



NOTRE DAME UNIVERSITY (NDU)

**BITCOIN ESTIMATION AND PREDICTION  
THROUGH NEURAL NETWORK AND MACHINE LEARNING**

By

**Marianne G. Hanna**

Supervised by

**Dr. Re-Mi Hage**

Submitted to the

**Faculty of Natural and Applied Science**

A senior project submitted to Notre Dame University

In partial fulfillment of the requirement

For the degree of Master of Actuarial Science

In the Department of Mathematics and Statistics

July, 2021

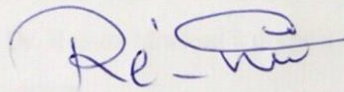
**Notre Dame University-Louaize,  
Zouk Mosbeh, Lebanon**

Faculty of Natural and Applied Sciences  
Department of Mathematics and Statistics

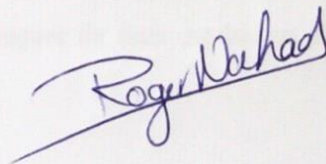
**Marianne Hanna**

We hereby approve the thesis titled Bitcoin Estimation and Predication through Neural Network and Machine Learning of the above candidate for the degree of Actuarial Science.

**Re-Mi Hage - Thesis Advisor**

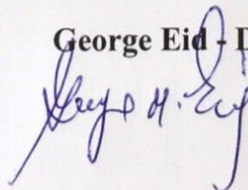


**Roger Nakad - Chair, First Reader**



This thesis is accepted by the Faculty of Natural and Applied Sciences

**George Eid - Dean**



## Acknowledgements

I would like to express my deepest appreciation to all those who gave me the opportunity to complete this Master thesis. I would like also to extend my sincere thanks to my advisor and professor Dr. Re-Mi Hage for her constant support, guidance and patience and for sharing with me her extensive knowledge during the taken courses and for providing me with the needed knowledge and skills, to help me continue and improve.

A great thank to Dr. Re-Mi Hage for her support and her guidance to all Actuarial Science students, and her help throughout all these years.

I would also like to extend my gratitude to Dr. George Eid: Dean of the natural and applied science, Dr. Roger Nakad: Chairperson of the faculty and every professor specially Dr. John Haddad, and Mrs. Claudia Freij Bou Nassif for their contribution and involvement to my thesis that could not have been successfully conducted without their efforts and experiences. To the spirit of Dr. Ramez Maalouf a deepest thanks for your empathy and every single information that you gave us.

I would also like to acknowledge and extend my heartfelt appreciation to my family, who believed in me, and supported me completely through my education, to all my close friends and colleagues for their continuous encouragement and feedback in completing this Master thesis.

Last but not least, without God's blessings I could not complete such work. Be grateful for all the obstacles in your life. They have strengthened you to pursue your journey.

## Abstract

Marianne G. Hanna: BITCOIN ESTIMATION AND PREDICTION THROUGH NEURAL NETWORK AND MACHINE LEARNING

(Under the supervision of Dr. Re-Mi Hage)

Bitcoin is the most popular purely digital cryptocurrency nowadays. It started in 2009 as a peer to peer payment system, decentralized, pseudo-anonymous and secure system of money. The main objective of this study is to estimate and predict the weekly close bitcoin price by including other variables like commodities, indexes and demand / supply variables through different kind of machine learning and deep learning such as multiple regression, time series ARIMA model, artificial neural network, combination of multiple regression and ARIMA model, and finally combination of multiple regression, ARIMA model and artificial neural network. The results show that the combination of multiple regression, Time Series ARIMA model, and artificial neural network is the most accurate model with an MAPE = 0.85% for the short term prediction of the close price of Bitcoin.

**Keywords:** Bitcoin, Machine Learning, Time Series, Multiple Regression, Neural Network.

# Table of Contents

Acknowledgements .....	iii
Abstract .....	iv
List of Figures .....	vii
List of Tables .....	ix
Chapter 1 .....	1
Introduction and definition .....	1
1.1 Introduction .....	1
Chapter 2 .....	7
Literature Review .....	7
2.1 Paper Review .....	7
Chapter 3 .....	14
Modeling and Methodology .....	14
3.1 Decision Tree .....	18
3.1.1 Gini Index .....	20
3.2 Multiple Regression .....	21
3.2.1 Regression Analysis .....	21
3.2.2 Least Square estimators .....	22
3.2.3 Fitted Values and Residuals .....	22
3.2.4 Analysis of variance .....	23
3.2.5 Model performance for multiple regression .....	25
3.3 Autoregressive Integrated Moving Average (ARIMA) Time Series .....	26
3.3.1 Autocorrelation and Partial autocorrelation function .....	27
3.3.2 Box-Jenkins (ARIMA) Models .....	28
3.4 Artificial Neural Network (ANN) .....	32
3.4.1 Type of Neural Network .....	36
3.5 Model performance .....	39
3.6.1 Box-Ljung test .....	39
3.6.2 Mean Absolute Percentage Error (MAPE) .....	39
3.6.3 The root Mean Squared Error (RMSE) .....	40
3.6.4 Sum of Squared Error (SSE) .....	40
3.6.5 Akaike Information Criterion (AIC) .....	40
3.6.6 Bayesian Information Criterion (BIC) .....	40

Chapter 4.....	41
Results and Discussion .....	41
4.1    Data presentation .....	41
4.2    Univariate and Multivariate variate analysis.....	43
4.2.1    Descriptive statistics and Boxplot.....	43
4.2.2    Bivariate analysis and Correlation Test .....	50
4.2.3    Multivariate analysis and Correlation plot .....	54
4.3    Modeling and machine learning.....	56
4.3.1    Regression Decision Tree .....	56
4.3.2    Multiple Regression .....	61
4.3.2.1    Assumptions for regression.....	61
4.3.3    Autoregressive Integrated Moving Average (ARIMA) or Time Series.....	75
4.3.4    Regression Time Series.....	84
4.3.5    Feedforward Artificial Neural Networks .....	89
4.3.6    Regression Time Series Neural Network .....	93
Chapter 5.....	101
Conclusion and Future Work .....	101
References.....	103

## List of Figures

Figure 1. 1: Blockchain Process.....	3
Figure 1. 2: How to get bitcoins.....	5
Figure 2. 1: Bitcoin daily Price Accuracy.....	13
Figure 2. 2: Bitcoin 5 minutes Price Accuracy.....	13
Figure 3. 1: Machine Learning Tree.....	15
Figure 3. 2: Machine Learning Steps.....	16
Figure 3. 3: Decision Tree Sample.....	19
Figure 3. 4: Forecasting steps in ARIMA.....	31
Figure 3. 5 : Brain shape of a Neural Network.....	33
Figure 3. 6: Sample representation of a Neural Network.....	34
Figure 3. 7: Cheat Sheet of Neural Network Shapes.....	35
Figure 3. 8: Flow of information FNN.....	36
Figure 3. 9: Feedback NN.....	36
Figure 3. 10: Recurrent NN.....	37
Figure 3. 11: Modular NN.....	37
Figure 3. 12: Convolution NN.....	37
Figure 3. 13: Bayesian Network.....	38
Figure 4. 1: Box plot of the weekly Close price and the weekly demand supply for bitcoin for the years 2019, 2020, and 2021. ....	46
Figure 4. 2: Histogram presenting the frequency distribution of the weekly Close price and the weekly demand supply for bitcoin for the years 2019 and 2020. ....	48
Figure 4. 3: Scatter plot presenting the weekly Close price and the weekly demand supply for bitcoin for the years 2019 and 2020.....	49
Figure 4. 4 presents the scatter plots of Close bitcoin with respect to the other variables are along with the coefficient of correlation and the P-value of the Pearson test.....	53
Figure 4. 5: Correlation Plot of each variable with the Close variables and the demand supply variables.....	55
Figure 4. 6: Tree of the variables.....	57
Figure 4. 7: The important variables of the Decision Tree.....	58
Figure 4. 8: Decision Tree Plot of the most important variables.....	60
Figure 4. 9: Correlation Plot of the weekly Close prices and demand/supply variables.....	63
Figure 4. 10: Cooks distance of the LMMOD (1).....	68
Figure 4. 11: Histogram of the residuals of selected variables.....	71
Figure 4. 12: The plot of the residuals vs the fitted value.....	71
Figure 4. 13: Partial fitted value plots.....	72
Figure 4. 14: Transformed Partial fitted value plots.....	73
Figure 4. 15 : Decomposition of multiplicative time series.....	75
Figure 4. 16: Time series plot (top), autocorrelation (left), and partial autocorrelation function (right) plots of variables.....	76
Figure 4. 17: Time series plot (top), autocorrelation (left), and partial autocorrelation function (right) plots of the first differenced variables.....	77

Figure 4. 18: Time series plot (top), autocorrelation (left), and partial autocorrelation function (right) plots of the second differenced variables.....	77
Figure 4. 19: Time series plot Residual of the ARIMA (2, 1, 2) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables of the first difference .....	79
Figure 4. 20: Forecast ARIMA (2,1,2) .....	80
Figure 4. 21: Arima future plots of the ARIMA (2,1,2) .....	80
Figure 4. 22: Time series plot Residual of the ARIMA (2, 1, 2) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables of the second difference.....	82
Figure 4. 23: Forecasting the ARIMA (4, 2, 2). .....	83
Figure 4. 24: Plots of the ARIMA (4,2,2) Future after forecasting the variables.....	83
Figure 4. 25: Time series plot Residual of the Regression ARIMA (2, 1, 2) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables for the first difference.....	85
Figure 4. 26: Regression Error versus the ARIMA Error (2,1,2) .....	86
Figure 4. 27: Time series plot Residual of the Regression ARIMA (4, 2, 2) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables for the second difference .....	87
Figure 4. 28: Regression Error versus the ARIMA Error (4,2,2) .....	88
Figure 4. 29: Neural Network of the weekly Close and demand/supply variables.....	90
Figure 4. 30: Neural Network of the most important variables of decision tree .....	91
Figure 4. 31: Neural Network of the significant variables of the multiple regression.....	92
Figure 4. 32: Histograms of the SSE and MAPE value for different NN.....	92
Figure 4. 33: Residual of the NNAR (1,1,2) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables.....	95
Figure 4. 34: NNAR (1,1,2) .....	95
Figure 4. 35: Residual of the NNAR (1,1,4) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables.....	97
Figure 4. 36: NNAR (1,1,4) .....	97
Figure 4. 37: Residual of the NNAR (1,1,6) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables.....	99
Figure 4. 38: NNAR (1,1,6) .....	99



## List of Tables

Table 1. 1: Pros of Bitcoin .....	4
Table 1. 2: Cons of Bitcoin .....	5
Table 1. 3: SWOT of Cross-Correlation .....	11
Table 3. 1: Decision Tree terminologies.....	19
Table 3. 2: Types of Neural Network .....	36
Table 4. 1: Brief variables definition .....	42
Table 4. 2: Minimum, First quartile (1st Qu), Median, Mean, Third quartile (3rd Qu), and Maximum value of the Close prices and for demand supply of bitcoins.....	47
Table 4. 3: Interval of the Coefficient Correlation .....	50
Table 4. 4: Pearson Correlation coefficient between the weekly Close Bitcoin price with the weekly Close variables and the weekly demand supply for bitcoin for the years 2019 and 2020.....	51
Table 4. 5: Pearson Correlation coefficient Test between the weekly Close Bitcoin price with the weekly Close variables and the weekly demand supply for bitcoin for the years 2019 and 2020.....	52
Table 4. 6: Correlation table of the weekly Close Price and demand/supply variables.....	62
Table 4. 7: the estimated coeficient, the test stastitcs and the p-value of the t-test of LMOD	65
Table 4. 8 R-squared, adjusted R-squared, MAPE, AIC and BIC of LMMOD .....	66
Table 4. 9: The results of the Close variables and demand/supply with the training data of the first trial.....	67
Table 4. 10: R-squared, adjusted R-squared, MAPE, AIC and BIC of the LMMOD (1) .....	67
Table 4. 11: The results of the Close variables and demand/supply of elimination of some values from the cooks distance .....	68
Table 4. 12: R-squared, adjusted R-squared, MAPE, AIC and BIC of the LMMOD (cook)..	69
Table 4. 13: The results of the Close variables and demand/supply with the new training data .....	69
Table 4. 14: R-squared, adjusted R-squared, MAPE, AIC and BIC of the LMMOD (S) .....	70
Table 4. 15: Table of the Studentized Breusch-Pagan test .....	71
Table 4. 16: Tukey test of the fitted value .....	73
Table 4. 17: The results of the Close variables and demand/supply with the new training data .....	74
Table 4. 18: R-squared, adjusted R-squared, MAPE, AIC and BIC of the LMMOD (F) .....	74
Table 4. 19: Box-Ljung test of ARIMA (2,1,2).....	78
Table 4. 20: Training set error of ARIMA (2, 1, 2) .....	78
Table 4. 21: Box-Ljung test of ARIMA(4,2,2).....	81
Table 4. 22: Training set error of ARIMA (4, 2, 2) .....	81
Table 4. 23: Box-Ljung test of Regression ARIMA (2,1,2) .....	84
Table 4. 24: Training set error of Regression ARIMA (2, 1, 2) .....	85
Table 4. 25: Box-Ljung test of Regression ARIMA (4,2,2) .....	86
Table 4. 26: Training set error of Regression ARIMA (4, 2, 2) .....	87

Table 4. 27: Box-Ljung test of Regression NNAR(1,1,2) .....	94
Table 4. 28: Training set error of NNAR (1, 1, 2) .....	94
Table 4. 29: Box-Ljung test of Regression NNAR(1,1,4) .....	96
Table 4. 30: Training set error of NNAR (1, 1, 4) .....	96
Table 4. 31: Box-Ljung test of Regression ARIMA (2,1,2) .....	98
Table 4. 32: Training set error of NNAR (1, 1, 6) .....	98
Table 4. 33: MAPE of the models .....	100

# Chapter 1

## Introduction and definition

### 1.1 Introduction

Financial experts acknowledged money as a tool for trading and exchange of goods, products and service and each nation has its own arrangement system of money (coins and paper cash). Bartering was the manner in which individuals traded merchandise and enterprises from the ancient periods on Earth but it wasn't always possible. If you grew rice, for instance, you could exchange sacks of rice for different merchandise and enterprises you required. People created commodities which are popular basic items as salt, tea, cattle and seeds to alleviate some of the bartering problems, but it raised other problems because it wasn't simple to ship and regularly they were transitory or hard to store. Money was created as precious metal to solve the issue of trading commodities. Later societies moved away from using precious metals to make money where the paper bills and coins made of non-precious metals represented certain values known as representative money (Fork, 2017). Credit cards were found to replace money with a chip technology that was created to store information and points. Using technology, mobile banking was created in the 90's. Recently, using most recent technology cryptocurrencies was introduced at around 2008 as an open source virtual currency with a peer to peer electronic cash system named "Bitcoin [฿]". In 2020-2021 cryptocurrency ATMs are now a part of our communities to enable people to convert cash into cryptocurrency, which is then stored in a crypto wallet on their phone. Using a QR codes, money can then be transmitted to someone else.

An anonymous programmer or a group of programmers known as Satoshi Nakamoto in 2009 claimed to be living in Japan, are the creator of Bitcoin (Bellis, 2019). The name of "Satoshi Nakamoto" was created depending on the four best ranking companies in 2009 SAMSUNG (Sa), TOSHIBA (Tochi), NAKAMICHI (Naka) AND MOTOROLA (moto). Nakamoto owns around one million Bitcoins, which value approximately \$3.6 billion as of September 2017.

Murray (2019) affirmed that Bitcoin [฿] is a mystery to many, but it might be the most discussed currency in the world defined as the first fully functional system of an open source distributed peer-to-peer programming software for the creation and exchange of a particular kind of cryptographic money. It's the new-age currency that only has an online presence and permits its user to be approximately anonymous.

Cointelegraph(website, 2017) claims that Bitcoin has some features restricted by being a fairly recent and somehow complex type of installment. It is considered as an extremely risky investment. Bitcoin turned into a prime open door for venture assignable to continuously fluctuating swapping rate. In spite of being unstable and to some extent unrecognized currency in the past years, it is being accepted nowadays from small local coffee shops to giants of industry.

Bitcoin is known to be the most valuable and expensive currency in the world . It is decentralized such as no banks or administration controls it. It is totally digital without any physical touch. Basically, it is tradable all through the world with no middle men, which implies it usage without the contribution of banks or clearinghouse. Each and every single trade is recorded in a public rundown called the blockchain. It is pseudo anonymous by the efficiency of its blockchains. Only your wallet ID are identified including the transactions stated (Murray (2019), Yellin, and al. (2018). Cointelegraph (website, (2017)) affirmed that bitcoins can't be printed and their number is limited to 21 millions ₿ only. It is designed to pay for merchandise and ventures, just like Euros or U.S. Dollars, but it is mined by computers performing complex mathematical equations.

Bitcoin was recorded in a public distributed ledger called “blockchain” presented as digital assets that serve the reward for a process known “mining” and can be exchanged for other currencies. They are divided into smaller units called satothis, where each satoshi worth 0.00000001 ₿. Bitcoin gives its users total control over their finances, as it is not being commanded by networks, by changing just one letter or number in a block of transactions, will also influence the mistake or fraud attempt that can be easily spotted and corrected by anyone of all of the following blocks as a result of being a public ledger. Figure 1.1 presents the representation of the blockchain process.

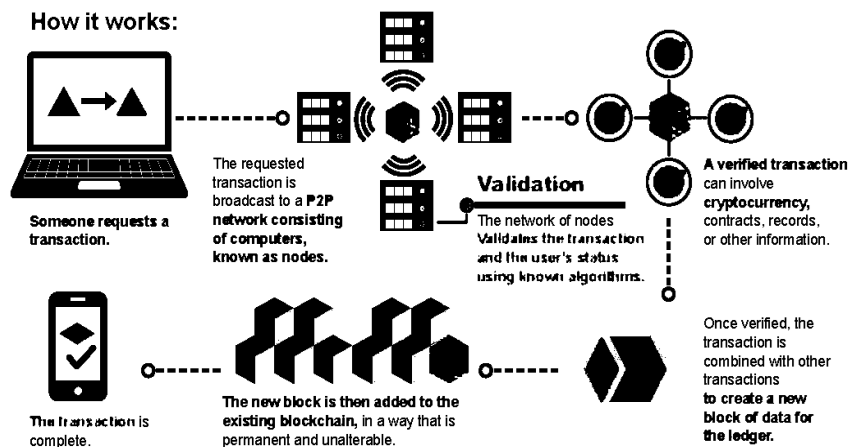


Figure 1. 1: Blockchain Process

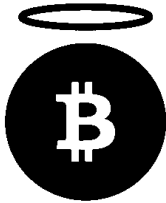
As described, first someone request a transaction, and then the requested transactions are broadcasted to a P2P network known as nodes. The validation of the nodes network is based on known algorithms where the verified transaction can involve contracts, records or other info of any cryptocurrency. Once verified, the transaction creates new square blocks of information for the record. The new squares are then added to the current blockchains in a manner that is lasting and fixed. After these steps, the transaction is completed.

Bellis (2019) stated that blockchains are designed in a way to be resistant to data modification where each block in the chain contains a cryptographic hash of the previous block, a timestamp, and transaction data. The Blockchains Revolution is in full swing and promises to shape our future fairly. Since then, cryptocurrencies have more than 1,600 unique names available online and the number continues to grow each day. Murray (2019) affirmed that bitcoin is a type of advanced digital currency that has a value with independence upon taxes collection, similarly like any normal cash money. Each one is a computer document file which is stored in a 'computerized wallet' application for smartphone or PC's. You can receive it through the digital app and send it to others. If you store bitcoins on your computer, you'll have to back up your computer regularly to prevent being hacked.

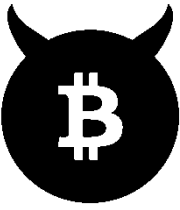
Yellin and al. (2018) and Murray (2019) stated that Bitcoin can be transferred between people, businesses or organizations by paying for goods and services at a lower charge rate anonymously all without the utilization of a bank. Small businesses may like Bitcoins because there are no credit card fees but some people just buy them as an investment, hoping that their value will increase.

As everything has its pros and cons, Bitcoin also has advantages and disadvantages detailed in table 1.1 and 1.2. The pros, listed and defined in the table 1.1, mainly are: freedom, high portability; commission, no Payment Card Industry (PCI), safety and control, transparent and neutral, and not counterfeit. Whereas, the cons listed and defined in table 1.2, mainly are: legal question, level of recognition, lost keys, volatility, and continuous development.

**Table 1. 1: Pros of Bitcoin**

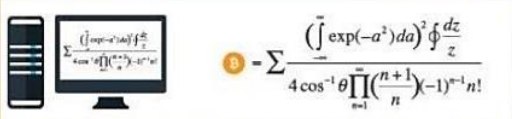
 <b>PROS</b>	<b>DEFINITION</b>
Freedom	Free of mind with no transaction
High Portability	Bitcoin is easy to carry and ready to use through online wallet
Commission	As a bitcoin owner you can choose the commission or choose not to pay the fee amount
No Payment Card Industry	As an owner you can't hold a card for bitcoin currency
Safety and Control	Every transaction is controlled and no one can steal your pay info
Transparent and Neutral	As the network is decentralized no one will control it
Not Counterfeit	You can't double spend as using the same money twice

**Table 1. 2: Cons of Bitcoin**

 CONS	DEFINITION
Legal Question	Legal status of bitcoin varies from country to another
Level of Recognition	How really the government know about and recognize the regulation of each country
Lost Keys	Essentially the key are for the password of the wallet, losing it means losing the wallet
Volatility	Due to the ups and downs the change is unpredictable
Continuous Development	The future of bitcoin is unclear and unregulated


## How to get Bitcoins


Bitcoin can get from digital world only, and it's have 3 ways for get it



$$\sum_{k=0}^{\infty} \frac{(\int_{-\infty}^{\infty} \exp(-a^2) da)^2 \oint \frac{dz}{z}}{4 \cos^{-1} \theta \prod_{n=1}^{\infty} \left(\frac{n+1}{n}\right) (-1)^{n-1} n!}$$

- 1** You can create Bitcoins through Bitcoins Mining, a process that involves running software on a computer to solve complex mathematical equations to generate a portion of the currency, if one of the equations is solved, then the payout is a Bitcoin.





- 2** You can get Bitcoins from selling something in online markets.
- 3** You can buy Bitcoins outright at various Bitcoin exchange markets.

**Figure 1. 2: How to get bitcoins**

Figure 1.2 present the three ways to get bitcoins:

1. People can make their transaction through computer process, than the latter solves incredibly difficult complex math puzzles. Once solved, the payout is a bitcoin, and transfers it just like money.
2. Sell things (let people pay you with bitcoin): Simply you sell anything and in return you get bitcoin not cash
3. Buy bitcoin: Buying a bitcoin ₿ depends on exchange markets

Bitcoin had a stormy year 2020-2021 that saw a meteoric rise in its value over the last six months, despite experts' critics, and regular warnings about its sustainability. It reached in mid-April, 2021 \$ 65,000, which has by far, been the best month by reaching a \$1 trillion market capital in just 12 years contrarily to Google which took 21 years, Amazon 24 years, Apple 42 years and Microsoft 44 years.

