finance sector. In order to be part of its success, Germany has to act now."
Handelsblatt, November 1, 2018

Regulators have fallen behind in creating a fruitful environment for cryptocurrency companies. Certainly, for the right reasons: to protect investors, to let the technology evolve and to understand the impact blockchain technology can have on the state, the financial system and the population. The cryptocurrency community is becoming impatient and regulators are seeking to provide the right balance between freedom and regulation.

Throughout the cryptocurrency hype in 2017 the new field became a hot topic in research and analysis. Scientists were battling with each other to build the best price prediction models and to estimate the future of the market. Although few people knew what one was talking about in early 2017 when mentioning Bitcoin - at the end of the year the topic was on everyone's lips. Yet the conversation was less about the technology, its vision and drawbacks than about the price, the possible value a few months from now and the potential returns people could make.

After the decline of the market in early 2018, the conversations seem to have vanished, except for a few comments about whether the value will drop to zero or not. There are ambitious entrepreneurs out there who believe in the technology and are seeking to build sustainable businesses. For them the calmness might in some cases be a relief and an opportunity to focus on the core: getting the basics right and understanding cryptocurrencies, their risks and values.

This work conducts an exploratory analysis of the top 795 cryptocurrencies[1] using data analytics tools and machine learning techniques. One valuable step in the

[1] Based on their market capitalization.

process of regulation and risk assessment is the classification of cryptocurrency tokens. In this research the empirical analysis of cryptocurrency market data is put into relation with the token classes: economic purpose, technological setup, legal claim and industry.

The following section will give a brief overview of blockchain technology and the cryptocurrency market with focus on the past two years. It then goes on to outline the Token Classification Framework to understand the basis of this work.

Thereafter, chapter 2 describes related work that has been done in the new research field of Blockchain Technology and especially on Cryptocurrency Tokens. Chapter 3 outlines an exploratory analysis of the cryptocurrency market with regards to correlations amongst tokens and insights that one can gain per token. Chapter 4 will explore the token classification done by the blockchain center with regards to numberic characteristics of the tokens. In Chapter 5 the findings are discussed and an outlook is provided.

1.1 Blockchain Technology and Cryptocurrency Tokens

The focus of this section is to provide a non-technical introduction into Blockchain Technology and Cryptocurrency Tokens.

In 2008 the Bitcoin Whitepaper, published under the acronym ”Satoshi Nakamoto”. Nakamoto (2008) proposed a system for electronic (cash) transactions without relying on trust. As opposed to existing systems such as SEPA or Paypal transactions, Satoshi suggested a peer-to-peer system that allows ”online payments to be sent directly from one party to another without going through a financial institution” (Nakamoto, 2008).

Satoshi’s proposal can be broken down into three elements: the system, electronic cash and trustless peer-to-peer transactions. The ”system” is the underlying blockchain technology platform. It is used as a database to store a ledger of all transactions. Unlike databases used in banks and other institutions, the database is public and decentralized. Decentralized means that individuals around the world store a copy of the database (Swan, 2015, chap. 1). Examples for blockchain technology platforms are Bitcoin and Ethereum. ”Electronic cash” is the mean of exchange on the blockchain technology platform and widely known as cryptocurrency, coin or token. Cryptocurrencies are different from fiat currencies as they are not issued

by central banks but by the companies who developed the underlying protocols. They are seen as privately issued currencies. The protocol is the third element. By implementing logic and incentive structures via code, protocols create trust environments for peers (Swan, 2015, chap. 1). The peer-to-peer transaction of value is one use case that a protocol can fulfill. Some other applications for blockchain protocols are transactions of financial assets such as stocks or private equity, land or property registries and identification of individuals by issuing passports or licenses on the blockchain technology platform.

Since the Bitcoin Whitepaper was published it took nearly five years for Bitcoin to be launched as the first cryptocurrency. Today (Nov 2018) there are more than 2000 cryptocurrencies listed on CoinMarketCap[2]. In 2017 the number of cryptocurrencies increased exponentially and so did the cryptocurrency market capitalization overall as figure 1.1 indicates[3]. In December 2017 the market reached its peak and thereafter fell back to less than half of the maximum market capitalization[4].

The development of the cryptocurrency market suggests a comparison with the Gartner Hype Cycle although it is not safe to say whether the negative hype has already reached its minimum. It is likely that the slope of enlightment (Linden and Fenn, 2003) will arrive eventually. For the market to mature the environment has to provide the necessary regulations to make it safe for entrepreneurs and investors to place their resources. As the technology is still new and constantly evolving regulators are taking various routes. Small countries such as Singapore and Gibraltar are at the forefront allowing blockchain companies to operate and issue licenses. Their main motivation is attracting new businesses and thereby tax income. In addition the volume of new companies registering is still manageable. Large countries are often more hesitant - not to say rigorous. China for instance banned all initial coin offerings (ICO) and cryptocurrency-to-fiat exchanges in September 2017[5]. However, China favours innovation and is at the same time experimenting with a state issued cryptocurrency in order to remain in control of funds flowing and in and out of the country. In England and South Africa regulators allow "sandbox projects" to operate under supervision of regulatory authorities. Thereby regulators assure

[2] https://coinmarketcap.com/ CoinMarketCap is a website that tracks cryptocurrency data. Each currency is listed with its real-time price (averaged across exchanges) and related data such as market capitalization and trading volumes.

[3] https://www.tradingview.com

[4] The cryptocurrency market capitalization is calculated by multiplying the price by the number of tokens in circulation. This calculation is controversial as the price decreases instantly with a decrease in demand.

[5] https://hackernoon.com/navigating-crypto-regulation-china-fbae88697a21

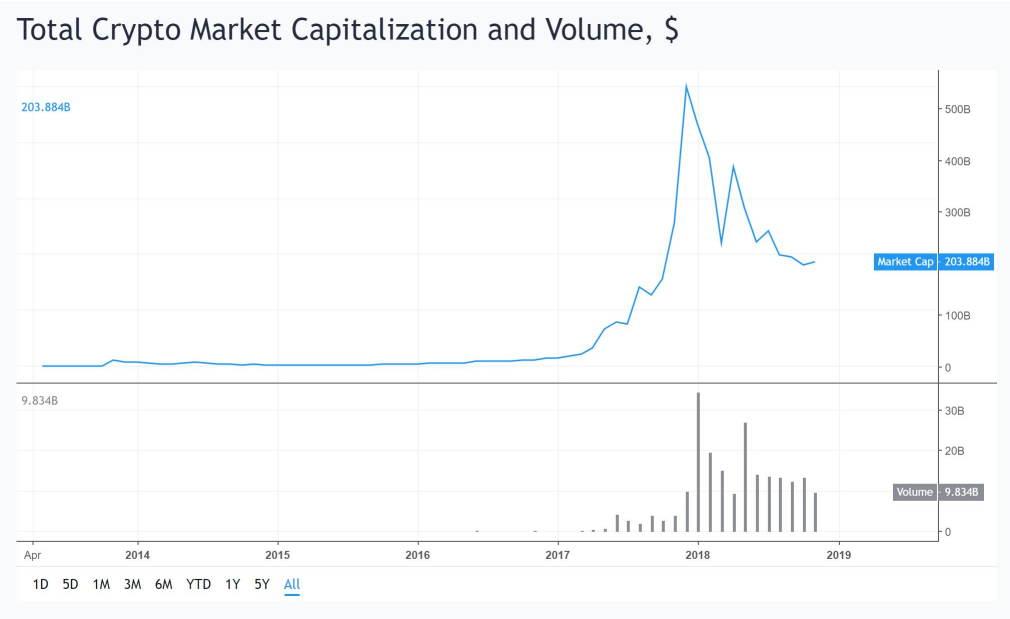

Fig. 1.1: Total Cryptocurrency Market Capitalization and Volume in USD

that new businesses operate within the countries laws and at the same time they learn to regulate companies with similar operations.

With the sheer volume of new companies entering the market since 2017, the need for measures to categorize and differentiate blockchain companies, to evaluate their risk and to define their legal status increases drastically together with the need to understand the market as a whole. As a starting point the Blockchain Center at the Frankfurt School of Management developed a cryptocurrency token classification framework. It is meant to serve as a foundation for blockchain regulation and taxation globally.

1.2 International Token Classification Framework

This section gives an overview of the token classification framework that has been developed by the Blockchain Center (BC) at the Frankfurt School of Management in Germany.

In the previous section we stated that a cryptocurrency is a privately issued currency (Mougayar, 2016, chap. 1). In this paper cryptocurrencies will be widely referred to as tokens.

In 2017 the Frankfurt School of Finance & Management in Germany launched a think tank and research center for blockchain technologies and emerging businesses. The Blockchain Center's intention is to be a knowledge platform for industry experts, start-ups and corporates[6]. Since its initiation, the Blockchain Center published research papers on blockchain technology in various contexts such as chemical industry, mobility, internet of things, sharing economy and manufacturing. In addition, the team performes due diligence on existing blockchains such as Ethereum. Through their research, webinars and conferences, the Blockchain Center has become a widely renowned institution in Europe and works closely with the German Federal Financial Supervisory Authority (BaFin[7]).

The motivation behind the International Token Classification (ITC) Framework is to provide a "tangible and holistic framework for the identification, classification and analysis of different token types"[8]. The framework is a starting point for regulators and tax authorities to consistently assess the legal and tax implications, associated risks and investment suitability of tokens[9].

The proposed framework has been developed over the course of several months by continuously testing and extending the logic until the outcome was suitable to classify the 800 highest ranked tokens by market capitalization. The result is a multi-dimensional approach and consists of four levels: Economic Purpose, Technological Setup, Legal Claim and Industry.

For a detailed description of the different layers of each category and their subcategories appendix A outlines all categories and definitions. The categorization caters for regulators, tax authorities, researchers and for investors who are seeking measures to evaluate risk and value of each token. Figure 1.2 indicates the basic sub-categories but the complete framework consists of a total of 70 categories and may be explored in more detail in A.1.

The economic purpose can be understood when considering certain examples. Bitcoin serves as an example for payment tokens as its primary use. Ethereum's token Ether is used as a fee to use the Ethereum platform to build upon. Utility tokens are cryptocurrencies with a pre-defined use, in this case Ethereum owners can make use of the Ethereum infrastructure to deploy and execute smart contracts. The least

[6] https://www.frankfurt-school.de/home/research/centres/blockchain
[7] Bundesanstalt für Finanzdienstleistungsaufsicht
[8] Frankfurt School Blockchain Center Newsletter November 2018
[9] https://www.mme.ch

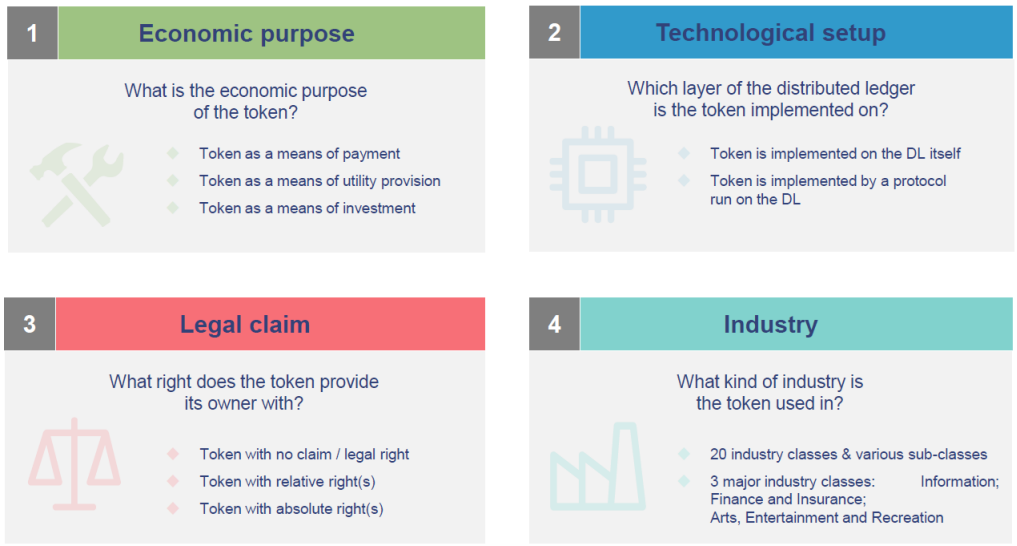

Fig. 1.2: Dimensions of the International Token Categorization Framework

frequent economic category is investment (or security) tokens. ICONOMI is a digital asset management platform that tokenizes digital assets into so-called Digital Asset Arrays (DAAs)[10]. DAAs are in essence similar to investment funds or ETFs.

The differentiation of the technological setup is comparable with the definition from section 1.1. When the token is implemented on the distributed ledger (DL) itself it refers to the underlying technology or a "system" and is considered to be the first layer. Any token that is not based on its own system but is implmented on top of another underlying technology is considered to be the second or third layer.

The legal claim and the industry categories are easier to grasp. The legal category refers to the rights that holding a token entitles one to. The determination of the legal rights has been a complex task for entrepreneurs and their law firms as there is little precedent to rely on. Industry tokens are purely based on the business sector the token issuing entity operates in.

[10] https://www.iconomi.net/

Chapter 2

Related Work

The following chapter outlines literature that has been published over the lifetime of cryptocurrencies. The first part elaborates on classification frameworks and has to rely on working publications as the concepts are still new - just like the cryptocurrency market itself. The second part describes papers that have used machine learning techniques to gain insights into cryptocurrency (mostly Bitcoin) price and market movements.

2.1 Classification Frameworks

This section provides an overview of token classification logics that have been published to date. The Blockchain Center has published the most indepth classification for the largest amount of tokens. This section begins by describing very basic classifications and ends by looking into frameworks that come close to the International Token Classification Framework.

Wu *et al.* (2018) distinguishes ”Coins” as cryptocurrencies that operate on their own independent network whilst ”Tokens” operate on top of such networks. The research shows the birthrate of coins has been relatively stable over the years. At the same time the number of tokens grew exponentially in 2017, together with the overall market capitalization of tokens (Wu *et al.*, 2018, p. 3 and 6). The authors concluded that the market capitalizations of coins grew at double the growth rate of tokens whilst their volatilities remain similar (Wu *et al.*, 2018, p. 6).

Swan (2015)’s book ”Blockchain” classifies blockchain applications into Blockchain 1.0, 2.0 and 3.0. The levels represent use cases for blockchain technology and is in line with the evolution of the technology to date. Blockchain 1.0 describes ”applications related to cash, such as currency transfer, remittance, and digital payment systems”. Blockchain 2.0 includes applications beyond payments to legally bind-

ing contracts related to value such as stocks, bonds, loans and property. Lastly, Blockchain 3.0 includes all applications that are beyond value transactions. The author considers governments, health, science, literacy as well as culture and art (Swan, 2015).

Mougayar (2016, chap. 1) who describes Blockchain as a layer on top of the internet, takes a more technical approach. He distinguishes blockchain applications based on their implementation as a layer on top of the internet network. The applications can be seen as "a trust layer, an exchange medium, a secure pipe [or] a set of decentralized capabilities" (Mougayar, 2016, chap. 1). As implementation layers he considers both private and public solutions for Blockchains as the first implementation layer, Blockchain Native Applications and Hybrid Blockchain applications which partly build upon (private) web applications and partly upon a Blockchain (Mougayar, 2016, chap. 1). Moreover, the author breaks down blockchain capabilities into ten subcategories, namely: cryptocurrencies, computing infrastructure, transaction platforms, decentralized databases, distributed accounting ledgers, development platforms, open source software, financial service marketplaces, peer-to-peer networks and trust services.

Kazan *et al.* (2015) investigated how Bitcoin companies configure value through digital business models, namely through the value chain, shops and networks as the three main configurations. In more detail the team in Copenhagen identified producers, transitioners, service providers, infomediaries, brokerages and disintermediators. The key finding was that value chain and value network driven businesses monetize their services for each transfer of a value unit, whereas value shop driven models commercialize through subsidized and revenue generating users (Kazan *et al.*, 2015). The study was conducted very early in the cryptocurrency timeline and is limited to five Bitcoin use cases only.

In line with blockchain, one of the first and widely used token classification frameworks was developed by a distributed economy think tank called "Untitled Inc". The think tank is an organized network of members who consider themselves as experienced professionals and domain experts in the cryptocurrency space[1]. Appendix A.2 shows the framework which classifies tokens into five dimensions: technical layer, (economic) purpose, underlying value, utility and legal status. Similarly to the ITC the collective at Untitled Inc developed their framework in order to make

[1] http://www.untitled-inc.com/ operating in Berlin, Frankfurt, Melbourne, Munich, Singapore, San Francisco, Tokyo, Vienna and Zurich

blockchain more accessible for regulators, politicians, investors and decision makers in business[2].

From a legal perspective, lawyers in Switzerland have published a widely used approach on categorizing tokens which is based on the Untitled Inc categorization and builds upon it. Moreover, it is an extension to the token categories[3] published by the FINMA[4] in early 2018. Whilst the FINMA used the econmic purpose as a primary category, the team of lawyers based their categorization on legal implications and made the economic purpose secondary. The three primary categories resulted in: native utility tokens, counterparty tokens and ownership tokens (Mueller, 2018). Appendix A.3 allows for more detail on the subcategories.

This paragraph indicates that research into cryptocurrency token classification frameworks is still in its infancy. The ITC as well as all the other mentioned categorizations have yet to prove their validity and applicability for regulators, investors and entrepreneurs.

2.2 Cryptocurrency Analysis

The following section dives into data analysis that has been conducted on cryptocurrencies with a focus on machine learning techniques. Due to the nature of machine learning a lot of the research focuses on predictions. We have structured the papers from broader research on various tokens to very specific research on individual tokens.

