
## Universidad Politécnica de Madrid

### Escuela Técnica Superior de Ingenieros Informáticos


Master in Digital Innovation

Master Thesis

# Crypto Token Price Prediction Based on Fundamental Blockchain Data


Author: **Ostap Kharysh**

Supervisors: **Mike Salo, Ihor Pidruchnyy**


Madrid, August 2022

This Master Thesis has been deposited in ETSI Informáticos de la Universidad Politécnica de Madrid.

*Master Thesis*
*Master in* Digital Innovation

*Title:* Cryptotoken Price Prediction Based on Fundamental Blockchain Data

August, 2022

*Author:* Ostap Kharysh

*Supervisor:*

Ihor Pirduchnyy

CEO

Applicature Inc.

*Co-supervisor:*

Mike Salo

Assistant Professor

UCU Business School
Ukrainian Catholic University

# Abstract

The blockchain industry is developing fast already shifting the paradigms of traditional economies introducing the new standards in business and society. Even though, the industry is novel it is not shifting the major aim of any business project, which is revenue generation. There are already communities of entrepreneurs and investors who largely dedicate their time and finances to bring up revolutionary products on the blockchain market. In blockchain world almost all the companies valuations depend on the price of the token issued by those companies. Even though, there are some advances in crypto token prediction approaches this field is relatively new and attracts a significant commercial interest due to the volatility of such assets. With our research we utilize the fundamental (blockchain) database information to predict the token price for the crypto advisory service. During the data processing part, we prove that there is a need for individual on-chain feature selection approach to predict the project token price. Based on the prediction results we believe that the neural network models are better suited to become advisory models to the digital finance investors as they demonstrate a higher resilience to token price spikes than the tree-based models.

# Acknowledgement

First of all, I am grateful to the Armed Forces of Ukraine that made it possible for me to complete my academic degree keeping my Motherland and the rest of free Europe safe against the russian aggression.

I would like to thank the EIT Digital for the opportunity to take part in a breathtaking academic journey towards the academic title of Master of Science. I am grateful to the Technical University of Madrid and the University of Rennes 1 for the comprehensive introduction to the world of Data Science, possibility to tackle the real-life business cases with the knowledge weaponry you supplied me during the 2 years of my education.

I am honoured to dedicate special thanks to my university colleague and now friend Lorenzo Framba with whom we managed to develop astonishing academic projects, to learn to cooperate and play as a team and support each other along the way of not only academic but also life challenges. There is still a lot waiting ahead of us in the future and I hope our ways will intersect again soon.

I want to express my warm gratitude to the whole Applicature team for forming a welcoming environment encouraging to deliver the highest results that ended up in a considerable value of deliverables that could be now utilized in commercial activities.

Finally, I am grateful to my family and friends who were always supporting and available to cheer me up in the time when I needed it.

# Contents

# List of Tables



# List of Figures

# 1 Introduction

## 1.1 Internship host organization

Applicature [1] - is a US-based service partner that provides a full service of IT service development, digital finance, marketing, funding, relationship management and advisory services to accelerate the digital finance projects growth. The company is in operation since 2017 engaging in different ranges of activities including MVP development, digital assets economy, launch strategy, security audits and introduction to the large community of business launchpads, advisors and investors.

Although, the company offers a broad range of services the focus of the engagement is the nature of distributed ledger transparency and possibility to exploit it as a fundamental data source for the token price predictions.

## 1.2 Structure

This report is build with several Sections Introduction 1 that provides an overview of birth of blockchain technology, formation of crypto economy, and utility of crypto project currency or token. The following section is dedicated to introduction of Token Economy 2, impact of demand and supply theory on crypto projects along with the launch approaches. In the Section of Motivation and Goal 3 we explain the content of the internship engagement and highlight the research goals. In the Related Works 4 we introduce the research activities and results of related studies. The part of Data Description and Feature Selection 5 includes the data source exploration, token selection, token price predictive power and feature selection. The Predictive Modelling 6 introduces the prediction models with their broad description and provides the result of token predictions. The document finishes with Conclusions 7 where we elaborate on the both parts of the internship engagements and discuss possible improvements.

## 1.3 The birth and expansion of blockchain technology

The history of blockchain starts over with introduction of Bitcoin in 2009 during the periods of economic recession which gave start to revolution in economic transactions activity. The so called "peer-to-peer cash system" drew attention of the investment community willing to open new financial opportunities and hedge against losses it experienced during the financial crisis of 2007 - 2008 years. The technology presented in the Bitcoin paper [1] identified the weakness of trust for the evolving amount of online financial transactions that were conducted through the centralized bank system. Satoshi Nakamoto introduced an idea of transferring trust handling from bank system to the complex cryptographic peer-to-peer system removing bank intermediation. The database inside the proposed system is called the chain of digital signatures or ledger. The system consists of users that conduct financial transactions iteratively constructing a public chain of transaction blocks with links from the previous to the next one. The mechanism that validates the transactions there is called "proof-of-work" which is a decentralized consensus mechanism that provides a complex hash computation for transaction verification. To conduct any miscellaneous hacks one needs to redo a completed block of transactions and all the previous once from the very first transactions. As the amount of the building blocks of the network is only increasing the difficulty to hack it is only rising.

Finally, the solution claimed in [1] was supported in academic reports as revolution in reliability for existing financial systems that avoids the need for payer and receiver centralized verification process [2], transaction system with immutable transaction blocks over the network

[1] https://applicature.com/

that eliminates the system hacking. Later on the benefits the decentralized system of transactions were expanded even broader to the a huge number of business problems creating the term *Crypto economy*.

## 1.4 Crypto economy versus Traditional economy

**Crypto economy** The injection of blockchain currency or cryptocurrency into the traditional business creates a new shift from the traditional economy to the crypto economy. The construction process imitates the tradition economies with a slight change in the paradigms. First of all, the economic agents that create value in the traditional world such as households, individuals, financial corporations, banks are replaced with crypto entities. In such coin-alike solutions such roles are given to blockchain miners or stake validators that ensure the trustless validation environment, investors and decentralized autonomous organizations (DAOs). DAO is decentralized autonomous organization which is a novel form of decentralized governance that offers a flat hierarchy of power and responsibilities discouraging centralization. The rules are encoded inside the system and cannot be broken. The initiation of such an economic structure also requires shifting from the centrally controlled governance of intermediary of exchange and provokes the initiation of a decentralized digital currency which is called cryptocurrency, coin or project token. In contrary to the traditional economy where the productive assets are factories, software or any machinery, the crypto economies are guided by the smart contracts as they play the role of major ecosystem rule builders that enable the interaction of any actors involved in the environment. It is also intriguing that goods in crypto economies cannot be any product in the traditional world, like cars, food, or clothes. The crypto world holds other tokens as goods. Those are called non-fungible tokens or NFTs. These unique tokens can be also represented as digital art and hold their value as art or as an enabler of some services, like access to the events, services, or simply lottery tickets.

**Monetary exchange** Now we come to the question how the exchange of cryptocurrency differentiates from the traditional "fiat" currency. In the traditional world, the exchange mechanism we are comfortable with uses a bank-regulated currency to pay either with banknotes or with the centralized digital payment system of the bank for any goods and services. In the case of the crypto economy, the intermediary of exchange mechanism is introduced as cryptocurrency. As it was previously mentioned, it has no connection to any central authority that can influence the price and needs to approve each transaction conducted from each payment account. On the contrary, here, the users can create a decentralized wallet connected to the blockchain and hold their cryptocurrency and assets on it. Each transaction is executable in case the applied rules are complied with. And those are simple: if the wallet contains enough value to conduct a transaction and pay the fees for transaction verification on the chain, it will be executed. As long as the transaction is approved, it will appear in the destination wallet after successful completion. Removal of centralized authority to supervise your activity as there is no central bank or any government in control results in freedom of financial activities. The business of the contemporary world sees such characteristics of the token economy as the possibility to develop projects not bounded by the regulations associated with the traditional economy. The evidence of such adoption is getting more an more noticeable day by the day. According to [3] it could be inferred that the amount of crypto coins created since 2015 till the beginning of 2022 is 9367, making a significant increase of 17,6 times over 5 years from 562 in 2015 till 9929 in Jan 2022.

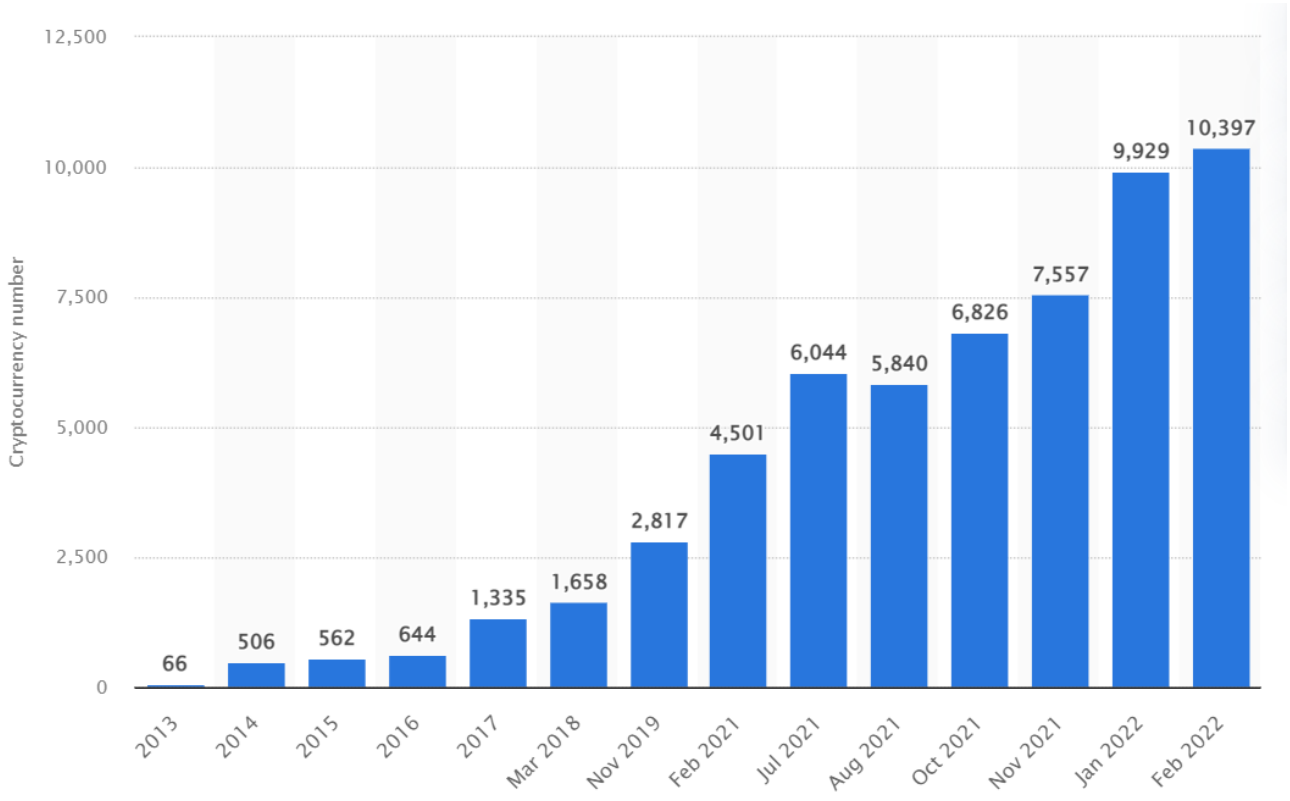



Figure 1: Number of cryptocurrencies worldwide from 2013 to February 2022, Statista

### 1.4.1 Token utility

To expand further how the crypto tokens can be used in the business projects we refer to the paper of [4] where it is explained which utility each token can hold. Normally, in general, each token can be used as currency that transmits some priced value. It could also be expanded to representation of asset ownership where token is seen as an project share that enables a % of profit claim from the business revenue. Apart from being a unit of account the token can also be used as a tool that enables governance rights for the holders, meaning that each holder can make the input in decisions on how the business should develop or which features should the project include. It is also commonly used practice to incentivize users for holding tokens like providing an early assess to the platform or rewarding for holding the project tokens. All in all, as this field is relatively new the ideas of how to use the crypto tokens are constantly appearing modifying the traditional assumptions of how the project economy can work. We would like to show you a couple of examples of business projects that introduced token in their ecosystem:

- MilestoneBased - a crypto project designed to improve the startup investment model. Consequenlty, the project is focused on investors and startups as its community. Each startup that is registered in MilesoneBased agrees with investors upon the funding for milestone towards a complete roadmap of development. This means, that after each finished milestone the startup has to prove to the project community and investors that the agreed milestone is successfully completed, thus request for the next funding. With the help of token that grants the governance ability the project community can approve the success of the startup milestones and issue the next level of investment. Apart from governance utility the token can be used as an intermediary of exchange for other crypto tokens.

- TalkAboat - is a project that aims to revolutionize the sphere of audio podcasting. It introduces the ecosystem token where podcast listeners can pay their subscription and content creators receive fees. Apart from payment activities, the users are incentivized to be active on the platform (leaving comments, likes, etc.) with a form of platform token rewards.

- StepN - is an example of the projects with 2 tokens in the ecosystem. The idea of the project is to enable users to earn from running in the real world. One of the token is an earning currency that the user earns from running. Another one is used for in platform activities and like increasing the running effectiveness with purchasing and improving project sneakers. Both tokens can be exchanged for other crypto tokens.

# 2 Crypto Economy and Market

## 2.1 Allocation

To provide a broader understanding of the crypto token economy, it could be viewed as an financial architecture of the project developed with injection of blockchain utility token that possess a financial value. Based on it the project explains what is a total amount of tokens available and also sorts out specific amount of tokens for different business needs. This notion is called token allocation. There is no clear rule which allocations should be included for which purpose. Normally, one could find an allocation for team, founders, and advisors as a share of tokens that they could use to profit from the project. Also, there are expense allocations like those dedicated to development, marketing, operational activities that the project plans to use for building the project. Commonly, the project also includes allocations for investors. There could be couple of them depending on the sales stage of the investment, like seed, private sale, public sale, etc. The investment is limited to the amount of tokens available to be purchased at the predefined price per token. For example, in the mentioned previously project StepN there is only private sale investment allocation of 16.3% of total amount of tokens, which equals to 6 billion multiplied by 16.3% = 978 million $. In the given project the presale price is not disclosed. This number is normally derived from the company valuation at the presale stage. As the price was not given, to provide an example, let's assume that the total company valuation at presale stage was 6 million $. Given the total amount of tokens (6 billion) we could derive that the price per token at presale was set to 0.001$ per token. Hence, the total amount of acquired investments at presale was 0.001$ multiplied by 978 millions = 978,000$. The needs and logic for acquiring investments resembles the needs of traditional economy series of investments.