Bitcoin was worth around \$29,000 on December 31, 2020. It surpassed the \$40,000 mark in January, 2021. On February 8<sup>th</sup>, the announcement of Tesla caused the price of Bitcoin to rise steeply in which the cryptocurrency reached approximately \$57,000 and continued it's grown on March. As stated before, in mid of April, bitcoin touched its highest value \$ 65,000 from its beginning of creation in 2009. From that point until around May 10, Bitcoin persisted above \$58,000. On May 13, Musk's, a former Bitcoin enthusiast, created a new cryptocurrency called 'Dogefather' and caused a massive drop in its price, brings it down to around \$49,000 on May 14. While it was getting better, the Chinese government's move allocated another massive disappointment to the currency, which fell to around \$30,000 on May 19, nearly its value on December 31, 2020. In June, 2021 this digital asset reached its lowest value approximately \$ 29,000 since the beginning of 2021, as CoinDesk data indicates. After failing, Bitcoin quickly recovered and once again reached \$34,000 approximately till the day writing this paper.

The bitcoin price prediction was a topic of many researchers. Many have tried to predict its price as a time series model, other have used machine learning model. Recently, deep learning has been used to predict the future price of bitcoin. The main challenge of bitcoin exchange rate is its high rate of price fluctuation described in the previous section. Since forecasting is necessary to tell about the future price of bitcoin, the latter is the objective of this study by improving the efficiency and the accuracy of the prediction using machine learning and deep learning.



## Chapter 2

### Literature Review

#### 2.1 Paper Review

God created Adam and Eve and from them people on earth are born with different minds and way of thinking. Each person will choose any subject according to his/her education, norms, culture and tradition, but a true researcher chooses the subject that add to his knowledge and information regardless of how it can affect on his society and environment.

In general, researches are putting efforts on having its paper published after getting enough information and data where every single paper has its own arrangement of philosophies and methodologies on a specific subject. In other words, this topic related to bitcoin uses many distinct methodologies by applying them on useful public data with different algorithms related to independent view and perspective.

As authors of Bitcoin price prediction using ARIMA model paper Fiaidhi, and al.(2020) from Lakehead university affirm that ARIMA is one of the most effortless and effective machine learning algorithms to perform time series estimating by mixing Auto Regression and Moving average to explore the conventions of stationary ARIMA model in forecasting Bitcoin costs and make it accessible to public by utilizing High charts library with the help of web services. To look over this study, we need to have the right choice of ARIMA (p, d, q), the length of time window of forecasting and the choice of the seasonality of bitcoin costs. Seven features are available to test the model by getting dataset from the crypto-currencies website: <https://www.coinmarketcap.com/> over the year 2013 till 2019 for daily and yearly basics where exactly date, open, high, low, close, volume and market capital are needed as features. As the series is not stationary where it varies with time, we need to make it stationary by using its log function to transform it. We take the observation of a specific interval of time and subtract the previous instant. To find the difference in seasonality and testing the stationary of the data we need:

$$\begin{aligned} & \text{timeseriesdifflgtransform} = \\ & \text{logtimeseriestransformed} - \text{logtimeseriestransformed.shift} \end{aligned} \quad (1)$$

Autocorrelation and partial autocorrelation function are needed to determine  $q$  and  $p$  respectively. Mean Square Error are needed to be at its lowest when choosing the ideal model by trying many models combinations like ARIMA (2, 1, 0) when the moving average is  $q=0$ ,  $d=1$  for stationarity and  $p=2$  as autoregressive function and another ARIMA model having (0, 1, 18) when the moving average is  $q=18$ ,  $d=1$  for stationarity and  $p=0$  as an autoregressive function to have a lesser MSE value. Finally, it is found that ARIMA (8, 1, 0) when the moving average is  $q=0$ ,  $d=1$  for stationarity and  $p=8$  as autoregressive function is the best fitting model using it on a seven day bitcoin closing dataset from January 1<sup>st</sup> to January 7<sup>th</sup> 2020 to test it. The MSE value is equal to 170,962.195 which are lower by trying also other order values of ARIMA model parameter. This paper used Python visual studio 2017 and apply it to find the MSE and the best fitting model on the dataset chosen.

The authors of Bitcoin price prediction using ensembles of neural networks Sin & Wang (2017) affirms the next day pricing level change of Bitcoin with its features through ANN approach called Genetic Algorithm based Selective Neural Network Ensemble (GASEN). This paper retrieved its 190 time series data from Bitcoinity.org as date range from May 2, 2015 till April 30, 2017 (730 days) and May 1<sup>st</sup>, 2017 till June 20, 2017 (50days) for a back testing with lots of features as Block size Votes on each exchange, confirmation time in each exchange, block version of each exchange, block size of each exchange, difficulty in each exchange, Hash Rate of each Exchange, Transaction Count in each Exchange, Arbitrage in each Exchange, Bid-Ask Sum in each Exchange, Book Value in each Exchange, Market's Capitalization in USD for each Exchange, Market's Capitalization for each Exchange, Price in each Exchange, Price-Volume of each Exchange, Rank of each Exchange, Spread of each Exchange, Trades per Minute in each Exchange, Volatility of each Exchange and Volume of each Exchange. As a result the accuracy is different for each training dataset with a range between 58 and 63 % with an 85 % return, whereas for the back testing 64% with 32 correct predictions out of 50.

The authors of Autoregressive Integrated Moving Average Model based Prediction of Bitcoin Close Price Anupriya and Garg (2018) stated that different types of analysis can be applied on any cryptocurrency as the statistical, empirical, SWOT and time series analysis. As going through the paper the ARIMA function is used under R programming languages where the data is between January 1<sup>st</sup>, 2015 and September 23, 2018 from the coindesk.com with four features: open, low, high and close price divided into two part training and testing data with a 60:40 ratio. Mean and percentage error is calculated by forecasting the data from the result that shows 60-70% as accuracy.

The authors of Bitcoin value analysis based on cross-correlations Papadopoulos & al. (2017) starts by defining bitcoin as a mining process which the money elements are based on mathematical properties listed as durability, portability, scarcity, and divisibility other than the physical properties as gold or silver or other central authorities as fiat currencies. The dataset is divided between many factors as the rapid growth of bitcoin transaction per month from year January 2009 till January 2015 retrieved from Wikipedia website and the evaluation of bitcoin with USD rates from July 2013 till July 2015 as weekly basis retrieved from

<http://bitcoincharts.com/charts/bitstampUSD#rg730zigWeeklyztgTzm1g10zm2g10zl>.

The economic factors also influence on bitcoin as the gold and crude oil or the major stock market indices by having a cross-correlation with the prices and differ between short and longtime series analysis. For a discrete time as inner product the cross-correlation applies on this equation:

$$C_{xy}(\tau) = \sum_{n=-\infty}^{\infty} x(n) \cdot y(n + \tau) \text{ for } \tau = 0, \pm 1, \pm 2, \dots \quad (2)$$

with a deterministic signals  $x(n)$  and  $y(n)$  with lag  $\tau$ . The discrete time for normalized stochastic processes of two processes  $x(n)$  and  $y(n)$  by the equation

$$C_{xy}(\tau) = \frac{E\{(x(n) - \mu_x) \cdot (y(n + \tau) - \mu_y)\}}{\sigma_x \sigma_y} \text{ for } \tau = 0, \pm 1, \pm 2, \dots \quad (3)$$

where  $E\{\}$  is the expected value and each variable  $x$  and  $y$  has its own mean  $\mu$  and standard deviation  $\sigma$ . The time averages across observation with an  $N$  the number of samples similar to equation 2

$$\tilde{c}_{xy} = \frac{\frac{1}{N} \sum_{n=1}^{N-\tau} (x(n) - \bar{x}) \cdot (y(n + \tau) - \bar{y})}{s_x s_y} \text{ for } \tau = 0, \pm 1, \pm 2, \dots, \pm(N - 1) \quad (4)$$

where the means and standard deviation are as follows:

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x(n) \quad \& \quad s_x = \frac{1}{N-1} \sqrt{\sum_{n=1}^N (x(n) - \bar{x})^2} \quad (5)$$

$$\bar{y} = \frac{1}{N} \sum_{n=1}^N y(n) \quad \& \quad s_y = \frac{1}{N-1} \sqrt{\sum_{n=1}^N (y(n) - \bar{y})^2} \quad (6)$$

For the dataset the URLs used are <http://bitcoincharts.com/> & <https://blockchain.info/pl/charts> using 1,623 observation per days from August 17, 2010 till January 25, 2015 and omitting the zero values by using Mat lab R2014a with three types of features listed respectively as the price of bitcoin, the number of bitcoin transactions and the bitcoin transaction fees.

- Price of bitcoin & number of transaction with a lag  $\tau=0$  gives a linear relation
- Price of bitcoin & number of transaction fees with a lag  $\tau=300$  gives a linear relation
- Number of transaction & number of transaction fees with a lag  $\tau=400$  gives a Strong connection.

Other factors are also applied for cross-analysis from some indexes NASDAQ, DAX, S&P500, the gold price and crude oil price has another dataset retrieved from <http://www.investing.com/> from August 29, 2010 till January 25, 2015 as a weekly basis and applying cross-correlation between the bitcoin price and each of the factors above for a  $c(0)=1$ . For the first 3 cases as the indexes the lag  $\tau=0$  and for a normalized value should be above 0.6 where it's revealed that a strong connection is found between the bitcoin price and the major stock market. For the gold price the lag  $\tau=95$  weeks and the crude oil has a lag  $\tau=12$  weeks with a 0.8 normalized and also a strong correlation is found between both the gold and oil prices. The SWOT analysis is presented on a table 1.3 and each sub-analysis contains different characteristics (few listed) and application on the methods factors above.

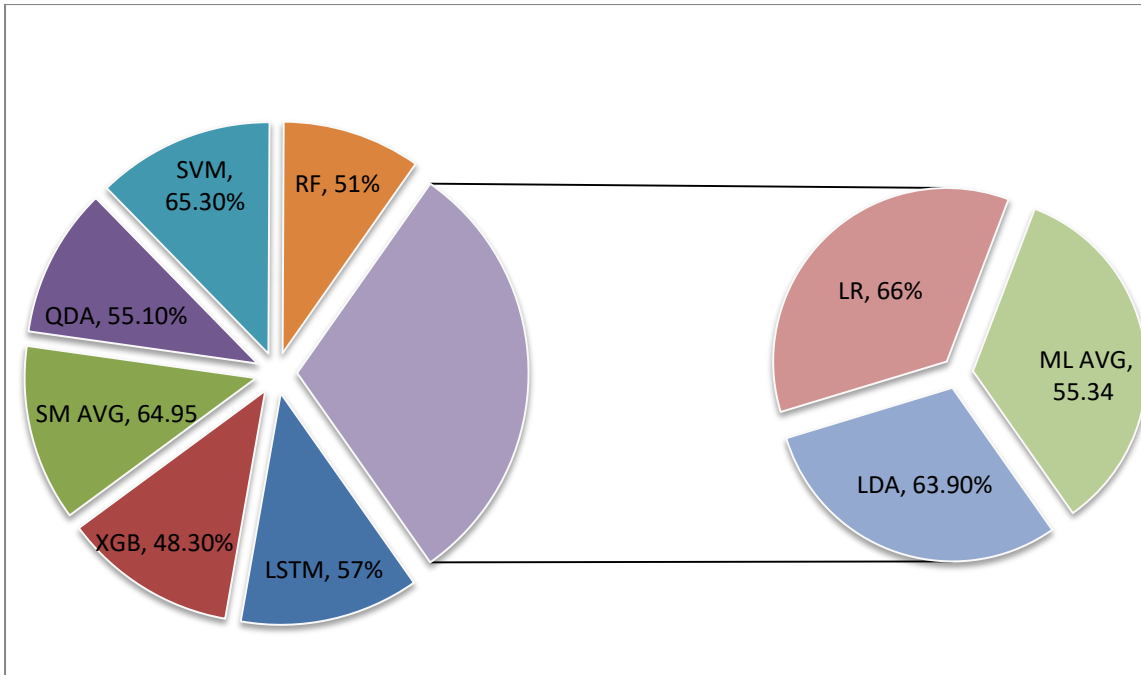
**Table 1. 3: SWOT of Cross-Correlation**

<p><b>Internal Origin</b></p>	<p><b>S</b></p> <ul style="list-style-type: none"> <li>• Worldwide use</li> <li>• Increasing number of users</li> <li>• Lack of brokers</li> <li>• Low transaction costs</li> <li>• Transaction speed</li> <li>• Protection of personal data of all participants</li> </ul>	<p><b>W</b></p> <ul style="list-style-type: none"> <li>• Highly dependent on participants trust in system</li> <li>• High value fluctuations</li> <li>• Decreasing reward for users providing computing power to the system</li> </ul>
<p><b>External Origin</b></p>	<p><b>O</b></p> <ul style="list-style-type: none"> <li>• Strong cross-correlations between the numbers of bitcoin transaction and transaction fees and the bitcoin price</li> <li>• Very good cross-correlations between the bitcoin price with gold and crude oil price</li> <li>• Correlation of bitcoin price with contemporary stock market indices NASDAQ, DAX and S&amp;P500.</li> </ul>	<p><b>T</b></p> <ul style="list-style-type: none"> <li>• Exchange rate is impossible to forecast with standard methods</li> <li>• Bitcoin value shows no signs of periodicity</li> <li>• Number of users</li> <li>• Transaction directly affect bitcoin price</li> </ul>

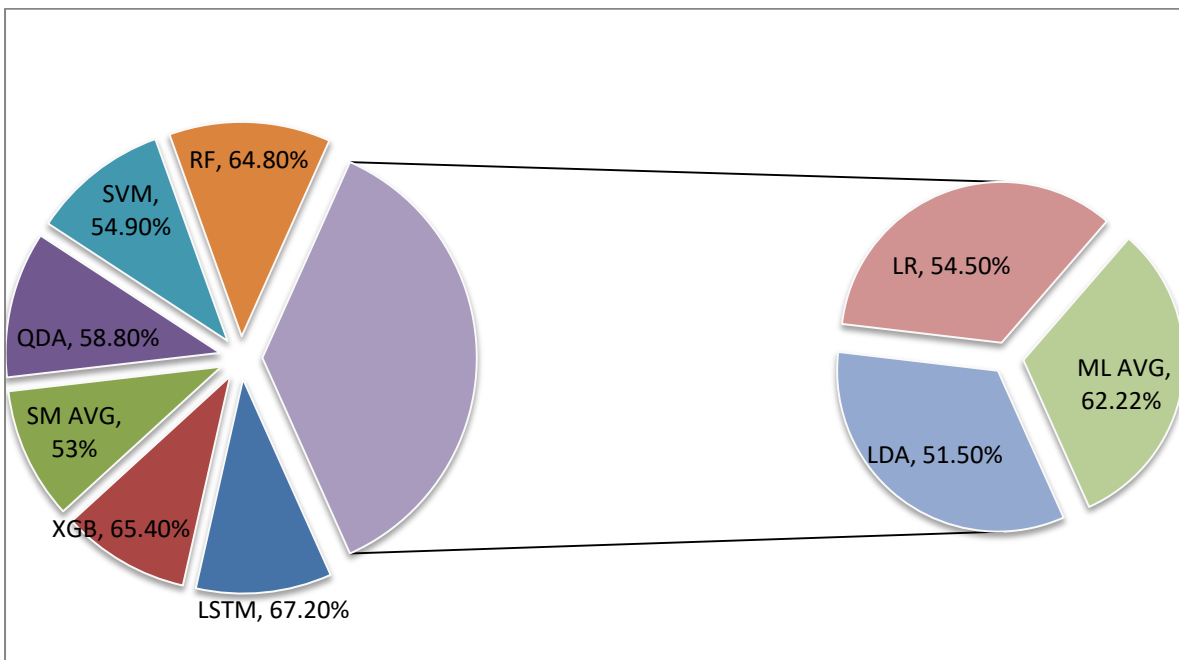
As shown, the prediction of the future period of bitcoin price can be based on standard models but the factor applied stays the same to forecasting the bitcoin price, number of transactions and transaction fees for the future.

The authors of Bitcoin price prediction using machine learning: An approach to sample dimension engineering Chen, Li and Sun (2019) affirms the definition of bitcoin use two of the analysis with their algorithms. Statistical Methods contains Logistic regression (LR) and Linear Discernment Analysis (LDA) for daily bitcoin price and Machine Learning contain Random forest (RF), decision tree boosted (XGBoost), Quadratic Discernment Analysis (QDA), Support vector machine (SVM) and Long short term memory (LSTM) used for bitcoin high frequency trading price. Each algorithm is defined and was applied to find the forecasted data where the daily price include more features divided into categories like property and network data, trading and market data, media and investor listed as block size, hash rate, mining difficulty, number of transaction confirmed transaction per day, mempool transaction count, mempool size, market capitalization, estimated transaction value, total transaction fees, google trend and gold spot price from a period of 2 years from February 2<sup>nd</sup>, 2017 till February 1,2019 from CoinMarketCap.com. Whereas, the 5-minute interval bitcoin from Binance as its data range from July 17,2017 till January 17,2018 with trading records as price, trading volume, open, close, high and low points. After calculating the accuracy, the fitting will be choosing between the two groups as SM and ML and through the accuracy of each it to have the perfect model for the two categories where in other words two best models for daily price and two for the 5 minute interval. As shown in the figure 2.1 the model chosen by the accuracy of the Daily bitcoin price are LR with 66% for SM category and SVM has 65.3% accuracy for ML category. For a 5 minute interval none of the algorithms fit as SM category whereas in the ML category LSTM is the fitting model as its accuracy equal 67.2%.

The data sets applied on the algorithm are divided into 2. The first data are based on bitcoin daily price and the second related to the 5 minute interval bitcoin. These equations differ between the bitcoin daily price and the 5-minute interval and are divided into 2 pie charts, to compare between the best models. From the figure 2.1 and 2.2 we can see the results of the best model where the average of a statistical method is 65% higher than the average accuracy of the machine learning is 55.3%. In the figure 2.2 the results for the best model with a 66% is the LR model in the SM and in ML the worst is XGB with 48.3% and the best is SVM with 65.3% accuracy competing with SM. Whereas in the figure 2.2 the results show the ML accuracy of 62.2% achieved a better accuracy than SM average of 53% by the LSTM with 67.2% accuracy regarding the ones in the SM that are 53% as exactly 54.5% for LR and 51.5% for LDA.



**Figure 2. 1: Bitcoin daily Price Accuracy**



**Figure 2. 2: Bitcoin 5 minutes Price Accuracy**

## **Chapter 3**

### **Modeling and Methodology**

Although there are several well-known Bitcoin price models but they are debatable. The best fitting model under same data is selected by comparing the accuracy and accepting the one with the highest accuracy.

Many models dealt with the bitcoin where some of them are defined and applied to get satisfying results. SAS Institute (2019) and Wakefield (2019) defines Machine Learning (ML) as a part and a sub type of artificial intelligence respectively referred to predictive modeling. SAS (2019), states that machine learning is a technique of diagnostic information. It's an automated diagnostic model structure that breaks down and analyses data based on the framework possibilities to gain info. It also distinguishes examples and settles on choices with a minimum human intercession, to foresee output values inside a satisfactory range.

Through the changing in technology, ML differs between today and past days by applying a hypothesis that computers can learn without being programmed through pattern recognition, to perform specific tasks and learn from the data. As new data are presented, ML can freely adjust and adapt with the iterative part of it by learning from previous calculations to create reliable choices and results. This science is not new, but as a result applying complex numerical mathematical calculations to large data in a faster and quicker way, and is the ongoing of recent developments. Moreover, to create a good ML system, some item is required as data preparation capabilities, algorithms in different types basic and advanced, automated and iterative processes, scalability and ensemble modeling. ML was applied to large data such as financial services, government, health care, retail, oil and gas, transportations etc.

As shown in figure 3.1 and as Wakefield (2019) affirmed in his article with the help of SAS Institute (2019), ML is divided into many methods: the supervised, unsupervised semi supervised and the reinforcement learning. Each one is divided into sub categories with many applicable algorithms. In this study, Decision Tree and Multiple Regression are used as supervised ML, Time Series as unsupervised, and Neural Network as reinforcement ML. A combination of the three methods is also applied. In this chapter, each ML method used in this study will be detailed and explained.