ElBahrawy *et al.* (2017) published one of the few studies focusing on the entire market and included 1469 tokens in their analysis. The team used the neutral model of evolution which is typically used in population genetics and ecology and are considering the view of a "cryptocurrency ecology" (ElBahrawy *et al.*, 2017). Although the market capitalization is increasing rapidly and tokens come and go, many properties of the market have been stable for years (ElBahrawy *et al.*, 2017). The analysis can be summarized in three major findings. Firstly, the market share distribution remains the same regardless of the total market capitalization. Secondly, the number of tokens did not change significantly as the token birth and death rate were

[2] http://www.untitled-inc.com/the-token-classification-framework-a-multi-dimensional-tool-for-understanding-and-classifying-crypto-tokens/

[3] https://www.finma.ch/en/news/2018/02/20180216-mm-ico-wegleitung/

[4] Swiss Financial Market Supervisory Authority

similar. Lastly, the time tokens remain in certain ranks[5] did not change. Bitcoin has always been number one and the lower the rank the shorter the time a token remains in the same position. However, the findings may not always hold true after the market grew exponentialy towards the end of 2017.

A study on factors influencing the price of five different cryptocurrencies was published by Sovbetov (2018). Autoregressive Distributed Lag (ARDL), a time series model used to predict currenct and lagged values of an exploratory variable, was used to determine the factors. Trading volumes and volatility turn out to be significant price drivers. Besides supply and demand, the research revealed that the cryptomarket attractiveness, marcro-financial and political factors play an important role in cryptocurrency price predictions (Sovbetov, 2018). Similar findings were published by Poyser (2017).

Phillips and Gorse (2017) used Reddit[6] data to predict cryptocurrency price bubbles. A hidden Markov model was built to detect epidemic and non-epidemic states of social media usage and trading volumes for Bitcoin, Litecoin, Ethereum and Monero. The model performance was enhanced by using a moving window approach. As a result, they developed a trading algorithm that performed better than a buy and hold strategy which was a great achievement in late 2017 when the market skyrocketed.

Correlation analysis and multiple linear regression allowed Abraham *et al.* (2018) to predict the direction of the price movements based on social media indicators. It turned out that twitter volumes and the google search volume index are more insightful than twitter sentiment which is overall neutral or positive.

Greaves and Au (2015) uses the "Union Find Algorithm" to group cryptocurrency accounts beloning to the same individual in order to evaluate the power of single players in the market on the price. In particular, the research investigated the influence of Mt. Gox on the price at the time. The two most informative variables were the net flow through Mt. Gox's account and the number of new addresses within an hour.

Saad and Mohaisen (2018) uses correlation analysis, multiple regression as well as a neural network and conjugate gradient algorithm with linear search in order to

[5] based on Market Capitalization
[6] Reddit is a social media platform that caters explicitly to subsets of users with particular interests

make Bitcoin price predictions. The the prediction accuracy of the regression model reaches 99.4%. The highly correlating features included the hash rate, the number of Bitcoins, the cost per transaction, the difficulty and the miner's revenue (Saad and Mohaisen, 2018).

More publications on bitcoin price behavior shall be mentioned for the sake of completeness. Amjad and Shah (2017) uses nonparametric time series prediction algorithms. Jiang and Liang (2017) builds a trading robot using conventional neural networks (CNN). After considering research based on models such as generalized autoregressive conditional heteroskedasticity (GARCH), recurrent neural networks (RNNs), long short-term memory (LSTM) and autoregressive integrated moving average (ARIMA), Jang and Lee (2018) implemented Bayesian Neural Networks (BNNs).

Chapter 3

Analysis

As stated in chapter 2, the purpose of this work is not to predict cryptocurrency prices but to investigate the relationship between token categories and token metrics such as price, velocity and trading volume. This chapter explains how the data points were collected, how and why additional metrics were computed and reveals the first insights of the exploratory data analysis (EDA).

3.1 Original Data

Initially the data was meant to be extracted solely from Santiment, a platform that makes cryptocurrency data accessible to traders and scientists. It turned out that the data was not yet fully available at the time of the research. Therefore, the research was conducted using data from CoinMarketCap paired with the ITC framework from the Blockchain Center in Frankfurt. To date (November 2018) 795 tokens have been classified. This section gives an overview of the original data.

The price, volume and market capitalization data was accessed through CoinMarketCap (CMC). The platform aggregates unconverted prices for each individual token-pair directly from the cryptocurrency exchanges. CoinMarketCap includes all exchanges that are: 1 operating for more than 60 days, 2 are accessible through an API and 3 provide a representative for any enquiries of CMC. Similar to stock data, prices are given as open, high, low and close prices. The metrics are converted to US dollars as a reference price based on current exchange rates. The volume is the total spot trading volume reported over the last 24 hours by all exchanges. The market capitalization is the price multiplied by the circulating supply of each cryptocurrency.[1]. Figure 3.1 shows one row of the CMC time series data. For the research we considered daily historical data from January 1, 2017 to November 16, 2018 which results in a maximum of 685 data points per token. All 795 tokens clas-

[1] https://coinmarketcap.com/methodology/#market-data

sified by the ITC were initially included in the analysis. Several data points turned out to be invalid due to a faulty conversion of negative exponential values and were excluded at a later stage.

Date	Open	High	Low	Close	Volume	MarketCap
2018-11-13	6373.19	6395.27	6342.67	6359.49	4503800000	110494466204

Fig. 3.1: Exemplary time series element from CoinMarketCap

The classification logic of the ITC was described in chapter 1. Figure 3.2 shows the information contained for each token. Besides the classification itself, the data includes helpful metrics such as URLs to Github, Twitter and Reddit, coin explorers as well as unique token names used to identify tokens on Santiment and CMC.

TokenID	TokenSymb	TokenLabel	TokenMining	SupplyCirc	SupplyTotal	SupplyMax	CMCID	urlCMC	SlugCMC	ClassCMC
1026	BNB	Binance Coin	Not mineable	95.512.523	192.400.000	-	1839	https://coinmarketcap.com/currencies/binance-coin/	binance-coin	Token

Economic	Tech	Legal	Industry	EconomicLabel	TechLabel	LegalLabel	IndustryLabel
EP22H	TS42D	LC32	IN10	EP22H: Utility Token > Settlement Token	TS42D: Non-native Protocol Token > ERC20 Token	LC32: Relative Rights Token	IN10: Exchange, Trading & Settlement

SlugSanbase	EthAddress	urlEtherscan	Website	Website2
binance-coin	0xB8c77482e45F1F44dE1745F52C74426C631bDD52	https://etherscan.io/token/0xB8c77482e45F1F44dE1745F52C74426C631bDD52	https://www.binance.com/	

Github	Reddit	Twitter	Explorer
	https://www.reddit.com/r/BinanceExchange/	https://twitter.com/binance	https://etherscan.io/token/0xB8c77482e45F1F44dE1745F52C74426C631bDD52

Explorer2	Explorer3
https://ethplorer.io/address/0xB8c77482e45F1F44dE1745F52C74426C631bDD52	

Fig. 3.2: Exemplary element of the token classification

For the analysis the ITC data was reduced to the token categories, the token name and the supply metrics. URLs and Ether addresses were excluded as they are not considered insightful features. Token categories serve as labels for the clustering in chapter 4 and are therefore renamed: once into the more general primary category and once into the secondary category which adds more detail to each class. For instance, a payment token (primary) can be an unpegged payment token or a stable coin (secondary).

The following sections describe the different steps taken to build a model. As Fayyad *et al.* (1996) and Wirth and Hipp (2000) stated, the data mining process is an iterative one. Nevertheless, the report shall follow a linear structure where possible. Iterations will be pointed out without significant changes to the structure.

Before exploring the data in more depth, additional features are computed in order to gain insights into the velocities and trends in the price data.

3.2 Feature Engineering

The planned machine learning models will be trained using a one dimensional dataframe of the 795 tokens classified by the Blockchain Center. The time series data consists of up to 685 data points per token. As a result, representative metrics to aggregate the time series data have to be identified. This section will describe some basic aggregation metrics, the conversion of binary metrics as well as some more complex financial indicators that were computed for the analysis.

3.2.1 Aggregated Metrics

Relative Volume: The relative volume puts the trading volume into perspective. Even though the trading volume of a token seems small in comparison with trading volumes of Bitcoin and Ethereum, with regards to its own market capitalization it might have a higher turnover. Therefore, the relative volume is calculated as the volume per token per day divided by the market capitalization of that token on the same day.

Differences Open-Close and High-Low: The section above outlines that each day token prices are captured by an open, high, low and close price. The prices can provide first insights into the volatility of the token. The differences between the high and low as well as the open and close price for each token and day are stored to get an understanding of the tokens' volatility.

Difference Maximal and Latest: Since the cryptocurrency market peaked in January 2018, the downfall has not stopped. Calculating the difference since each tokens' peak and the latest price shall be an indicator on how stable the coin was during the downfall.

Averages: Aggregation metrics helped to assign each token a single value for each metric. Over the past two years price, volume and market capitalization were fluctuating around their averages. Although the metric might not take into consideration the overall volatility of the market, it gives a first indication. If the average difference between the high and the low price is larger for one token, it is more volatile than another. If the relative volumes are higher on average, the token is

generally used more than another.

Standard Deviations: Another simple indication of volatilities is given by the standard deviations of the mean. The standard deviation has been calculated for volumes, close prices and market capitalizations for all tokens.

The aim of computing and aggregating these metrics was to find indicators that give more indepth information about financial market movements. Based on Murphy (1999) and Wilder (1978) indicators on trend and momentum are used in conjunction with volatility and volume movements.

3.2.2 Financial Indicators

Technical analysis of financial markets is a highly complex task. For the sake of interpretability, the guiding theme was to select simple metrics and complement them with more complex metrics where necessary. In our case we are adding more complex metrics to explain trends in the data as well as momentum. The TA-Lib[2] Python package was used to compute financial metrics for technical analysis.

Trend: The moving average is widely used for trend-following systems (Murphy, 1999, chap. 9). The simple moving average which was applied to the price data was computed for a series of time intervals. Time intervals between ten and 200 days[3] are commonly used. Because the cryptocurrency market is highly volatile, the analysis focused on shorter intervals, namely 10, 30 and 50 days. The 200 day interval was computed first but excluded at a later stage to assure more data points can be included.

Momentum: Wilder (1978) describes momentum as one of the most useful concepts but one of the hardest ones to understand. He suggests the Relative Strength Index (RSI) as a momentum oscillator. The RSI measures the magnitude and velocity of directional price movements. To keep the metrics aligned the daily intervals are chosen such that they mimic the moving averages.

Figure 3.3 visualizes the computed price metrics for the Ripple token as an example. Traders use the metrics as buy or sell signals. For example, when the shorter interval (e.g. 10 day) moving average falls below the larger interval (e.g. 50 day)

[2] https://mrjbq7.github.io/ta-lib/doc$_index.html$

[3] https://www.investopedia.com/ask/answers/122414/what-are-most-common-periods-used-creating-moving-average-ma-lines.asp

moving average, traders would see a sell signal. When the 10 day moving average crosses back above the 50 day moving average, traders would consequently see a buy signal (Murphy, 1999, chap. 9 and 10). Wilder (1978) explains RSI thresholds as signals for a stock to be overbought or oversold. When the RSI hits the threshold 70, traders would sell. When it falls below 30 traders would buy. Trendline analysis would take the technical analysis a step further but this exercise was left to a future, more in-depth research into price movements.

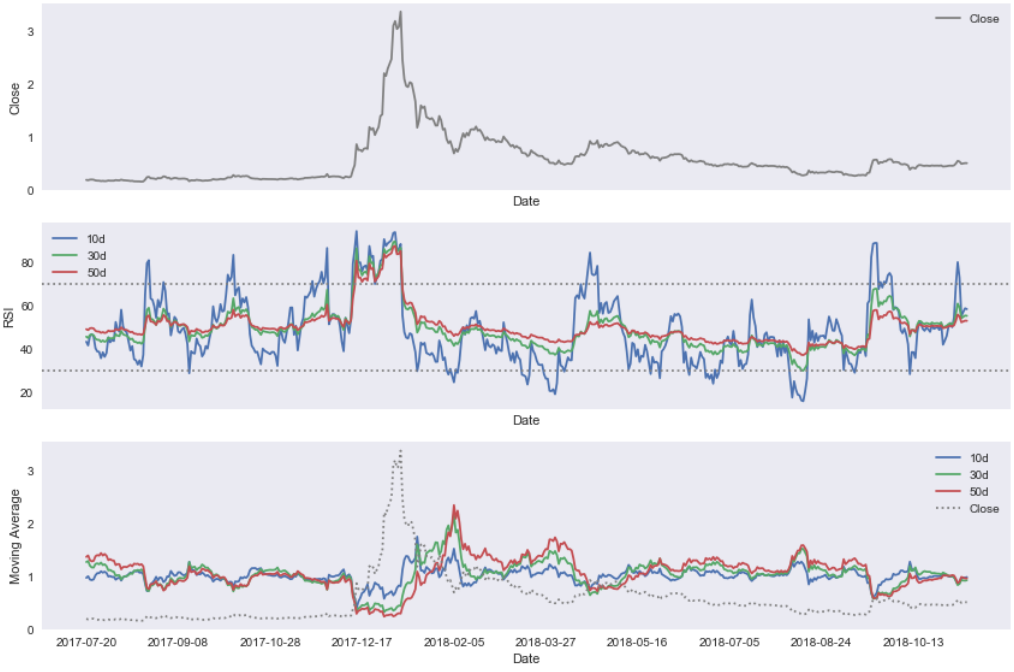

Fig. 3.3: Ripple Time Series Decomposition

3.2.3 Factor Metrics

After computing metrics about market movements we are considering further classes in the ITC data. Some of the features such as Token Mining which can take the values "Mineable" and "Not-Mineable", can be insightful but is not yet included in the feature set. As the variable can only take two values we are converting it into a factor variable that can take zero or one as a value (James *et al.*, 2013, chap. 3). Another similar variable is the "Maximal Token Supply". Some tokens have a maximal supply whilst others have an infinite supply. Instead of excluding the feature as a whole, a conversion into a factor variable can make it an insightful predictor. If a token has a maximal supply the factor will take value "1" and "0" if the supply is infinite.

The feature engineering leaves us with a dataset of 34 numeric features to be used for further exploration. Due to irregularities in the data further preprocessing steps need to be included to investigate outliers and to bring the data to a workable scale.

3.3 Exploratory Data Analysis

For the exploratory data analysis (EDA) we include the full feature set of 46 variables for 795 tokens. The data includes the numeric features, labels, token names and five day future close prices. The future close price metrics were included in case the data was used for predictions at some stage in the future. In this chapter the data will be preprocessed for model building. This includes identification of missing values, handling of outliers as well as data transformation. The data then undergoes a correlation analysis. The aim is to get an initial understanding for whether the feature space has redundancies. Lastly, the distribution of the token categorization labels will be quantified.

3.3.1 Preprocessing

Classification models are very sensitive to noise in the data. When classifiers are built from low quality data, the model will be less accurate (Teng, 1999). To enhance model performance the data will be pre-processed in three steps: identification of missing values, outlier detection and data transformation.

Table 3.1 shows all missing values in the dataframe, before the "Maximal Supply" variable was converted into a factor variable. The conversion was already the first step in handling missing values. For the financial metrics calculated the algorithm did not pick up on 165 tokens. This is because some tokens' time series includes less than 200 days. This means that the token has not been born 200 days before the data was captured. Nevertheless, the metrics can be computed for the 10, 30 and 50-day time intervals. Table 3.2 shows that nine tokens are still missing after the 200-day interval was excluded. This means that they have been in the market for less than 50 days. Those tokens will be excluded from the analysis so to keep the feature space broad enough. After the treatment and elimination of missing values, the dataset includes 786 tokens.

To gain a better understanding of the data and its features, pandas[4] and matlab-

[4] http://pandas.pydata.org/pandas-docs/version/0.15/index.html

Metric	Missing
SupplyMax	642
5dClosePct	165
5dFutureClose	165
5dCloseFuturePct	165
10dMovAv	165
10dRSI	165
30dMovAv	165
30dRSI	165
50dMovAv	165
50dRSI	165
200dMovAv	165
200dRSI	165

Tab. 3.1: Table of Metrics with Missing Values

plot[5] for Python offer useful visualization tools. As a start the data is expressed as a boxplot. The graphs are included in appendix B. Except for the "Average Relative Volume" and the calculated RSI values, the boxplots basically consist of a flat line. Figure 3.4 shows the boxplot for the circulating token supply. As Kokoska and Zwillinger (1999, chap. 2) suggest the values outside the box are mild or extreme outliers. In this case most values are very close to zero and all values that are above zero could be considered outliers. Some tokens are trading at an enormous scale compared to others. For example, one Bitcoin would be worth several thousand Dollars whereas one Ripple would range around (and mostly below) one Dollar.

Nevertheless, the extreme values must be investigated further. A first check is to see which values are equal or less than zero. In this case, values below zero only make sense for a few values such as the average differences between prices. In fact, there are several features that include values equal to zero. This might be caused by very small numbers that were not processed correctly when the data was extracted. More than 50 tokens have near zero and zero values. Due to their large number, the elimination of those tokens is not considered a valuable option. Next, the extreme values on the upper end need to be evaluated. As the boxplots are so distinct, a threshold of 20% is selected to check which variables cause the data to be imbalanced. Some of the values are outliers such as the first top ten tokens by market capitalization such as Bitcoin, Ethereum and Ripple. Others such as the Russian

[5] https://matplotlib.org/contents.html

Metric	Missing
5dClosePct	9
5dFutureClose	9
5dCloseFuturePct	9
10dMovAv	9
10dRSI	9
30dMovAv	9
30dRSI	9
50dMovAv	9
50dRSI	9
200dMovAv	163
200dRSI	163

Tab. 3.2: Table of Metrics with Missing Values

Mining Company token stands out due to high prices. Four other outlier tokens are eliminated as their outlier character was not explicable and faulty processing was suspected. Removing these outliers has not caused a significant improvement in the boxplot.