As it is explained in [5] the token economy could be viewed as not only economy of the project but also management system that is able to impact the user behavior by providing incetivization and rewarding in the form of receiving the token. For example, StepN 30% of total token supply to the "Move ot Earn" allocation to encourage users activity on the platform.

## 2.2 Demand and Supply

Based on the description in the previous Subsection 2.1 one can have an overview how the token economy is formed, what is an aim of the token allocations and formation of token price for investors. For the sake of keeping focus on the needs of this research we omit some of the facts that are not crucial for the token economy understanding in the scope of this research and dedicate some attention to the market price and how it is formed in crypto world. As on any market to maintain a value an asset needs to have demand and supply that regulates the price. This corresponds to the Market Equilibrium theory that claims: "The interaction

between consumers" demand curve and firms'supply curve determines the market price and quantity of a good or service that is bought and sold" [6]. Applying to the crypto world, every project spends significant amount of resources on marketing activities to attract investors and increase the demand for the project token in the community. Of course, if the project like StepN releases all the tokens straight away the demand will not be able to cover all the supply, hence, the price for token will fall. Same way, if the amount tokens issued in public use is low the price will increase and make the service inaccessible for the users as token price rises to unreasonably high level. As most of the crypto projects define their valuation by the initial crypto price they aim for issuing the amount of tokens enough for the token market price to resemble the market value of the company or exceed it. The mechanism that is used to bound the release of tokens is called vesting and is derived from the [7] option vesting mechanism of traditional finance. As a part of token economy development the project specifies a vesting period for each of the allocations so that the release of tokens is conducted smoothly corresponding to the demand for token, this way mitigating the risk of loosing its value due to high immediate supply and insufficient demand. For example, based on StepN one can find that the vesting period for presale investors starts in January 2023 after the launch in March 2022 which is also called cliff (period of without vesting) and than gets vested ruffly 3% per month until all 100% of the presale allocation is vested. This means that the private investors will receive the total tokens amount they purchased after the launch with a 11 month waiting period with no claim and starting from January 2023 until January 2026 with monthly claims. Likewise, the other allocations in the StepN token economy are also vested for a specific period of time. This way the project ensures specified amount of tokens issued in circulation and can adopt their marketing efforts to keep the demand level relative to the supply of tokens.

It is important to mention that each token economy and vesting model includes a financial projection of development expenses needed to be covered and revenue estimations the project aims to achieve. Furthermore, keeping community building, growth of buyers and token holders is a crucial to successfully build and maintain the token price of the project. Only in this case, the project can receive enough funds from selling tokens of allocations dedicated to development needs.

## 2.3 Token launch

Having discussed the token economy design along with allocation and vesting to maintain demand and supply we provide a short overview of the existing possibilities to launch the project token. The official launch happens on TGE or Token generation event, which is a technical generation of token on the blockchain-based network and publishing the token on the exchange. There are 2 ways to list the project token on exchange which is listing on CEX - centralized exchange or listing on DEX - decentralized exchange. This paper [8] explains the difference between the mentioned approaches. In scope of this thesis it is important to draw your attention that CEX use the approach of limit order books where personas of buyer and seller are matched by price. The price on CEX is changed according to the difference of buy-sell orders of the user. On contrary to the traditional exchange option there is an emerging DEX adoption for the token launch among crypto projects. The reason for it is flexibility. Anyone, no matter who it is can publish the token quote. In comparison to CEX the pricing mechanism is satisfied by forming a liquidity pool. This pool is a combination of project token and any other token, but often a stable coin - token equivalent to one of the national currencies like dollar, euro, etc. Let's assume StepN decides to launch the token at a price of 0.001$ per token. In order to set such launch price the project has to submit the appropriate amount of tokens and stablecoin equivalent. In our case the liquidity pool has to supply a proportion of 1000 tokens to 1 the USD-equivalent stable coin. After the pool is created anyone is able to purchase and sell tokens

from it. Due to the decentralized nature of DEX, the price variability depends on the size of liquidity pool. If one supplies more tokens and stablecoin equivalents keeping the proportion unchanged, the lower is the price fluctuation. This is due to the Automated Market Maker formula that is described in [9]. The idea is as follows: after setting the the liquidity pool at its initial conditions anyone could buy and sell tokens. The purchase process looks like swapping stablecoins for the tokens at the price set at the liquidity pool. The more tokens are purchased the more stable coins and the less tokens remain in the liquidity pool and vice versa. After each successful transaction (token purchase or sale) the price in liquidity pool changes depending on the proportion or remained tokens and stablecoins. The larger is the proportional difference the huger is the price difference for the next buyer or seller of the token.

Based on the general overview of the token launch approaches one could notice that exchanges are dependent on the constant activities in the project ecosystem. The project care to always have a constant and a bit higher demand to buy and hold tokens, than to sell them. That is why the cryptoproject should always care for keeping the community engaged, as the absence of active community results more users willing to sell than to buy those tokens, reducing the price and making the whole business unprofitable and devalued.

To sum up, this Section discusses the birth of blockchain and provides a general overview on how the business project in the traditional economy differentiates from the one in crypto (token) economy. In addition we discussed crypto project decide on token utility, design the economy and launch on the exchanges. Based on the overview one could infer that there is a life-worth significance for cryptoprojects to develop their community and pay a close attention to all the possible activities conducted with the project token to ensure its price health for a continuous and profitable development.

# 3 Motivation and Goal

As we already mentioned in the previous Sections 1 and 2 the crypto industry is only starting to rise making innovation symbiosis between the traditional industries and blockchain. There is a need to receive a proper domain knowledge to address problems in the field of blockchain based project or projects that use tokens as a the part of their business model. That's why the internship engagement is split in 2 parts:

- First part is to understand the nature of crypto projects. The approach selected to receive such knowledge is to develop the due diligence scoring metrics for crypto project maturity evaluation. This part of the internship covers the needs for Applicature to improve the onboarding process for the project in digital finance and reduces time spent preparing the service offering.

- The second part is dedicated to exploring the predictive potential of blockchain data and strategy how to infer the token price with it. This part is the focus of our research because it addresses the needs to understand the price dependencies for different tokens in order improve the token economy service offering and to explore the potential of automated token price advisory services for the investors that don't want to suffer random token price undervaluation and free up their portfolio before such event occurs and, vice versa, add the tokens to portfolio before the token price strikes.

## 3.1 Research goal

As we previously described there is a high importance of crypto projects to pay attention to how the community behaves not only on the platform but also how often the token is interchanged

between users or any other entities. There are some analytical solutions available that allow to track the wallet transaction activities like IntoTheBlock. However, the existing services do not allow a transparent understanding of blockchain (on-chain) characteristics that includes (user wallet activity, transaction characteristics) and how those can influence the project token price and predict it. The investors who conduct the long-term investments in several crypto projects maintain a constant interest in their token portfolios value. That is being said, they are interested in understanding if the value of the asset can radically change rather than engage in rapid constant buy/sell activities. To feel in control of their decisions investors need to have a transparent model that could infer if there is a need to expect a change in token price and how big is that change. In addition we want to provide a transparent clarification based on what characteristics the future token price is inferred to leave space for their personal analysis before taking any actions. This idea, makes an input to the goal of this thesis: utilizing the benefits of blockchain transparency analyse the predictive power of on-chain characteristics and forecast the future token price.

**Regression Problem** As we mentioned above, the business need this research aims to fulfill consists of predicting the token price itself and not the rise or fall movement. As it is explained, the given investors crypto portfolio management problem lies in the risk reduction due to the tendency of token price to fluctuate much actively than the share price of the public company in the traditional economy world. The ability to predict the future price should help estimate if particular crypto asset becomes undervalued or overvalued to take the corresponding portfolio management action, which is selling or purchasing in advance due to the predicted future token price. Hence, the business problem we face is the regression problem.

To narrow down the problem we focus on the projects that use native tokens as cryptocurrencies and explore how the fundamental blockchain value can predict the future price of such assets. We will focus on answering the following questions:

- What observation period should be taken into account while prediction the future token price?
- Does each crypto token share the same on-chain explanatory variables with other crypto tokens?
- Do the prices of other tokens posses a high explanatory power when predicting the token price?
- Considering the tree-based and neural network models can there be identified a favourite approach to predict the token price?

In the next Sections we build our research to receive answers to the stated questions.

# 4 Related works

## 4.1 Data selection

In this subsection we look through a selected amount of research papers related to the the cryptocurrency predictions. Firs of all, as we consider the topic of utilizing transparency of blockchain data one of the key interest of this investigation we believe that the research paper of [10] a valuable starting point for the related works description. The paper dedicates significant attention to on-chain data as explanatory variable for the Ethereum price predictions,

which includes blockchain characteristics, addresses information, intensity and congestion of transactions, wallet information, etc. The research finalized that, on-chain metrics are useful supplementary tool to increase the accuracy of the crypto token price. In addition, in paper of [11] the interconnectivity of cryptocurrencies prices and their performance is analyzed. This research shows a relative importance of cryptocurrencies news and as the impact on price of the cryptoasset and emphasises that the daily prices of one currency impact the price of another. This discovery we consider as an important input for our research and that advises for including the exploration of prices of other currencies to our set of explanatory variables. Moreover, while designing the investment portfolio [12] emphasized on the importance of wide variety of blockchain information and its combination to achieve meaningful results in cryptocurrency return predictions.

### 4.2 Prediction approaches

Based on the general overview of predictive models used in the papers for blockchain assets predictions we encounter several approaches contributive to our investigation. The research of [13] utilizes already a several machine learning techniques to maximize the intraday trading returns. The authors implement a logistic regression as a benchmark model for experimenting with selection of neural networks: Feedforward neural network (FNN) as a basic neural network, Long short-term memory (LSTM) and Gated recurring unit networks (GRU), which are the Recurrent Neural Network (RNN) based models that prioritize the importance of the closest observation timestamps for predictions), tree-based models (Random Forest (RF) and Gradient Boosting Classifier (GBC)) to research the decision building techniques inspired by root-cause analysis. As a result each of the models received higher than 50% accuracy on predictions. The best performing models appeared to be the RNN based models and GBC performing at range 51%-56% accuracy. It is also common to use the power of Support Vector Machines (SVM), which as [12] and [14] states can perform well under high market volatility. The improved SVM in [15] intoducing the data segmentation modelling prior to predictive learning. It is also important to mention, that the research outcomes RNN-based models can be seen among the most popular and modifiable when it comes to the time-dependent predictions as toke price. With their work [16] the performance of GRU, LSTM and bidirectional LSTM was taken into account. The special case of bidirectional LSTM (bi-LSTM), which is application of past and future input data sequences into separate LSTMs. Nonetheless, the regular LSTM and GRU outperformed bi-LSTM. However, one can encounter the research outcomes that do not advocate for much of difference between the performance of linear, neural network and tree-based models like in [17].

Having mentioned a selection of research papers we want to build up the the analysis logic relying on the approaches and outcomes authors presented us with. However, we dedicate more time to explore the data and receive a statistically explanatory variables in combination with reference to the blockchain domain understanding. In our case the data will be selected and not treated as a blackbox input for predicting the target as it was treated in [12].

## 5 Data description and feature selection

### 5.1 Source of data selection

As we previously mentioned, the goal of the research is: utilizing the benefits of blockchain transparency analyse the predictive power of on-chain characteristics and estimate the future token price. In addition, we don't want to treat the predictive feature as black-box values that

just correlate with the token price like in [12]. On the other hand we want to have a statistical reasoning of predictors selected feature-price prediction.

Based on our data search we identified a selection of sources like Yahoo Finance!, Gold Price, Crypto APIs. Such resources are applicable when treating basic "hyped" information for predicting the token prices. Those contain basic market information and small selection of generalized on chain-information which is crucial for our research. So all the given data sources are discharged. Going deeper into data source selection we found CoinGecko, CryptoCompare and CoinMetrics which are well known data sources in the blockchain world and often used for speed-crypto trading. Among those we selected CoinMetrics as it specifically focuses on extensive daily on-chain information (over 400 metrics) with some market information which is a good fit for our research.

## 5.2 Token selection

CoinMetrics provides a straightforward way for downloading the daily data feeds in a form of excel files for a selected token from the moment of creation till these days. Most of the popular tokens can be found on this service. In our research we decided to focus the top 3 tokens based on their market capitalization (as of June 2022) on CoinMarketCap. Those are Bitcoin (BTC), Ethereum (ETH), and Binance Coin (BNB). Apart from the main interest of our study, we want to check whether there is a need to select the on-chain explanatory variables for each token individually or there is no difference and those can be equally well used for each of the tokens. Accordingly, we will use 7 more tokens (Cardano (ADA), Ripple (XRP), Doge (DOGE), Polkadot (DOT), Tron (TRX), Uniswap (UNI), Litecoin (LTC)) to explore how similar is the selection of explanatory variables to each of the crypto assets.

## 5.3 General description on projects behind Bitcoin (BTC), Ethereum (ETH) and Binance Coin (BNB)

Having selected the tokens we reason how those can reflect the nature of Bitcoin, Ethereum and Binance Coin.