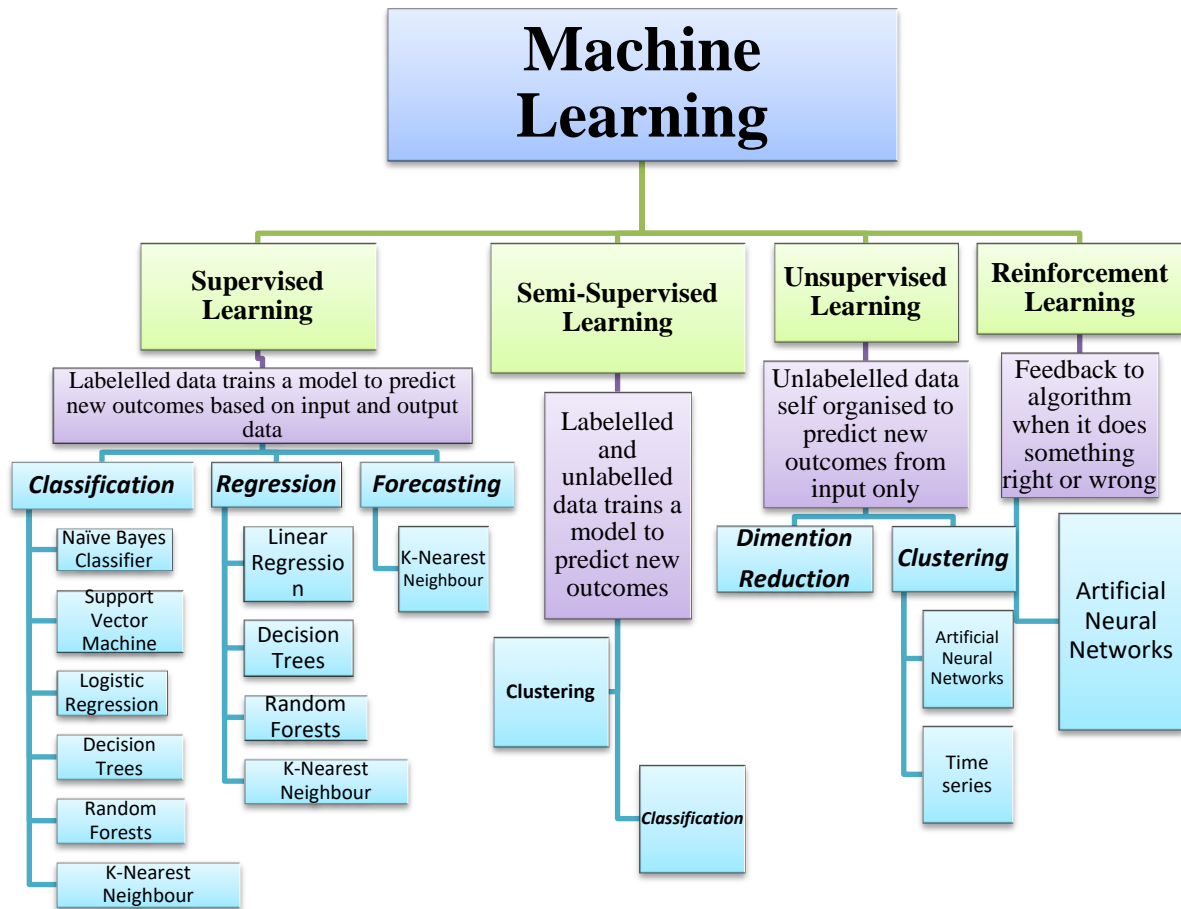
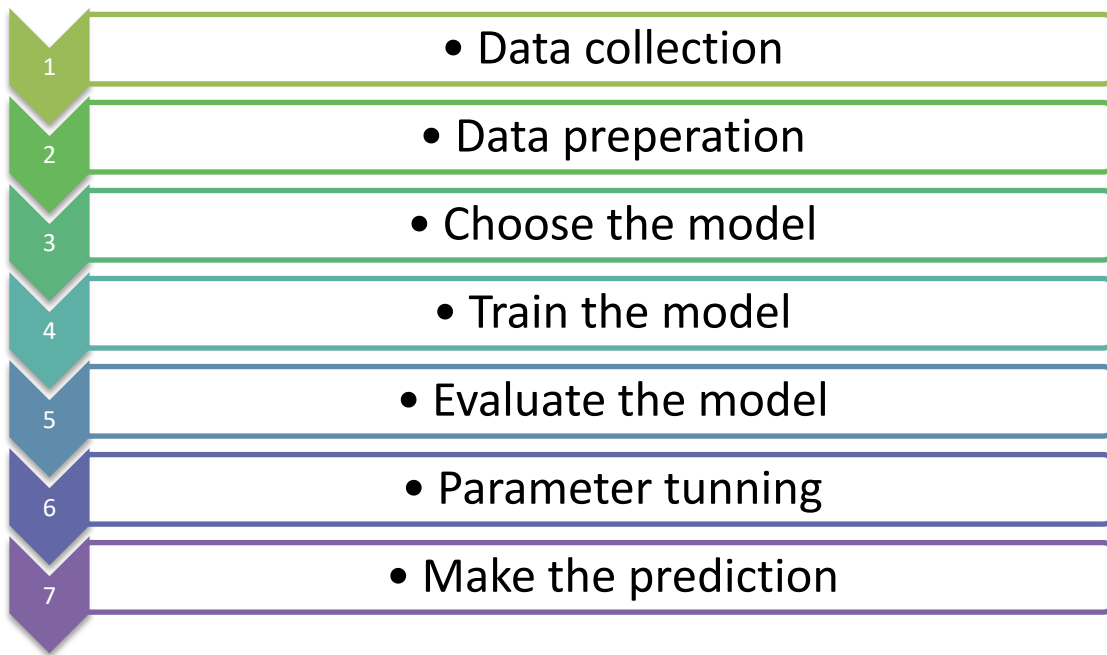


Figure 3. 1: Machine Learning Tree



**Figure 3. 2: Machine Learning Steps**

The Machine learning steps are:

1. Data Collection:

In which the outcome of this step is generally a representation of data .

2. Data Preparation:

First cleaning the data is require in order to remove duplicates, correct errors, deal with missing values, normalization, etc. Second, sometimes data are randomized or scaled, which erases the effects of the particular order. Finally the data are splitted into training and test data for evaluation

3. Choose a Model:

There different algorithms, thus it's important to choose the right one

4. Train the Model:

The goal of training is to answer a question or make a prediction correctly as often as possible

5. Evaluate the Model:

Use some metric or combination of metrics to "measure" objective performance of the chosen model by testing the model against previously unseen data.

6. Parameter Tuning:

Model parameters are tuned to improved performance and may include: number of training steps, learning rate, initialization values and distribution, etc.

7. Make Predictions:

The test set data which have are now used to test the model in order to determine of how the model will perform in the real world.

### 3.1 Decision Tree

Decision tree is a widely used data mining approach under supervised category of machine learning algorithms that can do both regressions and classification known as the Classification And Regression Tree - CART announced in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone.

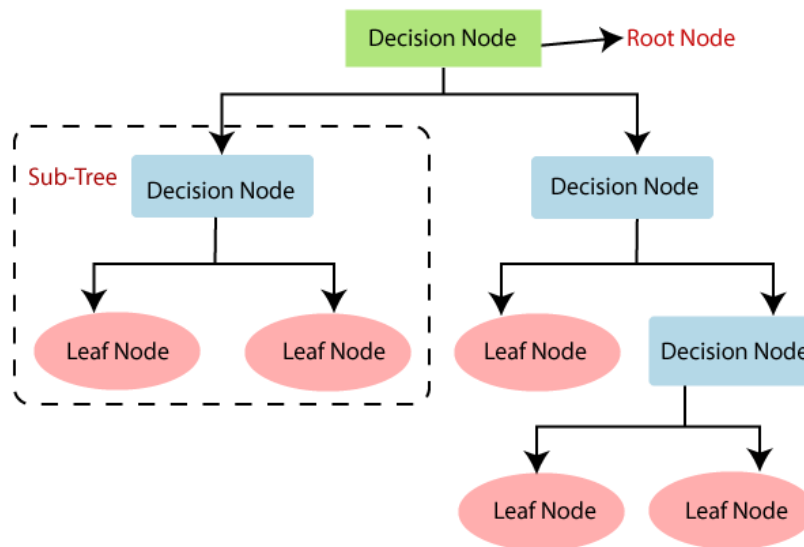
A classification tree is an algorithm with a numerical or categorical target variable used to predict which "class" a target variable belongs to. When a dataset needs to be divided into classes that correspond to the response variable, classification trees are used. The classes Yes or No are frequently used. A classification tree divides a dataset into subsets based on data homogeneity. In the other hand, when the target is to forecast the value of the variables, the regression tree technique is used. Each of the independent variables is used to fit the regression model. After that, each independent variable's data is separated at numerous locations. Finally, the difference between predicted and actual values is squared at each location to yield "A Sum of Squared Errors" (SSE). The SSE of the variables is compared, and the variable or point with the lowest SSE is selected as the split point. This process is repeated indefinitely.

This tree works with both categorical and continuous input and output variables. Categorical Variables have a limited target variables belonging to a specific group, whereas, the continuous variables have target variables with values found in variety of data types.

Decision tree is used to clarify, map out, and identify a solution to a difficult situation. In banking, investment, and business, decision trees are often used to determine a course of action. It is also known as tree diagrams in mathematics. A predicted consequence, possible option, or reaction is illustrated by the branches in a tree diagram. The forecast or result is displayed on the decision tree's last branch called "node". Each end product has a risk and reward amount assigned to it. When a person exploits a decision tree to make a decision, they can observe the benefits and cons of each final option. Decision trees are commonly employed to tackle problems that are too difficult to solve manually. It is comprised internal nodes in the training procedure representing the tests on each characteristic, branches demonstrating the conclusion of the method, and leaf nodes (terminal) signifying the class labels.

The root node is the tree's highest node. The branches usage indicates mutually exclusive possibilities where the structured model allows the viewer to comprehend how and why one choice may lead to the next. Users can custom the structure to display numerous alternative solutions to a problem in a clear and easy-to-understand arrangement in a way to demonstrate the relationship between distinct events or decisions.

The figure 3.3 illustrates the decision tree model with its terminologies.



**Figure 3. 3: Decision Tree Sample**

The decision tree terminologies are presented in the table 3.1 with their brief description.

**Table 3. 1: Decision Tree terminologies**

<b>Terminologies</b>	<b>Description</b>
Root node	Starting point of any decision tree and represents the entire data set.
Leaf node/Terminal node	Final output node
Splitting	Dividing the decision node/root into sub-tree
Sub-tree/branch	Formation of branches after splitting
Pruning/bagging	Processes of removing unwanted branches
Parent/ Child node	Parent node is the root node and the other nodes (decision/leaf) are child node

The pseudocode algorithm of the CART decision which will be our main concern in this study is as follows:

1. *Start by the root node.*
2. *Convert the ordered variables to unordered variable.*
3. *Perform a test to show the significant probability of the data node.*
4. *Choose the variable with the smallest significance probability*
5. *Find the split set that minimizes the sum of the Gini indexes and use it to split the node into child nodes.*
6. *If a stopping criterion is reached, exit. Otherwise apply more steps to each child node.*
7. *Prune or bag the tree with the CART.*

### **3.1.1. Gini Index**

The Gini index is a measure of misclassification that is used in the CART algorithm when the data contains multiple class labels. A low Gini index attribute should be preferred over a high Gini index. It only generates binary splits in the CART algorithm. The Gini index can be calculated using the equation:

$$Gini(y, S) = 1 - \sum_{c_j \in dom(y)} \left( \frac{|c_j|}{|S|} \right)^2 \quad (7)$$

Where S= total number of samples,

Y= target feature.

## 3.2 Multiple Regression

### 3.2.1 Regression Analysis

In statistics and machine learning, linear regression is one of the most well-known and well-understood algorithms because of its simple representation. The simple linear regression describes the linear relationship between the dependent variable  $y$  and one independent variable  $x$  represented as follow:

Regression analysis is a statistical tool that uses the relation between variables so that a response variable or dependent variable (often referred to as  $y$ ) can be predicted from the others variables called explanatory variables or independent variables (often referred to as  $x$ ), as well as their interactions (Neter, Kutner, Nachtsheim, & Wasserman, 1996). Only a certain number of independent or predictor variables (that are significant for the purpose of the analysis) should be included in a regression model. A multiple regression model may be defined as

$$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i \quad (8)$$

Where

- $\beta_0, \beta_1, \dots, \beta_{p-1}$  are model parameters:  $B_0$  and  $B_i$  called intercept and slope respectively
- $X_{i0} = 1$
- $X_{i0}, \dots, X_{i,p-1}$  are known constants that represent  $p - 1$  different independent variables
- $\varepsilon_i$  are independent  $N(0, \sigma^2)$  the error term (residual error).
- $i = 1, \dots, n$

$Y, X, \beta$  and  $\varepsilon$  could be written in matrix form

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{11} & X_{12} & 1 & \dots & 1 & X_{1,p-1} \\ 1 & X_{21} & X_{22} & 1 & \dots & 1 & X_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & 1 & \dots & 1 & X_{n,p-1} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \text{ and } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Where

- $Y$  is the responses vector
- $\beta$  is the vector of parameters

- $X$  is the matrix of constants
- $\varepsilon$  is a vector of independent normal random variables with expectation  $E\{\varepsilon\} = 0$  and variance-covariance matrix  $\sigma^2\{\varepsilon\} = \sigma^2 I$ . Consequently, the random vector  $Y$  has expectation  $E\{Y\} = X\beta$  and the variance-covariance matrix  $\sigma^2\{Y\} = \sigma^2 I$ .

### 3.2.2 Least Square estimators

Let the vector of estimated regression coefficients be  $b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$  then least square

estimators are given by  $b = (XX)^{-1}XY$ . They are the maximum estimated likelihood having the following properties: unbiased, maximum variance unbiased, consistent, and sufficient (Neter et al., 1996).

### 3.2.3 Fitted Values and Residuals

Examining the residuals, or the variations between the actual and expected values, is a good way to assess the model's fit efficiency. The goal is for the number of the residuals to be close to zero or as low as possible. Most situations will not follow a completely straight line in real life, so residuals are to be anticipated. Therefore, the median should be close to zero with a minimum and maximum approximately equal in absolute value. RSE represents the average variation of the observation points around the fitted regression line. It is known for model sigma or the standard deviation of residual errors. In comparing two models, the lowest RSE define the significance of the model.

The fitted values are represented by:

$$\hat{Y} = Xb \quad (9)$$

where the residual terms by:

$$e = Y - \hat{Y} = Y - Xb \quad (10)$$

The hat matrix ( $H$ ) is defined as

$$H = X(XX)^{-1}X \quad (11)$$

In the variance-covariance matrix of residual

$$\sigma^2\{e\} = \sigma^2(I - H) \quad (12)$$



and estimated by

$$s^2\{e\} = MSE (I - H) \quad (13)$$

where MSE designates the mean squared error.

### 3.2.4 Analysis of variance

The analysis of variance is performed using:

(1) Coefficient of determination  $R^2$

R-squared defines the variation of the data that can be clarified by the model and ranges from 0 to 1. The degrees of freedom are taken into account with the modified  $R^2$ . The  $R^2$  is a metric that indicates how well a model matches the results. An increase in  $R^2$  is a strong indicator. The more the  $R^2$  inclines to increase with more variables added to the model, the concerns will go to the Adjusted  $R^2$ . A  $R^2$  close to 1 means that the regression model has clarified a significant portion of the uncertainty, while a value close to 0 means that the regression model failed to clarify a significant portion of the variance in the result.

The Coefficient of determination is

$$R^2 = \frac{SSR}{SST} \quad (14)$$

(2) F test for regression relation

The F-statistic determines the model's overall significance. It determines if there is a non-zero coefficient for at least one predictor variable. When we use multiple predictors, such as in multiple linear regressions, the F-statistic becomes even more significant. A statistically significant p-value ( $p < 0.05$ ) corresponds to a high F-statistic.

Thus, in order to test whether there is a regression relation between the dependent variable Y and the set of X independent variables, two hypotheses are as follows:

$$H_0: \beta_1 = \dots = \beta_{p-1} = 0$$

$$H_1: \text{not all } \beta_k (k = 1, \dots, p - 1) = 0$$

The test statistic  $F^*$  is used where:

$$F^* = \frac{MSR}{MSE} \quad (15)$$

At  $\alpha$ , if  $F^* \leq F(1 - \alpha, p - 1, n - p)$ ,  $H_0$  is rejected.

### (3) Inferences about regression parameters

The Student T-test also known as T-test is an inferential statistic that permits to experiment an inference that applies to a population or a data sample. In our case it will assist us in determining the association between x and y, thus in order to test the significance of the independent variables, the two hypotheses are

$$H_0: b_k = 0$$

$$H_1: b_k \neq 0$$

And the  $t^*$  the test statistic used for

$$t^* = \frac{b_k}{s^2\{b_k\}} \quad (16)$$

Where  $s^2\{b_k\}$  is the estimated variance-covariance matrix is:

$$s^2\{b_k\} = MSE(XX)^{-1} \quad (17)$$

If  $|t^*| \leq t_{(1-\frac{\alpha}{2}), n-p}$ , then is rejected. This test statistic (and its corresponding p-value) is used as the criterion in the liner regression model.

In deduction of the last three parts, the F-statistic, R-squared  $R^2$ , and residual standard error (RSE) are metrics that are used to see how well the model matches our results.

RSE: Closer to zero the better

$R^2$  : Higher the better

F-statistic: Higher the better.

### 3.2.5 Model performance for multiple regression

#### 3.2.5.1 Studentized Breusch-Pagan test

The Breusch-Pagan test is used to detect whether or not a regression model has heteroscedasticity. The residuals are distributed with equal variance at each level of the predictor variable, which is one of the essential assumptions of a regression defined by homoscedasticity.. Whereas, when this assumption is broken, the residuals are said to have heteroscedasticity. When this happens, the regression's results become unreliable. Making a plot of the residuals versus the fitted values of the regression model is one technique to visually detect heteroscedasticity.

At higher values in the plot, the residuals become increasingly spread out, indicating the presence of heteroscedasticity. The following null and alternative hypotheses are used in the test:

$H_0$ : The residuals are spread with equal variance (homoscedasticity)

$H_1$ : The residuals are not spread with equal variance (heteroscedasticity)

The null hypothesis is not rejected. It is presumed that homoscedasticity exists in this scenario; the output of the original regression can now be interpreted. Otherwise, we reject the null hypothesis and conclude that heteroscedasticity exists in the regression model if the p-value of the test is less than some significance level ( $p < 0.05$ ), thus the regression's output table may not be defective. There are 2 ways to fix this issue:

1. Transform the response variable: Instead of using the original answer variable, you might use the log of the response variable. Taking the log of the answer variable is a common method for removing heteroscedasticity. The square root of the response variable is another typical modification.
2. Use weighted regression: Each data point is given a weight based on the variance of its fitted value in this sort of regression. This reduces the squared residuals of data points with higher variances by assigning tiny weights to them. Heteroscedasticity can be eliminated when the appropriate weights are employed.

### 3.3 Autoregressive Integrated Moving Average (ARIMA) Time Series

A time series is an ordered sequence of observations recorded at equally spaced sequentially through time intervals. Time series can be continuous when observations are made continuously through time or discrete when observations are taken only at specific times usually equally spaced. In this thesis we are concerned with discrete-time time series, which consist of observations made at discrete time weekly intervals.

ARIMA consists of 3 components AR, I and MA. AR component represents “Auto-Regressive” a model that uses the dependent connection between current information and its previous qualities i.e. the data is regressed on its past values. I component represents “**Integrated**” which implies that the information is fixed or in other terms stationary defined by deducting the data values from the previous values. Finally, MA represents “Moving Average” shows that the forecast, outcome or result of the model relies linearly and straightly upon the past qualities and joins the reliance between a perception and a remaining mistake know as error from a moving normal model applied to past errors.

The parameters of the ARIMA model are:

**p**: known as lag observations or a lag order to decide the result of the model by giving slacked information points.

**d**: known as the degree of differencing that demonstrates the occasions the slacked pointers have been deducted to make the information stationary.

**q**: known as the order or size of moving average or the number of forecast errors.

Examples of ARIMA (p, d, q) model:

- $d=0 \rightarrow$  ARMA model (no stationary data)
- $d=0, q=0 \rightarrow$  AR model (just autoregression, no stationary nor moving average)
- $p=0, d=0 \rightarrow$  MA model (moving average, no stationary nor autoregression)
- $p=1, d=0, q=0 \rightarrow$  AR(1) model,
- $p=0, d=0, q=0 \rightarrow$  Random Walk model,
- $p=0, d=0, q=1 \rightarrow$  MA (1) model .....

ARIMA model equation is as follows:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \dots + \phi_q \epsilon_{t-q} \quad (18)$$

AR (p) MODEL (first part of ARIMA):

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \epsilon_t \quad (19)$$

MA (q) MODEL (second part of ARIMA):

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \dots + \phi_q \epsilon_{t-q} \quad (20)$$

One the assumption of an ARIMA model is the concept of stationarity of a stochastic process where the mean and variance of a stationary process do not depend upon time. We can find two types of stationarity: Strongly Stationary and weak stationary.

- Strongly stationary: The combined distributions of any possible set of random numbers from the process are independent of time.
- Weakly stationary: It depends on time difference and not only on the occurrence of the data to estimate the values.

### 3.3.1 Autocorrelation and Partial autocorrelation function

To choose a suitable model, it is necessary to analyse the Autocorrelation and Partial autocorrelation function (ACF and PACF) where they reflect how the observations in a given time period are distributed. It is useful to plot them for modeling and forecasting purposes to measure the correlation of a stationary process. ACF and PACF help to determine the AR and MA in consecutive k time lags.

The autocorrelation coefficient ACF at lag k is given by:

$$\rho_k = \frac{\gamma_k}{\gamma_0} \quad (21)$$

Where  $\mu$  is the mean and  $\gamma_0$  is the variance of a time series.

The partial autocorrelation function is defined as a matrix:

$$\alpha_{kk} = \frac{\begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{1k-3} & \rho_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \dots & \rho_1 & \rho_k \end{bmatrix}}{\begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{1k-3} & \rho_{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \dots & \rho_1 & 1 \end{bmatrix}} \quad (22)$$

### 3.3.2 Box-Jenkins (ARIMA) Models

The basic idea behind self-projecting time series forecasting models is to find a mathematical formula that will approximately generate the historical patterns in a time series.

The Box-Jenkins methodology denotes a set of procedures for identifying and estimating time series models within the class of autoregressive integrated moving average (ARIMA) models which are regression models that use lagged values of the dependent variable and/or random disorder term as explanatory variables. These models were popularized by George Box and Gwilym Jenkins in the early 1970's.

ARIMA models are a class of linear models that is capable of representing stationary as well as non-stationary time series; they do not involve independent variables in their construction. They make use of the information in the series itself to generate forecasts. These models rely heavily on the autocorrelation pattern in the data. This method relates to both non-seasonal and seasonal data.

In an ARIMA model, the random disturbance term is typically assumed to “white noise”; i.e., it is identically and independently distributed with a mean of 0 and a common variance across all observations.

ARIMA methodology of forecasting is different from most methods because it does not assume any particular pattern in the historical data of the series to be forecast. It uses an interactive approach of identifying a possible model from a general class of models. The chosen model is then checked against the historical data to see if it accurately describes the series.

The Box-Jenkins methodology refers to a set of procedures for identifying, fitting, and checking ARIMA models with time series data. Forecasts follow directly from the form of fitted model.

The Box-Jenkins approach to modeling time series is a five-step process:

1. Time series stationarity:

The most important assumption for applying Box-Jenkins approach is to have a stationary time series process. If the series turns out to be non-stationary, difference transformation could be applied successively until reaching stationarity.

Stationarity condition can be tested using the ACF and PACF plots. The sample ACF and PACF of a stationary process cut off completely or decay gradually after a few lags. In contrast, a very slow decay in ACF and PACF plots indicate non-stationarity.

## 2. Model Identification:

The preliminary tool for model identification is the use of both sample ACF and PACF plots:

- A- If the sample ACF plot cuts off after a few lags and the sample PACF decays exponentially, then the stationary time series data follows a moving average MA process.
- B- If the sample ACF plot decays exponentially and the sample PACF cuts off after a few lags, then the stationary time series data follows an autoregressive AR process.
- C- If both ACF and PACF plots decay gradually, then the stationary time series data follows an autoregressive moving average ARMA process. With 95% confidence, the significant lags can be obtained from the correlograms where their corresponding auto-correlation coefficients lie outside the band given by:  $\pm 2/\sqrt{n}$ . Based on ACF and PACF plots, it could happen that more than one model can fit the time series data. Therefore, one should choose the most suitable model using the information criteria tool AIC and BIC explained in section 3.6.

## 3. Model Parameters Estimation:

The step right after identifying the time series model is to estimate its parameters based on the Least squares method and/or the Maximum likelihood estimation technique. In a linear regression model, the least squares estimates are also the maximum likelihood estimators when the errors belong to a normal distribution.

Therefore, it is sufficient in this paper to elaborate the maximum likelihood estimation for ARMA models.