The boxplots also indicate that the data is highly skewed. Skewness can be caused by outliers. Brys *et al.* (2003) describe skewness as the asymmetry of univariate continuous distributions. This means that the data is not normally distributed. For model building the aim is to find regularities in the distribution of the features. Therefore, the data has to be further processed. Appendix B.2 shows that scaling did not add significant variety to the data, most features still look the same. As an alternative to scaling data Benoit (2011) suggests a logarithmic transformation of variables using the natural logarithm. The logarithmic transformation is a useful tool to convert highly skewed variables into variables that are closer to a normal distribution (Benoit, 2011).

The log-transformation has certain implications on the analysis. For example, log transformation can only be applied to non-zero positive data. Changyong *et al.* (2014) indicate that one could replace zero values by near zero varibales but this can have significant impact on the outcome. For this analysis, log transformation will only include variables that are non-zero and positive. Table 3.3 displays the names of the log-transformed variables.

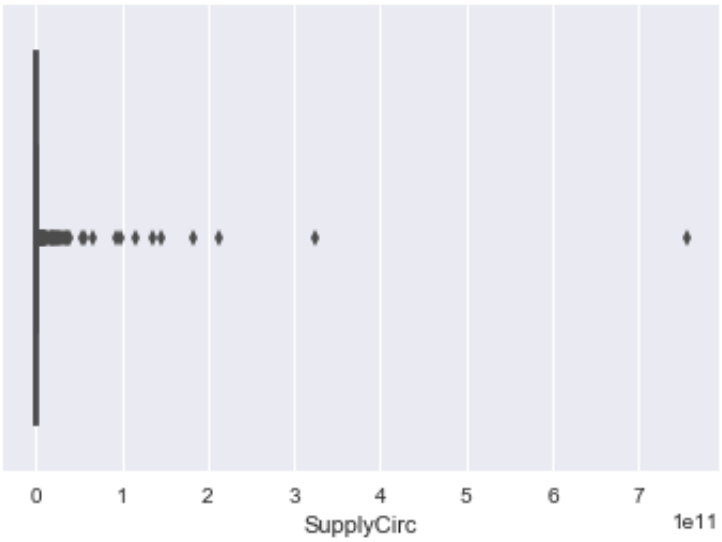

Fig. 3.4: Boxplot for Supply Circulation Feature

Non-zero, positive features	
SupplyCirc	10dMovAv
SupplyTotal	30dMovAv
AverageVolume	50dMovAv
MaximalVolume	AvMarketCap
LatestVolume	MarketCapStDev
DifferenceLMVolume	MaxMarketCap
VolumeStdDev	LatestMarketCap

Tab. 3.3: Non-zero, positive features for Log-Transformation

Appendix B.3 displays the new boxplot with the transformed data. The logarithmic transformation was effective and the transformed features suggest a normal distribution of the features. In the next step the data will undergo a correlation analysis.

3.3.2 Correlation Analysis

This subsection will elaborate on a multivariate correlation analysis of the numeric log-transformed dataset in order to detect relationships in the data. Yu and Liu (2004) describe the correlation analysis as a feature selection process. Features are investigated for relevance and redundancy. Furthermore, they describe correlations amongst features and a class (Yu and Liu, 2004).

Raschka and Mirjalili (2017, chap. 10) and James *et al.* (2013, chap. 3) describe important tools for correlation analysis. The aim is to gain a quick overview of whether there are correlations in the data and if so, whether they are positive or negative. A correlation matrix provides insight to detect if there are strongly correlating pairs of variables. Figure 3.5 shows the correlation value between -1 and 1 on a colour scale where the brightest colour represents a stong positive correlation and the dark blue represents a strong negative correlation. Based on Taylor (1990) a strong correlation has a correlation coefficient between 0.68 and 1 and accordingly -0.68 and -1.

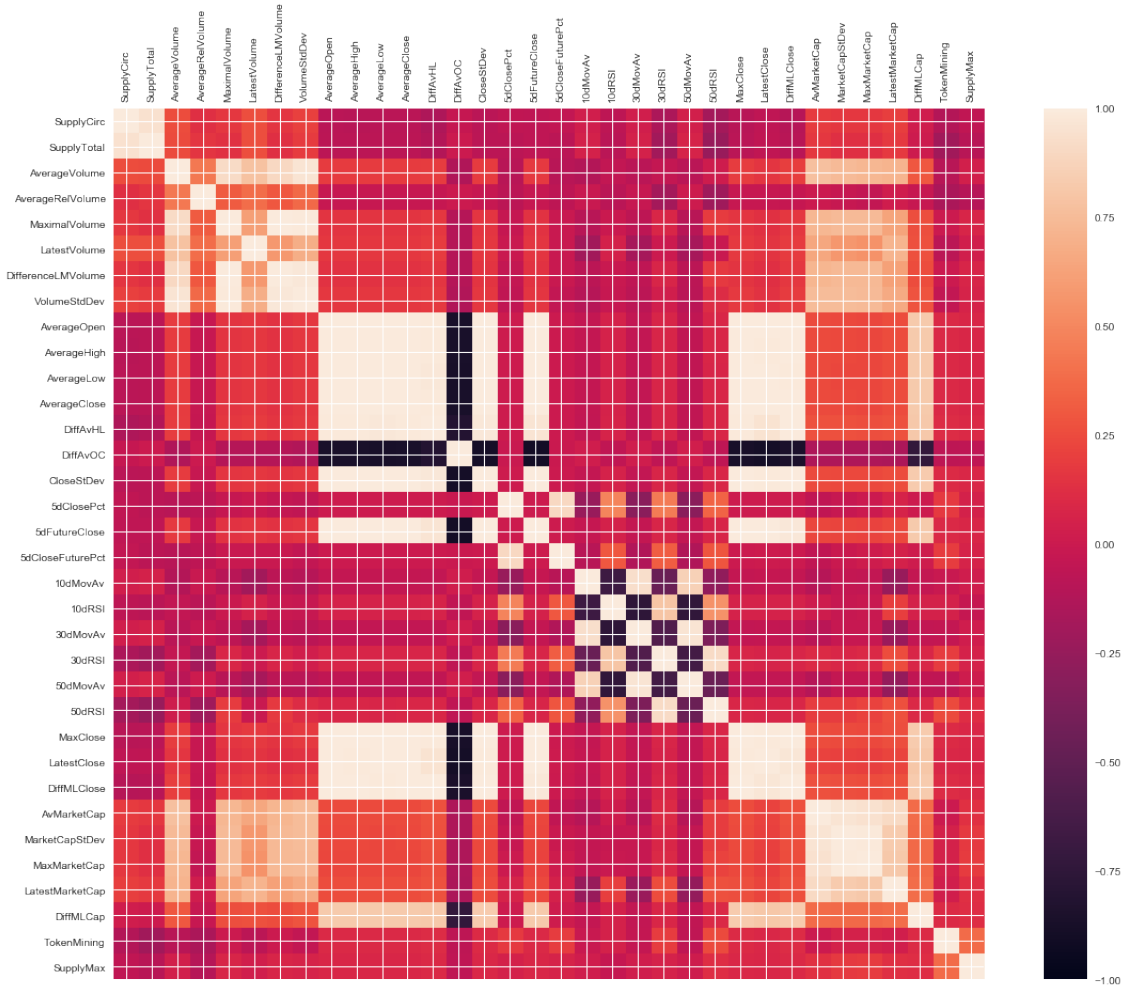

Fig. 3.5: Correlation Matrix Visualization for 34 Numeric Features

Strong positive correlations can be observed in certain variable groups. For example, the token supply circulation correlates with the total token supply. Similarly,

price and volume metrics are correlating positively. Regarding variables of different variable groups, the volume and the market capitalization are positively correlated. This relationship is not surprising because the more tokens are traded, the higher the price and the number of tokens released or mined.

The average relative volume sticks out from the volume metrics group. It is not correlated with the other volume features. This means that the relative proportion of the token volume traded does not increase and decrease with the overall trading volume. Even though the volume increased the proportion of the market capitalization did not change. This is a relevant insight when looking into token performance. Even though the volumes of Bitcoin are much higher as a number, the proportion traded in relation to its market capitalization is similar to many other tokens.

The average high, low, open and close price and the average standard deviation of the close price are positively correlated with the average difference between the maximal and the latest market capitalization. This is an unexpected relationship. We could think of it as the higher the rise the lower the fall. Highly volatile tokens whose average price has been high lost a large portion of their market capitalization over the past year. When considering the relationship in reverse, the tokens with a lower average price and standard deviation did not suffer as big a downfall.

The moving averages of 10, 30 and 50 days show a positive correlation. At the same time the RSIs for 10, 30 and 50 days are negatively correlated with the moving averages. This originates in the nature of the metrics. Whereas the moving average takes the mean of the prices, the RSI is making an estimate whether a token is overbought or oversold. A high RSI tells the trader to sell whereas a rising moving average (starting with the shortest interval) is a signal to buy. When analyzing the data, it is evident that the negative correlation of the 10-day Moving Average and the RSI is stronger than the 50-day Moving Average's.

Another negative correlation is the difference between the average open and close price and the rest of the price metrics. One possible interpretation is that the open and close prices do not rely on the general trend of the data but on other factors such as the activity of traders at a certain time in the day.

High correlations indicate redundancy in the data. Feature selection and variable importance will be addressed in chapter 4. Generally, it is recommended not to ex-

clude features too early in the analysis because some machine learning algorithms have their own mechanisms to prioritize variables.

3.3.3 Exploration of Classification Labels

Before diving into token classification modelling, this subsection gives a brief overview on the distribution of the different token categories.

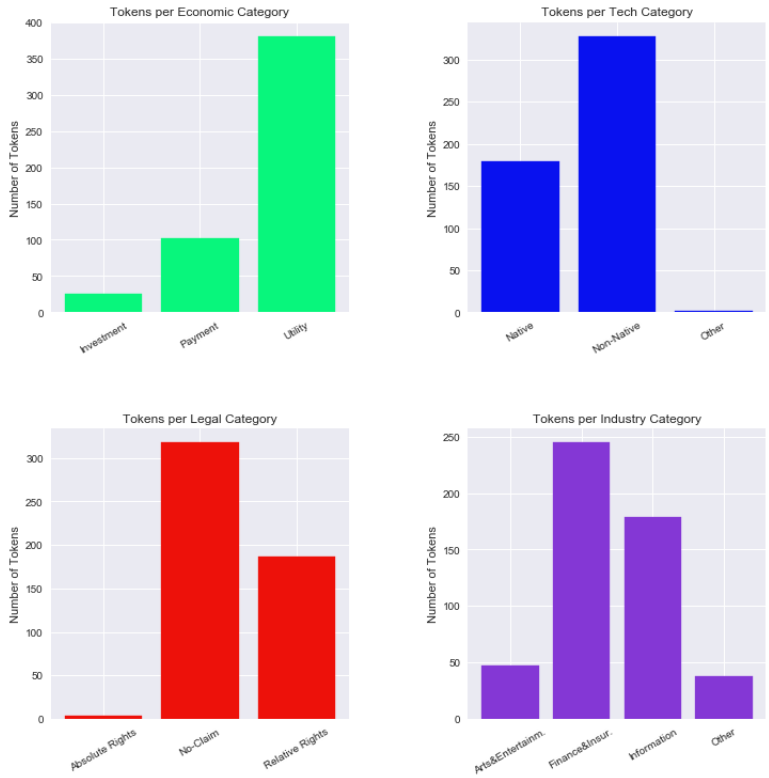

Fig. 3.6: Number of Tokens per Primary Label

Figure 3.6 shows the number of tokens per category and label. All four primary classification dimensions (economic, tech, legal and industry) are heavily imbalanced. López *et al.* (2012) describe class imbalance as one of the most persistent complications in supervised learning for real-world problems. Lessmann (2004) argues that support vector machines handle class imbalances well for business oriented classification problems and Chen and Wasikowski (2008) suggests a new ROC-based feature selection metric. After computing these metrics we learned that other ones outperformed them.

Appendices B.4 and B.5 show the distribution of the secondary, more detailed labels. Again, the imbalance is clear.

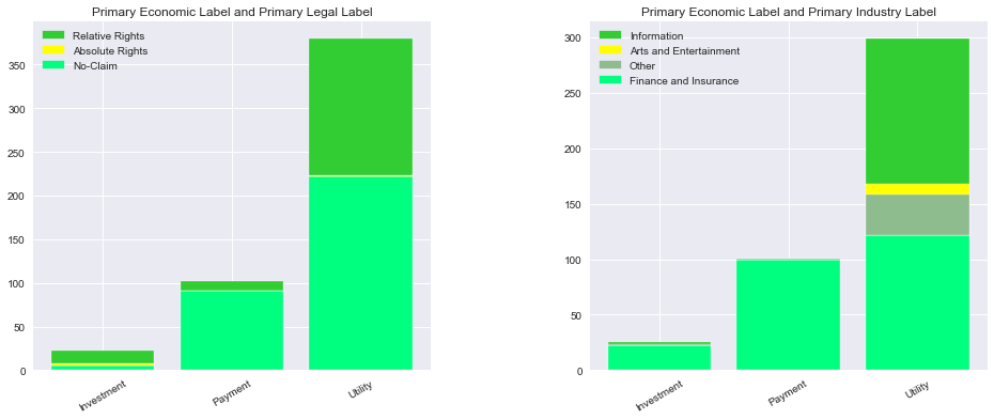

Fig. 3.7: Number of Tokens per Primary Economic, Legal and Industry Label

After understanding the distribution for each class individually, stacked bar plots help to understand how classes are related. Figure 3.7 shows the distribution of the legal rights primary label on the primary economic labels. One insight is that payment tokens are not absolute right tokens and with a very high probability, they are no-claim tokens. Investment tokens are in the majority relative right tokens. However, there are some no-claim tokens which is surprising for investment tokens. Investment tokens are securities by definition where the investor participates in returns and losses. Table 3.4 shows the six tokens which fall into that category. One reason could be that the companies issuing the tokens navigated through an uncertain legal environment and had to justify a different purpose than they originally intended in order to be able to operate. Another reason could be that the purpose of the token changes over time, or the tokens were mis-classified.

SlugCMC	**EconomicLabel**
c20	EP23M: Investment Token >Derivative Token
diamond	EP23Z: Investment Token >Other Investment Token
karma	EP23L: Investment Token >Debt Token
melon	EP23P: Investment Token >Fund Token
elixir	EP23L: Investment Token >Debt Token
bullion	EP23Z: Investment Token >Other Investment Token

Tab. 3.4: No-claim Investment Tokens

On the right graph in Figure 3.7 the primary economic label is plotted in relation with the primary industry label. Nearly 100% of the investment and payment to-

kens are rightly classified into the finance and insurance category. Utility tokens serve multiple different industries.

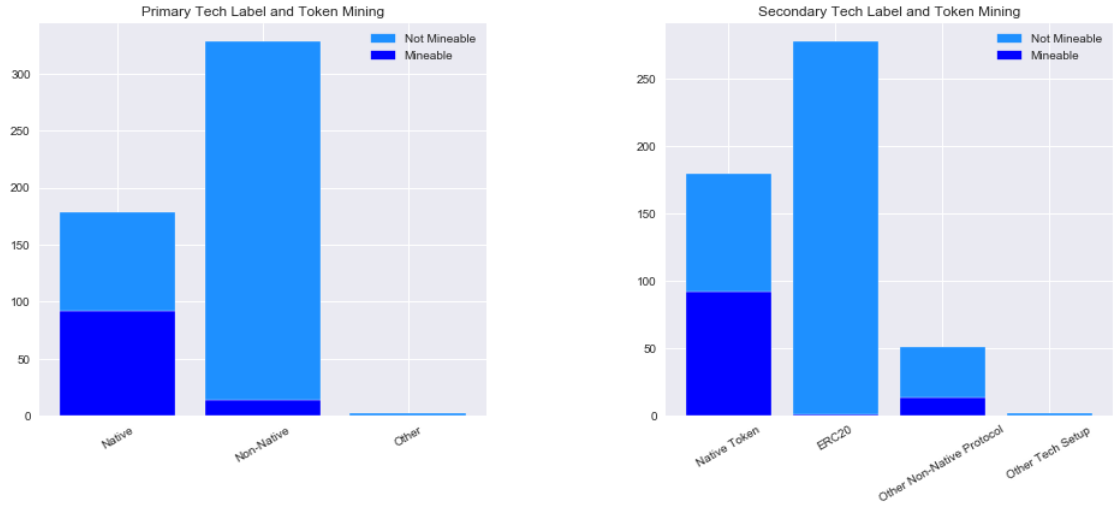

Fig. 3.8: Tech Labels in Relation to Token Mining

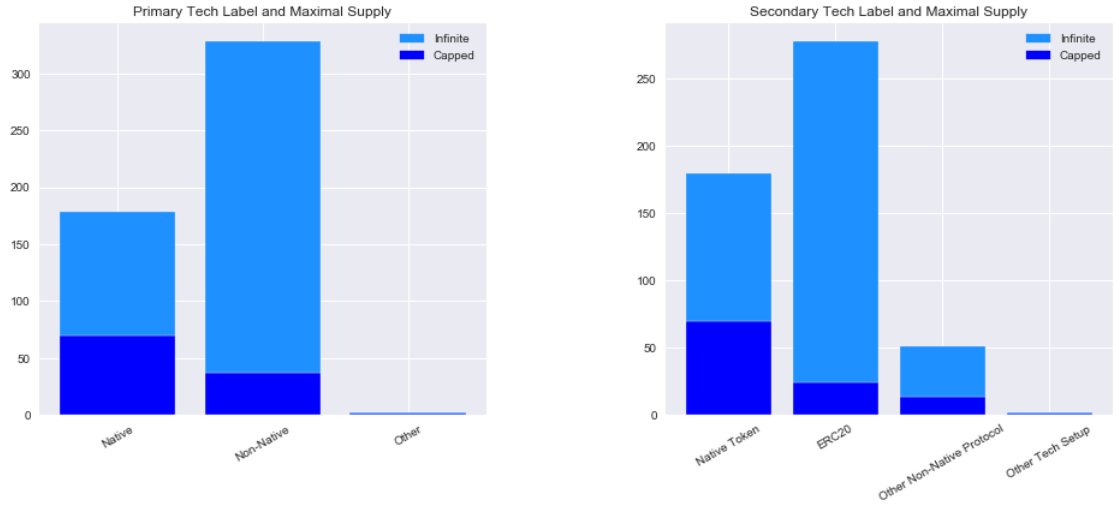

Fig. 3.9: Tech Labels in Relation to Maximum Supply

Appendices 3.8 and 3.9 investigate the relationship between the primary and secondary technological label with the token mining and the maximum token supply. Mineable tokens are typically protocol tokens but can be both native or non-native protocol tokens. ERC20 tokens together with most non-native tokens are typically not mineable. Native tokens are roughly half mineable and half not-mineable. Con-

sidering the token supply, the majority of tokens have an infinite supply. Each technological category has a proportion of capped tokens. For native tokens this proportion is 39.1% whilst for non-native tokens only 27.8% have a limited supply.