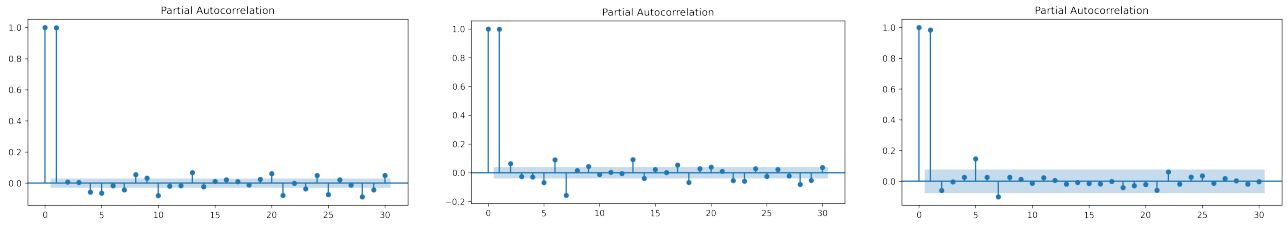
**Bitcoin** is not only a digital currency but also a first ever know blockchain. Its only utility is means of payment not pegged to any real asset. Blockchain of Bitcoin is not used as any building infrastructure for any other projects. The value of BTC is determined strictly by demand for it and the proof-of-work activity of validators [1] (the lower the price the lower amount of supply of validators and vice versa)

**Ethereum [18] and Binance coin [19]** are tokens of blockchains which are largerly used to upbuild crypto projects on them. Those projects can be referred to 2nd layer protocol projects. This means that such projects have internal structure driven by the 1st layer (Ethereum or Binance Smart Chain) respectively. The tokens of such chains are used to pay for transactions and posses an utility of architecture usage payment. Hence, such token utility uniqueness of BTC versus ETH and BNB can be considered as major philosophical difference in between those projects.

## 5.4 Asset price predictive power

Before we dive into feature selection we analyse the predictive power of token price as an explanatory variable to predict its future value. To explore the asset predictive power we

proceed with daily partial autocorrelation. For each price observations we take the previous day observations and apply Pearson partial correlation [20], which is a measure of linear tendency between the series of such observations. The difference between autocorrelation (ACF) and partial autocorrelation (PACF) that PACF considers the predictive impact of each time lag, individually, leaving only direct predictive effect discharging other indirect effects of other time lagged observations.




Bitcoin (BTC) Ethereum (ETH) Binance Coin (BNB)

Figure 2: Partial autocorrelation of the the top 3 tokens based on market capitalization

As a result of our analysis we plot the daily Partial autocorrelation plots for Bitcoin, Ethereum and Binance Coin (Figure 2). Based on our analysis one can see that there is a strong partial price correlation with the price of previous day. One could also observe that there are some statistically significant constant price observations of previous time lags for BTC and Ethereum that are decreasing with the time lag whereas the last significant observation for Binance Coin is 7 day lag.

Summing up our analysis, we found out that the previous price of the cryptocurrency possess an explantory power to predict the future price for Bitcoin, Ethereum, and Binance Coin. Consequently the token price has to candidate as explanatory variable for the future token price prediction.

## 5.5 Feature selection

Having selected 10 tokens we analyse the explanatory feature importance for each of the assets. With this analysis we decide whether we take into consideration the same explanatory variables for each token or there is a need to have individual selection for each of it. The on-chain data is solely numerical and is recorded in a daily manner. So, for each individual feature-to-price causality we apply Pearson correlation method [20], which is a measure of linear tendency between variables. The complete feature importance analysis consists of forming the groups of highly cross-correlated features (more than 0.5 Pearson coefficient threshold), than for each group we select the candidate which is the most correlated with the token price of the next day. Based on this analysis we identified that on average the price of the following day correlated with candidate 17.3 features (with standard deviation of 6.3) that are the least correlated between each other. In order to give a better understanding what features and how often they were selected as a result of our feature selection algorithm we visualize it in the Figure 3.

Here one can find 41 feature that are used to predict the token prices and their occurrence. One can find that the price of the previous day "*LastDayPrice*" is used in 8 out of 10 cases as an important correlation factor. The other once "TxTfrValMeanNtv" and "*TxTfrValMeanUSD*" which represent average transaction value in native tokens or with conversion to USD. Those features possess the similar logic. So, we could claim that the average transaction value is an important correlation factor (10 out of 10) with the token price. Applying the logic of blockchain architecture transaction it makes sense as it is a sign of increasing demand, hence, price of token. The increase in average transaction value is a result of increasing activity on the network,

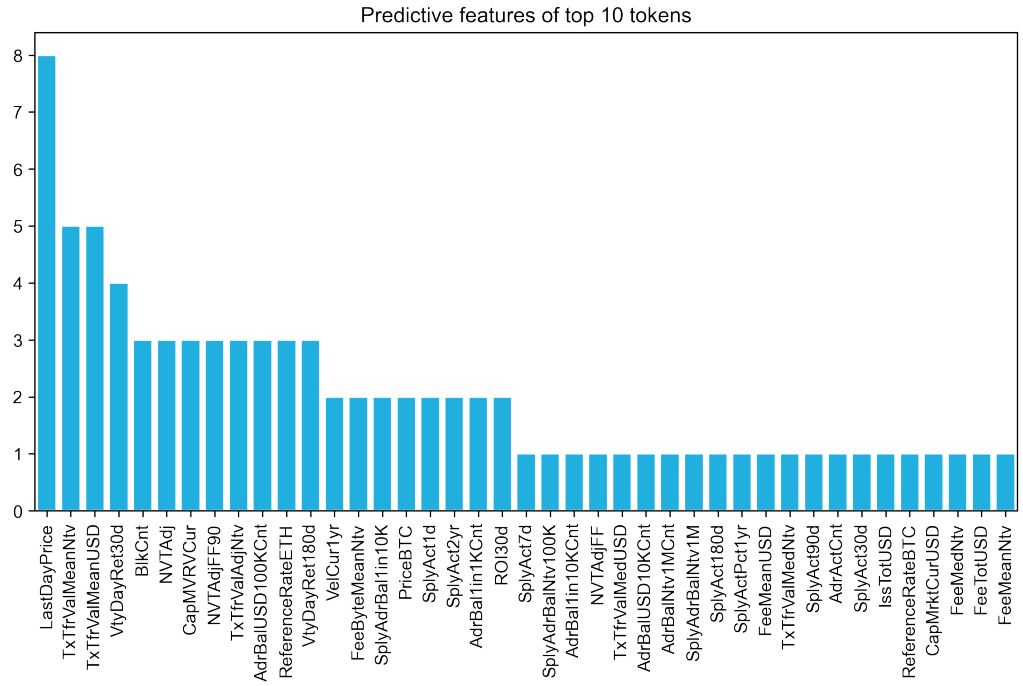



Figure 3: Predictive features of top 10 crypto assets based on market capitalization from Coinmarketcap

hence growth in usage demand. Based on the law of demand and supply (2.2) the value of the demanded asset shall increase, which in our case is a higher token price. The following "*VtyDayRet30d*", which is token price returns volatility for the previous 30 days, appeared to be an important feature for 4 tokens. The following features are less popular correlation elements, only 3 out of 10. Those are "*BlkCnt*" which represents a quantity of blocks created on chain (relatively corresponds to the amount of transactions created), "*NVTAdj*" which is network value to transaction (ratio between the amount of connections (addresses) in the network to the amount of transaction during the day), "*CapMVRVCur*" is the ratio of the sum USD value of the current supply to the sum "realized" USD value of the current supply (USD equivalent of token spend from the addresses), "*NVTAdjFF90*" is the showing of market capitalization as a moving average of 90 consecutive days, "*TxTfrValAdjNtv*" is the sum of native tokens transferred between distinct addresses, "*SplyAdrBalUSD10K*" is token supply held by unique addresses that hold at least 10,000 USD or greater equivalent of native tokens at the end of that day, "*ReferenceRateETH*", which is reference rate of Ethereum (ETH) token price which substituted "*LastDayPrice*" of the token, "*VtyDayRet180d*" is similar to "*VtyDayRet30d*" volatility but of 90 days period. The next features only appeared for 2 tokens our of 10: "*VelCur1yr*" ratio of the value transferred in the trailing 1 year divided by the current supply corresponding the same interval, "*FeeByteMeanNtv*", the mean transaction fee per byte of all blocks of chain per day, "*SplyAdrBal1in10K*" is token supply held by unique addresses that hold at least 1 to 10,000 of total token supply at balance, "*priceBTC*" the price of the BTC at previous day, "*SplyAct1d*" and "*SplyAct2yr*" supply of unique native tokens that were transacted at least once in the trailing of 1 day and 2 years, "*AdrBal1in1KCnt*" it is number of unique addresses that hold 1 to 1,000 of total token supply at balance, "*ROI30d*" return on investment on the token assuming it was purchased 30 days prior. The rest of the features are identified only once in 10 cases and can be considered a modification or an aggregation of the already described elements. The more detailed information regarding all the features can be found at data provider (CoinMetrics)

website[2] .

As the result of our correlation investigation one can reach to conclusion that in general, the features that can posses a predictive power to the future token price is the price of the previous day, network congestion and congestion metrics, returns of several consecutive days and amount of addresses holding more than certain threshold of tokens and total supply of tokens held by addresses with significant amount of tokens. This corresponds to the research of [10]. Surprisingly, only 2 times the price of BTC was identified as important correlation feature, despite BTC being mentioned as [21], [11] as a major driver of other cryptotoken price changes we see that BTC is mentioned in 2 out of 10 cases. In addition, even proposing other candidate tokens as predictors for the token price none of the tokens except from BTC was identified as an important predictive factor.

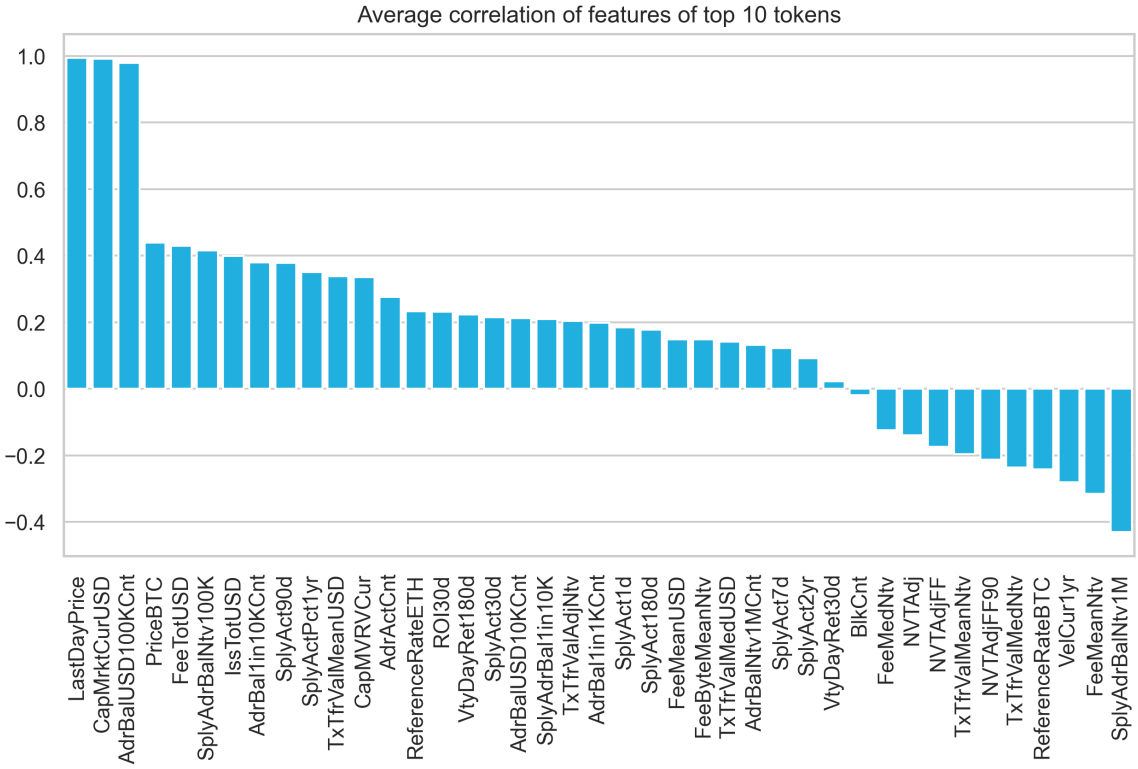



Figure 4: Correlation of features of top 10 crypto assts with the token price based on maket capitalization from Coinmarketcap

With the following Figure 4, one could find how the on-chain parameters correlate with the future (next day) token price on average. One can see that mostly the token price is positively correlated with on-chain features rather than negatively correlated. What is especially interesting here is that there is a clear division between those positively and negatively correlated features. One can identify the following positively correlated features: the last day price, market capitalization of the token (total amount of tokens multiplied by the token price), the supply in addresses with significant amount of native tokens, last day price of Bitcoin, transaction fees. Respectively the NVT (Network value to transfer value ratio) - ratio between amount of connections (addresses) in the network to the amount of transaction during the day and combination with it appeares to be always negatively correlated with the token price. The same holds for yearly velocity, which is a ratio of total token amount transferred to current trailing year of token supply.