Suppose ARMA (p,q) have a zero mean process given by:

$$Y_t = \alpha + \sum_{i=1}^p \beta_i Y_{t-i} + \sum_{i=1}^q \phi_i \epsilon_{t-i} + \epsilon_t \quad (23)$$

Where n observation  $y_1, y_2, \dots, y_n$  and  $\epsilon_t$  are i.i.d.  $N(0, \delta^2)$  with the parameter  $\beta_1, \beta_2, \dots, \beta_p, \phi_1, \phi_2, \dots, \phi_q$  and  $\delta^2$ .

A multivariate normal  $N_n(0_{nx1}, \Sigma_{n \times n})$  has likelihood function obtained by the joint distribution of  $Y_n = (y_1, y_2, \dots, y_n)^T$  where

$$\Sigma = [\gamma_{|i-j|}]_{i,j=1,2,\dots,n} = \begin{bmatrix} \gamma_0 & \cdots & \gamma_{n-1} \\ \vdots & \ddots & \vdots \\ \gamma_{n-1} & \cdots & \gamma_0 \end{bmatrix} \quad (24)$$

For instance, to evaluate the parameters a special case is applied to maximize the likelihood function AR (2).

AR (2) likelihood example:

Let  $Y_n = (y_1, y_2, \dots, y_n)^T$ , than the likelihood function is given by:

$$L(Y_n; \beta_1, \beta_2, \delta^2) = f(y_3, y_4, \dots, y_n; \beta_1, \beta_2, \delta^2 | y_1, y_2) * f(y_1, y_2; \beta_1, \beta_2, \delta^2) \quad (25)$$

However, when n is large, the conditional likelihood  $f(y_3, y_4, \dots, y_n; \beta_1, \beta_2, \delta^2 | y_1, y_2)$  is a good approximation of the exact likelihood  $L(Y_n; \beta_1, \beta_2, \delta^2)$ . Since AR (2) process is given by:

$$y_n = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \epsilon_t \quad (26)$$

$$\epsilon_t = y_n - \beta_1 Y_{t-1} - \beta_2 Y_{t-2} \quad (27)$$

Therefore the likelihood equation (25) can be written as:

$$L(Y_n; \beta_1, \beta_2, \delta^2) = f(\sigma_3, \sigma_4, \dots, \sigma_n; \beta_1, \beta_2, \delta^2 | \sigma_1, \sigma_2) \quad (28)$$

$(\sigma_3, \sigma_4, \dots, \sigma_n)$  are i.i.d.  $N(0, \delta^2)$ . Therefore the equation (28) can be written as:

$$L(Y_n; \beta_1, \beta_2, \delta^2) = \prod_{i=3}^n f(\sigma_i; \beta_1, \beta_2, \delta^2) \quad (29)$$

$$L(Y_n; \beta_1, \beta_2, \delta^2) = \prod_{i=3}^n \frac{1}{\sqrt{2\pi\delta}} e^{-\frac{1}{2\delta^2}\sigma_i^2} \quad (30)$$

$$L(Y_n; \beta_1, \beta_2, \delta^2) = \left(\frac{1}{\sqrt{2\pi\delta}}\right)^{n-3} e^{-\frac{1}{2\delta^2}\sum_{i=3}^n \sigma_i^2} \quad (31)$$

$$L(Y_n; \beta_1, \beta_2, \delta^2) = \left(\frac{1}{\sqrt{2\pi\delta}}\right)^{n-3} e^{-\frac{1}{2\delta^2}\sum_{i=3}^n (y_i - \beta_1 y_{i-1} - \beta_2 y_{i-2})^2} \quad (32)$$

Maximizing the equation is equivalent to minimizing the expression in  $\sum_{i=3}^n (y_i - \beta_1 y_{i-1} - \beta_2 y_{i-2})^2$  where  $y_n$  on  $y_{i-1}$  &  $y_{i-2}$  and hence  $\widehat{\beta}_1$  and  $\widehat{\beta}_2$  are obtained.

In the general ARMA case, one requires  $\Sigma$  to be invertible which is in some way complex.



#### 4. Testing

After the model parameters estimation step, the T-test is used to test whether or not the model selected suits best the data. In other word, we need to test whether or not each parameter chosen is significant.

The null hypothesis to be tested is:  $H_0: parameter_i = 0$  where  $i=1,2,\dots, \max(p,q)$

This is a two-sided test where its equation is:

$$OVTS = \frac{\widehat{parameter}_i - parameter_i}{standard\ error} \quad (33)$$

Since standard error is unknown, then the  $OVTS \sim t(n-m)$  where  $m = p + q$ . Assuming a degree of confidence of  $(1-\alpha)$ , we reject  $H_0$  if the p-value for the parameter is  $< \alpha$  (or the OVTS is in the rejection region).

The above steps are repeated for each parameter in order to assure that all the lags of the selected model are significant.

#### 5. Forecasting is represented in figure 3.4

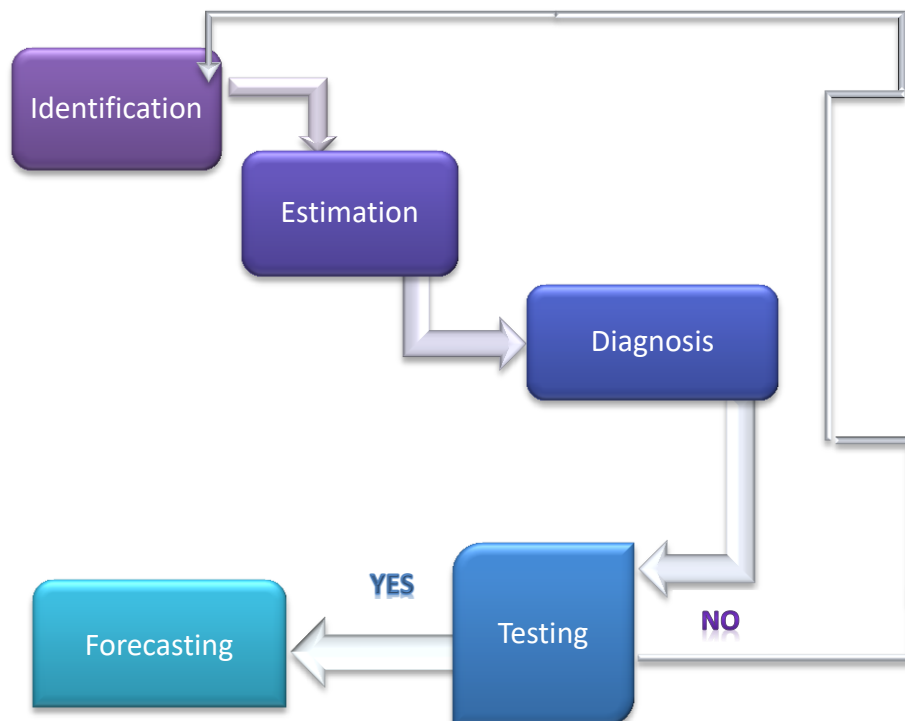


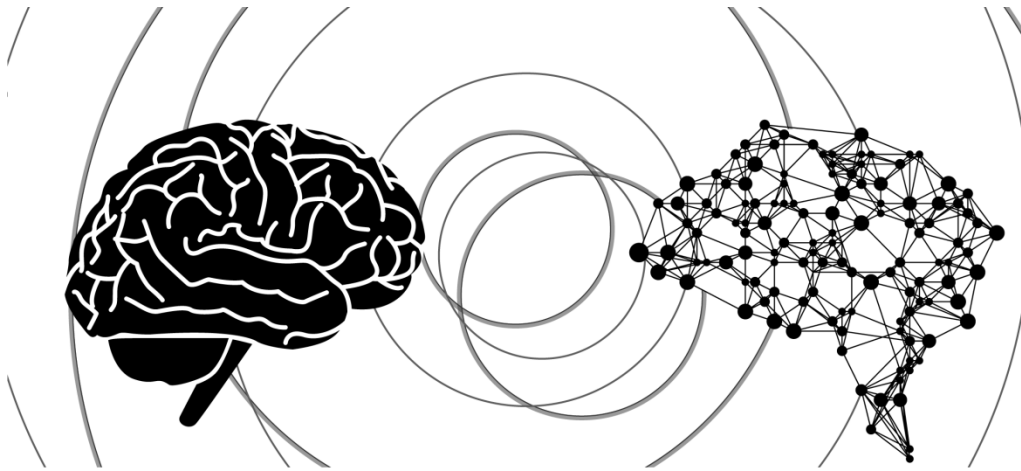
Figure 3. 4: Forecasting steps in ARIMA

### **3.4 Artificial Neural Network (ANN)**

Warren McCulloch, a Neurophysiologist, and Walter Pitts, a Mathematician, proposed the first artificial neural network in 1943 as a result of their research by applying mathematics to Boolean logic to model the function of neurons brain. Next, Frank Rosenblatt, a Psychologist, established a more advanced neural network called “perception” in 1958 which was able to produce numbers as inputs. Many other scientists utilized neural network and successfully created another part of the whole network. The last change is in 2006 by Geoffrey Hinton whom developed greedy layer pre-training.

Neural Networks are defined as universal mathematical models that use learning algorithms inspired by the brain to store information. They combine the input and output with nodes and neurons with a middle layer called hidden layers and presented through different types Feedforward, Feedback and so on. It is a huge necessity for our daily life of today. NN can be applied also on different programming software as Java, MATLAB, C, R or Python.

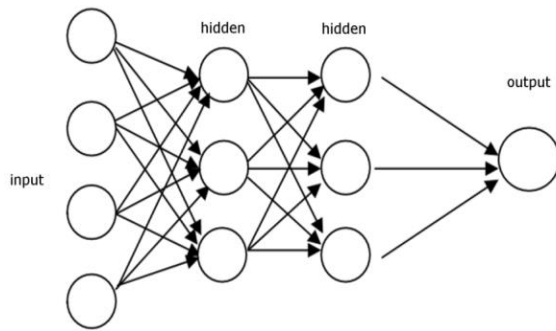
Neural Networks (NN) gain a huge popularity in the recent expansion related to the virtual world as the most mainstream machine learning algorithms used nowadays. Midrack (2019) stated that these NN are computer models of associated units or any physical device within a network of other tools. It is able to send, receive, or forward information called nodes intended to measure data likewise to how neurons (nerve cells) work with people either organic or artificial in nature. It comprises of thousands and millions of artificial "brain cells" or computational units that carry on and learn in an amazingly comparable manner to the human brain figure 3.5.



**Figure 3. 5 : Brain shape of a Neural Network**

Chen (2019) proclaims that NN structure the base of deep learning can learn from information and be prepared to perceive designs, patterns, order the data, and forecast future functions. In addition, it matches the tasks of a human brain to see associations between enormous proportions of data. Furthermore, NN are utilized in an assortment of applications in financial services from forecasting, anticipating, advertising and marketing research to misrepresentation identification and risk assessment. They are used for stock market price and financial exchange value in an expectation fluctuation. Chen (2019) added to the above information that NN can adjust to evolving input; so the network produces the most ideal outcome without expecting to update and redesign the yield standards and output criteria.

The idea of NN, which has its foundations in man-made brainpower (AI=Artificial Intelligence), is quickly picking up popularity in the improvement of trading frameworks. It may take hours or even long time to plan, yet time is a reasonable trade off when differed from its expansion and degree. Neural associations are especially fitting to perform plan affirmation to perceive and organize articles or signals in discourse, vision, and control framework. They can likewise be utilized for performing time-series expectation, modeling and displaying. NN are found in nowadays usage like image/pattern or facial recognition on smartphones or any gadgets that require this feature, self-driving vehicle trajectory prediction, data mining, forecasting, medical diagnosis, music composition and also found in the libraries of Google or Amazon. A sample of artificial neural network is presented in the figure 3.6.



**Figure 3. 6: Sample representation of a Neural Network**

NN consists of three essential layers: input, hidden and output layer as shown in the figure 3.6. First, the input layer is the main layer of a NN that gets the input data in the form of different writings, numbers, sound records files, picture pixels, and so forth. Secondly, the middle of a NN model is the hidden layers. There can be a solitary, hidden layer, as on account of a perceptron or different hidden layers. These concealed layers perform different sorts of numerical calculation on the information and perceive the examples that are essential for. Finally, in the output layer, we get the outcome result that we acquire through rigorous calculations and computations performed by the middle layer.

According to Nagwani(2016), the input nodes are connected with an activation function to transform them into outputs. Each node multiplies the input signal with a weight  $w_{ij}$  characteristic of the connection between nodes  $i$  and  $j$  of layers to relate the weighted input. The hidden layer node performs a single ‘sigmoid ’transformation implies by the equation:

$$z_j = g(\sum_i y_i w_{ij} - \beta_i) \tag{34}$$

Where  $z_j$  is the output of the  $j$ th node,  $y_i$  the input,  $\beta_i$  the bias of the  $j$ th node and  $g$  is the sigmoidal function given by

$$g(x) = \frac{1}{1+e^{-x}} \tag{35}$$

The figure 3.7 is a cheat sheet as it joins so many types of NN applied by on algorithms where some of them are already defined above. The types of NN are unique in their functioning.

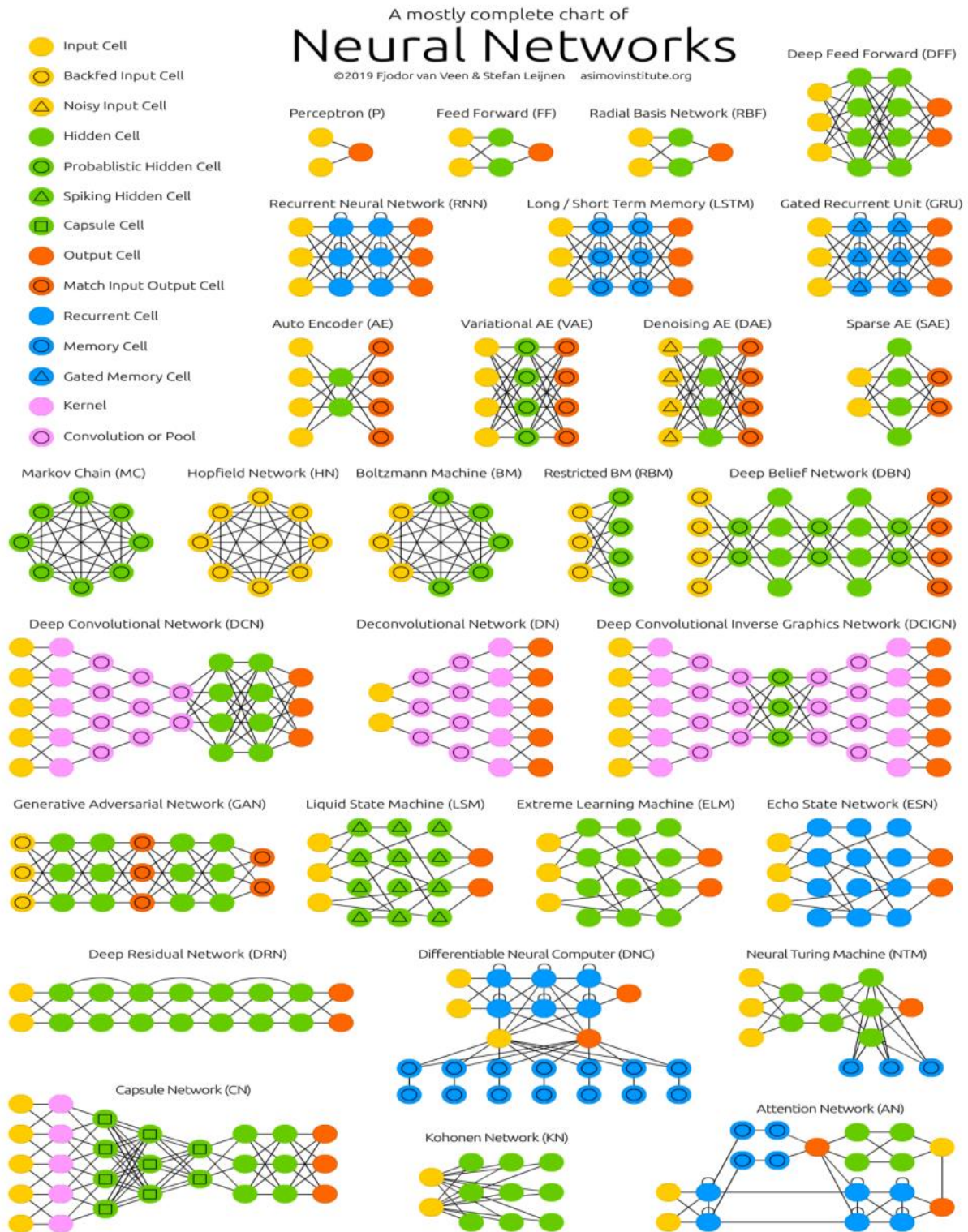
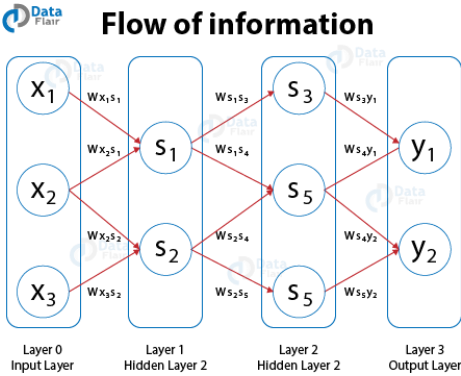
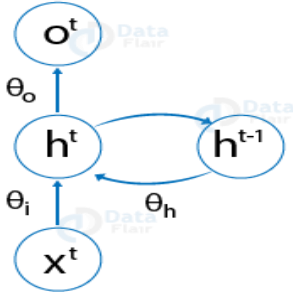


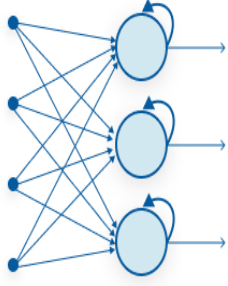
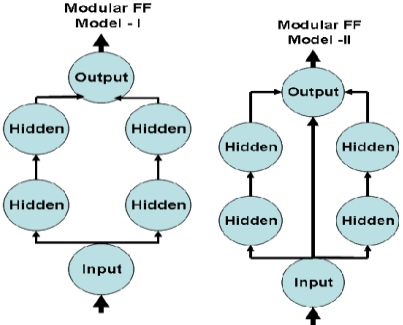
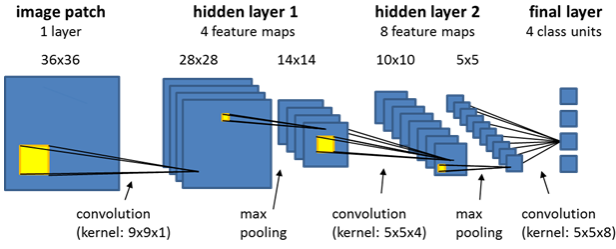
Figure 3. 7: Cheat Sheet of Neural Network Shapes

### 3.4.1 Type of Neural Network

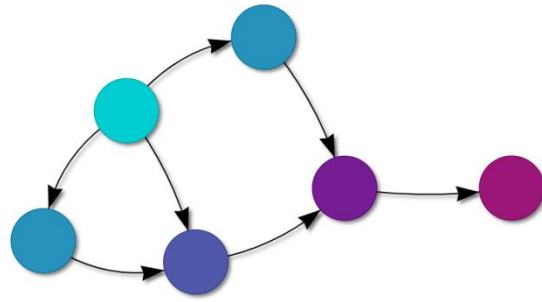
A Table could help to combine the most known and used types by defining them in a brief way:

Table 3. 2: Types of Neural Network

TYPES	PICTURE	DEFINITION
<p><b>Feedforward NN</b></p>	 <p><b>Flow of information</b></p> <p>Figure 3. 8: Flow of information FNN</p>	<p>This is the most simple of all varieties. The information moves in one direction only.</p> <p>The flow of information is from the input layer to the hidden layer and finally to the output and sent from input nodes directly to output nodes. There are no feedback loops or cycles in this network as shown in the figure 3.8.</p> <p>These types of neural networks are mostly used in <b>supervised learning</b> such as classification, image recognition etc.</p>
<p><b>Feedback NN</b></p>	 <p>Figure 3. 9: Feedback NN</p>	<p>In this type, the loops are a part of it. This is mainly for memory retention such as in the case of recurrent neural networks. These types of networks are most suited for areas where the data is sequential or time-dependent.</p>

<p><b>Recurrent NN</b></p>	 <p style="text-align: center;">Recurrent Neural Network</p> <p><b>Figure 3. 10: Recurrent NN</b></p>	<p>Unlike its feedforward cousin, the recurrent NN permits information to stream bi-directionally. This kind of organization is a well-known decision for design acknowledgment applications, such as speech recognition and handwriting solutions.</p>
<p><b>Modular NN</b></p>	 <p><b>Figure 3. 11: Modular NN</b></p>	<p>A modular NN is comprised of free NN. Each is given a lot of information sources and work together to finish sub-tasks. The last output of the measured NN is overseen by a middle step that gathers information from the individual organizations.</p>
<p><b>Convolutional NN</b></p>	 <p><b>Figure 3. 12: Convolution NN</b></p>	<p>Convolutional NN is fundamentally used to arrange pictures. For instance, they can group comparable photographs and recognize explicit articles inside a scene, including faces, road signs and people. We are seeing increasingly more of these organizations being used across numerous applications, from online media applications to medical services diagnostics arrangements.</p>

**Bayesian Network**



**Figure 3. 13: Bayesian Network**

This type of NN is a probabilistic graphical model for representing information about an unsure domain space where every node compares to an arbitrary variable and each edge speaks to the restrictive likelihood probability for the relating random factors. These kinds of Bayesian Networks are otherwise called Belief Networks. Due to the conditions and contingent probabilities, a BN relates to Directed Acyclic Graph (DAG) where no circle or self-association is permitted.



## 3.5 Model performance

The performance models in this study are: Box-Ljung test, MAPE, RMSE, SSE, AIC, and BIC.

### 3.6.1. Box-Ljung test

The Ljung-Box test is a statistical test that determines whether a time series has autocorrelation. It is commonly used in econometrics and other domains that deal with time series data. This test uses the hypothesis

$H_0$ : The residual are independently distributed

$H_1$ : The residuals are not independently distributed.

We'd wish to avoid rejecting the null hypothesis as much as possible. That instance, we want the test's p-value to be greater than 0.05 since this indicates that our time series model's residuals are independent, which is a common assumption when building a model.

The test equation is:

$$Q = \frac{n(n+2) \sum p_k^2}{n-k} \quad (36)$$

Where n = sample size,

$\sum$  = sum of 1 to h, where h is the number of lags being tested.

$p_k$  = Sample autocorrelation at lag k.

Q follows a chi-square distribution with h degrees of freedom. We reject the null hypothesis if  $Q > X_{1-\alpha, h}^2$ , so we can tell that the model are not independently distributed.

### 3.6.2. Mean Absolute Percentage Error (MAPE)

MAPE measure the accuracy of a forecast system as a percentage and can be calculated manually by

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - E_t}{A_t} \right| * 100 \quad (37)$$

where  $n$  is the sample size,  $A_t$  the actual value and  $E_t$  the expected or forecasted value.