Chapter 4

Token Classification Models

Chapter 3 gave a broad understanding of the data. Chapter 4 will provide the classification analysis using machine learning techniques. The section includes variable importance and feature selection, supervised learning methods for classification of the categories as well as unsupervised learning techniques to look beyond that predefined classification of tokens.

4.1 Variable Importance and Feature Selection

In this section the feature set will be analysed. For this exercise various different techniques are used and compared. The evaluation of variable importances is done before starting the model building as it is often insightful on its own. Relationships between certain input variables and the desired output classes are found initially. Feature selection has various benefits for prediction models. Models can be significantly faster and more cost-effective by reducing the feature space, for instance by removing redundant variables (Tuv *et al.*, 2009; Guyon and Elisseeff, 2003; Chandrashekar and Sahin, 2014). Feature selection can also be seen as an unsupervised learning exercise like in Hierarchical Clustering where the power of features is evaluated independently from the output variable (Talavera, 1999).

James *et al.* (2013, chap. 6) and Friedman *et al.* (2001, chap. 10) elaborate on various different feature selection methods. Feature selection techniques can be classified in subset selection, shrinkage and dimension reduction. Table 4.1 shows a selection of different methods, most of which were applied to the data. The different techniques vary in performance and interpretability. Appendix C.1 visualizes the scores from the chi-sqaured test on the best subset selection algorithm for the primary economic category. The test resulted in high values for the factor variables. The rest of the features scored much lower and it is not obvious which features are

the best to use for the analysis besides the factor variables: token mining and maximum token supply. I want to elaborate on two more techniques which are known to perform well feature selection.

Subset Selection	Shrinkage Methods	Dimension Reduction Methods
Best Subset Selection	Ridge Regression	Principle Component Analysis
Stepwise Selection	Lasso	Partial Least Squares
Recursive Feature Elimination	Gradient Boosting	

Tab. 4.1: Selected Feature Selection Techniques

Chen and Guestrin (2016) introduces the XGBoost algorithm[1], a tree-based model using boosting and the gradient descent algorithm to improve performance. Figure 4.1 shows the feature importances of all the metrics. The XGBoost model accuracy reaches 79.08% when all 34 features are included. When including only 25 features, the performance increases to 81.70% and the model is computationally less expensive. The features with the least importance for the model are: the average close, high and low prices, the difference between the average open and close prices and the difference between the average maximal and lowest prices, the 10-day moving average, the 5-day close percentage, the 30-day moving average and the difference between the latest and maximal volume. According to the XGBoost model, not only are those features less relevant, they also add noise so the model performs worse when they are included. The features with the highest importance are the latest volume, the total token supply and the 30-day RSI which is known as the momentum indicator. Consindering the total token supply averages per economic category, payment tokens on average have the highest value (6.4bn). With approximately 3.9 billion, utility tokens have the second highest supply and investment tokens have the lowest average at about 0.7 billion. Even more significant is the same distribution in the latest volume metric. For payment tokens, the average latest volume is at roughly USD 73 billion tokens, for utility tokens USD 8.0 billion and for investment tokens at approximately USD 0.2 billion. Interestingly, the momentum and velocity metric 30-day RSI differs only slightly amongst the economic categories. Per definition, one would suspect that utility and payment tokens have a significantly higher 30-day RSI than investment tokens. In reality, the metric values 47.5 for payment tokens, 45.5 for utility tokens and 46.2 for investment tokens.

Figures C.2 to C.4 in appendix C show the XGBoost feature importance plots for the primary labels tech, legal and industry. Table 4.2 summarizes the maximal ac-

[1] https://xgboost.readthedocs.io/en/latest/python/python_intro.html

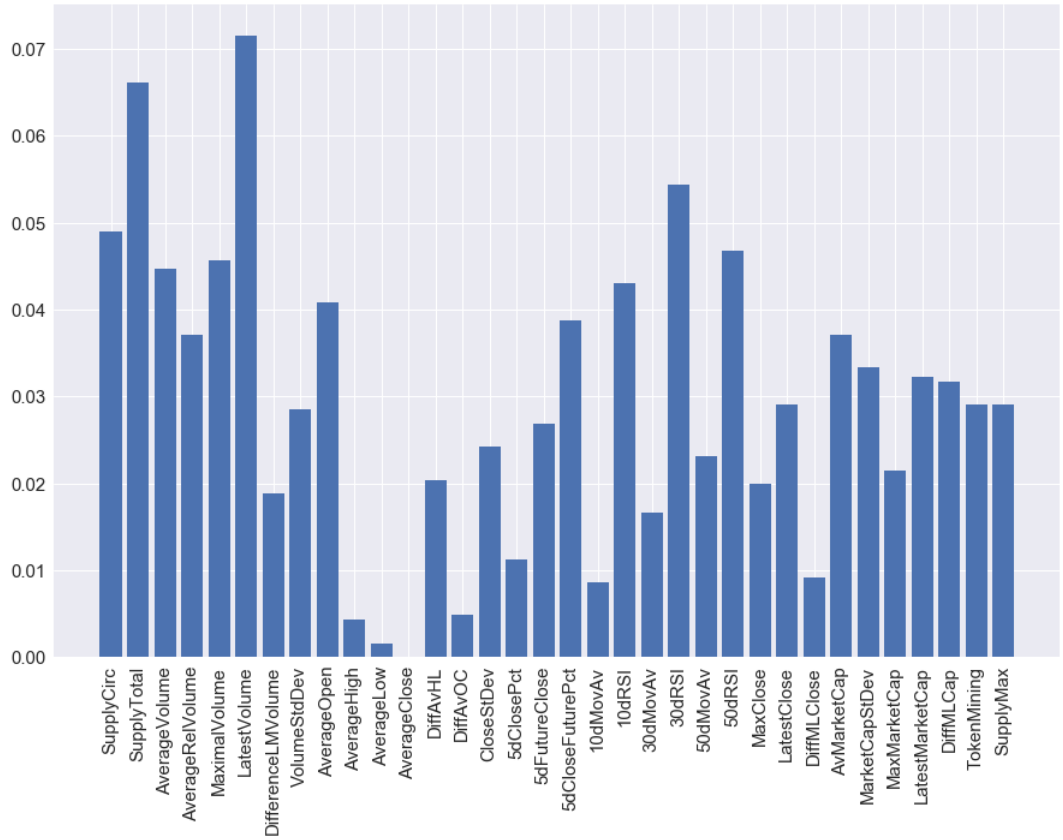

Fig. 4.1: Shrinkage Method XGBoost

curacies as well as the optimal number of features per prediction model. Whilst the economic and technological categories can be predicted with an accuracy of over 80%, the legal and industry labels are less predictable based on the available data.

Label	Maximal Accuracy	Number of Variables
Economic	81.70%	25
Tech	86.27%	21
Legal	65.36%	16
Industry	47.71%	19

Tab. 4.2: XGBoost Maximal Accuracies and Optimal Number of Features for Primary Labels

The XGBoost variable importance for the primary tech label in figure C.2 shows that slightly different variables are the most important in this prediction. Although the latest volume is the second most important again, this time the latest market

capitalization and the 50-day RSI also come into play. Native tokens have significantly higher recent exchange trading volumes (about USD 51 billion) and their average latest market capitalization remains at USD 788 million, whilst non-native tokens were at USD 28 million. This together with the USD 9.5 billions in recent trading volumes for non-native tokens is in line with what Wu *et al.* (2018) found when analyzing the difference between native tokens (coins) and non-native tokens. They considered tokens to be an "explosive immature ecosystem". The importance of the latest market capitalizations and volumes show that the static results during the downward trend in the market distinguishes the tokens from each other. This could mean that native tokens that remain at a higher level are more stable than non-native tokens. One potential reason is considered to be the unsustainable exponential growth in Initial Coin Offerings (ICOs). ICOs were often carried out without having Minimal Viable Products (MVPs) in place to prove the applications' value and viability (Wu *et al.*, 2018). When investors lost trust in the market, recently born tokens were more likely to be sold quickly whilst people who invested in the most popular tokens such as Bitcoin and Ethereum held.

Lastly for this section a principal component analysis (PCA) as indroduced by Wold *et al.* (1987) is used. This technique is a form of dimensionality reduction. The aim is to compute features that are representative of a combination of original features. The advantage is that large feature sets can often be reduced into a two or three dimensional feature space. Figure C.6 in the appendix shows a representation of the first two principal components computed for the token data. One dimension (x-axis) is representative of the two factor variables maximal token supply and mineability. The length of the arrow indicates the loading, i.e. how much of the information contained in the feature is included in the principal component. The second component is dominated by the average volume variables, followed by the market capitalization variables and by the supply metrics.

Figure 4.2 shows how much of the variance is explained by each principal component (blue bars) and the cumulatived variance explained (red line). One drawback of PCA is that beyond two components the interpretability of the data decreases. Eight principal components explain just over 95% of the variance of the whole dataset of 34 features. In order to maintain better interpretability, we plot the data points into the two-dimensional principal component space and colour the values based on their labels. Figure 4.3 shows the two dimensional principal component space for the technological labels. Appendices C.6 to C.8 display the same plot but colour coded based on the economic, legal and industry labels. The

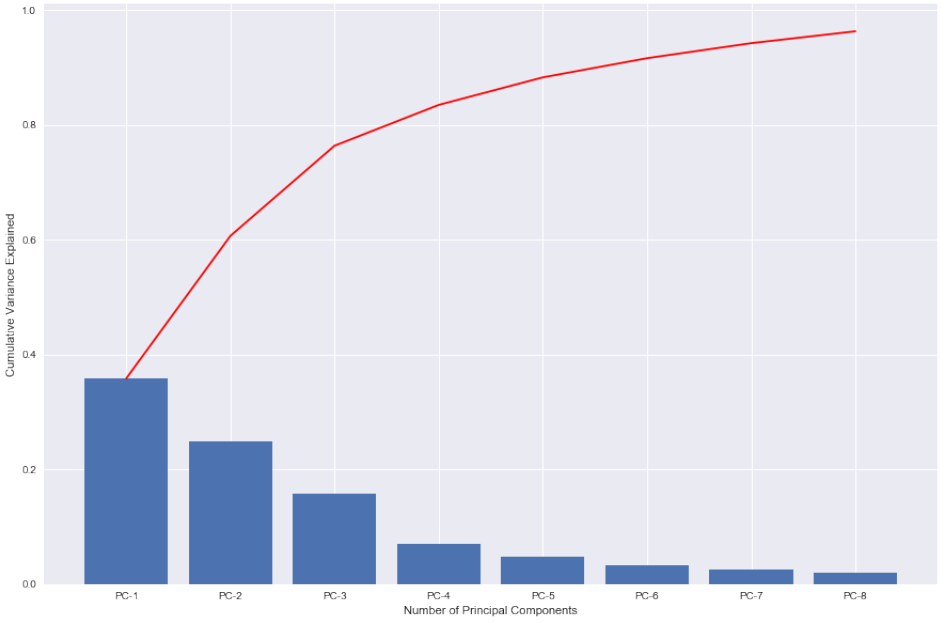

Fig. 4.2: Cumulative Variance Explained for First Five Components

plots show data in three different clusters but none of the labels mimics the clusters exactly. As the clusters are clearly separable based on their x-values one can confirm that the clusters are based on different combinations of maximal token supply and whether the tokens are mineable or not. There is no such categorization in the ITC, yet. In chapter 5 the possibility of introducing new categories is discussed.

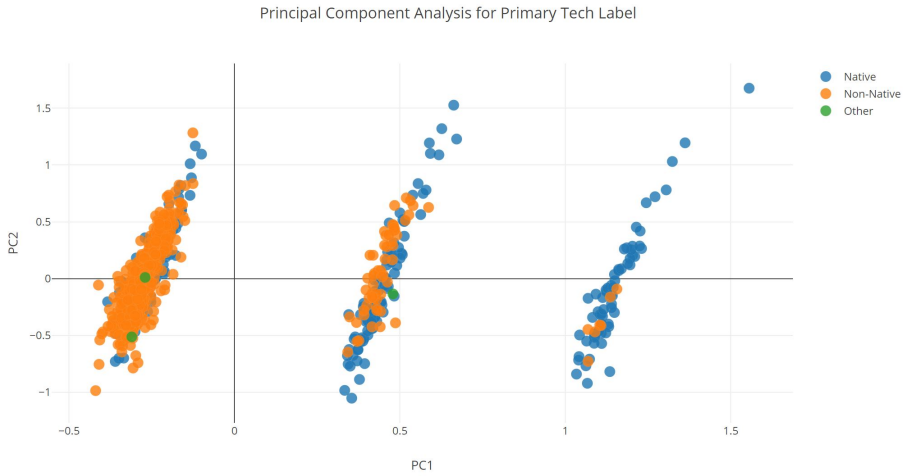

Fig. 4.3: 2-D Principal Components for Primary Tech Labels

In the following section on supervised clustering results of different models based on different feature selection methods will be compared. It is found that both the principal components and XGBoost feature variable importance can lead to better results.

4.2 Supervised Clustering

This section describes supervised learning techniques used to predict the economic, technological, legal and industry classes of each token. The aim of supervised learning techniques is to build classifiers based on a set of training data where the outcome class is known (Kotsiantis *et al.*, 2007). The classifiers are then used to predict the class of a test set. The outcome classes in the test set are first hidden and are then used to compare the predicted labels with the actual labels. There are a number of supervised learning classification models which Kotsiantis *et al.* (2007) evaluated and compared. The results of their work are summarized in table D.1 in the appendix. The following models are selected: K-Nearest Neighbors (KNN), Neural Networks, Support Vector Machines and tree-based models such as Decision Trees, Bagging, Random Forests and Gradient Boosting. Models like Support Vector Machines and Neural Networks are known for their high prediction accuracies, especially for large feature spaces. KNN and Decision trees are suitable for smaller feature spaces and are often used for better interpretability. The used feature set is small and therefore good performances for tree-based models and KNN are expected.

Before building each model the data is split into a training and a test set. Fan *et al.* (2008) developed a well-known package for classification problems suggesting an 80/20 split as in 80% for the training data and 20% for testing. As the data is imbalanced, the splitting was done in a stratified manner. This means that the proportion of classes was taken into account when splitting it into the test and training set. For KNN and the Neural Network, the target labels were encoded into numeric variables. For the economic label ”0” represents the investment token category, ”1” for payment tokens and ”2” for utility tokens. In the following paragraphs, each model will be briefly described by the example of the economic label using the three different feature sets. Each model is built on parameters that were specifically chosen to improve the accuracies. The results are summarized in table 4.3.

K-Nearest Neigbors is widely used in classification problems. It is a simpler and therefore a more understandable algorithm but computationally expensive (James

et al., 2013). The classification algorithm stores all different outcome categories and classifies each data point by a majority vote of its k neighbors (Cover and Hart, 1967). The category that appears most in one group of neighbors gets assigned to all data points in the group. The KNN algorithm can be modified by choosing the distance function that is applied to measure the distances between the data points and by choosing k, the number of neighbors. The KNNClassifier function allows for Manhatten, Euclidean and Minkowski distance. Amongst the suggested distances Walters-Williams and Li (2010) consider Euclidean as the most popular one and it will be used for the analysis. In order to define the optimal number of neighbors the accuracy for k=1 to k=50 was calculated. Figure 4.4 shows that the accuracy has two peaks (marked in red) at 26 and 28 neighbors. K=26 neighbors were chosen in order to achieve a test set accuracy of 81.7% when including all features and when using the XGBoost feature selection of 25 variables. For the principal component features the highest accuracy was achieved at 81.0% with k=24 neighbors. Figure 4.5 shows the resulting confusion matrix for the XGBoost features. It becomes clear that the algorithm is not picking up on the investment category although it is represented in the test and training set. Again, the reason is the imbalanced data. In the training data 32 data points are labeled investment tokens and 8 in the test set. Nevertheless, the algorithm does not manage to correctly classify the investment tokens.