[2]https://docs.coinmetrics.io/info/metrics

Having identified a correlation effect of the on-chain information we explore how these features are selected for the top 3 cryptotokens (BTC, ETH, BNB) based on market capitalization, as we are going to use those for our prediction analysis. For this candidates we will increase the feature-wise correlation threshold from 0.5 to 0.7 imposing even higher restrictions on what features can be considered as cross-correlated. We approach it this way in order to make a stronger correlation pairs and let define more features that can potentially improve the future prediction model. Here in the Figure 5 we identified the result of such feature candidates selected to be used as explanatory variables. As, one may notice only 3 features are identified for all 3 cryptotokens as significant correlation candidates:

- "*TxtfrValMeanNtv*", "*VtyDayRet30d*", "*LastDayPrice*"

Features that overlap between BTC and ETH only:

- "*SplyAct7d*", "*VtyDayRet30d*", "*LastDayPrice*", "*FeeMedNtv*"

Features that overlap between ETH and BNB only:

- "*TxtTfrValMedNtv*", "*AdrBalNtv1MCnt*"

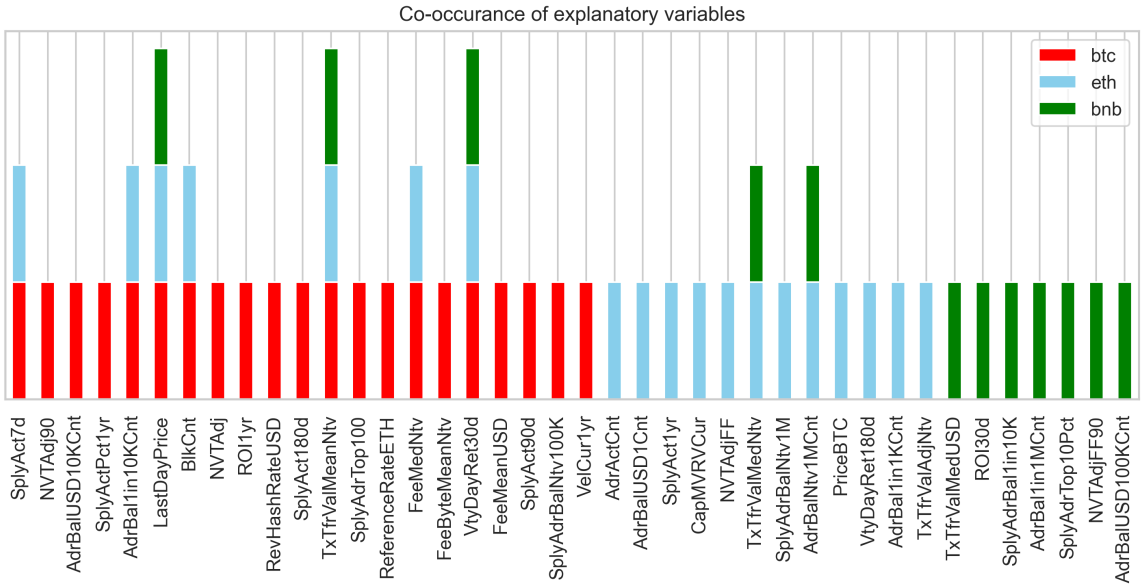



Figure 5: Co-occurance of features identified with correlation analysis for 3 token prices individually

To expand the investigation on the feature difference of the best correlation candidates for the token price we constructed the Table 1 of feature overlaps. The Table demonstrates a significant evidence for uniqueness of dependencies of each crypto token, hence contradicts and evidently challenges [12] that approaches treatment of each token equally using the on-chain information as a generalized black-box selection of features.

In addition we constructed the Table of person correlations between the features and token prices, which can be found in the Annex part of the report as Table 4 and Table 5 as continuation of the same Table. Surprisingly, one can notice that even prices of tokens can have a different correlation signs with the same features. Hence, one can observe that the decrease of transaction blocks created on chain and increase of amount of addresses holding 1 to 10000 of circulation supply positively correlates with the BTC price and negatively correlates with ETH price.

| Overlap of correlated features between crypto tokens | | | |
|---|---|---|---|
| Token name, (features) | BTC | ETH | BNB |
| BTC, (21) | 100% | 33% (7/21) | 14% (3/21) |
| ETH, (19) | 36% (7/19) | 100% | 26% (5/19) |
| BNB, (12) | 25% (3/12) | 41% (5/12) | 100% |

Table 1: The ratio of overlap of highly correlated features that were identified for each of the crypto tokens individually

#### 5.5.1 Feature selection research outcomes

To sum up, based on our broad feature investigation we can conclude that the token price of previous day lag is a significant predictive factor for the token price of the following day. Moreover, based on the partial autocorrelation plot 2 we evidence that the price of the previous time lags maintains a predictive power for the token price of the following day, opening up the possibility to use memory-based machine learning approaches of prediction in the upcoming section. We also discovered that the most selected and highly correlated with token price on-chain features are the previous day token prices and metrics of transaction values, network congestion, holders possessing tokens on their balances. There was no evidence found of strict dependency on BTC price of other crypto token prices as BTC price "*PriceBTC*" was mentioned only once as a selected predictive feature. It is also important to point out that some feature can possess completely opposite correlation signs for different crypto tokens.

# 6 Predictive Modelling

## 6.1 Settings

The aim of this section is to predict the token price utilizing on-chain features discovery in Section 5. Here we focus on previously identified tokens of interest, BTC, ETH, BNB and apply several prediction models in order to estimate how well the token price can be predicted. All of the mentioned tokens were launched on the different dates, hence appeared in the different timeslots of the information provider we obtain on-chain data from. The end date for all the token price records is 14th of June 2022. The start date for BTC price observations is 18 July 2010, for ETH 8 August 2015, BNB 15 July 2017. As we already defined in Section 5 the token price possess different predictive nature and cannot be treated equally for the prediction purpose. That is why we do not take the same set of features for predictive modelling enabling the learning with the significant features for each individual token as it is identified in Section 5.

#### 6.1.1 Prediction models

As one of the aims of this study is to explore the feasibility of applying models for the regression prediction we provide a selection of regression models that are applied for BTC, ETH and BNB token price that were often used in the related works (Section 4). We could group the selected models as tree-based models: Decision Tree, Random Forest, XGBoost and Neural network-based models: Feedforward neural network (FNN), Long Short-Term memory (LSTM), Gated Recurrent Unit (GRU). having created the models we evaluate and compare their performance and discuss the results. The research of [16] specifies that GRU and LSTM to be the best predictive models for the given problem, whereas [17] claims that there is no significant difference

in the neural network-based and tree-based models performing token price predictions. In scope of our work we expect to receive better result with GRU or LSTM setting Decision Tree model as a baseline.

To evaluate the models performance we analyse the last 80 days of the daily predicted vs. actual tokens prices for BTC, ETH and BNB. As we are dealing with regression problem we apply the root mean squared error (RMSE) as the metric used to compare the performance on the test set.

### 6.1.2 Model fitting

When constructing the machine learning models one can tune the parameters with an aim to reduce the prediction error, hence increase the prediction accuracy on the test set. In most cases we are looking for omitting the overfitting, underfitting and bias towards specific values and outliers in our models.

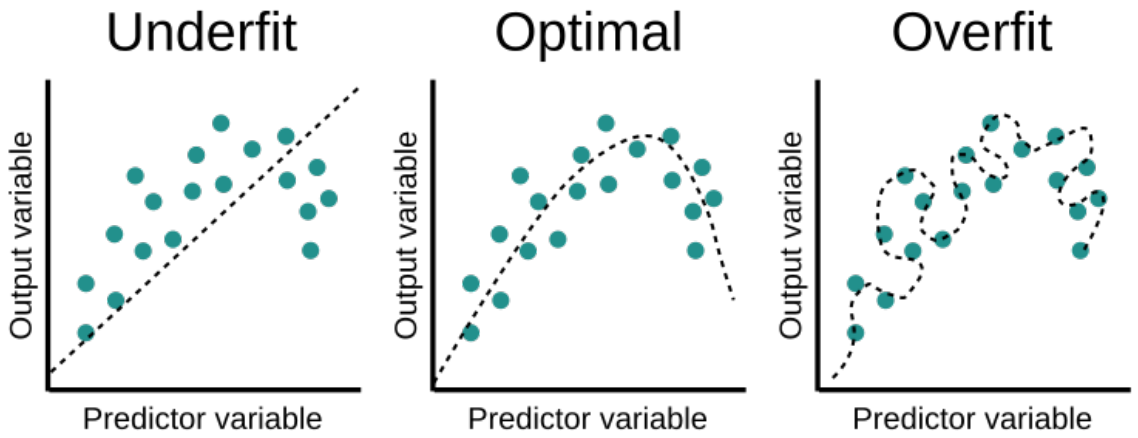



Figure 6: Model fitting issues, educative.io

**Model underfitting** is the result of poor training performance in which the model is not able to capture the variability of the train set. This way it cannot derive the prediction results that approximate relatively well the true results of the target variable. There are multiple reasons for underfitting, but we underline the most common once: noisy training dataset, too simple model, not enough training time, small dataset.

**Model overfitting** is also a result of poor training performance which is skewed towards exploiting the variability of the training set. As a result of overfitting the models provide low errors in predicting the target variable of the training set and high errors on the test set. That means, that the model is overtrained on the training examples and cannot perform well on the new previously unseen data. The popular reasons for overfitting is noisy training dataset, high model variability, small dataset, too complex model, long training time.

When constructing the predictive models a special attention should be given to the bias-variance trade-off when solving the prediction task. That's why we preprocess the datasets before using it in the prediction to avoid outliers, misleading empty values, etc. Moreover each machine learning model undergoes the parameters tuning procedure in order to deliver the best possible prediction results.

## 6.2 Data Normalization

For the given datasets of BTC, ETH and BNB we conduct data cleaning removing the features with missing data that is randomly ommited or is empty in 15% of observations or higher. That process was initially conducted in Data description and Data description and Feature selection in Section 5.

As each of the features selected possess different ranges of maximal and minimal values throughout the observation. If the features follow the different scaling it results in longer convergence time, hence increase the computation time while fitting the training set. In some cases it might also increase a bias towards specific feature variability potentially damaging the model learning performance. That's why we apply *Min-Max Normalization* which is feature standardization technique that redistributes the values of each features to the common scale in range from 0 to 1 keeping the distribution unchanged and preserving the relationships among the original feature values. In addition, the bounded range results in lowering the standard deviation suppressing the outliers effect. The *Min-Max Normalization* algorithm looks as follows:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

,where $x$ is a value in the feature column

After the mentioned normalization procedure the dataset is ready to be used for the price prediction studies.

## 6.3 Competing machine learning models

In this section we discuss briefly the machine learning models that are used in our predictions study. We start with tree-based and then are followed by neural-network based models.

### 6.3.1 Decision Tree

Decision tree is non-parametric supervised machine learning model that can be used for classification and regression problems. The idea behind this model is to construct a decision boundaries inferred from the dataset explanatory variables (predictors) to receive the best accuracy predicting the target variable (predicted), which in our case is token price.

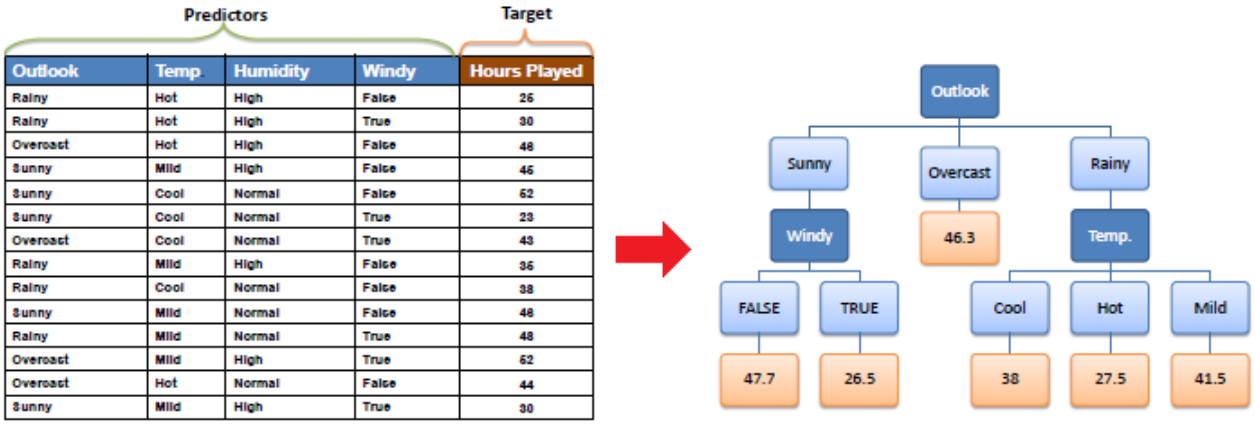



Figure 7: Decision tree regressor example, [22]

The decision tree follows the logic of top-down greedy sample partitioning. Starting from the top (root node) the algorithm defines which of the explanatory variables to use as the first

partition node and which to use for the following once. The order should be defined based on how well each feature type can separate the classes or values of the response variable keeping the smallest amount of response variables of other type or significantly different value in the same subsample created by the node separation. In other words we are looking for the most homogeneous separations. To evaluate the homogeneity of the sample the *Entropy* is calculated:

$$E(S) = \sum_{j=1}^{c} -P(x_i) log_2 P(x_i) \tag{2}$$

,where $P(x_i)$ is the probability of value occurrence of given class in the subsample divided as a result of node separation.

The decision tree aims to lower the entropy level which is presented in the range between 0 and 1. In order to understand which explanatory variable to use to construct the next node and if there is possible improvement in the tree building the *Information Gain* is used:

$$IG(T, A) = E(T) - \sum_{v \epsilon A} \frac{|T_v|}{T} E(T_v) \tag{3}$$

,where $T$ is the target variable, $A$ feature (explanatory variable) we are testing as the next possible node, $v$ each value in feature column $A$. When we calculate the information gain for each of the feature columns we select the next node to produce next subsamples, which results in the highest information gain. The processes continues interatively for each tree branch until there is no more information gain or any other restrictions are applied. The last nodes to be produces are called leaves as they do not have any more child nodes. When the model is learned the frequency table constructed by the decision tree is used to identify the target output for the new previously unknown instance.

**Hyperparameters tuning** As a part the decision tree prediction we also use the the parameters tuning techniques to improve the performance of the decision tree. We apply *Grid Search*, which is a tuning technique that attempts to compute the optimum values of hyperparameters. In our case, we decided to focus on experimenting with feature selection where we move with to options for the tree nodes selections: a random feature selection as the next node and best node by the highest information gain, longest path between the root node and leaf nodes, the minimum amount of samples required in a leaf node, the minimum weighted fraction of all the input samples required to be in a leaf node, the number of features to consider when looking for the best split: square root of all the feature elements, maximum amount of leafs to be created in best-first fashion. The hyperparameters that can be considered the best for the given training problem are selected based on the performance on cross-validation that are repeated 3 times on 10 randomly selected validation samples (K-Fold validation). The combination of hyperparameters that result in the lowest validation error are selected to form a final model setting.

### 6.3.2 Random Forest

The Random Forest regression is a supervised learning algorithm that uses ensemble learning. Ensemble learning is a technique that combines prediction from multiple machine learning algorithms to make a more accurate prediction than a single model. In case of Random forest we construct multiple Decision Tree Regressors and average their outputs to form the output, which in our case is a mean token price of all estimators. The description of such process can be found in Figure 8.