### 3.6.3. The root Mean Squared Error (RMSE)

RMSE is the squared root of the mean square error (MSE) which processes the deviation of the forecasted value. Its equation is

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2} \quad (38)$$

### 3.6.4. Sum of Squared Error (SSE)

SSE has the same properties as the MSE and RMSE but measures the total squared deviation of forecasted observations from the actual values. Its equation is:

$$SSE = \sum_{t=1}^n e_t^2 \quad (39)$$

### 3.6.5. Akaike Information Criterion (AIC)

As Browlee (2019) defined, Akaike information criterion is a method for scoring and selecting a model developed by Hirotugu Akaike as an information theory and frequentist based inference. Applying to a dataset, the lowest AIC the better the model is.

### 3.6.6. Bayesian Information Criterion (BIC)

Browlee (2019) affirmed that the BIC is also another method for scoring and picking a model. Like AIC it fit under the maximum likelihood estimation framework. Although the quantity measured differs from the AIC, it can be seen to be proportional to it. Unlike the AIC, the BIC penalizes models for their difficulty, which means that more complex models will have a lower (larger) score and, as a result, will be less likely to be chosen. The Bayesian probability method implies that if a collection of candidate models contains a true model for the dataset, the probability that BIC will choose the true model increases as the size of the training dataset grows. It is the same as AIC, the lowest is better.

## **Chapter 4**

### **Results and Discussion**

#### **4.1 Data presentation**

The purpose of this study is to model and predict the close value of bitcoins. Thus the following Close price variables (A) of Bitcoin, Oil, Gold, Euro, CNYuan, LA, EI1, EI2, EI3, CI1, CI2, CI3, AI1, AI2, AI3 and demand/supply variables (B) which are Bitcoin addresses, Block Size, Blockchain addresses, Blockchain transaction, Cost per transaction, Difficulty, Hash rate, Miners rewards, Mining commissions, Number of bitcoin, Transaction value, Transaction volume and Unspent transaction described in the table 1 are considered. Data are retrieved weekly from <https://www.investing.com/> , <https://www.quandl.com/> and <https://www.blockchain.com/> for the years 2019 and 2020. All the descriptive statistics, inferential statistics, machine learning for modeling and prediction was conducted on R studio version 4.0.5.

**Table 4. 1: Brief variables definition**

<b>Variables</b>	<b>Description</b>
<b>A. Close</b>	
Bitcoin	Digital currency Price
Crude Oil (Oil)	Oil Price
Gold	Gold Price
Euro	Monetary unit and currency of the European Union
Chinese Yuan(CNYuan)	Currency of Republic of China
Latin America S&P 40 index (LA)	Latin America stock market index
Euro Stoxx 50 index (EI1)	European blue-chip index
Euro FTSE 100 index ( EI2),	European blue-chip index
Europe S&P 350 index (EI3)	European market-cap-weighted index
Chinese SSE composite Index (CI1)	Shanghai Stock Exchange Stock Market Composite Index
Chinese SZSE component Index (CI2)	Shenzhen Component Stock Exchange Index
Chinese CSI 300 Index (CI3)	Capitalization weighted Stock Market Index
Asia S&P 50 Index(AI1)	Asian Stock Index
Asia Dow Jones Titans 50 index (AI2)	Dow Jones Asia-Pacific Index
Asia FTSE ASEAN 40 index (AI3)	FTSE ASEAN region Stars Index
<b>B. Demand and Supply</b>	
Bitcoin addresses	Number of unique Bitcoin addresses used per day
Block Size	Average block size expressed in megabytes
Blockchain addresses	Number of unique addresses used in blockchain
Blockchain transaction	Number of transactions on blockchain
Cost per transaction	Miners' income divided by the number of transactions
Difficulty	Difficulty mining a new blockchain block
Hash rate	Times a hash function can be calculated per second
Miners rewards	Block rewards paid to miners
Mining commissions	Average transaction fees (in USD)
Number of bitcoin	Number of mined Bitcoins circulating on the network
Transaction value	Value of daily transactions
Transaction volume	Number of transactions per day
Unspent transaction	Number of valid unspent transactions

First Univariate analysis was conducted on each variable included in this study:

- Descriptive statistics was conducted to calculate the measure of central tendency (mean and median), measure of percentile(first and third quartiles) and measure of dispersion(standard deviation and coefficient of variation)
- Boxplot was used for outliers detection

Second Bivariate analysis was conducted to check if there is a significant relation between the close price of bitcoin and all the variables included in this study.

Third Multivariate analysis was conducted to check if there is a significant relation between all variable included in this study.

Finally the following machine learning models were used and their accuracy was calculated:

- Regression decision tree
- Multiple regression
- Time series
- Regression time series
- Artificial Neural network
- Regression time series Neural network

## **4.2 Univariate and Multivariate variate analysis**

### **4.2.1 Descriptive statistics and Boxplot**

A boxplot is used to illustrate the spreading of the variables using quartiles. The graph is shaped by a box with a horizontal line inside representing the median or the second quartile. The lower and the upper horizontal border line of the box represent the first and the third quartile respectively. The vertical lines extended from the top and the bottom called whiskers. Its limits are the maximum and the minimum observation respectively. The shape of the box plot indicates its spreading (Tall) or compactness (Short). In addition, when the median line is in the middle of the box, the data are considered to be normally distributed. On the other hand, if the median line is closer to the first quartile, the distribution of the data is considered to be positively skewed and if it's closer to the third quartile, the distribution is considered to be negatively skewed. Moreover, the boxplot is employed to detect the presence of outliers: data that are plotted outside the whiskers are considered as outliers.

The boxplots of the Close price and the demand supply variables are presented in figure 4.1. Their minimum value, first quartile (1st Qu), median, mean, third quartile (3rd Qu), maximum value, standard deviation and the coefficient of variation are presented in table 4.2. Their histograms are plotted in figure 4.2. The following observations are concluded from figure 4.1 and 4.2:

- Close Bitcoin shows short shape showing the compactness of the data and a positively skewed distribution. Several data are plotted outside the whiskers, thus Bitcoin contains several outliers. Close Bitcoin ranges between 5,000 & 10,000 having the highest frequency equal to 60.

- Close Oil has no outliers with a tall shape that shows the spreading of the data with a negatively skewed distribution. The highest frequency of Close AI3 ranges between 55 & 66 with a frequency equal to 28.
- Similar to Close Oil, Close Gold has no outliers. A tall shape shows the spreading of the data and presents a positively skewed distribution. The highest frequency of Close Gold ranges between 1,500 & 1,600 with a frequency equal to 25.
- Close Euro shows outliers above the upper whisker with a normal shape. It presents the compactness of the data and positively skewed distribution. The highest frequency of Close Euro ranges between 1.1 & 1.125 with a frequency equal to 33.
- Close CNYuan has no outliers with a tall shape that shows the spreading of the data and presents a positively skewed distribution. The highest frequency of Close CNYuan ranges between 0.140 & 0.1425 with a frequency equal to 25.
- Close LA has no outliers with a tall shape that shows the spreading of the data with a negatively skewed distribution. The highest frequency of Close LA ranges between 32.5 & 35 with a frequency equal to 32.
- Close EI1 and Close EI2 show a short shape with few outliers under the bottom whisker showing a negatively skewed distribution. The highest frequency of Close EI1 and Close EI2 range between 3,750 & 4,000 and 1,400 & 1,500 respectively with frequencies equal to 33 & 35 respectively.
- Similar to LA, no outliers are found in Close EI3. The spreading of the data defines by the tall box shape presenting the spread of the data and negatively skewed distribution. The highest frequency of Close EI3 ranges between 1,800 and 1,900 with a frequency equal to 31.
- Close CI1 and Close CI2 have no outliers with a tall shape that shows the spreading of the data and present a positively skewed distribution. The highest frequency of Close CI1 and Close CI2 range between 425 & 450 and 0.14 & 0.16 with frequencies equal to 30 & 26 respectively.
- Close CI3 shows short shape showing the compactness of the data and a positively skewed distribution. Several data are plotted outside the whiskers, thus Bitcoin contains several outliers. Close CI3 is between 550 & 600 with a frequency equal to 43.

- Close AI1 shows outliers above the upper and lower of the whisker with a normal shape. The highest frequency of Close AI1 is between 3,300 & 3,400 with a frequency equal to 32.
- Close AI2 shows outliers above the upper whisker with a normal shape. It presents the compactness of the data and showing a negatively skewed distribution. The highest frequency of Close AI2 is between 160 & 170 with a frequency equal to 36.
- Close AI3 has no outliers with a tall shape that shows the spreading of the data with a negatively skewed distribution. The highest frequency of Close AI3 ranges between 10,500 & 11,000 with a frequency equal to 31.
- Bitcoin addresses have no outliers with a tall shape that shows the spreading of the data and presents a positively skewed distribution. The highest frequency of Bitcoin addresses ranges between 400,000 & 450,000 with a frequency equal to 35.
- Block Size has no outliers with a tall shape that shows the spreading of the data and presents a positively skewed distribution. The highest frequency of Block Size ranges between 0.8 & 0.9 with a frequency equal to 28.
- Blockchain addresses, Blockchain transaction, Cost transaction, Miners rewards and Number of bitcoin has all no outlier with a tall shape that shows the spreading of the data and presents a positively skewed distribution. The highest frequency of Blockchain addresses, Blockchain transaction, Cost transaction, Miners rewards and Number of bitcoin range between 45,000,000 & 45,500,000 , 500,000,000 & 550,000,000 , 50 & 60 , 14,000,000 & 16,000,000 , and 18,400,000 & 18,500,000 with frequencies equal to 28,24, 22,17,and 15 respectively.
- Difficulty, Hash rate and Unspent transaction have no outliers with a tall shape that shows the spreading of the data with a negatively skewed distribution. The highest frequency of Difficulty, Hash rate and Unspent transaction range between 5,500,000,000,000 & 6,000,000,000,000, 90,000,000 & 100,000,000 and 66,000,000 & 68,000,000 with frequencies equal to 22, 28, and 25 respectively.
- Mining commissions, Transaction value and Transaction volume demonstrate the compactness of the data due to its short shape with outliers above the upper whisker with a positively skewed distribution. The highest frequency of Mining commissions, Transaction value and Transaction volume range between 0 & 1, 260,000 & 300,000 and 50,000 & 75,000 with frequencies equal to 43, 36, and 25 respectively.

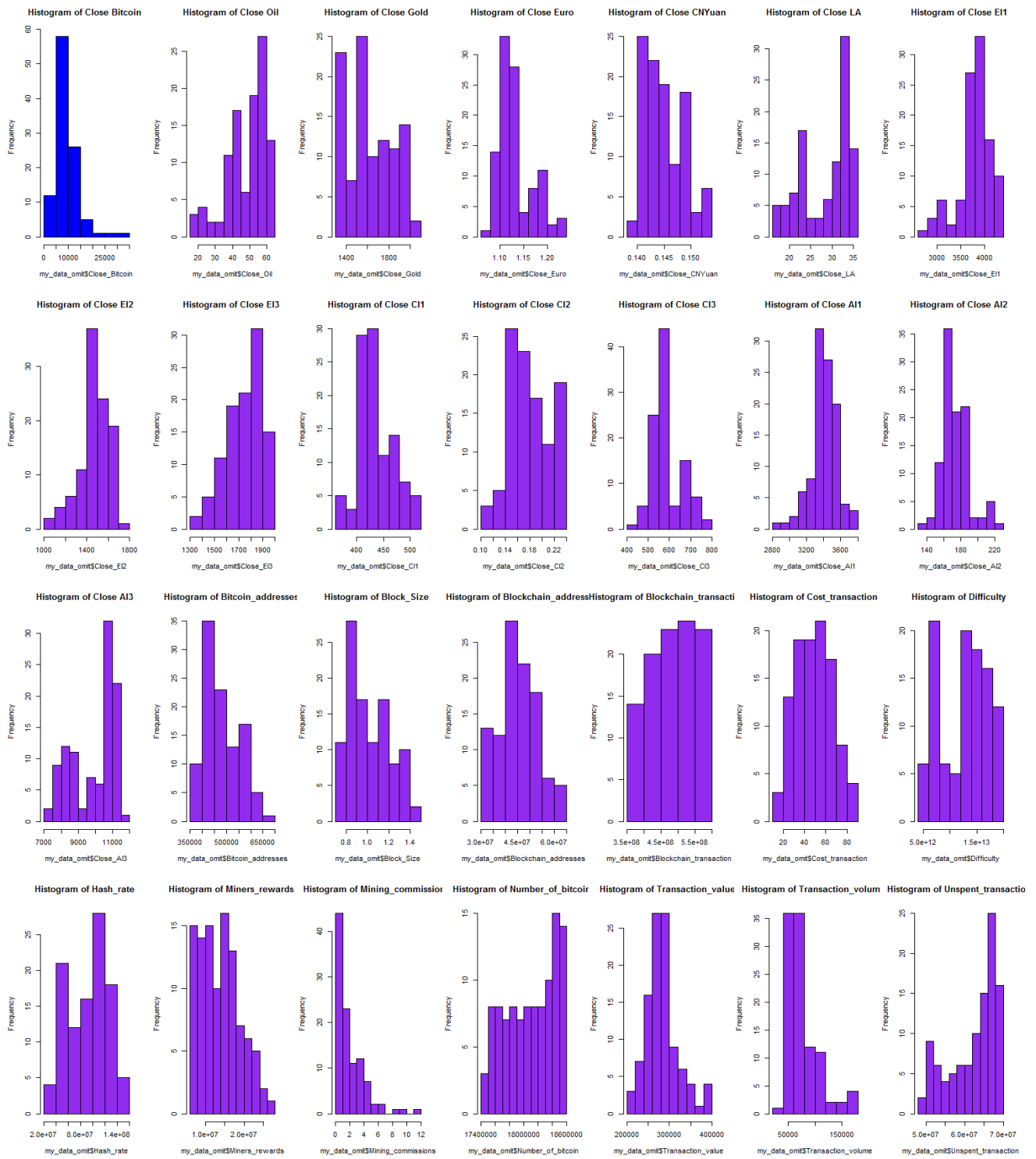


**Figure 4. 1: Box plot of the weekly Close price and the weekly demand supply for bitcoin for the years 2019, 2020, and 2021.**



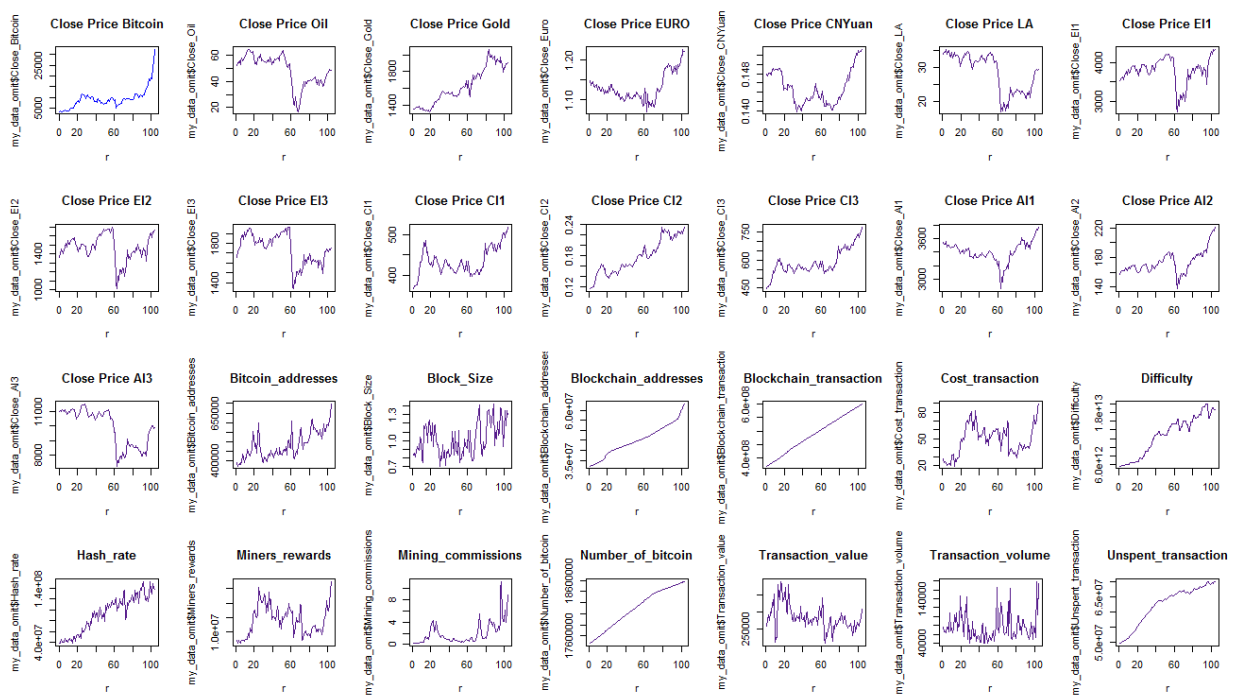
**Table 4. 2: Minimum, First quartile (1st Qu), Median, Mean, Third quartile (3rd Qu), and Maximum value of the Close prices and for demand supply of bitcoins.**

<i>Variables</i>	<i>Min.</i>	<i>1<sup>st</sup> Quartile</i>	<i>Median</i>	<i>Mean</i>	<i>3<sup>rd</sup> Quartile</i>	<i>Max.</i>	<i>Standard deviation</i>	<i>Coefficient of variation</i>
<b>Bitcoin</b>	3,502	7,255	9,174	9,481	10,723	32,193	4,527.36	0.48
<b>Oil</b>	16.94	40.58	53.09	48.39	57.24	64	11.54	0.24
<b>Gold</b>	1,322	1,450	1,562	1,627	1,814	2,061	215.64	0.13
<b>Euro</b>	1.07	1.11	1.12	1.13	1.14	1.23	0.04	0.03
<b>CNYuan</b>	0.14	0.14	0.14	0.15	0.15	0.15	0	0.03
<b>LA</b>	16.85	22.85	31.47	28.52	33.41	35.31	5.70	0.20
<b>EI1</b>	2,726	3,686	3,840	3,798	4,003	4,345	334.56	0.09
<b>EI2</b>	1,009	1,409	1,467	1,468	1,572	1,701	143.73	0.10
<b>EI3</b>	1,339	1,629	1,761	1,744	1,867	196	151.21	0.09
<b>CI1</b>	367	413.50	428.80	436.10	459.40	519.20	33.01	0.08
<b>CI2</b>	0.12	0.15	0.18	0.18	0.21	0.24	0.03	0.18
<b>CI3</b>	444.70	545.20	566.70	587.10	626.60	779.10	68.65	0.12
<b>AI1</b>	2,848	3,333	3,411	3,404	3,504	3,783	156.84	0.05
<b>AI2</b>	137.70	161.90	170.20	173.60	182.20	221.60	15.67	0.09
<b>AI3</b>	7,214	8,554	10,574	9,900	10,979	11,531	1,283.06	0.13
<b>Bitcoin addresses</b>	372,206	430,656	462,891	481,724	539,052	696,527	69,275.35	0.14
<b>Block Size</b>	0.70	0.85	0.97	1.01	1.17	1.42	0.19	0.19
<b>Blockchain addresses</b>	32,136,516	40,209,522	44,720,985	45,222,606	50,699,378	63,057,974	7,683,209	0.17
<b>Blockchain transaction</b>	370,814,223	431,756,827	489,152,606	487,467,833	543,643,406	599,940,609	67,030,200	0.14
<b>Cost transaction</b>	19.08	35.97	48.95	48.74	61.16	89.10	16.74	0.34
<b>Difficulty</b>	5,620 x 10 <sup>9</sup>	7,930 x 10 <sup>9</sup>	13,700 x 10 <sup>9</sup>	12,800 x 10 <sup>9</sup>	16,220 x 10 <sup>9</sup>	20,000 x 10 <sup>9</sup>	4,498,232 x 10 <sup>6</sup>	0.35
<b>Hash rate</b>	38,264,362	62,142,324	98,945,079	93,456,210	117,796,451	152,205,060	33,236,680	0.36
<b>Miners rewards</b>	6,044,025	9,763,526	13,805,468	13,853,702	17,259,722	27,533,620	5,093,997	0.37
<b>Mining commissions</b>	0.19	0.60	1.14	1.99	2.92	11.29	2.04	1.02
<b>Number of bitcoin</b>	17,465,763	17,801,241	18,136,319	18,098,863	18,419,125	18,582,738	344,858	0.02
<b>Transaction value</b>	209,154	260,795	279,706	283,662	299,006	390,735	37,091.40	0.13
<b>Transaction volume</b>	39,980	56,293	68,967	76,432	86,712	178,998	29,515.56	0.39
<b>Unspent transaction</b>	49,631,380	58,071,123	64,517,713	62,353,455	66,742,984	69,991,706	6,131,960	0.10



**Figure 4. 2: Histogram presenting the frequency distribution of the weekly Close price and the weekly demand supply for bitcoin for the years 2019 and 2020.**

Figure 4.3 presents the scatter plot of the weekly Close price and the weekly demand supply for bitcoin for the years 2019 and 2020. The x-axis represent the week number (time) and the y-axis represents the value the variables for its corresponding week. We can notice the linear relationship between time and blockchain address, blockchain transaction, number of bitcoins, and unspent transaction. Close Bitcoin, Gold, CI1, CI2, CI3, Bitcoin addresses, Difficulty and Hash rate has an increase trend with respect to time. Close Oil, Euro, LA, EI1, EI2, EI3, AI1, AI2, and AI3 have a continuous up and down line with respect to time at the beginning but on March 2020 the values decreases and than start to re-increase again after. Similar to the previous variables Close CNYuan has an increase trend at the end but has two deacing part at Jun 2019 and Jun 2020 with an increasing part in between. The other variables Block size, Cost transaction, Miners rewards, Mining commissions, Transaction value and Transaction volume have a fluctuating trend through increasing and decreasing all along with the time.



**Figure 4. 3: Scatter plot presenting the weekly Close price and the weekly demand supply for bitcoin for the years 2019 and 2020**

#### 4.2.2 Bivariate analysis and Correlation Test

There are several types of correlation coefficients such as Pearson, Kendall, and Spearman but the most common one is the Pearson correlation.