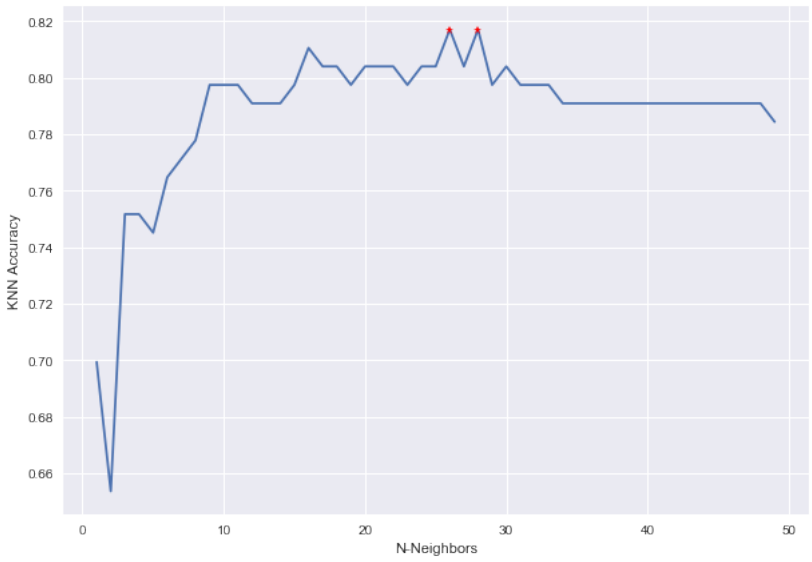

Fig. 4.4: KNN-Accuracy for XGBoost Feature Selection based on Number of Neighbors

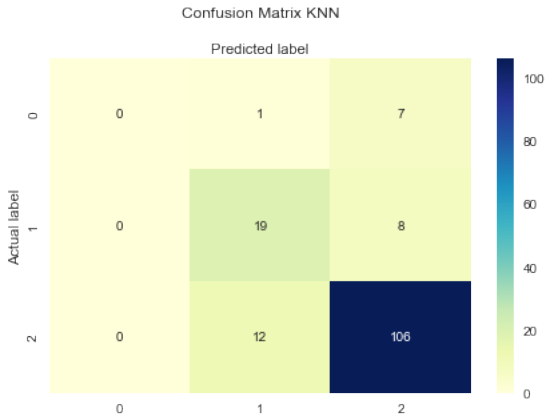

Fig. 4.5: KNN-Confusion Matrix for XGBoost Feature Selection Test Set

Tree-based models include Decision Trees, Bagging, Random Forests and the used Gradient Boosting method XGBoost. The Decision Tree builds on the basis of the other three models. The algorithm splits the data into different groups based on the most significant features. This step is repeated until the data is partitioned into a sufficient number of subgroups (Breiman, 2017). The splits are estimated through techniques such as the Gini Index which measures the pureness of each group or the Entropy which defines the degree of disorganization in a system (Friedman *et al.*, 2001, chap. 9). The test set accuracy using Entropy to estimate the split outperforms the Gini Index by 0.6%. The accuracy of 79.7% is achieved for all three feature sets using a maximum depth of two for the principal component features and of four for the XGBoost feature selection and all features.

The Bagging algorithm improves the performance of the Decision Tree. The aim is to reduce the variance of the predicted values. This is achieved by combining several tree classifiers modeled on randomly chosen subsets of the training data (Breiman, 1996, 1999). The most important modification in the Bagging model is the number of base estimators, in this case the number of trees to be combined. The Bagging algorithm improved all decision tree results. The PCA feature set was improved by 0.7% (n = 6) whilst the XGBoost feature set improved by 1.3% (n = 37) and the original feature set by 2.0% (n = 10) in their accuracies.

Lastly, Breiman (2001) introduced the Random Forests algorithm which is considered to be a *panacea* for data science problems. This algorithm is used for predictions for both classification and regression problems, for dimension reduction, for treatment of missing values and outlier detection. Instead of building one tree, the

classification algorithm builds multiple trees. To classify a new data point, each tree returns a classification value and the decision is made based on a majority vote by all trees (Breiman, 2001). The optimal number of estimators was computed by comparing accuracies for numbers between one and 400. The optimal numbers are 358 (XGBoost features selection), 20 (PCA features) and 39 (all features). Random Forests perform best when using the original feature set and achieves an accuracy of 81.0%. As the algorithm has its own feature selection method, the previous feature selection would not add additional value.

The last tree-based algorithm is the Gradient Boosting algorithm. The goal of the Boosting algorithm is to identify so-called *weak rules* and combine them into one single *strong rule*. This happens in an iterative process where all features (rules) are first equally weighted, then after several interations, the weak rules are identified based on their lower accuracies (Chen and Guestrin, 2016). XGBoost is considered ten times faster than the Generalized Boosted Models algorithm GBM (Chen and Guestrin, 2016). For this model, various parameters had to be adjusted: the maximal tree depth, the number of estimators (trees), the alpha rate and the learning rate which represents the Boosting learning rate. The alpha rate is the regularization term on the weights which are assigned to the rules[2]. Even after choosing the optimal parameters, the XGBoost did not outperform the other models.

Chang and Lin (2011) introduced a library for support vector machines. The idea is to create a space with as many dimensions as features in the dataset. Each output variable gets assigned one coordinate in the n-dimensional space. Next, the different groups in the data get separated such that the line is the furthest away from the two closest points of each group. Support Vector Classification (SVC) also tolerates outliers which are data points that are located on the wrong side of the line i.e. apart from their category (Chang and Lin, 2011). The parameter to regulate the error term is C. The other two parameters that were tuned are the kernel which defines the shape of the separating line and the kernel coefficient gamma which regulates the risk of overfitting on the training data. The Support Vector Classification model reached the maximal accuracy of 81.7% for the PCA feature set with prediction error C equal to 10 using the RBF kernel with gamma eqaul to 1.

Friedman *et al.* (2001, chap. 11) describes the central idea of neural networks as the extraction of ”linear combinations of the input variables as derived features [and then, the target is modeled] as a nonlinear function of these features”. The

[2] https://xgboost.readthedocs.io/en/latest/python/index.html

structure of a neural network includes an input layer, one or more hidden layers and an output layer. Each hidden layer consists of a number of perceptrons. Perceptrons weigh the input features at each layer to improve the performance of the model (Friedman *et al.*, 2001). Although the cross-validated accuracy achieved by the neural network is 81.1%, the test set accuracy remains below the best performing model at 79.1% for the XGBoost feature set. The confusion matrix of the neural network allows more insight into the distribution of the predicitons. The neural network classified all tokens in the test set as utility tokens and thereby misclassified 27 payment and eight investment tokens out of a total of 153 tokens in the test set.

Test Set Accuracy	**KNN**	**Decision Tree**	**Bagging**	**Random Forests**	**XGBoost**	**Support Vector Machines**	**Neural Networks**
XGBoost	**81.7%**	79.7%	81.0%	79.7%	78.4%	79.7%	79.1%
PCA	81.0%	79.7%	80.4%	78.4%	75.8%	**81.7%**	76.5%
All Features	**81.7%**	79.7%	**81.7%**	81.0%	78.4%	79.7%	77.1%

Tab. 4.3: Economic Label Prediction Accuracies based on Model Feature Selection

Table 4.3 is an overview of the performances of all computed models for the economic label. The different rows show the results for differently selected features. The models that include all variables performed very well, whereas the models built on the principal component features were the least accurate. The accuracy of the XGBoost feature selection came close to the selection of all features and is computationally less expensive. Each feature selection achieved 81.7% accuracy at its maximum. On average, the best performing classification method for the economic label is K-Nearest Neighbors.

The same models with accordingly adjusted parameters were used to predict the technological labels native (”0”), non-native (”1”) and other technological setup (”2”). Overall, the performance was between 5% and 12% above the accuracies for the economic label. One possible reason for the predictions being more accurate is that one class (”Other”) only consists of three tokens. None of the algorithms predicted these three outliers correctly. But neither did any algorithm for the outlier in the economic label (”Investment”) which consists of 40 tokens. As a result, the predictions for the tech category were more accurate as the remaining two classes include more tokens (native: 247 and non-native: 512). In two out of the three feature selection methods the Random Forests model outperformed all others. Only for the original feature set the Bagging algorithm outperformed Random Forests.

All prediction accuracies for the technological label are summarized in table 4.4. The principal component feature selection method achieved the highest test set prediction accuracy at 90.2% compared to 81.7% for the economic label.

Test Set Accuracy	KNN	Decision Tree	Bagging	Random Forests	XGBoost	Support Vector Machines	Neural Networks
XGBoost	86.3%	85.6%	88.2%	**88.9%**	85.0%	85.6%	85.0%
PCA	85.6%	83.7%	87.6%	**90.2%**	86.3%	85.6%	84.3%
All Features	86.3%	85.6%	**88.9%**	88.2%	85.6%	86.9%	85.6%

Tab. 4.4: Tech Label Prediction Accuracies based on Model Feature Selection

The test set prediction accuracies for the legal and the industry labels are summarized in tables D.1 and D.2. The overall accuracies remain under 80%. The legal label predictions are maximal 75.2% accurate. The industry label predictions are only 55.6% accurate at their maximum. As variables included in the models are mostly market and technology related, the legal and industry labels could not be predicted very accurately. Chapter 5 outlines a number of variables that could potentially be gathered to make predictions more accurate, especially for market unrelated labels.

4.3 Unsupervised Clustering

Unsupervised learning algorithms cluster unlabeled data based on (dis-) similarities (Zhu and Goldberg, 2009). This is the key difference to the supervised learning algorithms that were used in the previous section. Unsupervised learning algorithms are mainly used for clustering and outlier detection (Zhu and Goldberg, 2009). The most popular unsupervised clustering algorithms are K-Means Clustering and Hierarchical Clustering (James *et al.*, 2013). In this section, the two algorithms are applied to the token classification data to evaluate if additional categories could be used to distinguish tokens.

The first step in unsupervised learning is to define the number of clusters the data should be split into. Figure 4.6 shows the cost of adding a new cluster plotted against the number of clusters. To determin the optimal number of clusters the Elbow rule (Kodinariya and Makwana, 2013) is applied. The elbows are marked by red dots in the plot. They indicate where adding a new cluster reduced the costs by less than the previously added cluster (Kodinariya and Makwana, 2013). The two most obvious elbows in the graph are at k=2 and k=5 clusters and will be used for further modelling.

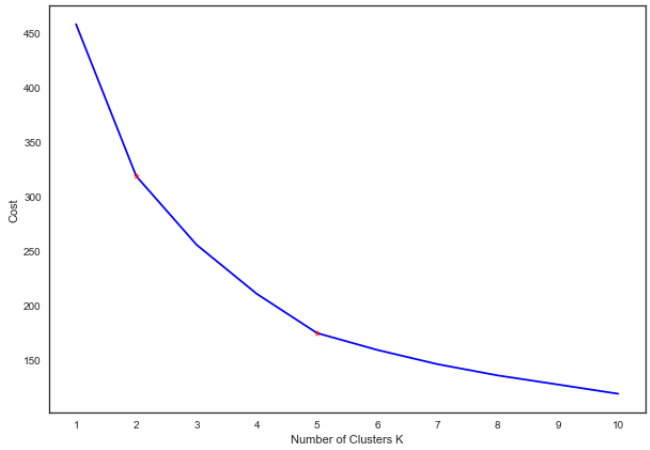

Fig. 4.6: Cost-Cluster Plot to Determine Optimal Number of K

K-Means Clustering partitions a dataset into a pre-defined number of distinct clusters. The algorithm iterates through different cluster allocations until it finds an optimal (local) solution (James *et al.*, 2013, chap. 10). The cluster allocation forms by measuring the distances between the datapoints (Likas *et al.*, 2003; James *et al.*, 2013). Figures 4.7 and 4.8 show the distributions of the clusters using K=2 and K=5 clusters. To plot the data, the first two principal components were used. Another option is to select two dominant features to plot the data but the principal components explain more of the variance in the data. In chapter 3, we found that the first principal component is dominated by the two factor variables; token mining and maximal token supply.

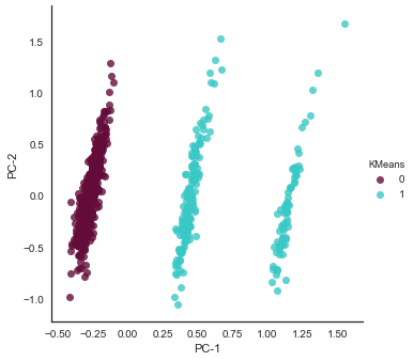

Fig. 4.7: K-Means Scatterplot based on First Two Principal Components K=2

When evaluating the different categories for K=2 it becomes clear that the data

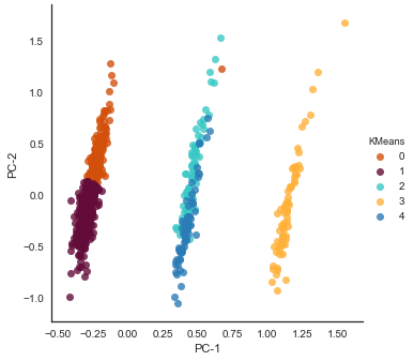

Fig. 4.8: K-Means Scatterplot based on First Two Principal Components K=5

points marked in purple in Figure 4.7 are all tokens that are both not mineable and have an infinite token supply. The cluster gathers the majority of the 545 tokens. Examples for tokens in the cluster are Stellar, Tether, Tron or the Binance Coin. The second cluster marked in turquoise gathers 217 tokens that are either mineable regardless of whether their token supply is capped or infinite, or not mineable and have a capped token supply. Examples for the second cluster are the top five tokens by market capitalization; Bitcoin, Ethereum, Riple, Bitcoin Cash and EOS. Other well-known tokens such as IOTA or ZCash are also included. The unsupervised learning clusters do not match any of the clusters of the ITC.

Next, the clusters of K=5 will be evaluated. The classification results in two large clusters with 341 (purple) and 205 (orange) tokens and three smaller clusters with 75 (turquise), 71 (yellow) and 70 (blue) tokens. Again, the tokens are separated based on the first principal component and therefore by their token supply and whether they are mineable or not. Additionally, the clustering into five clusters takes the second principal component into account. It separates the purple cluster from Figure 4.7 into two clusters in Figure 4.8. Figure C.5 in the appendix shows that the second principal component is dominated by volume and market capitalization metrics. The tokens included in the orange cluster, located above the purple cluster, have an average relative volume of 10.1%, whilst the purple cluster has an average relative volume of 4.7%. Moreover, the average market capitalization of the orange cluster is roughly 25 times larger than the average market capitalization of the purple cluster. The orange cluster is considered to be the high volume, non-mineable tokens with infinite supply. The cluster includes tokens such as Stellar, Tether, Tron, Binance Coin but also Ethereum. Ethereum is a mineable token and therefore an outlier in the category. In Figure 4.8 Ethereum is the only orange data point amongst the turquoise data points. The purple cluster has the lowest overall

average volume and is considered the low volume, non-mineable tokens with infinite supply. Table 4.5 summarizes the clusters and gives token examples for each cluster. Whether such a classification should be added to the ITC will be discussed further in chapter 5.

	Token Mining	Supply Maximal	Token Examples
Orange (0)	Not mineable*	Infinite	Ethereum*, Stellar, Tether
Purple (1)	Not mineable	Infinite	MOAC, Paypex, C20
Turquoise (2)	Not mineable	Capped	Ripple, EOS, Cardano, IOTA
Yellow (3)	Mineable	Capped	Bitcoin, Litecoin, Dash
Blue (4)	Mineable	Infinite	Monero, ZCash, Dodgecoin

Tab. 4.5: K-Means Clustering Overview for K=5

The second unsupervised clustering method used in this work is Hierarchical Clustering. It is sometimes chosen over K-Means clustering as the optimal number of tokens does not have to be initially specified (Johnson, 1967). The underlying algorithm is based on pairwise dissimilarities of the data points. Müllner (2011) defined the following methods that are used to calculate the distances: single, complete, average, weighted and centroid, median and ward. After trying all options the most balanced clusters were achieved using the ward distance meassure. Another advantage of hierarchical clustering is the visualization called dendrogram, which is a tree-based representation of the clusters (James *et al.*, 2013, chap. 10). To determine the number of clusters, a horizontal axis is drawn over the vertical tree. Thereby, the dendrogram is cut into a number of clusters (Wilks, 2011; Johnson, 1967).