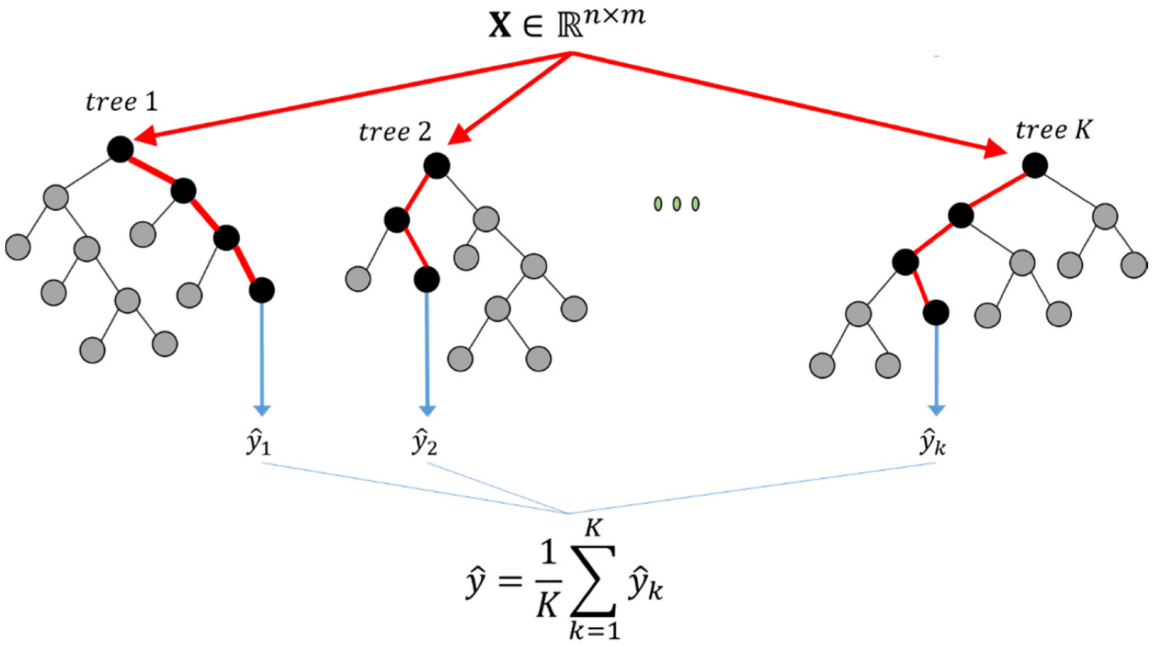



Figure 8: Random Forest regression, [23]

**Hyperparameters tuning** The Random Forest is complex model and requires more computational power than Decision Tree alone as it combines a number of decision trees. As it inherits the decision trees nature we have the same amount of parameters to tune: a random feature selection as the next node and best node by the highest information gain, longest path between the root node and leaf nodes, the minimum amount of samples required in a leaf node, the minimum weighted fraction of all the input samples required to be in a leaf node, the number of features to consider when looking for the best split: square root of all the feature elements, maximum amount of leafs to be created in best-first fashion. In addition to the mentioned hyperparameters we also add an option to enable and disable *Bootstrapping*. This process conducts a random data sampling for a given number of iterations and given number of variables. This way the decision trees in the random forest obtain a distinctive training nature on the same dataset. The output of multiple randomly drawn decision trees from the dataset is then averaged and presented as an output of the Random Forest Regressor.

### 6.3.3 eXtreme Gradient Boosting

Extreme Gradient Boosting or XGBoost is a gradient boosted decision trees algorithm [24]. The idea behind the algorithm that each decision tree tries to correct the predecessor's decision tree error. This means that the trees form a sequential chain of self-correcting predictors that take into account the predecessors residuals.

The XGBoost is a loss function optimized model that supports parallel processing, maintains efficient memory management and regularization which helps in reducing the overfitting.

The process for the XGBoost tree construction can be explained in several steps:

1. We set an initial prediction (base score) or move with the default score which is 0.5

2. Then we start building a tree from the single node. As we start from the prediction of 0.5 we calculate the residual of each instance (observation). Having calculated the residuals we move towards the calculation of *Similarity Score*:

$$Similarity = \frac{(\sum_{i=1}^{N} y_i - base)^2}{N + \lambda} \tag{4}$$

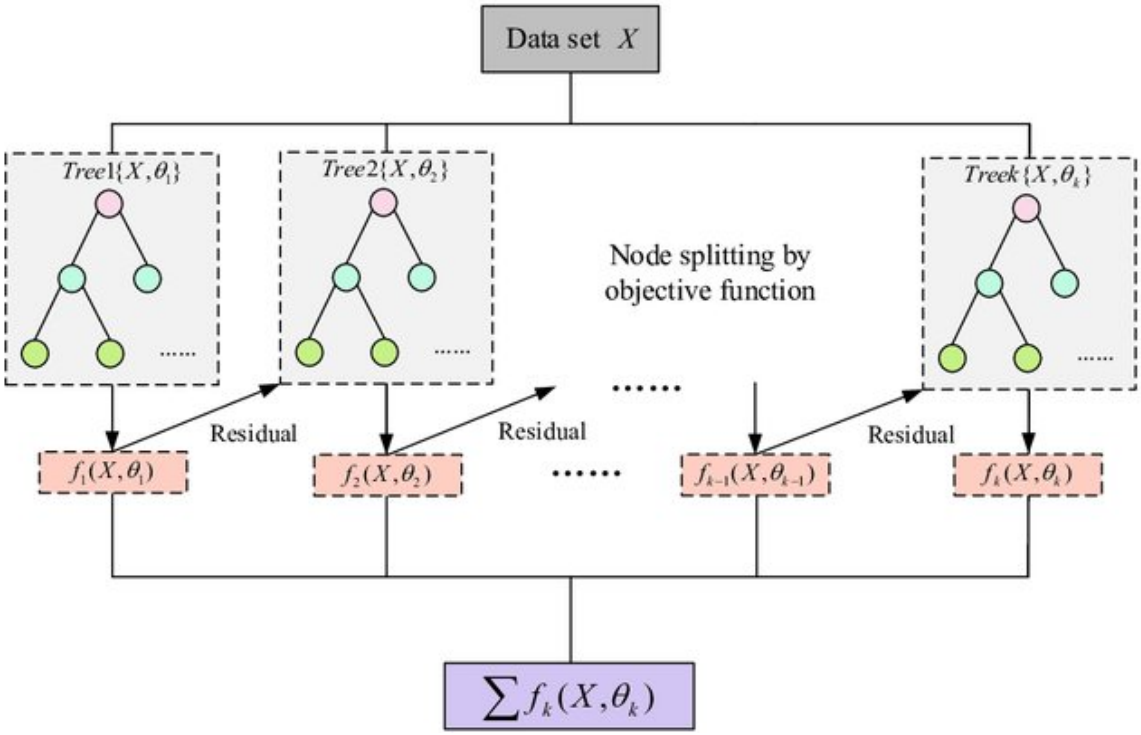



Figure 9: eXtreme Gradient Boosting, [25]

where $N$ - number of instances, $y_i$ is the true target value, *base* is base score, and $\lambda$ is the regularization parameter (1 by default) that reduces the prediction's sensitivity to individual observations (the value of *Gain* 5 decrease is inversely proportional to the number of residuals in the node.)

3. Then we conduct the next separation to proceed with the child node. We iteratively conduct the threshold split among all the observation to find out which of the separations can produce 2 groups that clusters best the residuals. We can estimate that with *Gain*:

$$Gain = Left_{Similarity} + Right_{Similarity} - Root_{Similarity} \tag{5}$$

where $Left_{Similarity}$ is the *Similarity* of the left leaf (child node), $Right_{Similarity}$ is the *Similarity* of the right leaf (child node) and $Root_{Similarity}$ is similarity of the root node (father node). The node separation that results in the highest gain wins the selection.

The process is conducted iteratively until there is no positive gain or other limitations are applied like $\gamma$. The higher $\gamma$ we initiate the higher requirements for *Gain* we set to construct a leaf. If the difference between *Gain* and $\gamma$ is negative then we prune the branch (we remove the node). Based on the description one could notice that both $\gamma$ and $\lambda$ prevent from overfitting the model.

Having constructed the tree we use the residual values of leafs to produce the output error results for out future predictions:

$$OutputValue = \frac{(\sum_{i=1}^{N} y_i - base)}{N + \lambda} \tag{6}$$

This equation is quite similar to the *Similarity* score, except the *Similarity* takes a square of sum of residuals. $\lambda$ plays here as regularization parameter that reduces the prediction sensitivity to the individual observation. This means that the higher is amount of residual values that appear in the leaf the lower is the regularization impact of $\lambda$ on predicted residual output value.

Each time we build a new tree we can make a new prediction for each instance:

$$\hat{y} = base + \eta \times OutputValue \tag{7}$$

where $\eta$ is the Learning Rate of the XGBoost. As the XGBoost constructs a chain of decision trees the residuals from the newly predicted target values and the true target values are used to train the next decision tree. The amount of trees to construct can be artificially set. Another way to how the algorithm stops training is when it gets to the part where the residuals become smaller than a predefined threshold value. This process can be graphically observed in Figure 9.

**Hyperparameters tuning** In the XGBoost we experiment with setting *alpha* and *gamma* parameters for the model training and also set some boundaries on the decision trees depth. In addition we tune the learning rate of the XGBoost regressor.

### 6.3.4 Feedforward Neural Network

This subsection opens a new type of machine learning models we implement for the task of the token price prediction, which is called Neural Networks.

**Neural network** Artificial Neural Networks (ANN) is the type of machine learning models that are inspired by the idea of imitating the human brain work, especially neurons signaling communication. The artificial neural networks consist of node layers that form a network. If we drop down the ANN it consists of the *Input Layer* with the amount of nodes that correspond to the number of features in the neural network, then it can be followed by the any amount of *Hidden Layers* that can contain any number of neurons with weights that need to be adjusted during the learning process and *Output Layer* with the number of neurons that correspond to the amount of the prediction outputs the neural network has to produce. Here, in Figure 12 one can observe that the neural network contains 1 input layer, 2 hidden layer, and 1 output layer with 1 neuron, hence 1 output.

**Neuron** To have a better understanding what is the typical architecture of the neuron lets take a look at look at Figure 10. Here we see that the input of the node receives 3 outputs $O_i$ from the previous layer with the weight $w_{ij}$ applied to those outputs. Then the the sum of the output weights is processed via the $ActivationFunction$ which is a mathematical formula for value transformation. The output of the activation function is then considered as a output of the neuron. Then the output of the neuron can be transferred to other neurons connected to it applying specific weights for each connection (if this neuron belongs to the *Hidden Layer*) or the output of the neuron is the prediction output in case the neuron belongs to the *Output Layer*.

As we already mentioned, the neural network can be seen as an imitation of human brain with neurons and synapses that transfer information. The learning of the neural network is inspired by the human brain memory. The way it is designed here is by tuning the weights of synapses that connect the neurons. The weights are those elements that are iteratively adjusted during the learning process to better fit the training data and reduce the loss. As we are talking here about the regression problem the loss we try to tune for each of the models is the mean squared error (MSE) between the predicted token price and the actual token price.

**Feedforward** The Feedforward neural network is the basic type of the neural networks. It can be considered as a simple model due to the fact that the information is processed in only

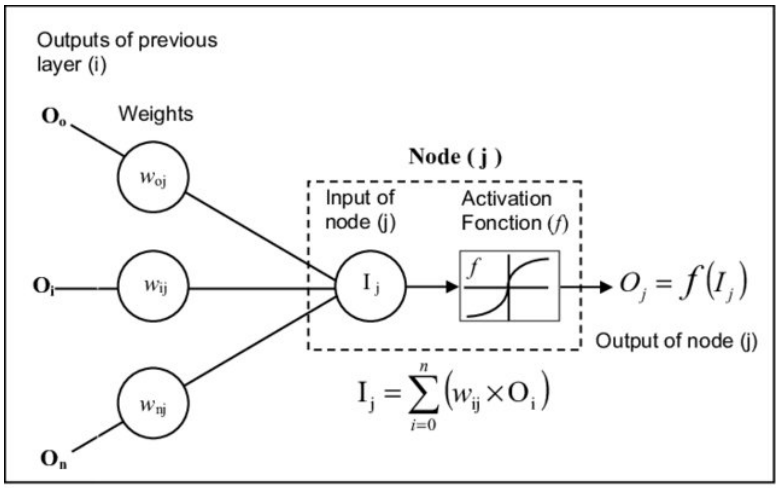



Figure 10: Typical neuron architecture, [26]

one direction from the input nodes to the output nodes. This learning flow can be also called as *Forward Propagation*. In Figure 12 one can see an example of Feedforward neural network. Each neuron has a directed connection to each neuron from the previous and the next layers. After each training iteration the loss function is calculated and the weights of each neuron are tuned with the aim to reduce loss.