The **Pearson's correlation coefficient** assesses the association between two continuous variables. As it is based on the method of covariance, it is known as the best approach for quantifying the relationship between variables of interest. The Pearson correlation coefficient ( $r$ ) is calculated as follows:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (40)$$

With:

- $r$  is the Pearson coefficient of correlation
- $x_i$  is the values of the x-variable in a sample
- $\bar{x}$  is the mean of the values of the x-variable
- $y_i$  is the values of the y-variable in a sample
- $\bar{y}$  is the mean of the values of the y-variable

The value of  $r$  range from -1 till 1 and its interpretation is detailed in a table 4.3

**Table 4. 3: Interval of the Coefficient Correlation**

<i>Interval of Coefficient Correlation</i>	<i>Interpretation</i>
<b>+ 1 or -1</b>	Perfect uphill(+) or downhill (-) linear relationship
<b>+0.9 and 1 or -0.9 and 1</b>	Very High uphill(+) or downhill (-) linear relationship
<b>+0.7 to 0.9 or -0.7 to 0.9</b>	Strong uphill(+) or downhill (-) linear relationship
<b>+0.5 to 0.7 or -0.5 to 0.7</b>	Moderate uphill(+) or downhill (-) linear relationship
<b>+0.3 to 0.5 or -0.3 to 0.5</b>	Weak uphill(+) or downhill (-) linear relationship
<b>+0.01 to 0.3 or -0.01 to 0.3</b>	Very Low uphill(+) or downhill (-) linear relationship
<b>0.0</b>	No relationship or negligible correlation

The table 4.4 presents the Pearson correlation coefficient between Bitcoin and the 27 other assets. A strong uphill correlation is noted with CI3, Blockchain addresses, AI2, Bitcoin addresses and Mining commissions with values of 0.80, 0.79, 0.78, 0.74 and 0.74 respectively. A moderate uphill is considered with Gold, Euro, CI1, CI2, Blockchain transaction, Cost transaction, Difficulty, Hash rate, Miners rewards, Number of bitcoin, Unspent transaction with value of respectively. A weak uphill correlation is considered with EI1, EI3 and CI2 which range between 0.61, 0.64, 0.67, 0.64, 0.69, 0.60, 0.60, 0.60, 0.55, 0.63 and 0.62. The rest of variable have weak and very low uphill and downhill CNYuan, LA, EI1, EI2, EI3, AI1, AI3, Block Size, Transaction value and Transaction volume with absolute value of  $r$  lower than 0.3.

**Table 4. 4: Pearson Correlation coefficient between the weekly Close Bitcoin price with the weekly Close variables and the weekly demand supply for bitcoin for the years 2019 and 2020.**

<i>Variables</i>	<i>Pearson coefficient of correlation “r”</i>
<b>Bitcoin and Oil</b>	-0.18
<b>Bitcoin and Gold</b>	0.61
<b>Bitcoin and Euro</b>	0.64
<b>Bitcoin and CNYuan</b>	0.38
<b>Bitcoin and LA</b>	-0.21
<b>Bitcoin and EI1</b>	0.40
<b>Bitcoin and EI2</b>	0.24
<b>Bitcoin and EI3</b>	-0.15
<b>Bitcoin and CI1</b>	0.67
<b>Bitcoin and CI2</b>	0.64
<b>Bitcoin and CI3</b>	0.80
<b>Bitcoin and AI1</b>	0.45
<b>Bitcoin and AI2</b>	0.78
<b>Bitcoin and AI3</b>	-0.20
<b>Bitcoin and Bitcoin addresses</b>	0.74
<b>Bitcoin and Block Size</b>	0.45
<b>Bitcoin and Blockchain addresses</b>	0.79
<b>Bitcoin and Blockchain transaction</b>	0.69
<b>Bitcoin and Cost transaction</b>	0.60
<b>Bitcoin and Difficulty</b>	0.60
<b>Bitcoin and Hash rate</b>	0.60
<b>Bitcoin and Miners rewards</b>	0.55
<b>Bitcoin and Mining commissions</b>	0.74
<b>Bitcoin and Number of bitcoin</b>	0.64
<b>Bitcoin and Transaction value</b>	-0.11
<b>Bitcoin and Transaction volume</b>	0.27
<b>Bitcoin and Unspent transaction</b>	0.62

The null hypothesis of the Pearson test is: the correlation coefficient is not significantly different from 0 i.e. there is no significant linear relationship (correlation) between x and y in the population. The alternative hypothesis is the population correlation coefficient is significantly different from 0 i.e. there is a significant linear relationship (correlation) between x and y in the population. If the p-value is smaller than  $\alpha$  (usual set to 0.05), we reject the null hypothesis and conclude that there is significant a linear relationship (correlation) between x and y. The p-value of Pearson test for the correlation between Bitcoin and the other 27 variables as well as the 95% confidence interval are presented in table 4.5.

**Table 4. 5: Pearson Correlation coefficient Test between the weekly Close Bitcoin price with the weekly Close variables and the weekly demand supply for bitcoin for the years 2019 and 2020.**

<i>Variables</i>	<i>P-value</i>	<i>95% CI</i>
<b>Bitcoin and Oil</b>	0.07	[-0.35, 0.01]
<b>Bitcoin and Gold</b>	0.00	[0.48, 0.72]
<b>Bitcoin and Euro</b>	0.00	[0.51, 0.74]
<b>Bitcoin and CNYuan</b>	0.00	[0.20, 0.53]
<b>Bitcoin and LA</b>	0.03	[-0.39, -0.02]
<b>Bitcoin and EI1</b>	0.00	[0.23, 0.56]
<b>Bitcoin and EI2</b>	0.01	[0.05, 0.41]
<b>Bitcoin and EI3</b>	0.12	[-0.33, 0.04]
<b>Bitcoin and CI1</b>	0.00	[0.55, 0.76]
<b>Bitcoin and CI2</b>	0.00	[0.51, 0.74]
<b>Bitcoin and CI3</b>	0.00	[0.72, 0.86]
<b>Bitcoin and AI1</b>	0.00	[0.28, 0.59]
<b>Bitcoin and AI2</b>	0.00	[0.69, 0.84]
<b>Bitcoin and AI3</b>	0.03	[-0.38, -0.01]
<b>Bitcoin and Bitcoin addresses</b>	0.00	[0.64, 0.82]
<b>Bitcoin and Block Size</b>	0.00	[0.28, 0.59]
<b>Bitcoin and Blockchain addresses</b>	0.00	[0.70, 0.85]
<b>Bitcoin and Blockchain transaction</b>	0.00	[0.57, 0.80]
<b>Bitcoin and Cost transaction</b>	0.00	[0.47, 0.71]
<b>Bitcoin and Difficulty</b>	0.00	[0.46, 0.71]
<b>Bitcoin and Hash rate</b>	0.00	[0.47, 0.71]
<b>Bitcoin and Miners rewards</b>	0.00	[0.40, 0.67]
<b>Bitcoin and Mining commissions</b>	0.00	[0.63, 0.81]
<b>Bitcoin and Number of bitcoin</b>	0.00	[0.50, 0.73]
<b>Bitcoin and Transaction value</b>	0.25	[-0.30, 0.08]
<b>Bitcoin and Transaction volume</b>	0.01	[0.08, 0.44]
<b>Bitcoin and Unspent transaction</b>	0.00	[0.49, 0.72]

Since the p-value of the Pearson correlation test between Bitcoin and Transaction value, Bitcoin and EI3, Bitcoin and Oil is 0.25, 0.12 and 0.07 respectively greater than 0.05 so we fail to reject the null hypothesis. Whereas, the others assets Gold, Euro, CNYuan, LA, EI1, EI2, CI1, CI2, CI3, AI1, AI2, AI3, Bitcoin addresses, Block Size, Blockchain addresses, Blockchain transaction, Cost transaction, Difficulty, Hash rate, Miners rewards, Mining commissions, Number of bitcoin, Transaction volume and Unspent transaction have a p-value approximately to 0 thus there is a significant relationship between the latter variables and bitcoin.

Figure 4.4 shows the linearity or nonlinearity association between the variables. None of the variables are linear where the data are scattering at both sides of the linear line with some outliers.

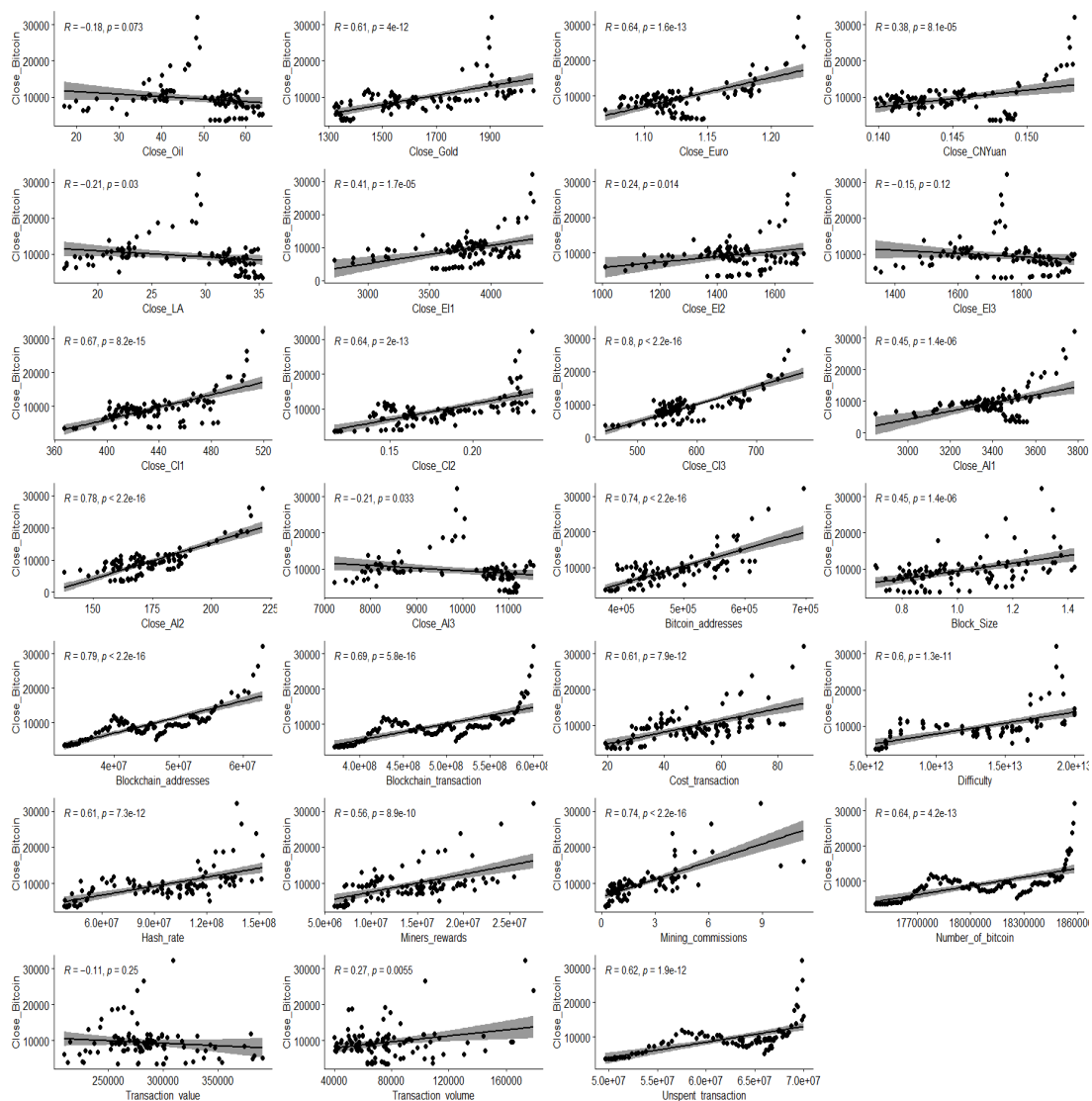
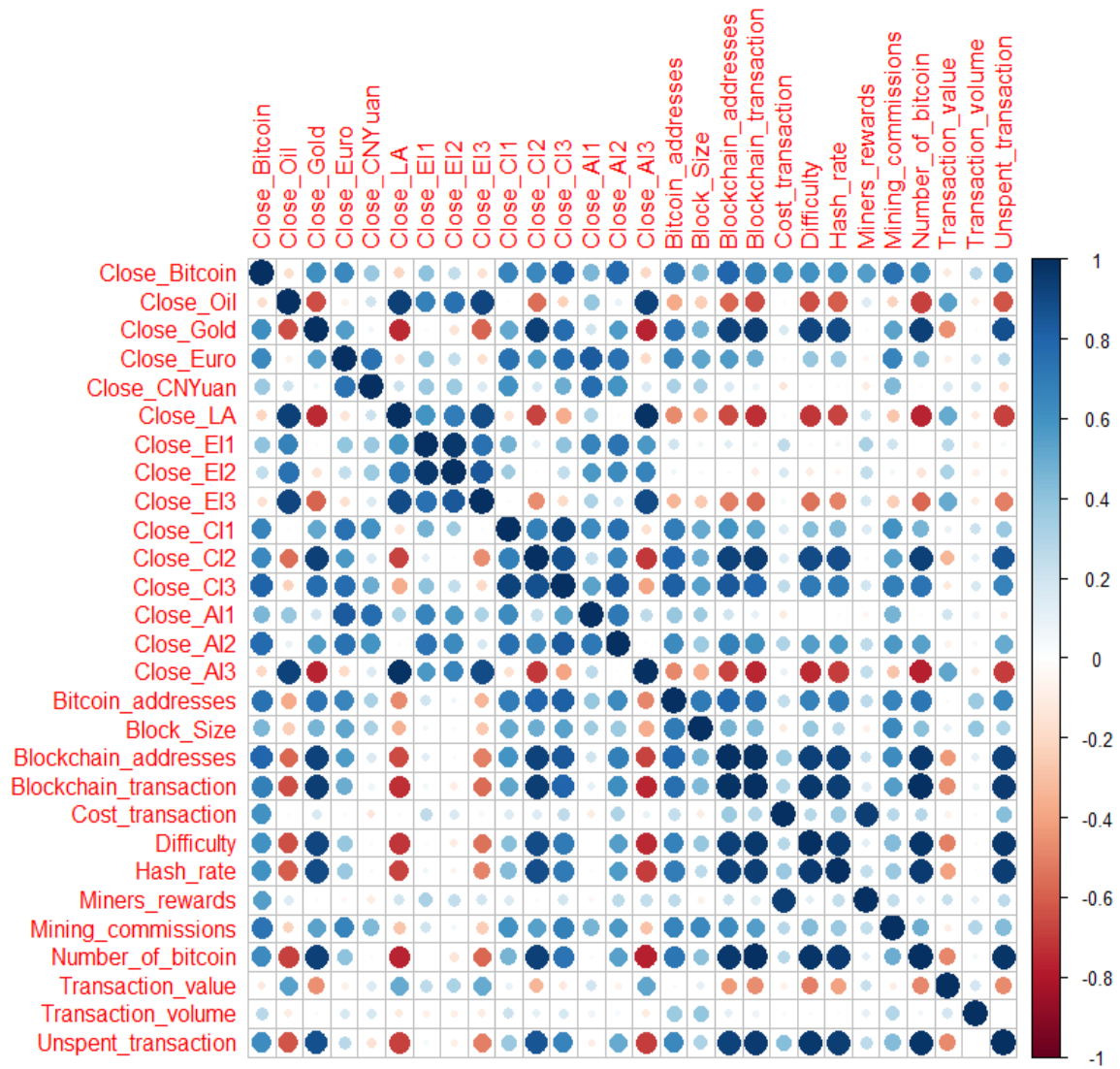


Figure 4. 4 presents the scatter plots of Close bitcoin with respect to the other variables are along with the coefficient of correlation and the P-value of the Pearson test.

### 4.2.3 Multivariate analysis and Correlation plot

The figure 4.5 shows the level of correlation between the variables. As a matter of fact, correlation coefficients are shaded in this graph according to their values where the highest correlations positive or negative are designed as dark blue or red respectively and the lowest one with brighter colors. The highly correlated variables are the variables that has the darkest dot between each other. Close Oil is correlated with LA, EI3, AI3 and Number of bitcoin. Close Gold is correlated with LA, CI2, Blockchain addresses, Blockchain transaction, Difficulty, Hash rate, Number of bitcoin and Unspent transaction. Close LA is correlated with EI3, AI3, Blockchain transaction, Difficulty, Number of bitcoin and Unspent transaction. Close EI1 is correlated with EI2. Close CI2 is correlated with AI3, Blockchain addresses, Blockchain transaction, Difficulty, Hash rate, Number of bitcoin and Unspent transaction. Close AI3 is correlated with Blockchain transaction, Difficulty, Hash rate, Number of bitcoin and Unspent transaction. Blockchain addresses are correlated with Blockchain transaction, Difficulty, Hash rate, Number of bitcoin and Unspent transaction. Blockchain transaction is correlated with Difficulty, Hash rate, Number of bitcoin and Unspent transaction. Cost transaction is correlated with Miners rewards. Difficulty is correlated with Hash rate, Number of bitcoin and Unspent transaction. Hash rate is correlated with Number of bitcoin and Unspent transaction. Number of bitcoin is correlated with Unspent transaction.





**Figure 4. 5: Correlation Plot of each variable with the Close variables and the demand supply variables**

## 4.3 Modeling and machine learning

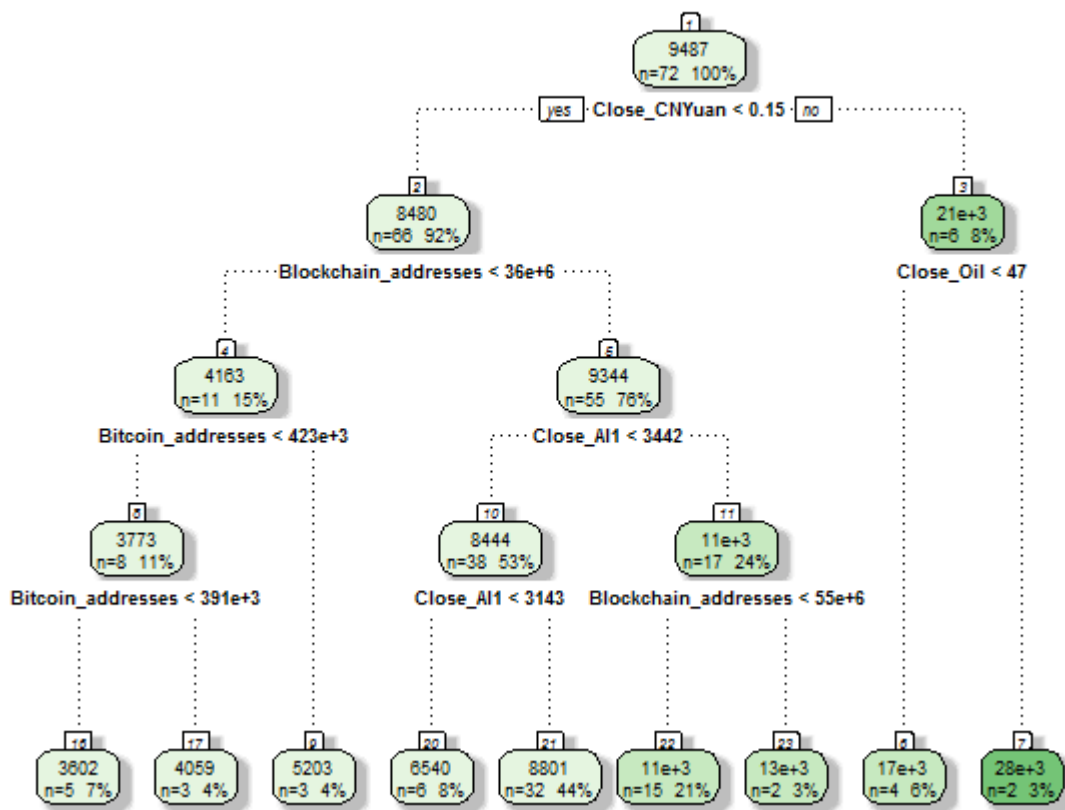
### 4.3.1 Regression Decision Tree

First the decision tree was conducted including all the independent variables. Below are the results of the conducted decision tree with close bitcoin as the predictor through training the data by assigning a min-split of 5 and max-depth of 4:

- 1) Root 72 /1.449653e+09 / 9486.818
- 2) Close\_CNYuan < 0.15035 / 66 / 4.365771e+08 / 8480.247
- 4) Blockchain\_addresses < 3.643691e+07 / 11 / 4.923193e+06 / 4163.109
- 8) Bitcoin\_addresses < 423361.5 / 8 / 4.312531e+05 / 3773.288
- 16) Bitcoin\_addresses < 391384 / 5 / 2.016449e+04 / 3601.960 \*
- 17) Bitcoin\_addresses >= 391384 / 3 / 1.971381e+04 / 4058.833 \*
- 9) Bitcoin\_addresses >= 423361.5 / 3 / 3.442129e+04 / 5202.633 \*
- 5) Blockchain\_addresses >= 3.643691e+07 / 55 / 1.856365e+08 / 9343.675
- 10) Close\_AII < 3441.952 / 38 / 6.956731e+07 / 8443.661
- 20) Close\_AII < 3143.153 / 6 / 3.646715e+06 / 6540.267 \*
- 21) Close\_AII >= 3143.153 / 32 / 4.010737e+07 / 8800.547 \*
- 11) Close\_AII >= 3441.952 / 17 / 1.648375e+07 / 11355.470
- 22) Blockchain\_addresses < 5.490356e+07 / 15 / 6.239539e+06 / 11075.230 \*
- 23) Blockchain\_addresses >= 5.490356e+07 / 2 / 2.312680e+05 / 13457.250 \*
- 3) Close\_CNYuan >= 0.15035 / 6 / 2.106330e+08 / 20559.100
- 6) Close\_Oil < 47.025 / 4 / 8.842934e+06 / 16829.320 \*
- 7) Close\_Oil >= 47.025 / 2 / 3.485541e+07 / 28018.650 \*

The star at the end of some line presents the terminal node. We started with 72 observations at the root node with the first predictor we split on (the first variable that enhances a decrease in SSE) is Close\_CNYuan. All the observation related are in the 2<sup>nd</sup> node with  $Close\_CNYuan < 0.15035$  with 66 total observations followed by an average Close Bitcoin price of 8,480.247, and an SSE of 436,577,100. In the 3<sup>rd</sup> branch ( $Close\_CNYuan \geq 0.15035$ ) 6 observation are found with average Close Bitcoin price of 20,559.1 and SSE of 210,633,000. This tells us that the most important variable has primarily the biggest reduction of SSE is the Close. This decision tree presents an RMSE of 1,095.264 and an MAPE of 9.1%.

Figure 4.6 presents the split of the decision tree of all the variables. For illustration and to explain the figure by looking at the extreme right side of this figure, when the Close CNYuan is greater than 0.15 and the Close oil is greater than 47 the average Close price of Bitcoin is 28,000.