In appendix E.2 the uncut dendrogram is displayed. Figure 4.9 shows the cut dendrogram for K=5 clusters. The equivalent for K=2 clusters is appended in E.1 together with the cluster plots based on the principal components in figure E.3 for K=2 and figure E.4 for K=5 clusters. The Hierarchical Clustering into K=2 clusters mimics the K-Means Clustering for K=2 clusters. This means that one cluster includes all tokens that are not mineable and have an infinite supply. The cluster size is 545. The second cluster gathers the other 217 tokens. The clustering into five clusters is slightly different from the K-Means clustering. It seems that the Hierarchical Clustering puts a stronger emphasis on other metrics such as volume and market capitalization. Table 4.6 shows the differences amongst the clusters. Although some clusters are very similar to the K-Means clusters, some changes are made such that the largest average volume and market capitalization tokens are all in one cluster. Ethereum for example, moved from being in one cluster with tokens

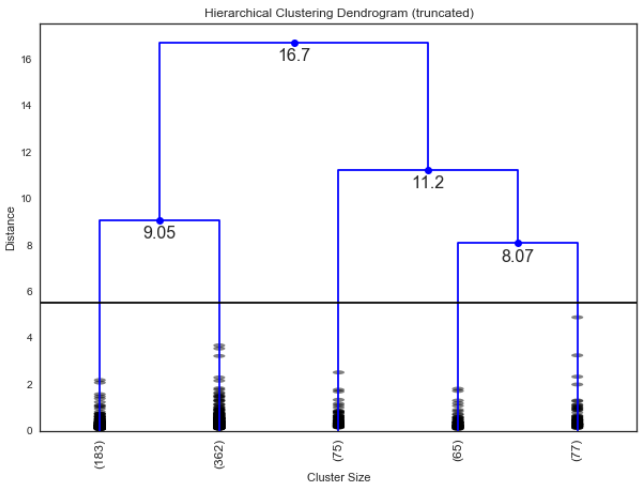

Fig. 4.9: Hierarchical Clustering Dendrogram K=5

	Token Mining	Supply Maximal	Token Examples
Orange (1)	Not mineable	Infinite	Tether, Stellar, Tron, Binance Coin
Purple (2)	Not mineable	Infinite	Maker, MOAC, Paypex, C20
Turquoise (3)	Not mineable	Capped	Ripple, EOS, Cardano, IOTA
Yellow (4)	Mineable	Infinite	Komodo, Monacoin, Emercoin
Blue (5)	Mineable	Capped*	Bitcoin, Ethereum, Litecoin

Tab. 4.6: Hierarchical Clustering Overview Token Mining and Max Supply for K=5

like Stellar and Tether to being grouped with Bitcoin and Litecoin in the blue cluster. This cluster mostly consists of native, no-claim tokens with a few exceptions. All tokens in the blue cluster have high volumes and high market capitalizations. However, the average relative volume is fairly low. The reason for this could be that people hold on to their tokens, hoping their value would increase. Another possibility is that a vast amount of the mined tokens have been distributed but are not accessible in the sense that they are sitting on a lost harddrive. A third reason could be that the tokens are traded directly more than through exchanges. Considering that the tokens in this cluster have very strong communities of developers and early adopters, this is a valid and probable reason .

The orange cluster has the highest average relative volume, meaning that on average more than 10% of its average market capitalization is traded. Compared to all other clusters this is high and indicates that the underlying applications are actually used. Binance for example is a well known cryptocurrency exchange and is used as a gateway to buying cryptocurrencies that are not available on other major

	Av. Volume	Av. Rel. Vol.	Av. MarketCap	KMeans
Orange (1)	Medium	11.1%	Medium	4 (0)
Purple (2)	Lowest	4.6%	Lowest	0 (4)
Turquoise (3)	Medium	4.9%	High	1
Yellow (4)	Low	2.7%	Medium	3
Blue (5)	Highest	3.4%	Highest	2 (3, 4)

Tab. 4.7: Hierarchical Clustering Market Metrics for K=5

exchanges. Tether is a USD pegged cryptocurrency and used when the market is experiencing a downfall. Traders use Tether as a consistent store of value instead of trading back to fiat currencies. In figure E.4 the differentiation between the purple and the orange cluster is similar to the K-Means plot. Both clusters include tokens with infinite token supply that are not mineable but the orange cluster has higher volumes and market capitalizations on average. The purple hierarchical cluster includes 362 tokens and it included 341 tokens in the K-Means clustering. One well-known token that was clustered differently was the Maker token. The turquoise cluster remained the same in both unsupervised clustering methods. The yellow cluster became diminished by strong tokens such as Monero, Ethereum Classic, ZCash and Dodgecoin. As a result, the cluster is lower in volume and the cluster consists of less popular tokens such as Komodo and Monacoin.

As such, the Hierarchical Clustering appears to be an indicator of token performance. The clusters are not mimicing any of the ITC categories exactly, although the blue cluster in hierarchical clustering involves mostly native tokens.

Chapter 5

Conclusion

The task of classifying cryptocurrency tokens is two-sided. Firstly, a suitable framework needs to be designed. Secondly, the classification needs to be constantly back-checked. This report is a general analysis of cryptocurrency market data and a back-check of the ITC framework with real-world data. We intended the work to be a broad exploration rather than a narrow one. In the following section, three fields of research will be described that we believe are worth further exploration going forward.

5.1 Outlook

Firstly, the collected time series data on 795 cryptocurrencies together with the computed financial metrics of moving averages and the RSIs can be used for modelling with machine learning. One applicable technique is the Long Short-Term Memory (LSTM) algorithm which achieved a prediction accuracy of over 99% for the Bitcoin time series using standard averaging or the exponential moving average. Currently, the time series data is on a daily basis. For building automated trading algorithms, shorter time intervals need to be used.

The second path for future analysis is to take the feature engineering to the next level. This can be done by finding ways to include more detail about each time series into the feature space. For example, one way is to use state transitions for better representation of the time series. Additionally, new data points could be added to improve prediction accuracies of categories and add insightful information on them. For example, the token birthdate could provide insight into the token age, i.e. if older tokens have a higher market capitaliztation than younger tokens. Or, if younger tokens are mainly non mineable with infinite supply, as this would require less planning and community building prior to an ICO. For the supervised learning models in chapter 4.2 the prediction accuracies varied largely amongst

categories. Whilst the accuracy for the industry classification reached 55.6% at its maximum, the maximal accuracy for the technological label was achieved at 88.9%. To improve the accuracy for the industry label a sentiment analysis on Twitter or Reddit data can be pursued to identify predominant topics that allow us to derive the industry label of the applications. To increase the prediction accuracy of the legal label, one could include the country of legislation as a data point. Different legislations favour different token purposes and therefore the country information would add value. To get more accuracy for the technological label, data on the Github activity is considered relevant and can be gathered through the Santiment API in the future.

Thirdly, the most extensive field of further research is to build additional token categories based on real-world data. In this study one additional categorization was presented. When using additional data such as google search trends or twitter volmes, a popularity metric can be included into the token categorization. The market data on its own can be further developed into performance metrics. Machine learning models for prediction can give an indication on whether a token is more or less predictable. Furthermore, this predictability metric can indicate how stable a token is compared to other tokens. Other categories could gather information about the underlying organizations. For example, is there an uninstitutional community like the Bitcoin Community or is there an profit-driven company behind an application? For both investors and regulators such metric adds valuable insight.

5.2 Implications

The ITC will make it easier for regulatores and investors to evaluate and distinguish tokens. The tokens have been manually classified by Blockchain experts. In this research, it was discovered that the classification for primary and secondary labels can not yet be automated by using algorithms on market data, as the prediction accuracies are not sufficient. However, once more data is included, the token classification models may help to assign economic and technological labels to new tokens. For the legal and the industry label, more analysis needs to be done using additional metrics in order to evaluate whether the process can be automated. A manual back-check by Blockchain experts should still be mandatory at this stage.

The unsupervised clustering is a proof of concept that algorithms will pick up information that is beyond the current classification of the ITC. We believe that the more

data is used to explain the behavior of tokens, the more the market can be demystified. When looking at listed equity profiles today, we see information from various perspectives on each stock. From charts with trend lines over the market sentiment (bearish vs. bullish) to recent news about the stock. For cryptocurrency tokens there is still very little information published and investors have to gather their information across a large number of different sources. These sources can often be informal chat groups that have access to first hand knowledge. This mechanism only works for a small community of people and is not accessible to people outside these communities. Therefore, an extensive token profile accessible for every investor, regulator and entrepreneur will allow a fairer market for a larger circle of investors. The ITC as well as the Santiment database are promising starting points for such token profiles and we believe that we are not far from establishing more transparency for cryptocurrency tokens.

Bibliography

Abraham, J., Higdon, D., Nelson, J. and Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis, *SMU Data Science Review* **1**(3): 1.

Amjad, M. and Shah, D. (2017). Trading bitcoin and online time series prediction, pp. 1–15.

Benoit, K. (2011). Linear regression models with logarithmic transformations, *London School of Economics, London* **22**(1): 23–36.

Breiman, L. (1996). Bagging predictors, *Machine learning* **24**(2): 123–140.

Breiman, L. (1999). Pasting small votes for classification in large databases and on-line, *Machine learning* **36**(1-2): 85–103.

Breiman, L. (2001). Random forests, *Machine learning* **45**(1): 5–32.

Breiman, L. (2017). *Classification and regression trees*, Routledge.

Brys, G., Hubert, M. and Struyf, A. (2003). A comparison of some new measures of skewness, Springer, pp. 98–113.

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods, *Computers & Electrical Engineering* **40**(1): 16–28.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines, *ACM transactions on intelligent systems and technology (TIST)* **2**(3): 27.

Changyong, F., Hongyue, W., Naiji, L., Tian, C., Hua, H., Ying, L. *et al. (2014). Log-transformation and its implications for data analysis*, Shanghai archives of psychiatry **26***(2): 105.*

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, ACM, pp. 785–794.

Chen, X.-w. and Wasikowski, M. (2008). Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems, ACM, pp. 124–132.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification, IEEE transactions on information theory **13***(1): 21–27.*

ElBahrawy, A., Alessandretti, L., Kandler, A., Pastor-Satorras, R. and Baronchelli, A. (2017). Evolutionary dynamics of the cryptocurrency market, Royal Society open science **4***(11): 170623.*

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J. (2008). Liblinear: A library for large linear classification, Journal of machine learning research **9***(Aug): 1871–1874.*

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, AI magazine **17***(3): 37.*

Friedman, J., Hastie, T. and Tibshirani, R. (2001). The elements of statistical learning, *Vol. 1, Springer series in statistics New York, NY, USA.*

Greaves, A. and Au, B. (2015). Using the bitcoin transaction graph to predict the price of bitcoin. Published on snap.stanford.edu.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, Journal of machine learning research **3***(Mar): 1157–1182.*

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An introduction to statistical learning, *Vol. 112, Springer.*

Jang, H. and Lee, J. (2018). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information, IEEE Access **6***: 5427–5437.*

Jiang, Z. and Liang, J. (2017). Cryptocurrency portfolio management with deep reinforcement learning, pp. 905–913.

Johnson, S. C. (1967). Hierarchical clustering schemes, Psychometrika **32***(3): 241–254.*

Kazan, E., Tan, C.-W. and Lim, E. T. (2015). Value creation in cryptocurrency networks: Towards a taxonomy of digital business models for bitcoin companies.

Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering, International Journal **1***(6): 90–95.*

Kokoska, S. and Zwillinger, D. (1999). CRC standard probability and statistics tables and formulae, *Crc Press.*

Kotsiantis, S. B., Zaharakis, I. and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques, Emerging artificial intelligence applications in computer engineering **160***: 3–24.*

Lessmann, S. (2004). Solving imbalanced classification problems with support vector machines., Vol. 4, pp. 214–220.

Likas, A., Vlassis, N. and Verbeek, J. J. (2003). The global k-means clustering algorithm, Pattern recognition **36***(2): 451–461.*

Linden, A. and Fenn, J. (2003). Understanding gartner's hype cycles, Strategic Analysis Report R-20-1971. Gartner, Inc .

López, V., Fernández, A., Moreno-Torres, J. G. and Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics, Expert Systems with Applications **39***(7): 6585–6608.*

Mougayar, W. (2016). The business blockchain: promise, practice, and application of the next Internet technology, *John Wiley & Sons.*

Mueller, Glarner, L. M. F. G. H. (2018). Conceptual framework for legal and risk assessment of crypto tokens. Published on www.mme.ch.

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms, arXiv preprint arXiv:1109.2378 .

Murphy, J. J. (1999). Technical analysis of the financial markets: A comprehensive guide to trading methods and applications, *Penguin.*

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Published on www.bitcoin.org.

Phillips, R. C. and Gorse, D. (2017). Predicting cryptocurrency price bubbles using social media data and epidemic modelling, pp. 1–7.

Poyser, O. (2017). Exploring the determinants of bitcoin's price: an application of bayesian structural time series, arXiv preprint arXiv:1706.01437 .

Raschka, S. and Mirjalili, V. (2017). Python machine learning, *Packt Publishing Ltd.*

Saad, M. and Mohaisen, A. (2018). Towards characterizing blockchain-based cryptocurrencies for highly-accurate predictions, p. 704709.

Sovbetov, Y. (2018). Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero, Journal of Economics and Financial Analysis **4***(2): 001027.*

Swan, M. (2015). Blockchain, *O'Reilly Media, Inc.*

Talavera, L. (1999). Feature selection as a preprocessing step for hierarchical clustering, Vol. 99, Citeseer, pp. 389–397.

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review, Journal of diagnostic medical sonography **6***(1): 35–39.*

Teng, C.-M. (1999). Correcting noisy data., Citeseer, pp. 239–248.

Tuv, E., Borisov, A., Runger, G. and Torkkola, K. (2009). Feature selection with ensembles, artificial variables, and redundancy elimination, Journal of Machine Learning Research **10***(Jul): 1341–1366.*

Walters-Williams, J. and Li, Y. (2010). Comparative study of distance functions for nearest neighbors, Springer, pp. 79–84.

Wilder, J. W. (1978). New concepts in technical trading systems*, Trend Research.*

Wilks, D. S. (2011). Cluster analysis, Vol. 100, Elsevier, pp. 603–616.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining, Citeseer, pp. 29–39.

Wold, S., Esbensen, K. and Geladi, P. (1987). Principal component analysis, Chemometrics and intelligent laboratory systems **2***(1-3): 37–52.*

Wu, K., Wheatley, S. and Sornette, D. (2018). Classification of cryptocurrency coins and tokens by the dynamics of their market capitalizations, Royal Society Open Science **5***(9): 180381.*

Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy, Journal of machine learning research **5***(Oct): 1205–1224.*

Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning, Synthesis lectures on artificial intelligence and machine learning **3***(1): 1–130.*

Appendix A

Token Classification Frameworks

A.1 International Token Classification, BC Frankfurt

Fig. A.1: International Token Classification
A flexible and extendable framework for the classification of all kinds of cryptographic tokens. *Source: itsa.global*

A.2 Token Classification Framework, Untitled Inc

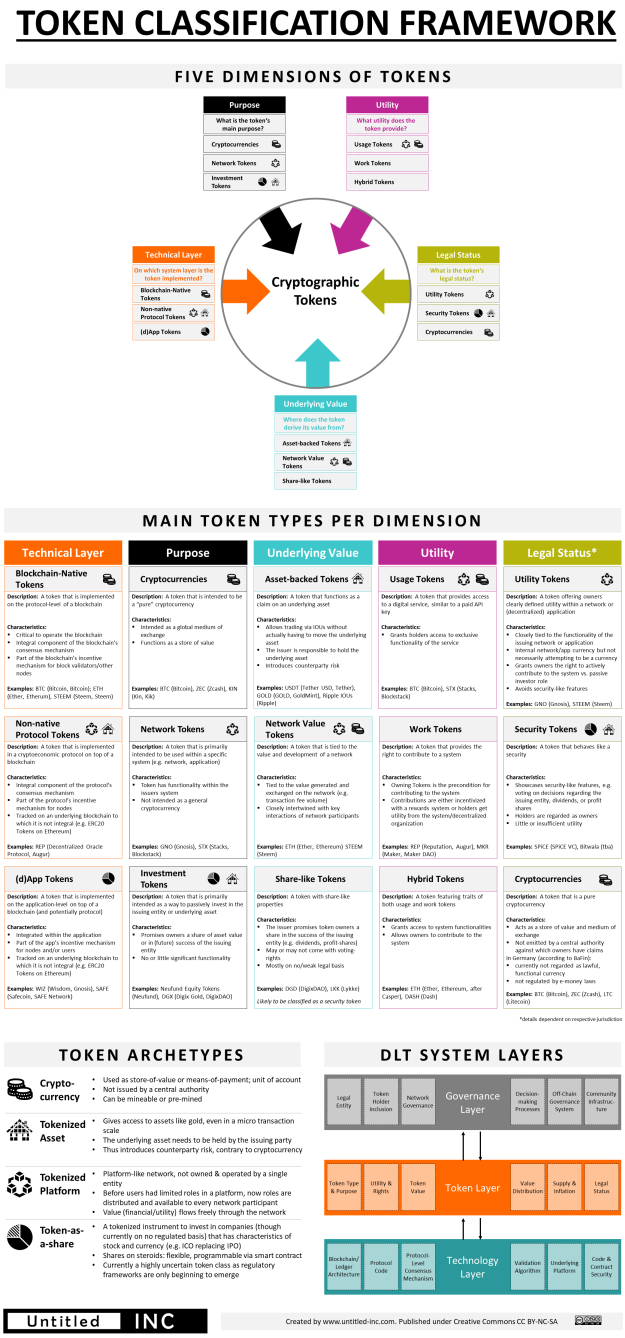

Fig. A.2: Token Classification Framework
A multi-dimensional tool for understanding and classifying crypto tokens.
Source: www.untitled-inc.com

A.3 Functional BCP Classification, MME

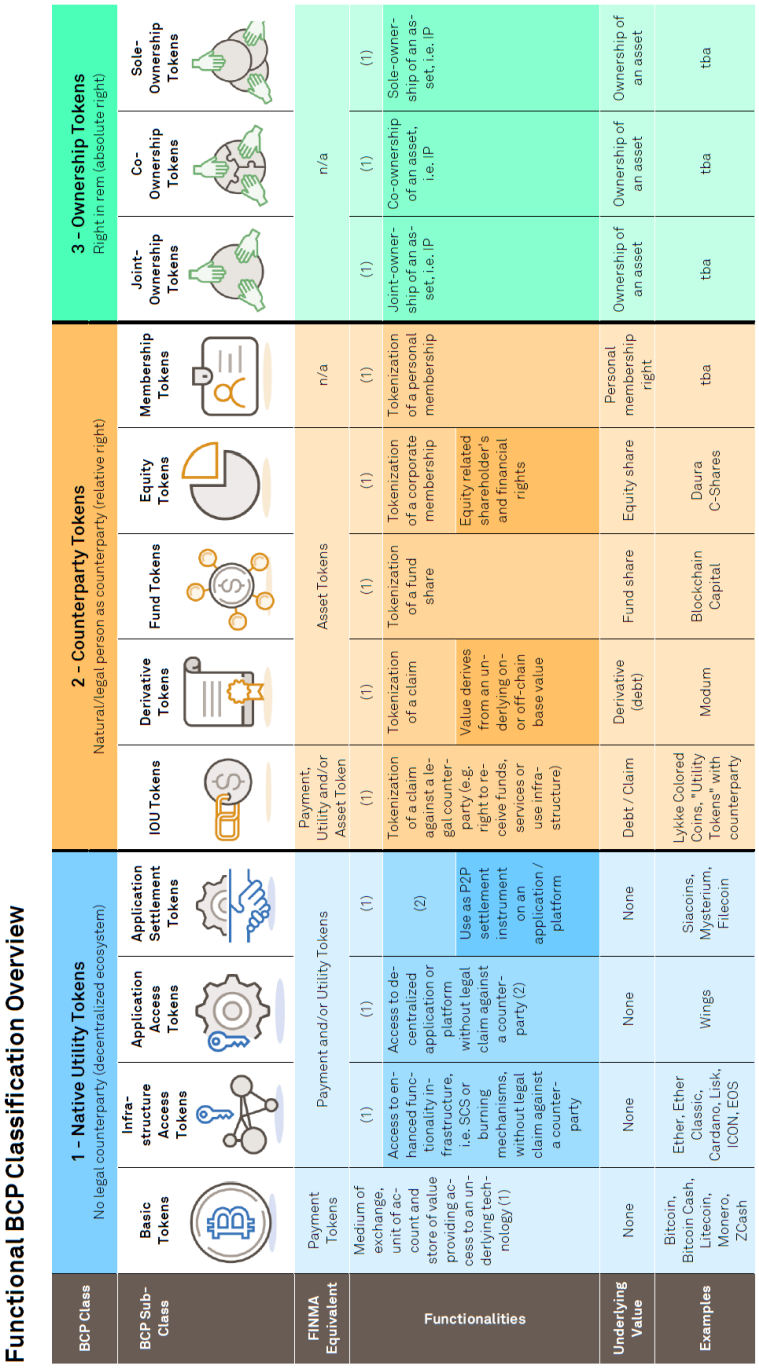

Fig. A.3: Functional Blockchain Crypto Property Classification.
Source: www.mme.ch

Appendix B

Exploratory Data Analysis

B.1 Boxplots

Fig. B.1: Unscaled Data Boxplots
Plots for 34 numeric features including 774 tokens.