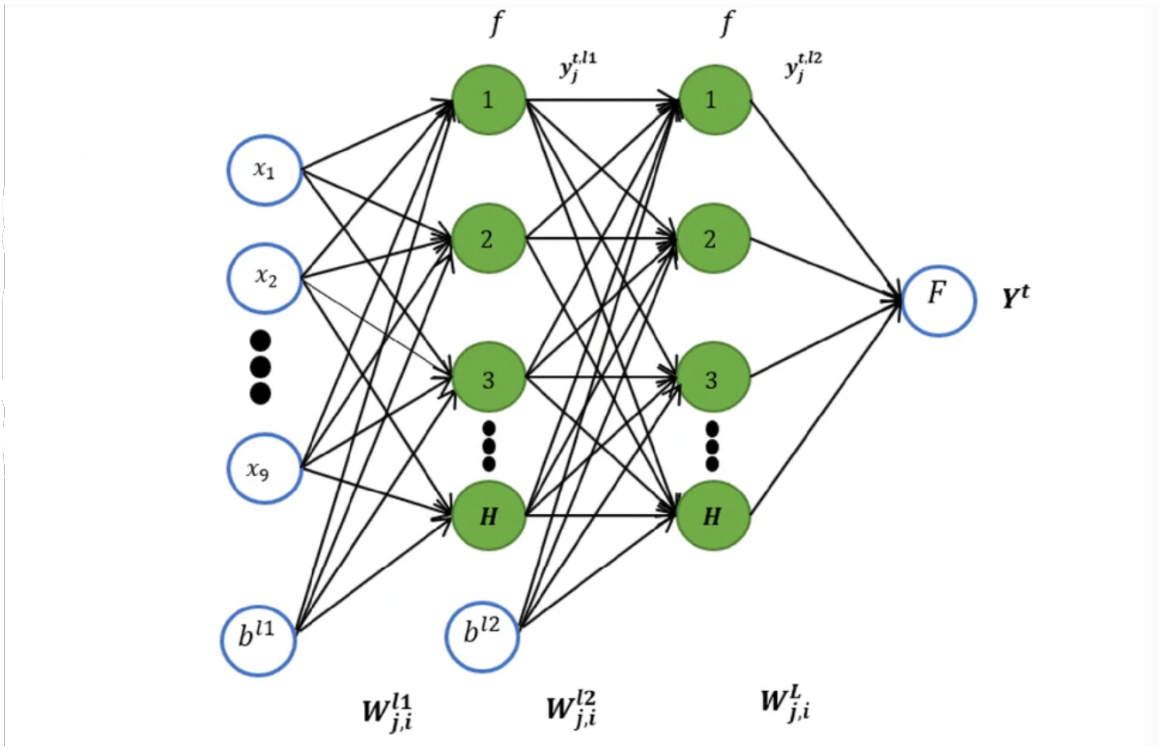


Figure 11: Feedforward neural network, [27]

During the phase of FNN model building we built 3 distinctive FNN neural network models for each of the tokens (BTC, ETH and BNB) as we proved in Section 5 that each of them requires an individual approach in feature selection and cannot be treated equally. In all 3 cases we attempt to produce different architectures, forming different numbers of neurons and hidden layers. The only thing that remains in common is that the activation function for hidden layers with which we able to receive the best results is ReLu:

$$f(x) = max(0, x) \tag{8}$$

which means that if this function receives any negative input it returns 0 and if it receives any positive input in will return that input. This function is famous by its sparsity as it can produce 0 values this way switching off the output of certain neurons, which reduces the chances of overfitting. As a result, the sparse network increases chances of neurons to process important features that help fit the training data.

### 6.3.5 Long Short-Term memory

The Long Short-Term memory (LSTM) is a type of the neural networks which belong to the Recurrent neural networks. This types of neural networks include the architecture of feedback connections. With this approach the model is designed to tackle sequence-related problem. So we start with recurrent neural network and then proceed with explaining LSTM.

**Recurrent Neural Network problem** Here we provide the overview of the Recurrent Neural network (RNN) and reason why we don't select it as a competing model. The specialty of RNN is in the hidden state. The neuron takes the information from both the input at time $t$ and result performed by the activation function from the hidden state neurons at time $t-1$ to calculate the output for time $t$. This feature provides "memory" to the model that "knows" the previous inputs and their outputs. This feature comes especially handy for Natural language Processing and time-dependent series as blockchain information we deal with in scope of this report. Here we are not bounded with how much we dig to the past with our sequential training data but use the shared weights of each activation output up to the predefined lag to predict the future output.

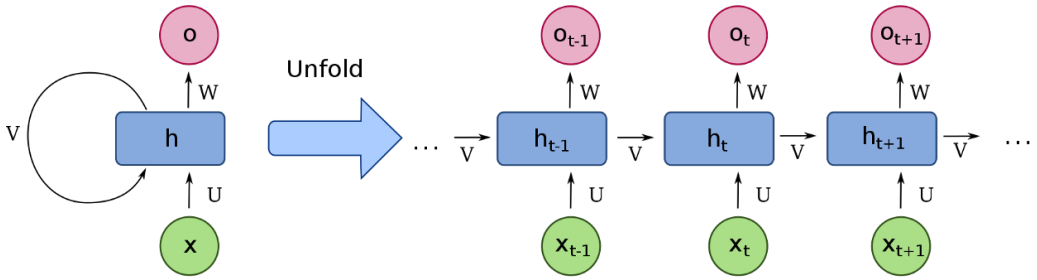



Figure 12: Recurrent Neural network, [28]

It appeares to be that RNN can maintain low amount of lagged activation results in memory before the *gradient explodes* or *vanishes*. This means that even if the architecture of RNN enables to handle as long sequence of explanatory data as one requires on practice there is a risk of losing the no memory effect. The reason for *vanishing gradient* is the following: considering a large network with 100 layers, at each step the matrix of weights is multiplied with activations from the previous layers. In case activations at each layer are less then 1 they will constantly descrease to a small number approximated to 0. On the other hand if the activations at each layer are greater then 1 they will result in enormously high values that follow to infinity. This is the reason why RNN is often considered a transition solution to more advanced and robust once. One of such is the Long Short-Term memory (LSTM).

**LSTM** The LSTM was designed as the recurrent neural network algorithm that can solve the problem of *vanishing gradient* and provide a more robust solution to the memory-based neural networks. The solution lies in the architecture of LSTM which implements the gating mechanism that controls a memorizing mechanism. There are 3 gates in the LSTM: *input gate*, *forget gate* and *candidate gate (output gate)* . The output of LSTM is completely dependent on the mentioned gates.

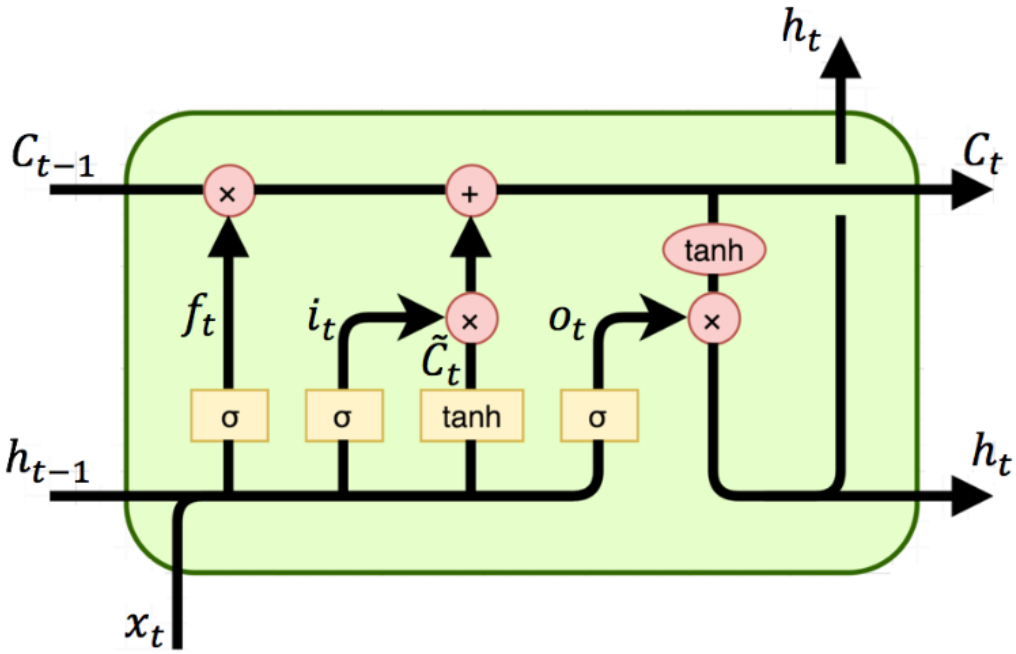



Figure 13: Long Short-term memory, [29]

Here we provide a short walk-through how the LSTM works supporting our explanation with the Figure 13. The general setting are as follows

- The cell state $C_t$ is the upper horizontal line in the Figure 13. It conducts linear pointwise iteractions going along the whole LSTM architecture. This way the LSTM adds and removes the information from the cell state.
- The mentioned before 3 gates are described combined in the Figure. The activation function are presented in a nature of tangent function and sigmoid function. The sigmoid function decides the weight of the information to let through by assigning the value between 0 and 1, which is why they are considered as "gates".

  The tangent function looks the following way:

$$T(x) = \frac{\epsilon^x - \epsilon^{-x}}{\epsilon^x + \epsilon^{-x}}; T(x) = [-1, 1] \tag{9}$$

  The sigmoid function looks the following way:

$$S(x) = \frac{1}{1 + \epsilon^{-1}}; S(x) = [0, 1] \tag{10}$$

Now let's make the LSTM gates walkthrough:

1. The model typical starts from the *forget gate* which is used to decide what information to remove and what to keep in the cell state. In this case it decides the weight of the $(h_{t-1})$ output of the previous LSTM at lag $t$ and the feature input that corresponds to lag $t$ $(x_t)$. Here the previous cell state $C_{t-1}$ is updated with pointwise multiplication with the *forget gate* output and creates a new cell state $C_t$. The above mentioned description can be formalized here:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{11}$$

2. The second element is the *Input gate* which decides which new information to store in the cell state. The *sigmoid* layer decides which values to update $i_t$, whereas *tanh* layer creates a vector of candidate values $\tilde{\mathrm{C}}$ that can be added to the cell state. The result of their pointwise multiplication is added to the cell state $C_t$, this way we decide the value of update for each state value.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{12}$$

$$\tilde{C}_t = tanh(W_C[h_{t-1}, x_t] + b_C) \tag{13}$$

3. The output of LSTM is conducted as a part of *output gate*. It is based on the cell state $C_t$ with additional filtration. We run the sigmoid function that provides the weight $O_t$ to each cell state $C_t$ value for the output. Then we transform the cell state $C_t$ values in the range between -1 and 1 and multiply it by the output of the sigmoid function. This way we output the parts decided by the LSTM $h_t$.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{14}$$

$$h_t = o_t * tanh(C_t) \tag{15}$$

The update of the current state of LSTM can be described as follows as it combines the outputs from the part 1 and part 2 of the LSTM algorithm description 6.3.5.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{16}$$

As we already mentioned we build 3 distinctive LSTM neural network models for each of the tokens (BTC, ETH and BNB). In all 3 cases we attempt to find the the lowest loss on validation, building distinctive architectures, setting learning rates and optimizers, forming different numbers of neurons and hidden layers.

### 6.3.6 Gated Recurrent Unit

**GRU** Gated Recurrent Unit is the newer generation of Recurrent Neural network which resembles some nature of LSTM. The major difference is that GRU has no *cell state* and has *hidden state* instead to transfer the information.

As one can find in Figure 14 GRU architecture consists of 2 gates, whereas LSTM has 3 gates as described above. It can be easily identified as the amount of gates equals to the number of sigmoid functions. These 2 gates are called *reset gate* and *update gate*:

1. The *reset gate* works as a discriminator on the amount of past information that should be "forgotten" which is similar to forget gate in LSTM:

$$r_t = \sigma(W_r[x_t, h_{t-1}] + b_r) \tag{17}$$

when $r_t$ is close to 0 almost no past information is taken into account for the mentioned calculations

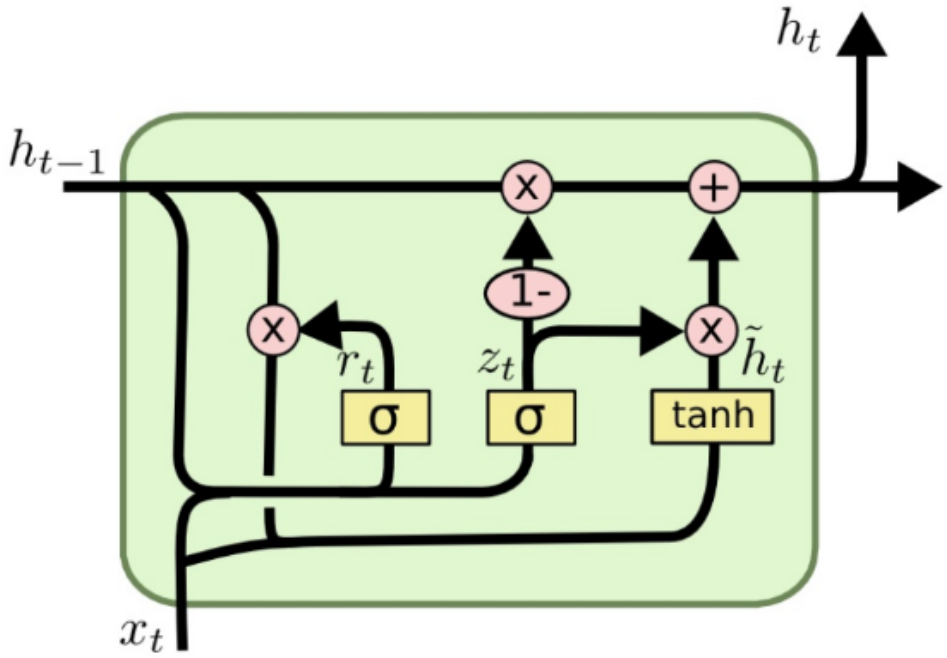



Figure 14: Gated Recurrent Unit, [30]

2. The *update gate* decides the weight of the input updates to consider for the activation:

$$z_t = \sigma(W_z[x_t, h_{t-1}] + b_z) \tag{18}$$

3. We then compute a candidate activation:

$$\tilde{h}_t = tanh(W_{\tilde{h}}[x_t, r_t * h_{t-1}] + b_{\tilde{h}}) \tag{19}$$

4. And finally the activation $h_t$ is linear combination of the previous activation $h_{t-1}$ and the candidate activation function:

$$h_t = (1 - z_t)h_{t-1} + z_t\tilde{h}_t \tag{20}$$

The major difference one might observe in here is that unlike LSTM the GRU cannot control the degree of state exposure as it is updated each time completely.

We will develop 3 distinctive GRU architectures to find the lowest loss on validation while tuning the learning rates and optimizers, forming different numbers of neurons and hidden layers.

## 6.4 Results

In this part we discuss the results we obtained from the token price predictions. As we previously identified (Figure 5) we predict BTC price for the next day utilizing 21 explanatory variables, ETH with 19 and BNB with 12. In addition, based on partial price partial autocorrelation observations (Figure 2) we estimate the most appropriate sequence of past values for Bitcoin to be up to 10th lag, and for Ethereum and Binance Coin up to 7th lag. The LSTM and GRU will fit the mentioned sequential data. The baseline for our predictions is the prediction performance of the Decision Tree.

Here we provide the Table 2 prediction result on the test set of the 80 daily token price predictions in the MSE scale.