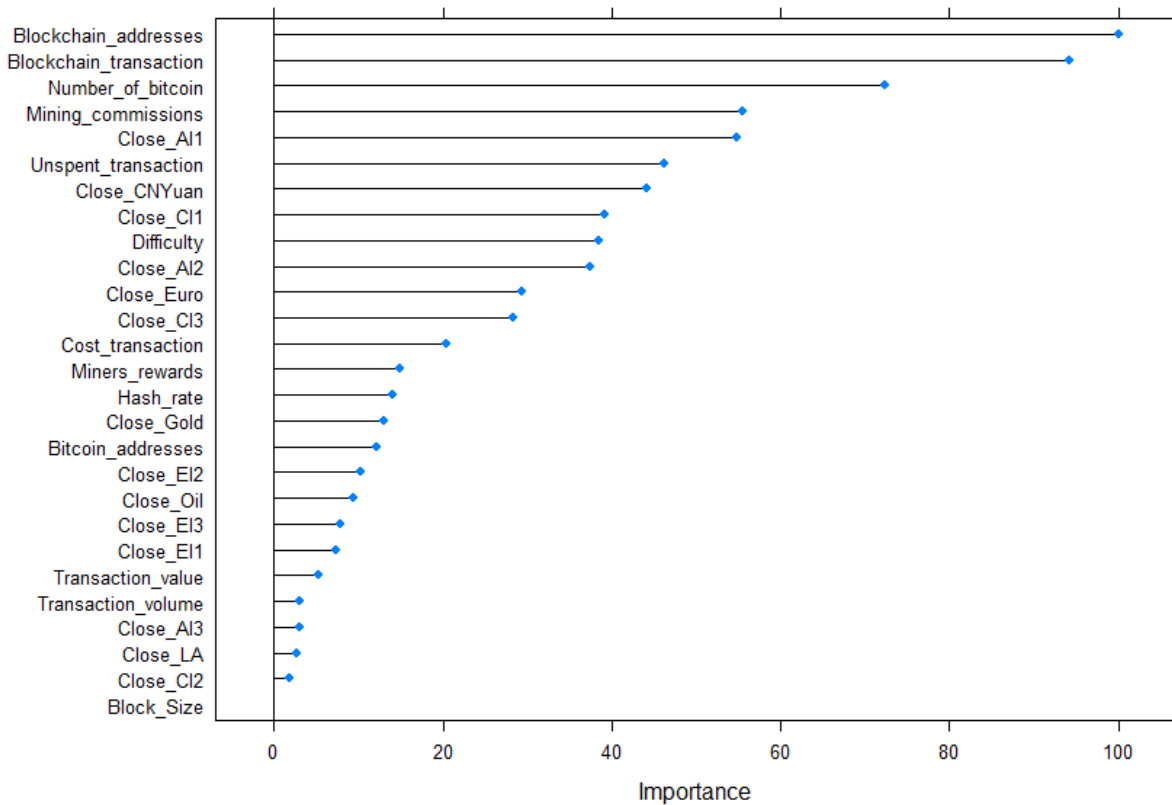


**Figure 4. 6: Tree of the variables**

An optimization algorithm was conducted to calculate the min-split and max-depth required to reduce the error. The results showed that min-split of 10 and max-depth of 12 was reduced the error from 0.3 to 0.28. The RMSE and MAPE of the decision tree with that min-split of 10 and max-depth of 12 are 1,589.19 and 10.26% respectively.

Choosing the right variables to be plotted and reach the minimum MAPE, bagging which is a technique for combining and averaging numerous models was applied. Averaging across many trees minimizes the variability of each tree and reduces overfitting, resulting an improved predictive performance. Through it, variables are ranked from the most to the least by importance indices presented in the figure 4.7. The cross-validated RMSE,  $R^2$  and MAE calculated and associated with the 27 variables in the model are as follows 2,072.96, 0.81 and 1,382.623 respectively. The sum of squared errors (SSE) for the observations in the testing set is used to evaluate the performance of a decision tree for regression. This decision tree has an RMSE = 1,810.049, which will serve as a benchmark against which advanced algorithm. The most essential predictors are those that have the greatest average impact on SSE.

The variable importance for regression trees is calculated as the average over all  $m$  trees of the overall amount of SSE reduced by splits over a given predictor. Thus the importance variable is scaled over 100 i.e. 100 is the score of the most important variable. Figure 4.7 represents the importance scale of all variables. The Blockchain addresses have the highest score where close AI3 have the lowest score. Thus, the most important variables used are in the following machine learning methods are: Close Bitcoin, Blockchain addresses, Blockchain transaction, Number of bitcoin, Mining commissions, Close AI1, Unspent transaction, Close CNYuan, Close CI1, Difficulty, and Close AI2.



**Figure 4. 7: The important variables of the Decision Tree**

Thus a new decision tree was conducted with the most important variables. The results are as follows:

- 1) Root 72 /1.449653e+09 /9486.818
- 2) Close\_CNYuan < 0.15035 / 66 / 4.365771e+08 / 8480.247
- 4) Blockchain\_addresses < 3.643691e+07 / 11 / 4.923193e+06 / 4163.109
- 8) Blockchain\_addresses < 3.472359e+07 / 8 / 4.312531e+05 / 3773.287 \*
- 9) Blockchain\_addresses >= 3.472359e+07 / 3 / 3.442129e+04 / 5202.633 \*
- 5) Blockchain\_addresses >= 3.643691e+07 / 55 / 1.856365e+08 / 9343.675
- 10) Close\_All < 3441.952 / 38 / 6.956731e+07 / 8443.661
- 20) Close\_All < 3143.153 / 6 / 3.646715e+06 / 6540.267 \*
- 21) Close\_All >= 3143.153 / 32 / 4.010737e+07 / 8800.547
- 42) Mining\_commissions < 0.9686661 / 15 / 8.504373e+06 / 8036.493
- 84) Difficulty < 1.335e+13 / 10 / 1.347217e+06 / 7571.950
- 168) Close\_CNYuan >= 0.14165 / 7 / 1.297246e+05 / 7346.629 \*
- 169) Close\_CNYuan < 0.14165 / 3 / 3.286496e+04 / 8097.700 \*
- 85) Difficulty >= 1.335e+13 / 5 / 6.831403e+05 / 8965.580 \*
- 43) Mining\_commissions >= 0.9686661 / 17 / 1.511986e+07 / 9474.712
- 86) Unspent\_transaction < 5.671347e+07 / 4 / 5.543252e+05 / 8321.750 \*
- 87) Unspent\_transaction >= 5.671347e+07 / 13 / 7.612168e+06 / 9829.469
- 174) Difficulty >= 9.035e+12 / 10 / 2.506817e+06 / 9493.120
- 348) Blockchain\_addresses >= 4.25243e+07 / 7 / 9.172852e+05 / 9269.100 \*
- 349) Blockchain\_addresses < 4.25243e+07 / 3 / 4.185492e+05 / 10015.830 \*
- 175) Difficulty < 9.035e+12 / 3 / 2.030164e+05 / 10950.630 \*
- 11) Close\_All >= 3441.952 / 17 / 1.648375e+07 / 11355.470
- 22) Blockchain\_addresses < 5.45751e+07 / 14 / 6.151368e+06 / 11054.740
- 44) Difficulty >= 1.71e+13 / 8 / 3.172494e+06 / 10711.700 \*
- 45) Difficulty < 1.71e+13 / 6 / 7.822112e+05 / 11512.130 \*
- 23) Blockchain\_addresses >= 5.45751e+07 / 3 / 3.157704e+06 / 12758.870 \*
- 3) Close\_CNYuan >= 0.15035 / 6 / 2.106330e+08 / 20559.100 \*

The star at the end of some line presents the terminal node. We started with 72 observations at the root node with the first predictor we split on (the first variable that enhances a decrease in SSE) is Close\_CNYuan. All the observation related are in the 2<sup>nd</sup> node with  $Close\_CNYuan < 0.15035$  with 66 total observations followed by an average Close Bitcoin price of 8,480.247, and an SSE of 436,577,100. In the 3<sup>rd</sup> branch ( $Close\_CNYuan \geq 0.15035$ ) 6 observation are found with average Close Bitcoin price of 20,559.1 and SSE of 210,633,000. This tells us that the most important variable has primarily the biggest reduction of SSE is the Close. The RMSE and the MAPE after applying the new decision tree by choosing only the 10 important variables with a min-split (10) and max-split (12) are 1,280.9 and 6.92% respectively. Thus, this new tree has reduced the MAPE by approximately 3.34%.

The figure 4.8 is the new decision tree with the most important variables. For illustration and to explain the figure by looking at the extreme right side of this figure, when the Close\_CNYuan is greater than 0.15 the average Close price of Bitcoin is 21,000.

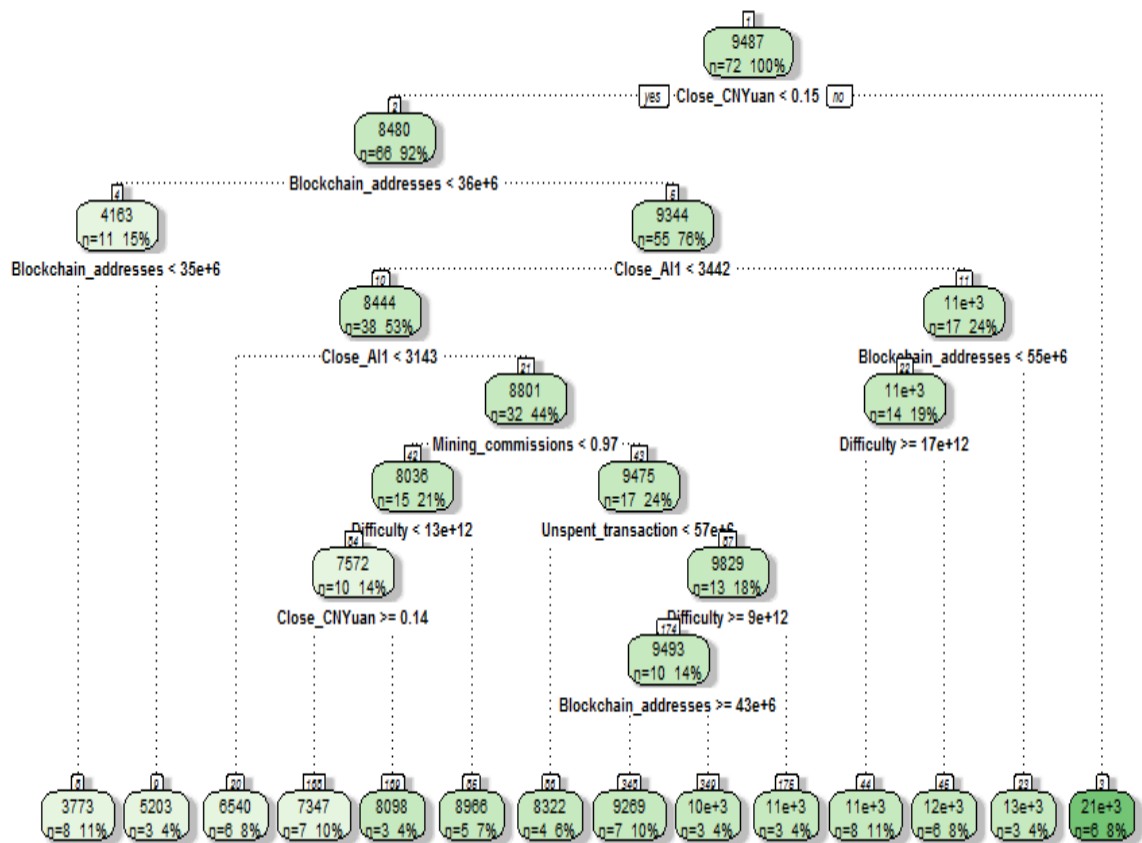


Figure 4. 8: Decision Tree Plot of the most important variables

### **4.3.2 Multiple Regression**

To predict the value of the future values of bitcoin, the first step is to use the multiple regression with  $y$  the dependent variable as the Close value of the Bitcoin and  $x$  the independent variable as the Close value of the other variables and the demand/supply variables (27 variables in total). The second step is to use the multiple regression with  $y$  the dependent variable as the Close value of the Bitcoin and  $x$  the most important independent variable deduced from the decision tree (Close Bitcoin, Blockchain addresses, Blockchain transaction, Number of bitcoin, Mining commissions, Close AI1, Unspent transaction, Close CNYuan, Close CI1, Difficulty, and Close AI2). The third step to be more precise is to adjust the cook distance of the values and apply the decision tree model values with new training data to obtain the most significant variables. The fourth step is to use the significant variables (Close Bitcoin, Blockchain addresses, Blockchain transaction, Number of bitcoin, and Close CNYuan). The final step is to apply the tukey test value (Close Bitcoin, Blockchain addresses, Blockchain transaction, log (Number of bitcoin), and Close CNYuan)

#### **4.3.2.1 Assumptions for regression**

In this section, the multicollinearity between the variables is conducted as part A shows, whereas Part B shows the application of the regression model in different parts by applying all the 27 variables and eliminating throughout the process to get the most significant variables where  $P\text{-value} < 0.05$ . This process starts by applying all the variables, apply the important variables of decision, checking their cooks distance and adjusting it with a new model and adjusted training dataset. Finally after performing the new model with a new training data, we clean and choose the significant variables to be able to get a model with only significant variables.

In this part,  $P$ -value is calculated in addition to the  $R$ -squared, Adjusted  $R$ -squared, MAPE, AIC and BIC.

## A- Multicollinearity

A multicollinearity problem is caused by a highly related two independent variables, which leads to problems with the analysis and interpretation. To investigate possible multicollinearity, first look at the correlation coefficients for each pair of continuous variables. Correlations of 0.8 or above suggest a strong relationship and only one of the two variables is needed in the regression analysis.

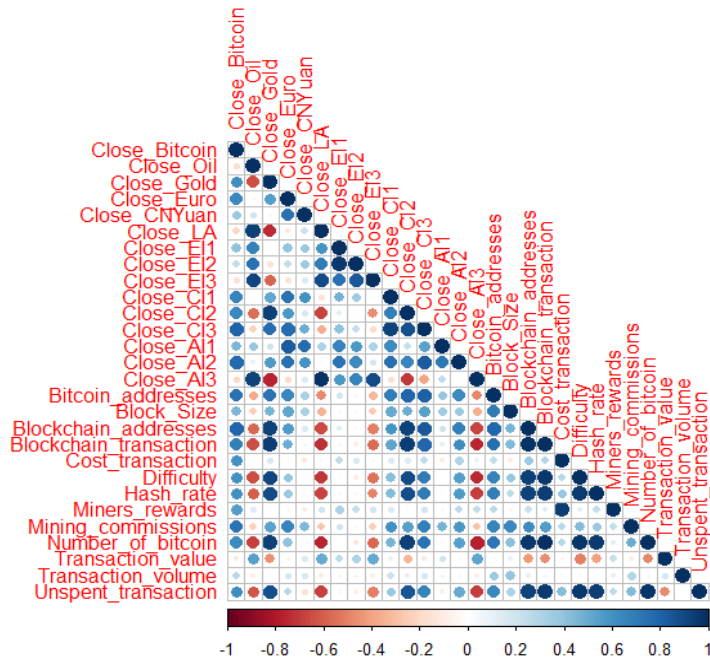
The Pearson's correlation coefficients rounded to two decimal places are presented in table 4.6.

**Table 4. 6: Correlation table of the weekly Close Price and demand/supply variables**

Variables	Bitcoin	Oil	Gold	Euro	CNYuan	LA	E1	E2	E3	C11	C12	C13	A11	A12	A13	Blockchain Transaction	Cost Transaction	Difficulty	Hash_rate	Miners Rewards	Mining Commission	Number Of Bitcoin	Transaction Value	Transaction Volume	Unspent Transaction	
Close_Bitcoin	1.00																									
Close_Oil	-0.18	1.00																								
Close_Gold	0.61	-0.65	1.00																							
Close_Euro	0.64	-0.07	0.56	1.00																						
Close_CNYuan	0.38	0.20	0.06	0.75	1.00																					
Close_LA	-0.21	<b>0.94</b>	-0.74	-0.14	0.21	1.00																				
Close_E1	0.41	0.67	0.02	0.39	0.37	0.60	1.00																			
Close_E2	0.24	0.74	-0.15	0.24	0.36	0.70	<b>0.96</b>	1.00																		
Close_E3	-0.15	<b>0.91</b>	-0.59	-0.16	0.16	<b>0.90</b>	0.74	<b>0.85</b>	1.00																	
Close_C11	0.67	0.02	0.52	0.74	0.61	-0.14	0.48	0.35	0.02	1.00																
Close_C12	0.64	-0.55	<b>0.93</b>	0.57	0.18	-0.68	0.13	-0.01	-0.46	0.69	1.00															
Close_C13	<b>0.80</b>	-0.23	0.76	0.77	0.50	-0.37	0.40	0.25	-0.20	<b>0.93</b>	<b>0.88</b>	1.00														
Close_A11	0.45	0.39	0.20	<b>0.83</b>	0.76	0.32	0.67	0.57	0.32	0.64	0.24	0.54	1.00													
Close_A12	0.78	0.10	0.56	0.75	0.60	0.03	0.74	0.63	0.18	0.76	0.65	<b>0.84</b>	0.71	1.00												
Close_A13	-0.21	<b>0.93</b>	-0.77	-0.19	0.16	<b>0.99</b>	0.57	0.66	<b>0.90</b>	-0.16	-0.70	-0.39	0.27	-0.01	1.00											
Blockchain Transaction	0.69	-0.65	<b>0.95</b>	0.49	0.07	-0.73	0.05	-0.10	-0.56	0.53	<b>0.94</b>	0.79	0.11	0.61	-0.75	1.00										
Cost Transaction	0.61	-0.02	0.17	-0.03	-0.13	0.04	0.25	0.16	0.05	0.14	0.14	0.26	-0.11	0.31	0.09	0.29	1.00									
Difficulty	0.60	-0.64	<b>0.91</b>	0.39	-0.03	-0.71	0.02	-0.10	-0.55	0.42	<b>0.90</b>	0.71	0.00	0.55	-0.74	<b>0.97</b>	0.28	1.00								
Hash_rate	0.61	-0.61	<b>0.89</b>	0.38	-0.03	-0.68	0.06	-0.07	-0.50	0.43	<b>0.89</b>	0.70	0.02	0.56	-0.69	<b>0.95</b>	0.38	<b>0.95</b>	1.00							
Miners Rewards	0.56	0.15	0.02	-0.05	-0.10	0.19	0.33	0.25	0.19	0.15	0.03	0.20	-0.04	0.26	0.25	0.14	<b>0.95</b>	0.12	0.24	1.00						
Mining Commission	0.74	-0.23	0.54	0.66	0.44	-0.28	0.21	0.06	-0.25	0.61	0.54	0.69	0.47	0.58	-0.28	0.55	0.27	0.43	0.38	0.25	1.00					
Number Of Bitcoin	0.64	-0.68	<b>0.94</b>	0.41	-0.02	-0.76	0.00	-0.14	-0.58	0.46	<b>0.93</b>	0.74	0.02	0.55	-0.78	<b>0.99</b>	0.29	<b>0.97</b>	<b>0.95</b>	0.13	0.50	1.00				
Transaction Value	-0.11	0.55	-0.45	-0.07	0.15	0.51	0.26	0.32	0.51	0.09	-0.34	-0.13	0.20	-0.07	0.52	-0.47	-0.07	-0.50	-0.40	0.23	-0.10	-0.48	1.00			
Transaction Volume	0.27	-0.10	0.04	0.17	0.18	-0.11	-0.07	-0.08	-0.10	0.21	0.11	0.17	0.07	0.05	-0.11	0.07	0.04	0.03	0.02	0.11	0.30	0.06	0.19	1.00		
Unspent Transaction	0.62	-0.62	0.88	0.28	-0.16	-0.69	0.04	-0.09	-0.51	0.38	<b>0.86</b>	0.67	-0.09	0.51	-0.69	<b>0.96</b>	0.42	<b>0.97</b>	<b>0.94</b>	0.27	0.44	<b>0.97</b>	-0.48	0.00	1.00	

The correlations between the variables are presented in the figure 4.9 as a lower triangular shape where the correlation coefficients are shaded according to their values. The highest correlations positive or negative are designed as dark blue or red respectively and the lowest one with brighter colors.





**Figure 4. 9: Correlation Plot of the weekly Close prices and demand/supply variables**

A strong relationship was observed between:

- Close LA & Oil
- EI2 & EI1
- EI3 & Oil, LA, EI2
- CI2 & Gold
- CI3 & Bitcoin, CI1, CI2
- AI1 & Euro
- AI2 & CI3
- AI3 & Oil, LA, EI3
- Blockchain Transaction & Gold, CI2
- Difficulty & Gold, CI2, Blockchain Transaction
- Hash rate & Gold, CI2, Blockchain Transaction, Difficulty
- Miners rewards & Cost per Transaction
- Number of Bitcoin & Gold, CI2, Blockchain Transaction, Difficulty, Hash rate
- Unspent Transaction & CI2, Blockchain Transaction, Difficulty, Hash rate, Number of Bitcoin

## B- Application of the LMMOD

The multiple regression used with  $y$  the dependent variable as the Close value of the Bitcoin and  $x$  the independent variable as the Close value of the other variables and the demand/supply variables (27 variables in total) is denoted by LMMOD

The first step is to run the entire variable in the multiple regression to display the significant and non-significant variables.

Table 4.7 presents the estimated value of the coefficient  $B_i$ , the test value of the t-test, and the p-value of the t-test by applying all the variables. Thus the model is:

(41)

$$\begin{aligned} \text{Close}_{\text{Bitcoin}} = & -117,900 + 85.66 * \text{Close}_{\text{Oil}} + 10.69 * \text{Close}_{\text{Gold}} - 13,820 * \text{Close}_{\text{Euro}} \\ & + 171,100 * \text{Close}_{\text{CNYuan}} - 284.1 * \text{Close}_{\text{LA}} - 3,58 * \text{Close}_{\text{EI1}} + 11.42 \\ & * \text{Close}_{\text{EI2}} - 4.69 * \text{Close}_{\text{EI3}} - 27.74 * \text{Close}_{\text{CI1}} + 49,170 * \text{Close}_{\text{CI2}} \\ & + 4.26 * \text{Close}_{\text{CI3}} - 1.92 * \text{Close}_{\text{AI1}} + 62.08 * \text{Close}_{\text{AI2}} + 0.23 * \text{Close}_{\text{AI3}} \\ & + 0 * \text{Bitcoin}_{\text{addresses}} - 473 * \text{Block}_{\text{Size}} + 0 * \text{Blockchain}_{\text{addresses}} + 0 \\ & * \text{Blockchain}_{\text{transactions}} - 60.77 * \text{Cost}_{\text{transaction}} + 0 * \text{Difficulty} + 0 \\ & * \text{Hash}_{\text{rate}} + 0 * \text{Miners}_{\text{rewards}} + 100.3 * \text{Mining}_{\text{commissions}} + 0.01 \\ & * \text{Number}_{\text{Bitcoin}} - 0.01 * \text{Transaction}_{\text{value}} + 0.01 * \text{Transction}_{\text{volume}} \\ & + 0 * \text{Unspent}_{\text{transaction}} \end{aligned}$$

**Table 4. 7: the estimated coefficient, the test statistics and the p-value of the t-test of LMOD**

<b><i>LMMOD</i></b>	<b>Estimate</b>	<b>T-test</b>	<b>P-Value</b>
<b>(Intercept)</b>	-117,900.00	-0.41	0.69
<b>Close_Oil</b>	85.66	1.63	0.11
<b>Close_Gold</b>	10.69	3.11	0.00
<b>Close_Euro</b>	-13,820.00	-0.74	0.46
<b>Close_CNYuan</b>	171,100.00	1.13	0.27
<b>Close_LA</b>	-284.10	-1.23	0.23
<b>Close_EI1</b>	-3.58	-0.83	0.41
<b>Close_EI2</b>	11.42	1.46	0.15
<b>Close_EI3</b>	-4.69	-0.73	0.47
<b>Close_CI1</b>	-27.74	-0.62	0.54
<b>Close_CI2</b>	49,170.00	1.33	0.19
<b>Close_CI3</b>	4.26	0.10	0.92
<b>Close_AI1</b>	-1.92	-0.51	0.62
<b>Close_AI2</b>	62.08	0.77	0.44
<b>Close_AI3</b>	0.23	0.23	0.82
<b>Bitcoin_addresses</b>	0.00	-0.31	0.76
<b>Block_Size</b>	-473.00	-0.40	0.69
<b>Blockchain_addresses</b>	0.00	6.65	0.00
<b>Blockchain_transaction</b>	0.00	-2.60	0.01
<b>Cost_transaction</b>	-60.77	-0.67	0.51
<b>Difficulty</b>	0.00	0.63	0.53
<b>Hash_rate</b>	0.00	-2.58	0.01
<b>Miners_rewards</b>	0.00	1.24	0.22
<b>Mining_commissions</b>	100.30	0.94	0.35
<b>Number_of_bitcoin</b>	0.01	0.46	0.65
<b>Transaction_value</b>	-0.01	-0.61	0.54
<b>Transaction_volume</b>	0.01	1.83	0.07
<b>Unspent_transaction</b>	0.00	0.47	0.64

From table 4.7 only the following variables are reported to be significant (p-value <0.05) Close Gold, Blockchain addresses, Blockchain transaction, Hash rate, and Transaction volume, which is due to multicollinearity.