Fig. B.2: Scaled Data Boxplots
Plots for 34 numeric features including 774 tokens.

Fig. B.3: Log-Transformed Data Boxplots
Plots for 34 numeric features including 774 tokens.

B.2 Token Distribution

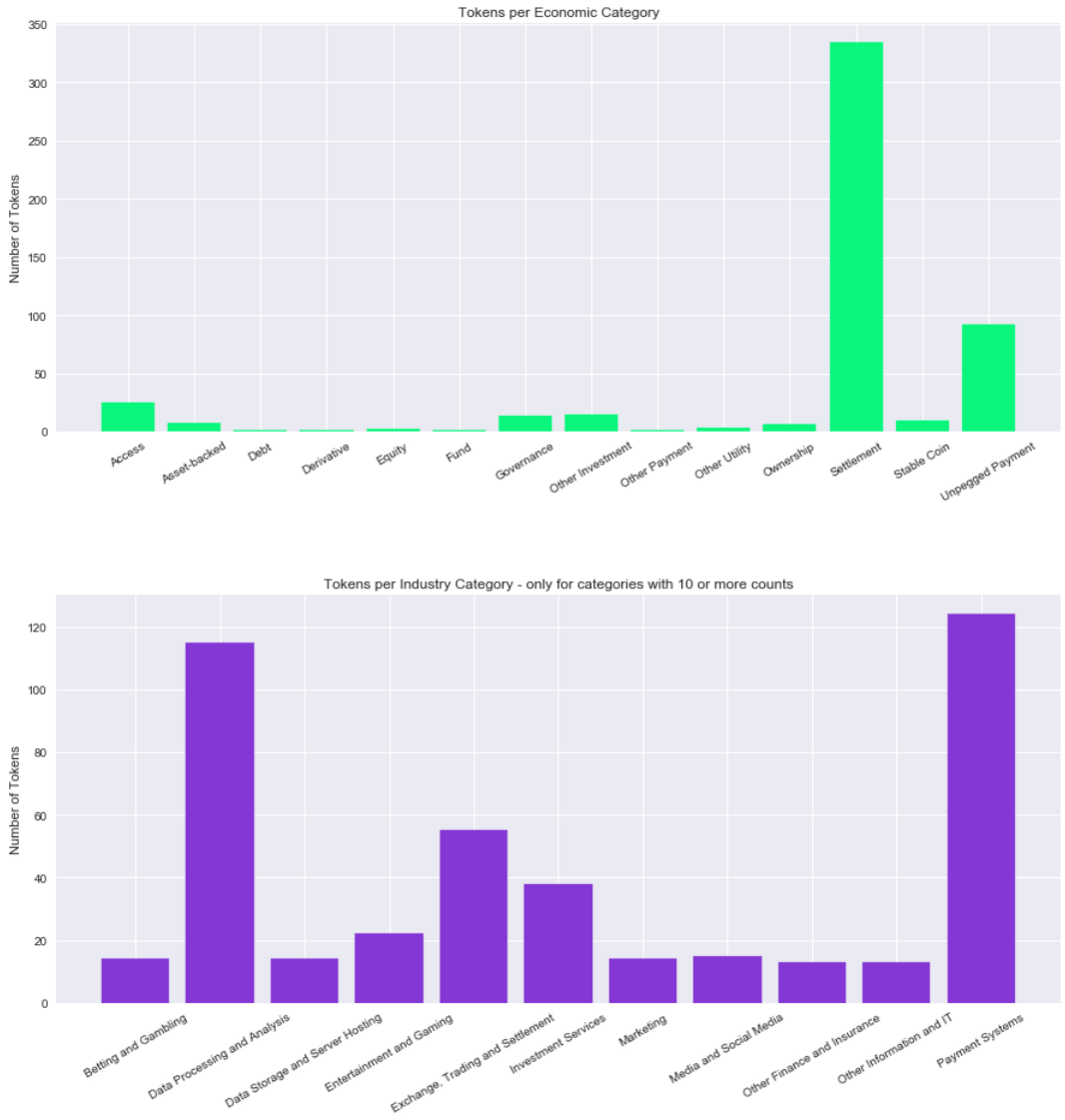

Fig. B.4: Number of Tokens per Secondary Labels Economic and Industry

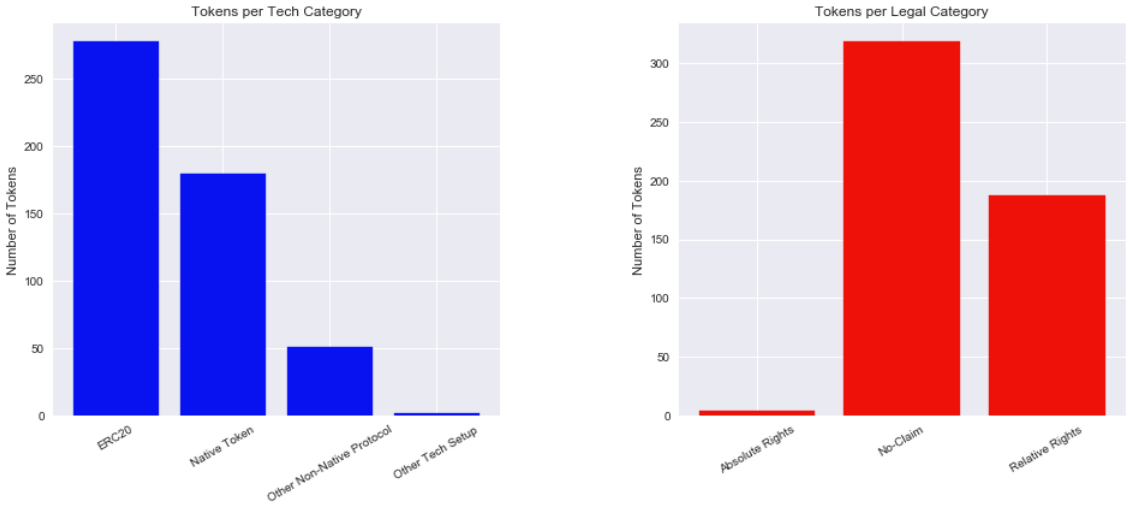

Fig. B.5: Number of Tokens per Secondary Labels Tech and Legal

Appendix C

Variable Importance and Feature Selection

C.1 Subset Selection

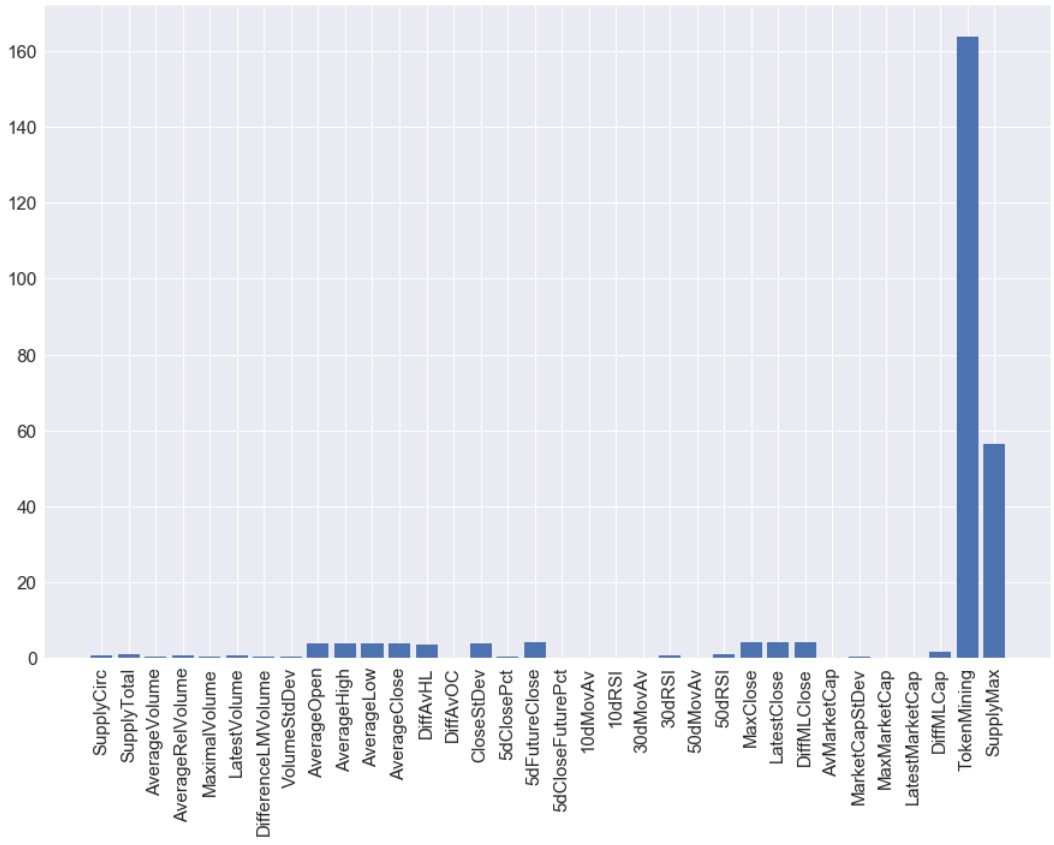

Fig. C.1: Best Subset Selection Chi Squared Test Primary Economic Label

C.2 Shrinkage

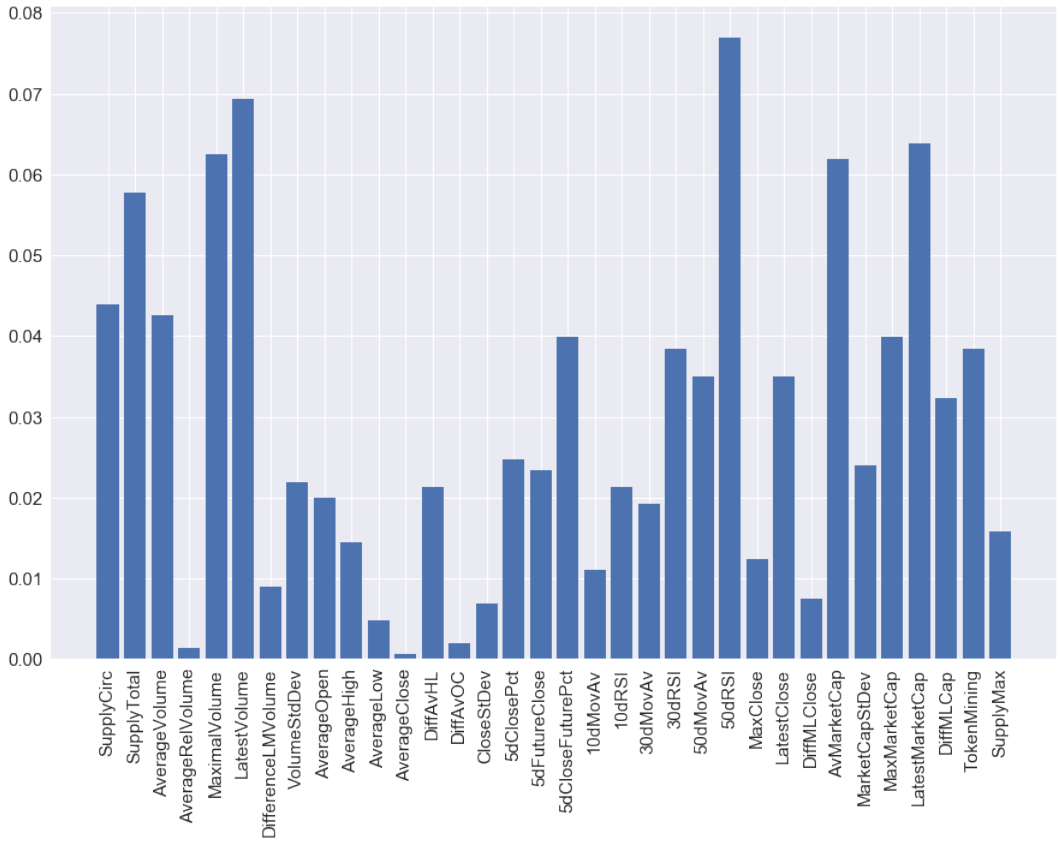

Fig. C.2: XGBoost Feature Importance Primary Tech Label

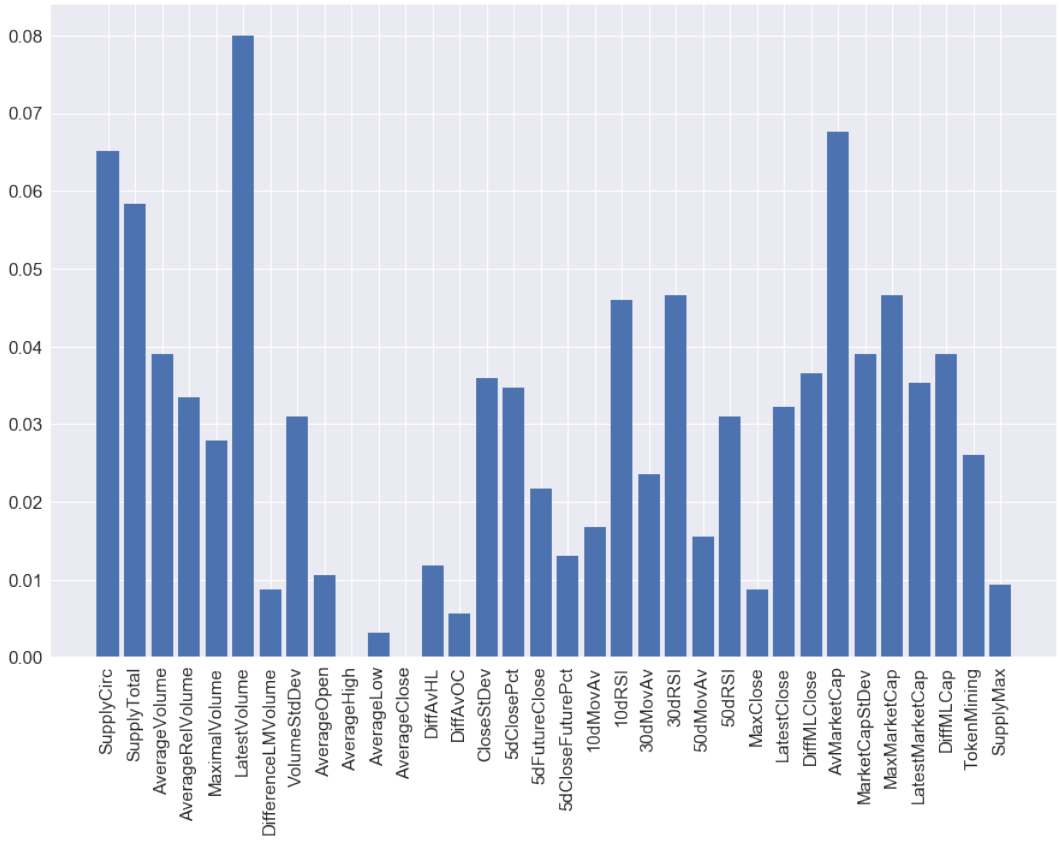

Fig. C.3: XGBoost Feature Importance Primary Legal Label

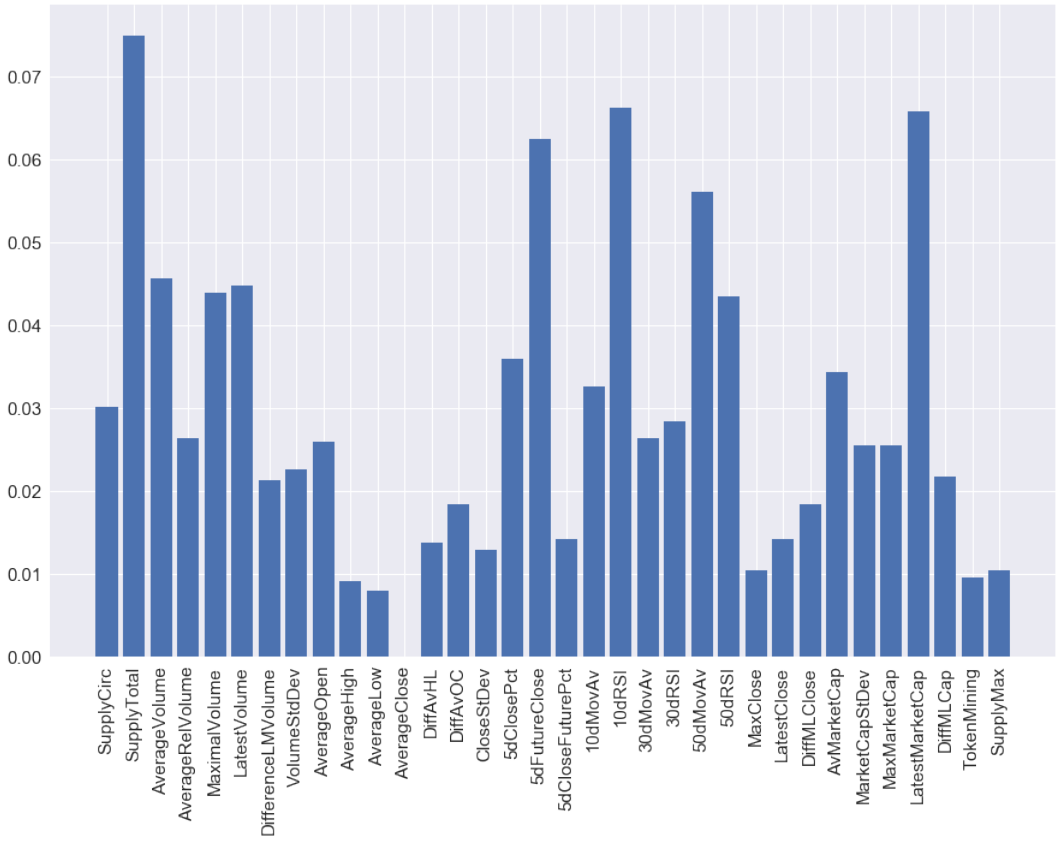

Fig. C.4: XGBoost Feature Importance Primary Industry Label

C.3 Dimension Reduction

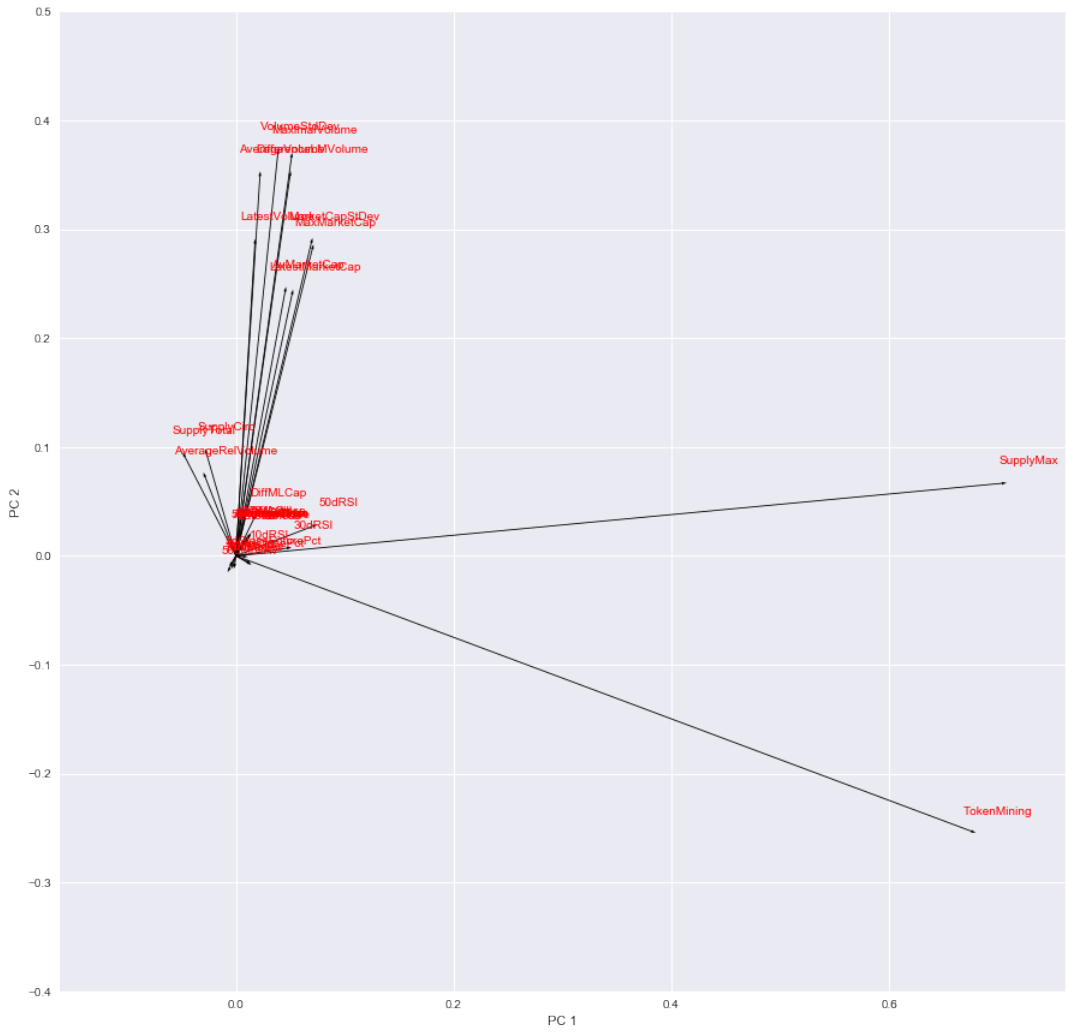

Fig. C.5: 2-D Principal Components Feature Directions

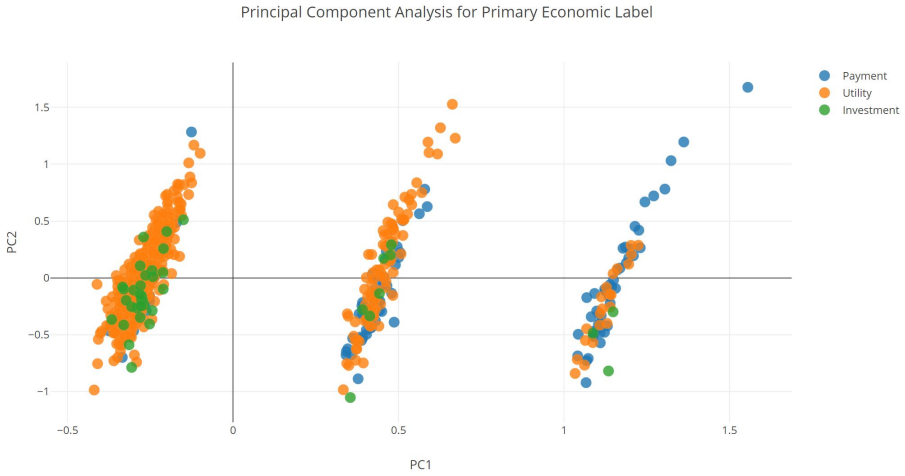

Fig. C.6: 2-D Principal Components for Primary Economic Labels

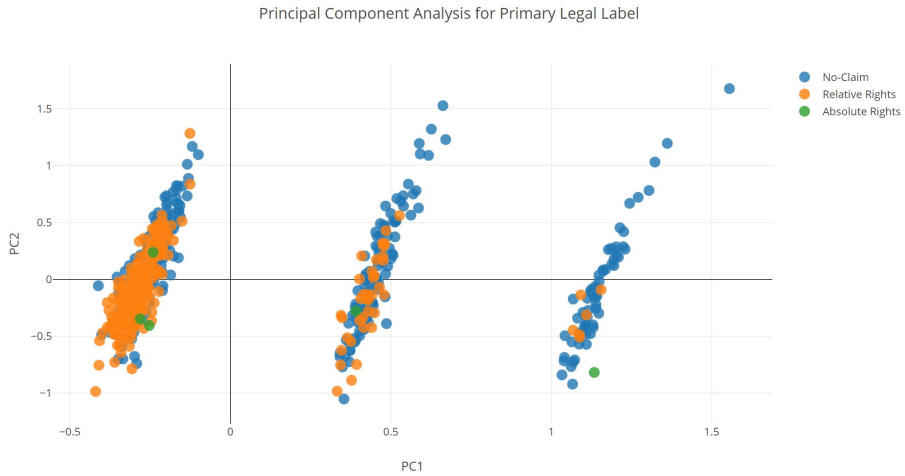

Fig. C.7: 2-D Principal Components for Primary Legal Labels

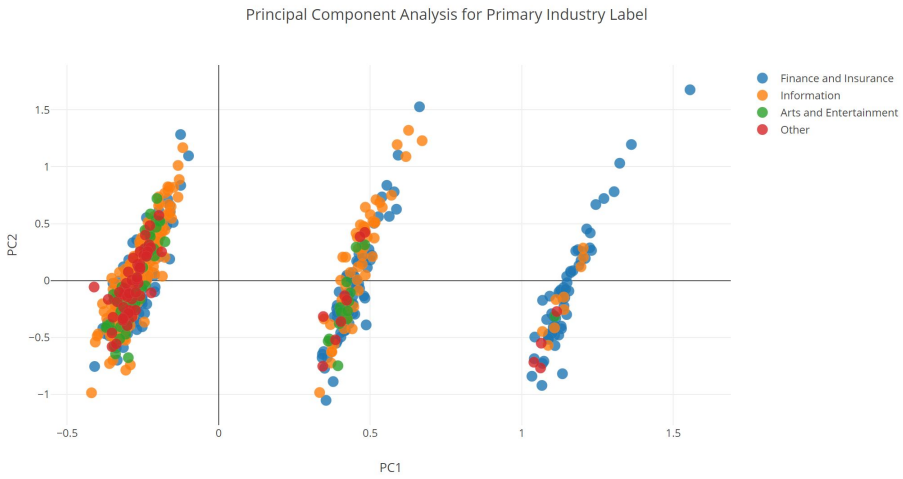

Fig. C.8: 2-D Principal Components for Primary Industry Labels

Appendix D

Supervised Classification

D.1 Classification Techniques

	Decision Trees	Neural Networks	Naïve Bayes	kNN	SVM	Rule-learners
Accuracy in general	**	***	*	**	****	**
Speed of learning with respect to number of attributes and the number of instances	***	*	****	****	*	**
Speed of classification	****	****	****	*	****	****
Tolerance to missing values	***	*	****	*	**	**
Tolerance to irrelevant attributes	***	*	**	**	****	**
Tolerance to redundant attributes	**	**	*	**	***	**
Tolerance to highly interdependent attributes (e.g. parity problems)	**	***	*	*	***	**