**Price prediction performance**

| Token Name | Decision Tree | Random Forest | XGBoost | FNN | LSTM | GRU |
|---|---|---|---|---|---|---|
| BTC | 0.00111 | 0.00141 | 0.00115 | 0.00110 | 0.00116 | 0.00074 |
| ETH | 0.00059 | 0.00057 | 0.00106 | 0.00215 | 0.00317 | 0.00523 |
| BNB | 0.00994 | 0.00951 | 0.01344 | 0.01545 | 0.03756 | 0.02138 |

Table 2: Mean squared error (MSE) of the daily price predictions on 90 consecutive days

**BTC** One can notice that there is a significant dominance of GRU in the BTC price prediction, overperforming all the other models results. The rest of neural network models (FNN and LSTM) do possess slightly lower loss than tree-based models.

**ETH** With Ethereum price prediction the decision trees (baseline) and Random Forest possess significantly lower losses with little dominance of the Random forest. In this case the neural networks showed themselves as much worse performance than the tree-based algorithms.

**BNB** The results of Binance Coin price predictions also advocate for better prediction performance of the Random Forest and Decision Tree models, with a slight dominance of Random Forest. The neural networks seem to underperform again with the significantly higher loss results.

Let us have a look at the BTC price prediction performance of models. For visual convenience we separate the prediction results of the tree-based algorithms (Figure 15) and neural-networks (Figure 16). As one might observe the prediction of the tree-based algorithms seems to overperform the predictions of the neural network models until the approximate 48th prediction day. After we observe a significant fall in the BTC price the tree-based algorithms lose their relatively low loss and increase it in several times, whereas the lost of neural-network based models, and especially GRU is kept unchange after the radical price shift.

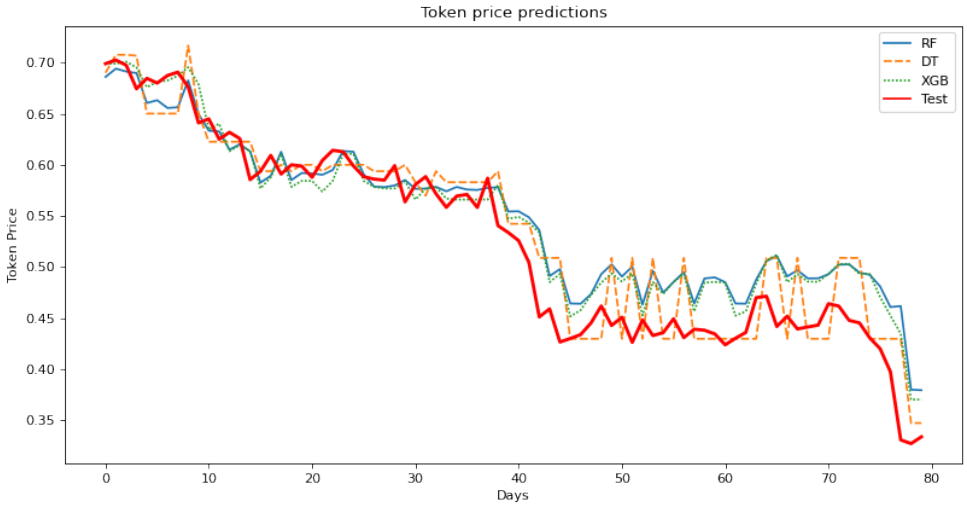



Figure 15: Bitcoin prediction results for Decision Tree (DT), Random Forest(RT), XGBoost (XGB)

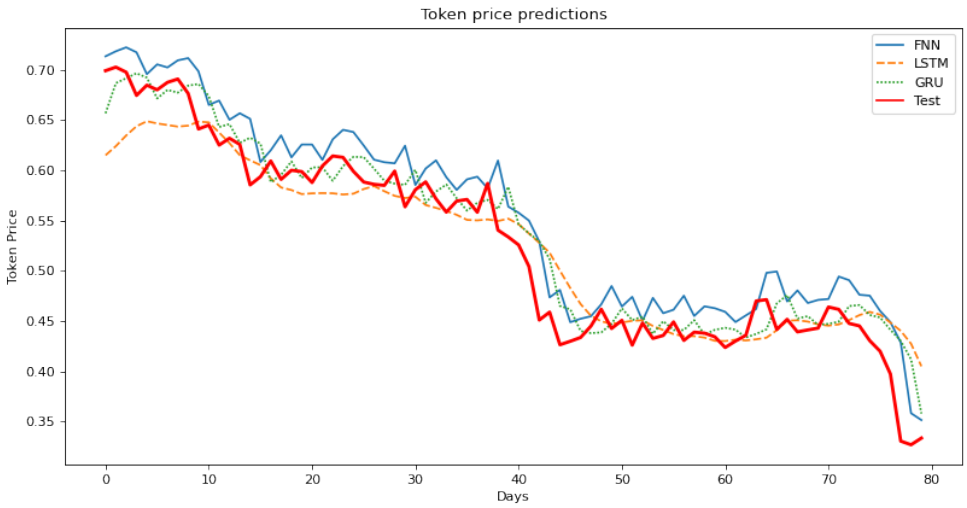



Figure 16: Bitcoin prediction results for Feedforward Neural network (FNN), Long short-term memory (LSTM), Gated Recurrent Unit (GRU)

Another interesting observation to mention is Binance Coin prediction performance Figures 17 and 18. Starting from the 57th observation one can observe that the decision trees maintain the increasing loss while predicting the last 23 observations. At the same time the neural networks maintain their prediction loss at the relatively stable range.

Based on the results described above one can observe that there is no sole favourite machine learning method identified for the given token price prediction problem, which corresponds to the outcomes of [17]. Based on our observation the tree-based algorithms are able to perform relatively well in terms of token price prediction if not experiencing radical price changes, whereas the neural network models provide relatively stable prediction loss that does not increase in case of the radical price change. Overall, the outcomes of the experiments advise us to focus on Random Forest, LSTM and GRU as the best-performing solutions for the given prediction problem.

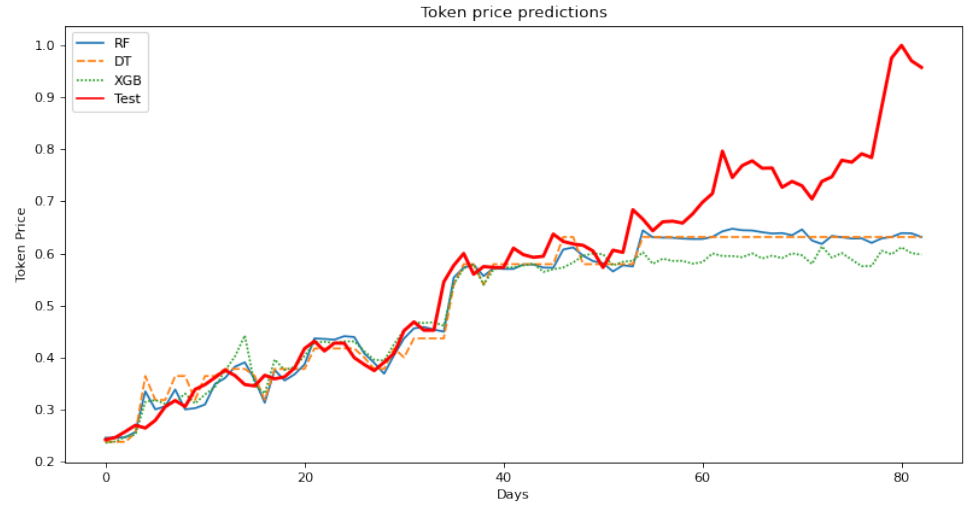



Figure 17: Binance Coin prediction results for Decision Tree (DT), Random Forest(RT), XGBoost (XGB)

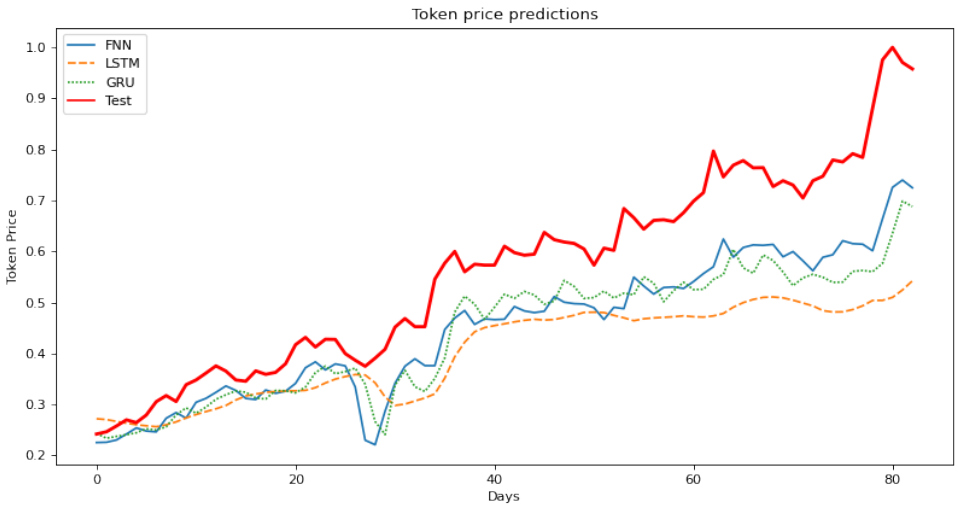



Figure 18: Binance Coin prediction results for Feedforward Neural network (FNN), Long short-term memory (LSTM), Gated Recurrent Unit (GRU)

In addition, to the main aim of this research we decided to analyse the accuracy of guessing the direction of token price movements based on the token price predictions below (Table 3). Here one might observe that tree-based algorithms outperform the neural network based algorithms. All in all, it can be inferred that the accuracy derived by price prediction results relatively corresponds to the models price forecasts.

**Accuracy of token price movement prediction**

| Token Name | Decision Tree | Random Forest | XGBoost | FNN | LSTM | GRU |
|---|---|---|---|---|---|---|
| BTC | 42.6% | 49.4% | 43.8% | 50.5% | 49.3% | 49.3% |
| ETH | 53.9% | 48.3% | 51.6% | 51.6% | 48.7% | 47.5% |
| BNB | 48.3% | 55% | 53.9% | 49.4% | 45.1% | 52.4% |

Table 3: Share of correct movement guess out of the daily price prediction for 80 consecutive days

All in all, we believe that the machine learning approaches taken in scope of this research showed promising results in token price prediction. The tree-based algorithms demonstrate relative low prediction loss in comparison to the neural networks. However, the neural network-based model demonstrates the ability to capture the spikes in price changes over unseen data whereas tree-based algorithms do not follow the trend which is crucial for the scope of our business problem. As, by design, such predictions results are to be used for the portfolio management activities and inability to capture the price spikes makes such tree-based model an infeasible approach. It is also worth mentioning that during our research we faced with problem of limited computation resources to experiment with more complex neural network architectures, which means that there is still space left for prediction improvements of LSTM and GRU prediction models in scope of the given business problem.

# 7 Conclusions

As it was already described in the Motivation and Goal Section 3 the internship engagement consists of 2 parts. The fist part consists of getting acquainted with blockchain and crypto projects that are born as a result of blockchain adoption in the business wold.

**1st Part: Due Diligence** In the first part of the engagement prior to conducting the prediction work we made a research on the digital finance projects and designed a scoring metrics to evaluate the maturity of the company. The scoring metrics included the team, marketing, business model, token economy and the financial analysis. This way we designed a due diligence model as a preparatory part of this report. It is taken in commercial use for onboarding a significant amount of projects brought to the company. Only in the time of conducting this research around 15 projects already went through due diligence designed by us.

As a result of this part of the internship engagement we provided a solution to the company needs of having a proper valuation metrics for onboarding startups in digital finance. As the service offering became inferrable from the due diligence report the business development department reduced time spent preparing the service offering letter to the clients. In addition, the due diligence provides an intuitive and transparent scores explanation to the projects evaluated with the scoring technique. This way, we reduced the requests for scores clarifications and smoothed out the onboarding process.

**2nd Part: Token Price Prediction** The second part of the engagement with the company was to explore the blockchain information to put more clarity on the token prices and their dependency. This part of activity was a pure research and not dedicated to any of the company's ongoing activities. As the blockchain is still emerging there is a limited resources available that can contribute to both the token economy development and investors decisions for crypto investment portfolio management. With our research we provided a selection of features that are identified as important correlation parameters for the token price predictions. The conducted work on feature selection provided an insights to the host company on which blockchain parameters to take into account when designing the token economy for the crypto projects. As a result of token price predictions we identified that the neural network models,(GRU and LSTM) are better suited for the business problem of token price forecasting as those are able to provide a stable prediction loss throughout the test period maintaining themselves robust to the price spikes. The application of such models could be used as an automated advisory system for the crypto investors and their crypto portfolio management activities.

As a result of our prediction studies we can conclude the following:

- What observation period should be taken into account while prediction the future token price?
  - Based on the result obtained in this report the daily token price resembles statistically significant prediction power up to 7-10th day in the past. That recommends that the prediction range of up to 10 days can be considered significant for predicting the future token price.
- Does each crypto token share the same on-chain explanatory variables with other crypto tokens?
  - Based on our research only previous day token price, mean transaction value and price volatility of last 30 days can be considered as a shared explanatory variables for

BTC, ETH and BNB. The other explanatory variables that are significant for some tokens almost do not overlap with the others. Hence, the on-chain feature selection shall be conducted individually for each token. We consider this to be an important investigation as there are research in this field ([12]) that use on-chain feature data equally as a black-box for each crypto token.

- Do the prices of other tokens posses a high explanatory power when predicting the token price?
  - As the result of the research there is a lack of evidence that the prices of other tokens can play a significant role in token price predictions.
- Considering the tree-based and neural network models can there be identified a favourite approach to predict the token price?
  - The report shows no favourite algorithm for the token price predictions in terms of constant forecast loss dominance. Even though the tree-based algorithms demonstrate relative low prediction loss in comparison to the neural network they appeared to be not resilient to the price spikes whereas the neural network models maintained stable loss along all the test period predictions. As the aim of the predictions to provide the advisory solution for the crypto portfolio management, the neural networks (LSTM and GRU) appear to be a more optimal decision for a given business need as the tree-based algorithms.