Dealing with discrete/binary/continuous attributes	****	***(not discrete)	***(not continuous)	***(not directly discrete)	**(not discrete)	***(not directly continuous)
Tolerance to noise	**	**	***	*	**	*
Dealing with danger of overfitting	**	*	***	***	**	**
Attempts for incremental learning	**	***	****	****	**	*
Explanation ability/transparency of knowledge/classifications	****	*	****	**	*	****
Model parameter handling	***	*	****	***	*	***

Table 4. Comparing learning algorithms (**** stars represent the best and * star the worst performance)

Fig. D.1: Evaluation of Supervised Learning Techniques for Classification by Kotsiantis *et al.* (2007)

D.2 Prediction Accuracies

Test Set Accuracy	KNN	Decision Tree	Bagging	Random Forests	XGBoost	Support Vector Machines	Neural Networks
XGBoost	73.9%	73.2%	73.2%	**74.5%**	67.3%	72.5%	56.9%
PCA	**75.2%**	66.7%	72.5%	74.5%	67.3%	72.5%	73.2%
All Features	**73.9%**	73.2%	**73.9%**	73.2%	68.6%	72.5%	72.5%

Tab. D.1: Legal Label Prediction Accuracies based on Model Feature Selection

Test Set Accuracy	KNN	Decision Tree	Bagging	Random Forests	XGBoost	Support Vector Machines	Neural Networks
XGBoost	46.4%	49.0%	50.3%	**54.2%**	51.6%	51.0%	49.0%
PCA	48.4%	47.1%	**50.3%**	47.7%	46.4%	49.0%	46.4%
All Features	46.4%	51.6%	51.0%	**55.6%**	50.3%	51.0%	48.4%

Tab. D.2: Industry Label Prediction Accuracies based on Model Feature Selection

Appendix E

Unsupervised Clustering

E.1 Hierarchical Clustering

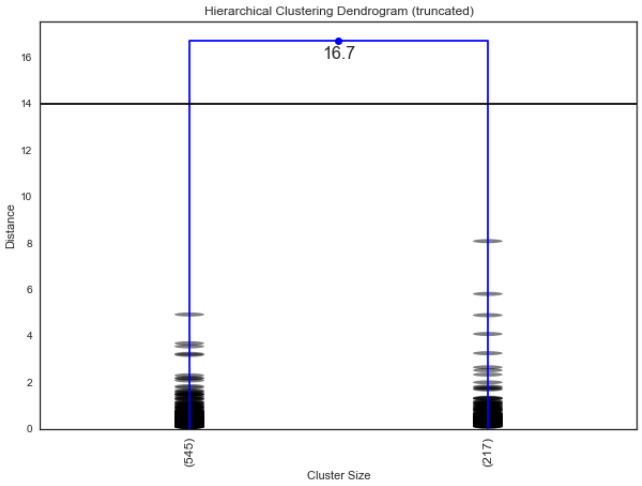

Fig. E.1: Hierarchical Clustering Dendrogram K=2

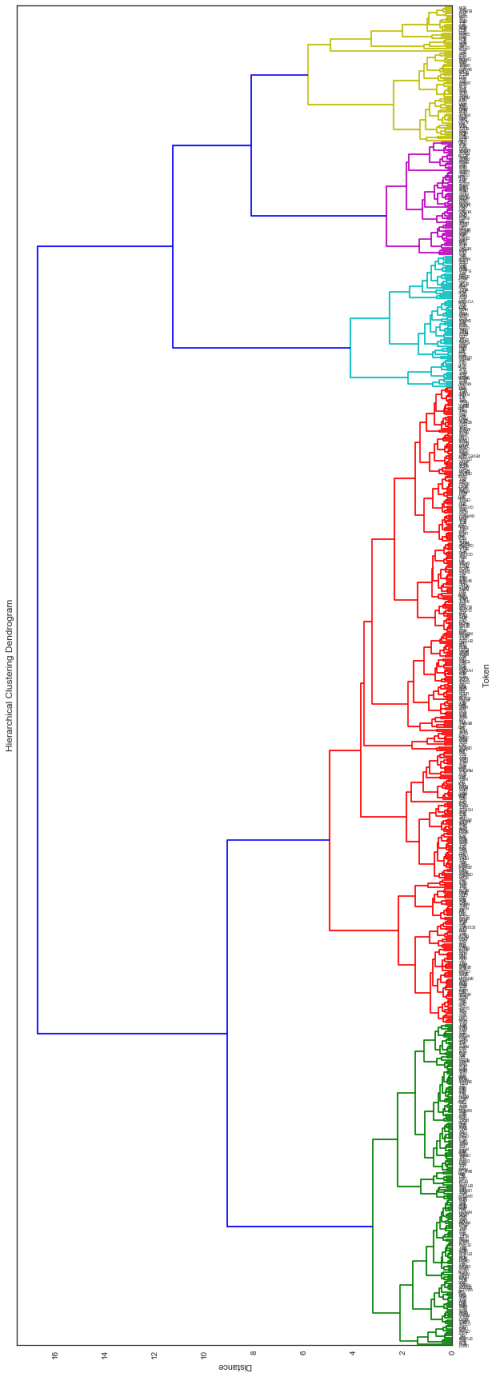

Fig. E.2: Complete Hierarchical Clustering Dendrogram K=5

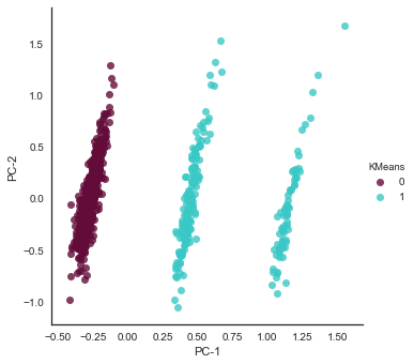

Fig. E.3: Hierarchical Clustering Scatterplot based on First Two Principal Components K=2

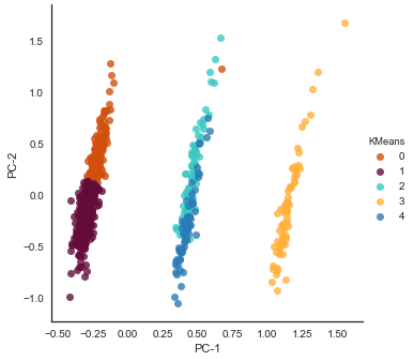

Fig. E.4: Hierarchical Clustering Scatterplot based on First Two Principal Components K=5