## 7.1 Possible improvements

Although we reached promising results with the token price prediction, we believe that there is a possibility for improvements, especially, with the neural network-based models. As some researches claim the news related to crypto project can be used as predictive feature and increase the prediction accuracy. In addition, we also believe that increasing the resource computational power can enable utilization of more complicated neural network models which could potentially increase the token price prediction accuracy.

# 8 References


[1] Satoshi Nakamoto. *Bitcoin: A Peer-to-Peer Electronic Cash System*. Accessed: 2015-07-01. 2008. URL: https://bitcoin.org/bitcoin.pdf.

[2] Ana Reyna et al. "On blockchain and its integration with IoT. Challenges and opportunities". In: *Future Generation Computer Systems* 88 (2018), pp. 173–190. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2018.05.046. URL: https://www.sciencedirect.com/science/article/pii/S0167739X17329205.

[3] Raynor de Best. *Number of crypto coins 2013-2022*. 2022. URL: https://www.statista.com/statistics/863917/number-crypto-coins-tokens/.

[4] Luis Oliveira et al. "To Token or not to Token: Tools for Understanding Blockchain Tokens". In: Oct. 2018. URL: https://www.zora.uzh.ch/id/eprint/157908/1/.

[5] Moon Soo Kim and Jee Yong Chung. "Sustainable Growth and Token Economy Design: The Case of Steemit". In: *Sustainability* 11.1 (2019). ISSN: 2071-1050. DOI: 10.3390/su11010167. URL: https://www.mdpi.com/2071-1050/11/1/167.

[6] Jeffrey M. Perloff. *Microeconomics (the addison-wesley series in economics) (February 2003 edition)*. 2003. URL: https://openlibrary.org/books/OL9335492M/Microeconomics_%28The_Addison-Wesley_Series_in_Economics%29.

[7] Volker Laux. "Stock option vesting conditions, CEO turnover, and myopic investment". In: *Journal of Financial Economics* 106.3 (2012), pp. 513–526. ISSN: 0304-405X. DOI: https://doi.org/10.1016/j.jfineco.2012.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0304405X12001183.

[8] Andrea Barbon and Angelo Ranaldo. "On the quality of cryptocurrency markets: Centralized versus Decentralized Exchanges". In: *SSRN Electronic Journal* (2021). DOI: 10.2139/ssrn.3984897.

[9] Vijay Mohan. "Automated market makers and Decentralized Exchanges: A defi primer". In: *Financial Innovation* 8.1 (2022). DOI: 10.1186/s40854-021-00314-5.

[10] Nishant Jagannath et al. "An On-Chain Analysis-Based Approach to Predict Ethereum Prices". In: *IEEE Access* 9 (2021), pp. 167972–167989. DOI: 10.1109/ACCESS.2021.3135620.

[11] Ana Todorovska et al. "Analysis of Cryptocurrency Interdependencies". In: (Nov. 2021). DOI: 10.7566/JPSCP.36.011004.

[12] David Zhao, Alessandro Rinaldo, and C. Brookins. "Cryptocurrency Price Prediction and Trading Strategies Using Support Vector Machines". In: *arXiv: Trading and Market Microstructure* (2019).

[13] Patrick Jaquart, David Dann, and Christof Weinhardt. "Short-term bitcoin market prediction via machine learning". In: *The Journal of Finance and Data Science* 7 (2021), pp. 45–66. ISSN: 2405-9188. DOI: https://doi.org/10.1016/j.jfds.2021.03.001. URL: https://www.sciencedirect.com/science/article/pii/S2405918821000027.

[14] Shaik Javed and Abishek. R Parvez. *Bitcoin price prediction using random forest regression*. 2022. URL: https://journalppw.com/index.php/jpsp/article/view/4115.

[15] Yiming Li. "The Price Prediction of Virtual Currency Base on Improved Support Vector Regression". In: *2021 4th International Conference on Information Systems and Computer Aided Education*. ICISCAE 2021. Dalian, China: Association for Computing Machinery, 2021, 2587–2591. ISBN: 9781450390255. DOI: 10.1145/3482632.3487476. URL: https://doi.org/10.1145/3482632.3487476.

[16] Mohammad Hamayel and Amani Owda. "A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms". In: *AI* 2 (Oct. 2021), pp. 477–496. DOI: 10.3390/ai2040030.

[17] Azim Muhammad Fahmi et al. "Regression based Analysis for Bitcoin Price Prediction". In: *International Journal of Engineering  Technology* 7.4.38 (2018), pp. 1070–1073. ISSN: 2227-524X.

[18] Vitalik Buterin. "A next generation smart contract  decentralized application platform". In: (2013). URL: https://www.weusecoins.com/assets/pdf/library/Ethereum_white_paper-a_next_generation_smart_contract_and_decentralized_application_platform-vitalik-buterin.pdf.

[19] Binance. In: *Binance whitepaper* (2020). URL: https://whitepaper.io/document/725/binance-whitepaper.

[20] David Freedman, Robert Pisani, and Roger Purves. "Statistics (international student edition)". In: *Pisani, R. Purves, 4th edn. WW Norton & Company, New York* (2007).

[21] Abhinandan Kulal. "Followness of Altcoins in the Dominance of Bitcoin: A Phase Analysis". In: *Macro Management  Public Policies* 3 (Sept. 2021). DOI: 10.30564/mmpp.v3i3.3589.

[22] Hossein Hamooni and Abdullah Mueen. "Dual-domain hierarchical classification of phonetic time series". In: *2014 IEEE international conference on data mining*. IEEE. 2014, pp. 160–169.

[23] Chris Aldrich. "Process variable importance analysis by use of random forests in a shapley regression framework". In: *Minerals* 10.5 (2020), p. 420.

[24] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[25] Rui Guo et al. "Degradation state recognition of piston pump based on ICEEMDAN and XGBoost". In: *Applied Sciences* 10.18 (2020), p. 6593.

[26] Hosni Ghedira and Monique Bernier. "The effect of some internal neural network parameters on SAR texture classification performance". In: *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*. Vol. 6. IEEE. 2004, pp. 3845–3848.

[27] Faisal Mohammad and Young-Chon Kim. "Energy load forecasting model based on deep neural networks for smart grids". In: *International Journal of System Assurance Engineering and Management* 11.4 (2020), pp. 824–834.

[28] Younes Ed-Doughmi, Najlae Idrissi, and Youssef Hbali. "Real-Time System for Driver Fatigue Detection Based on a Recurrent Neuronal Network". In: *Journal of Imaging* 6.3 (2020). ISSN: 2313-433X. DOI: 10.3390/jimaging6030008. URL: https://www.mdpi.com/2313-433X/6/3/8.

[29] Polash Dey et al. "Comparative Analysis of Recurrent Neural Networks in Stock Price Prediction for Different Frequency Domains". In: *Algorithms* 14 (Aug. 2021), p. 251. DOI: 10.3390/a14080251.

[30] Adnan Riaz et al. "SBAG: A Hybrid Deep Learning Model for Large Scale Traffic Speed Prediction". In: *International Journal of Advanced Computer Science and Applications* 11 (Jan. 2020), pp. 287–291. DOI: 10.14569/IJACSA.2020.0110135.

# 9 Annex

## 9.1 Correlation of features with next day token price

| Feature | BTC | ETH | BNB |
|---|---|---|---|
| SplyAdrTop100 | 0.242 | N/A | N/A |
| FeeMedNtv | -0.145 | 0.605 | N/A |
| NVTAdj90 | 0.192 | N/A | N/A |
| NVTAdj | 0.146 | N/A | N/A |
| LastDayPrice | 0.999 | 0.998357 | 0.990 |
| FeeByteMeanNtv | -0.139 | N/A | N/A |
| TxTfrValMeanNtv | -0.152 | -0.131 | -0.227 |
| AdrBalUSD10KCnt | 0.981 | N/A | N/A |
| ReferenceRateETH | -0.294 | N/A | N/A |
| SplyAct180d | 0.338 | N/A | N/A |
| RevHashRateUSD | -0.116 | N/A | N/A |
| VelCur1yr | -0.347 | N/A | N/A |
| BlkCnt | -0.175 | 0.333 | N/A |
| SplyActPct1yr | -0.289 | N/A | N/A |
| SplyAct90d | 0.267 | N/A | N/A |
| ROI1yr | -0.114 | N/A | N/A |
| AdrBal1in10KCnt | 0.379 | -0.209 | N/A |
| SplyAct7d | 0.121 | 0.565 | N/A |
| SplyAdrBalNtv100K | 0.416 | N/A | N/A |
| FeeMeanUSD | 0.516 | N/A | N/A |

Table 4: Correlation of features with next day price of BTC, ETH, BNB (I)

| Feature | BTC | ETH | BNB |
|---|---|---|---|
| VtyDayRet30d | -0.100 | -0.201 | -0.365 |
| AdrBalNtv1MCnt | N/A | 0.487 | 0.131 |
| TxTfrValAdjNtv | N/A | 0.174 | N/A |
| CapMVRVCur | N/A | 0.175 | N/A |
| NVTAdjFF | N/A | -0.174 | N/A |
| SplyAdrBalNtv1M | N/A | 0.663 | N/A |
| SplyAct1yr | N/A | 0.348 | N/A |
| PriceBTC | N/A | 0.575 | N/A |
| VtyDayRet180d | N/A | -0.274 | N/A |
| AdrActCnt | N/A | 0.483 | N/A |
| AdrBal1in1KCnt | N/A | -0.205 | N/A |
| TxTfrValMedNtv | N/A | -0.293 | -0.319 |
| AdrBalUSD1Cnt | N/A | 0.917 | N/A |
| SplyAdrTop10Pct | N/A | N/A | 0.237 |
| SplyAdrBal1in10K | N/A | N/A | 0.106 |
| AdrBalUSD100KCnt | N/A | N/A | 0.564 |
| TxTfrValMedUSD | N/A | N/A | 0.141 |
| AdrBal1in1MCnt | N/A | N/A | -0.125 |
| ROI30d | N/A | N/A | -0.201 |
| NVTAdjFF90 | N/A | N/A | 0.694 |

Table 5: Correlation of features with next day price of BTC, ETH, BNB (II)

## 9.2 Ethereum prediction results

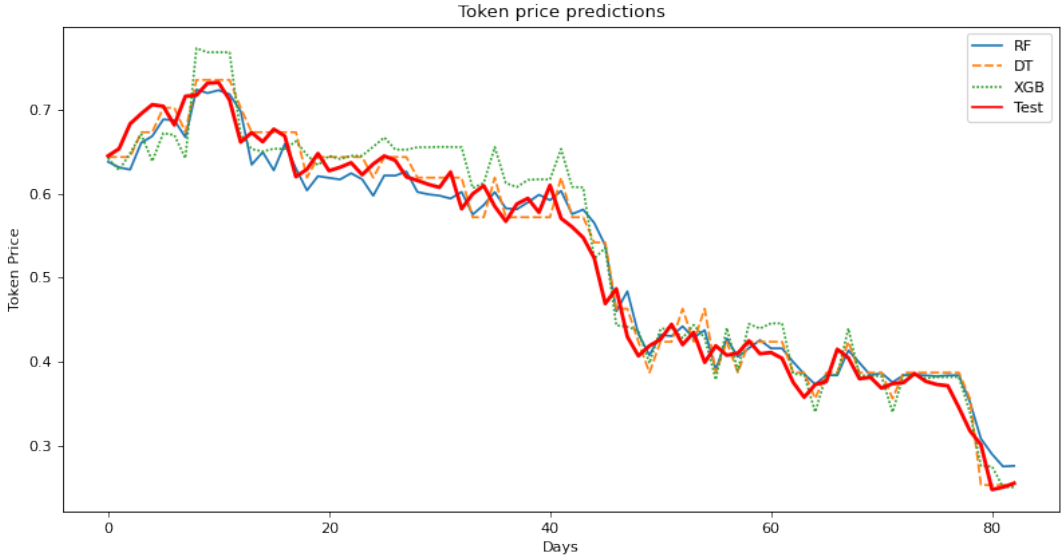



Figure 19: Ethereum prediction results for Decision Tree (DT), Random Forest(RT), XGBoost (XGB)

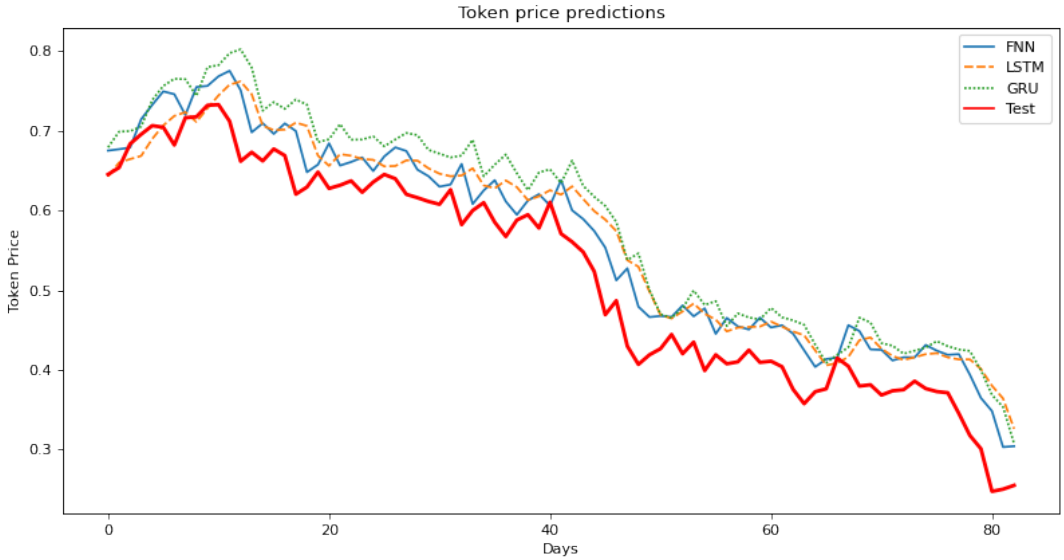



Figure 20: Ethereum prediction results for Feedforward Neural network (FNN), Long short-term memory (LSTM), Gated Recurrent Unit (GRU)