R-squared, Adjusted R-squared, MAPE AIC, and BIC are reported in table 4.8:

**Table 4. 8 R-squared, adjusted R-squared, MAPE, AIC and BIC of LMMOD**

<b>Multiple R-squared</b>	<b>Adjusted R-squared</b>	<b>MAPE</b>	<b>AIC</b>	<b>BIC</b>
0.98	0.97	8.8%	1,189.49	1,255.51

98% (r-squared) of the variation in the close bitcoin is explained by this model and 2% is due to other variables not included in the model. The AIC and BIC are 1,189.49 and 1,255.51 respectively. The MAPE is approximately 8.8% indicating that the accuracy of the model is approximately 91.2%.

The second step is to apply the important variables found through bagging the Close and demand/supply variables by Decision Tree. LMMOD was adjusted to LMMOD (1) by including only the following variables: Close Bitcoin, Blockchain addresses, Blockchain transaction, Number of bitcoin, Mining commissions, Close AI1, Unspent transaction, Close CNYuan, Close CI1, Difficulty, and Close AI2. The fitted equation is as follows:

(42)

$$\begin{aligned}
 Close_{Bitcoin} = & -582,400 + 0 * Blockchain_{addresses} + 0 * Blockchain_{transactions} + 0.04 \\
 & * Number_{Bitcoin} + 311.5 * Mining_{commissions} + 3.93 * Close_{AI1} + 0 \\
 & * Unspent_{transaction} + 217,400 * Close_{CNYuan} + 3.08 * Close_{CI1} + 0 \\
 & * Difficulty - 19.63 * Close_{AI2}
 \end{aligned}$$

Table 4.9 presents the estimated value of the coefficient  $B_i$ , the test value of the t-test, and the p-value of the t-test by applying all the variables.

**Table 4. 9: The results of the Close variables and demand/supply with the training data of the first trial**

<b>LMMOD (1)</b>	<b>Estimate</b>	<b>T-test</b>	<b>P-Value</b>
<b>(Intercept)</b>	-582,400.00	-3.26	0.00
<b>Blockchain_addresses</b>	0.00	15.19	0.00
<b>Blockchain_transaction</b>	0.00	-6.61	0.00
<b>Number_of_bitcoin</b>	0.04	3.32	0.00
<b>Mining_commissions</b>	311.50	3.35	0.00
<b>Close_AI1</b>	3.93	1.39	0.17
<b>Unspent_transaction</b>	0.00	1.27	0.21
<b>Close_CNYuan</b>	217,400.00	1.81	0.07
<b>Close_CI1</b>	3.08	0.48	0.63
<b>Difficulty</b>	0.00	2.01	0.05
<b>Close_AI2</b>	-19.63	-0.54	0.59

Close AI2, Close CI1, Unspent transaction, and Close AI1 are reported not significant to the model since its p-value >0.05.

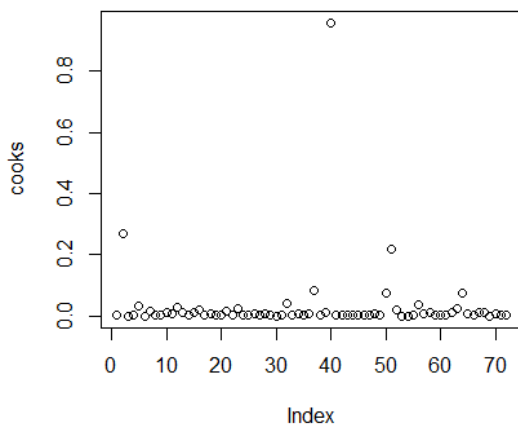
R-squared, Adjusted R-squared, MAPE, AIC, and BIC are reported in table 4.10:

**Table 4. 10: R-squared, adjusted R-squared, MAPE, AIC and BIC of the LMMOD (1)**

<b>Multiple R-squared</b>	<b>Adjusted R-squared</b>	<b>MAPE</b>	<b>AIC</b>	<b>BIC</b>
0.96	0.95	9.15%	1,206.95	1,233.97

96% (r-squared) of the variation in the Close Bitcoin is explained by this model and 4% is due to other variables not included in the model. The AIC and BIC are 1,206.95 and 1,233.97 respectively. The MAPE is approximately 9.15% which slightly greater than the LMMOD indicating that the accuracy of the model is approximately 90.85%.

The third step is to adjust the cook distance of the LMMOD (1). As the figure 4.10 shows many point are above the constant point line approximately on 0 which need to be removed. Thus, LMMOD (cooks) include the same variables as LMMOD (1) (Close Bitcoin, Blockchain addresses, Blockchain transaction, Number of bitcoin, Mining commissions, Close AI1, Unspent transaction, Close CNYuan, Close CI1, Difficulty, and Close AI2 by eliminating the influential points.



**Figure 4. 10: Cooks distance of the LMMOD (1)**

Table 4.11 presents the estimated value of the coefficient  $B_i$ , the test value of the t-test, and the p-value of the t-test of the LMMOD(cooks) and table 4.12 represent the R-squared, Adjusted R-squared, MAPE AIC, and BIC.

**Table 4. 11: The results of the Close variables and demand/supply of elimination of some values from the cooks distance**

<i>LMMOD (cooks)</i>	<b>Estimate</b>	<b>T-test</b>	<b>P-Value</b>
<b>(Intercept)</b>	-552,800.00	-2.86	0.01
<b>Blockchain_addresses</b>	0.00	8.98	0.00
<b>Blockchain_transaction</b>	0.00	-5.24	0.00
<b>Number_of_bitcoin</b>	0.03	2.83	0.01
<b>Mining_commissions</b>	155.70	1.81	0.08
<b>Close_AI1</b>	2.43	0.98	0.33
<b>Unspent_transaction</b>	0.00	0.93	0.36
<b>Close_CNYuan</b>	226,700.00	2.13	0.04
<b>Close_CI1</b>	2.20	0.40	0.69
<b>Difficulty</b>	0.00	1.45	0.15
<b>Close_AI2</b>	-4.72	-0.15	0.88

Close AI1, Close AI2, Close CI1, Unspent transaction, Difficulty and Mining commissions are reported not significant to the model since its p-value >0.05.

**Table 4. 12: R-squared, adjusted R-squared, MAPE, AIC and BIC of the LMMOD (cook)**

<b>Multiple R-squared</b>	<b>Adjusted R-squared</b>	<b>MAPE</b>	<b>AIC</b>	<b>BIC</b>
0.94	0.93	9.67%	1,134.84	1,161.64

94% (r-squared) of the variation in the Close Bitcoin is explained by this model and 6% is due to other variables not included in the model. The AIC and BIC are 1,134.84 and 1,161.64 respectively. The MAPE is approximately is 9.67%.

The Fourth step is to include the significant variables in the LMMOD (S): Close Bitcoin, Blockchain addresses, Blockchain transaction, Number of bitcoin, and Close CNYuan. Table 4.13 presents the estimated value of the coefficient  $B_i$ , the test value of the t-test, and the p-value of the t-test by applying all the variables. The fitted equation is as follows:

(43)

$$Close_{Bitcoin} = -389,400 + 0 * Blockchain_{addresses} + 0 * Blockchain_{transactions} + 0.02 * Number_{Bitcoin} + 166,100 * Close_{CNYuan}$$

**Table 4. 13: The results of the Close variables and demand/supply with the new training data**

<b>LMMOD(S)</b>	<b>Estimate</b>	<b>T-test</b>	<b>P-Value</b>
<b>(Intercept)</b>	-389,400.00	-2.78	0.01
<b>Blockchain_addresses</b>	0.00	14.93	0.00
<b>Blockchain_transaction</b>	0.00	-6.98	0.00
<b>Number_of_bitcoin</b>	0.02	2.88	0.01
<b>Close_CNYuan</b>	166,100.00	2.36	0.02

All the variables are now significant to the model.

R-squared, Adjusted R-squared, MAPE, AIC, and BIC are reported in table 4.14:

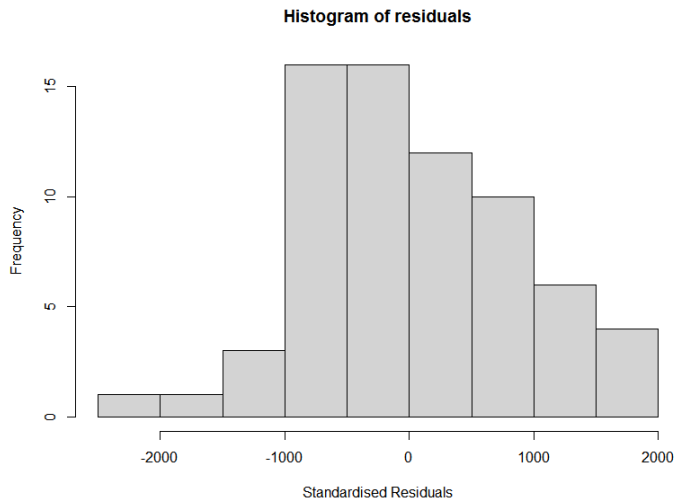
**Table 4. 14: R-squared, adjusted R-squared, MAPE, AIC and BIC of the LMMOD (S)**

<b>Multiple R-squared</b>	<b>Adjusted R-squared</b>	<b>MAPE</b>	<b>AIC</b>	<b>BIC</b>
0.92	0.92	11.12%	1,135.5	1,148.85

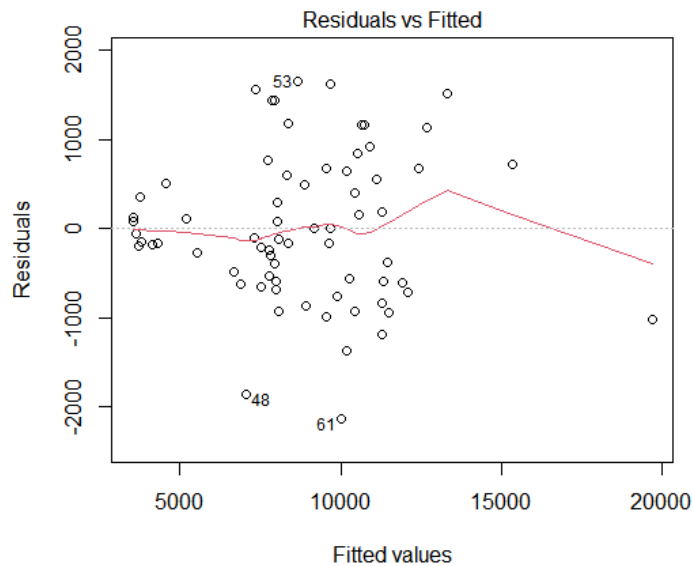
92% (r-squared) of the variation in the Close Bitcoin is explained by this model and 8% is due to other variables not included in the model which is similar to the LMMOD (cooks). The AIC and BIC are 1,135.5 and 1,148.85 respectively and they are less than the LMMOD (1). The MAPE of the Final LMMOD (S) is approximately 11.12 % higher than the LMMOD (1) but include only the significant variables.

The histogram of the residual is presented in figure 4.11 shows a positively skewed shape whereas the figure 4.12 represents the residuals vs fitted value. It demonstrates whether or not the residuals have non-linear patterns. If the model fails to capture the non-linear relationship between predictor variables and outcome variables, the pattern will appear in this plot. If the residuals are evenly distributed along a horizontal line with no discernible patterns, you don't have non-linear relationships. In our case, the residual vs fitted is distributed with the red line where some variables far from the line. The non-linear relationship appears in this plot and indicates dependency between the residuals and the fitted values, which suggest a different model.





**Figure 4. 11: Histogram of the residuals of selected variables**



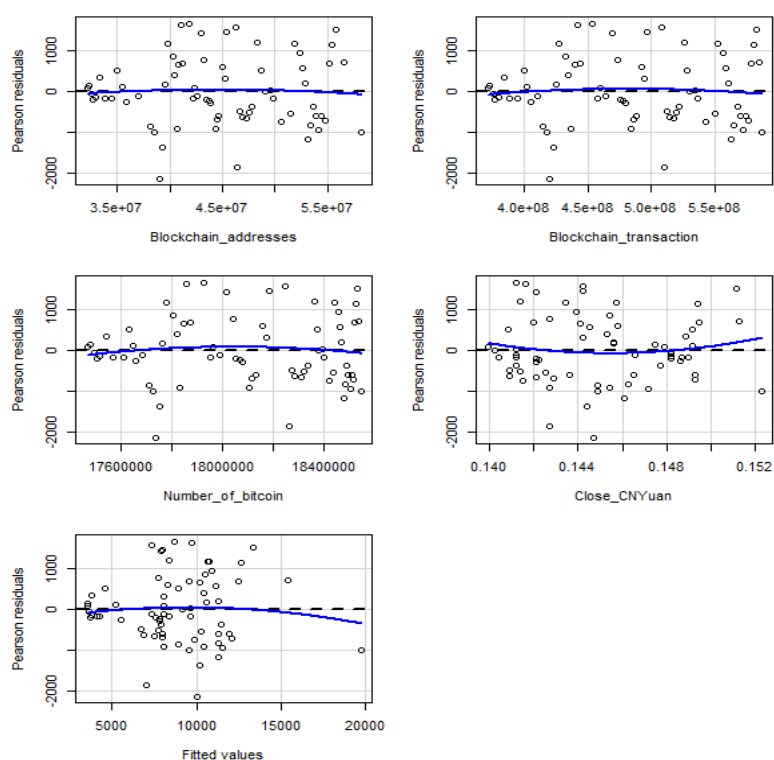
**Figure 4. 12: The plot of the residuals vs the fitted value**

The Studentized Breusch-Pagan test or BP-test indicate the values on the Residual vs Fitted Plot. In this case, from the table 4.15 we can determine that  $p\text{-value} = 0.17 > 0.05$  thus we fail to reject  $H_0$  and homoscedasticity is present. Thus, the model is accurate.

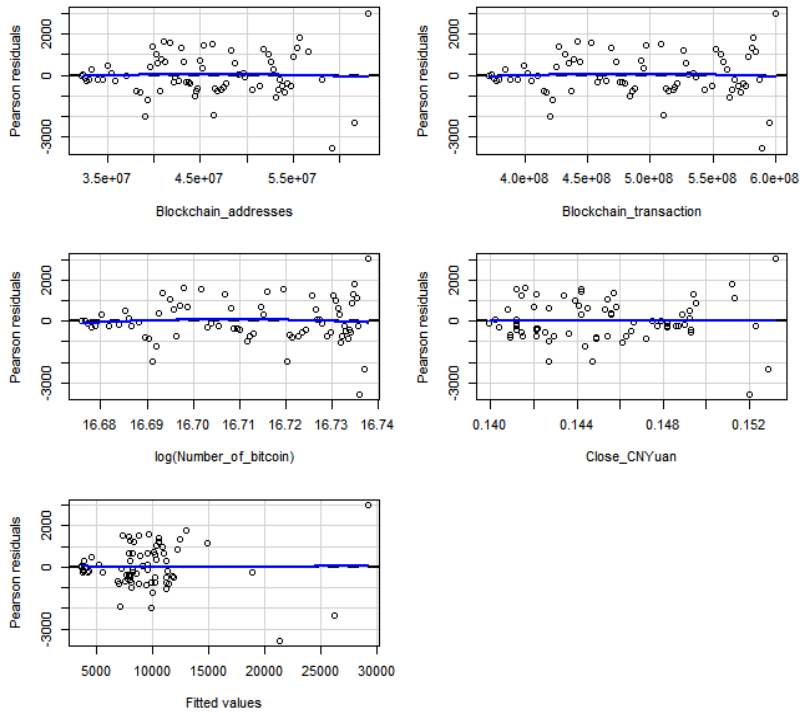
**Table 4. 15: Table of the Studentized Breusch-Pagan test**

Studentized Breusch-Pagan test	
BP =6.36	P-value =0.17

The Figure 4.13 known as partial regression plots or added variable plots where the y-axis represents the response variable, while the x-axis depicts a single forecaster variable. The blue curve symbolizes the relationship between the predictor variable and the response variable by keeping the values of all other predictor variables constant. As it shows the Number of bitcoin is logarithmic. Figure 4.14 present the partial regression after transforming the number of bitcoin to log number of bitcoins. Now, all the blue curve are aligned with the x-axis.



**Figure 4. 13: Partial fitted value plots**



**Figure 4. 14: Transformed Partial fitted value plots**

The table 4.16 shows the values of the significant variables of the partial fitted values plots in the figure 4.14 and indicates that the  $p\text{-value} > 0.05$  which indicates the linearity of the fitted value.

**Table 4. 16: Tukey test of the fitted value**

	<b>Test stat</b>	<b>P-Value</b>
<b>Blockchain_addresses</b>	-0.53	0.59
<b>Blockchain_transaction</b>	-0.71	0.48
<b>log(Number_of_bitcoin)</b>	-0.73	0.47
<b>Close_CNYuan</b>	-0.03	0.97
<b>Tukey test</b>	0.15	0.88

The Final step is to include the tukey test variable in the LMMOD (F): Close Bitcoin, Blockchain addresses, Blockchain transaction, log (Number of bitcoin), and Close CNYuan. Table 4.17 presents the estimated value of the coefficient  $B_i$ , the test value of the t-test, and the p-value of the t-test by applying all the variables. The fitted equation is as follows:

(44)

$$Close_{Bitcoin} = -5,577,000 + 0 * Blockchain_{addresses} + 0 * Blockchain_{transactions} + 336,400 * \log(Number_{Bitcoin}) + 116,500 * Close_{CNYuan}$$

**Table 4. 17: The results of the Close variables and demand/supply with the new training data**

<i>LMMOD(F)</i>	<b>Estimate</b>	<b>T-test</b>	<b>P-Value</b>
<b>(Intercept)</b>	-5,577,000.00	-2.08	0.04
<b>Blockchain_addresses</b>	0.00	19.06	0.00
<b>Blockchain_transaction</b>	0.00	-6.55	0.00
<b>log(Number_of_bitcoin)</b>	336,400.00	2.09	0.04
<b>Close_CNYuan</b>	116,500.00	1.43	0.016

R-squared, Adjusted R-squared, MAPE, AIC, and BIC are reported in table 4.18:

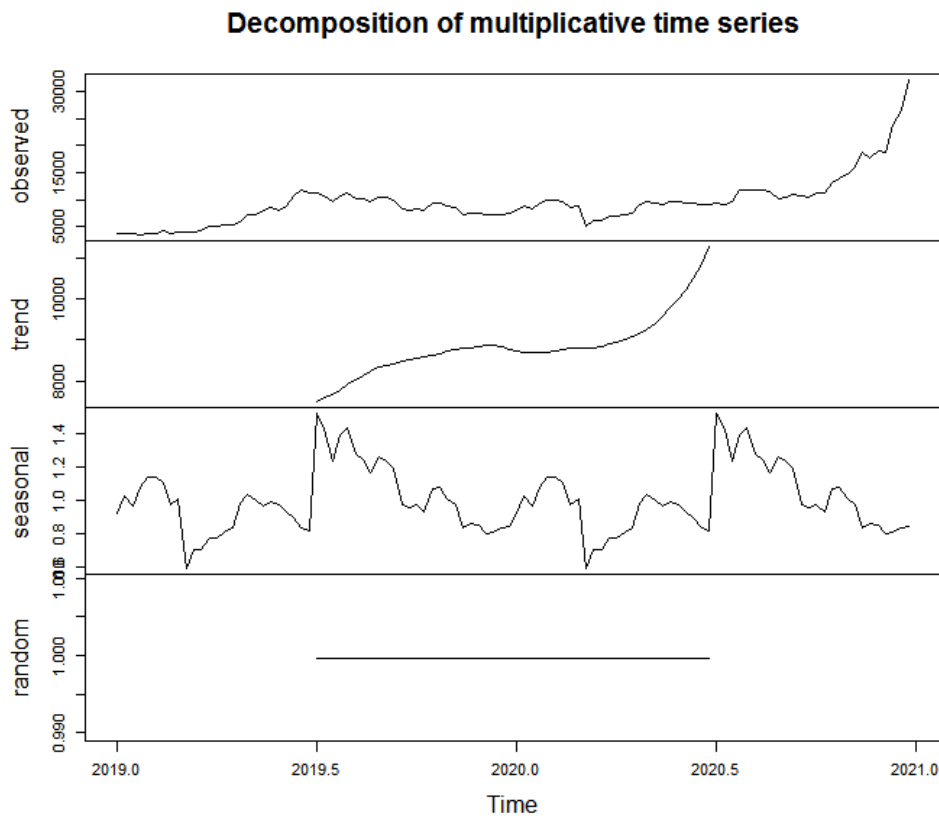
**Table 4. 18: R-squared, adjusted R-squared, MAPE, AIC and BIC of the LMMOD (F)**

<b>Multiple R-squared</b>	<b>Adjusted R-squared</b>	<b>MAPE</b>	<b>AIC</b>	<b>BIC</b>
0.95	0.94	10.76%	1,215.08	1,228.74

95% (r-squared) of the variation in the Close Bitcoin is explained by this model and 5% is due to other variables not included in the model. The AIC and BIC are 1,215.08 and 1,228.74 respectively. The MAPE of the Final LMMOD (F) is approximately 10.76% lower than the LMMOD (S) but include non-significant variables.

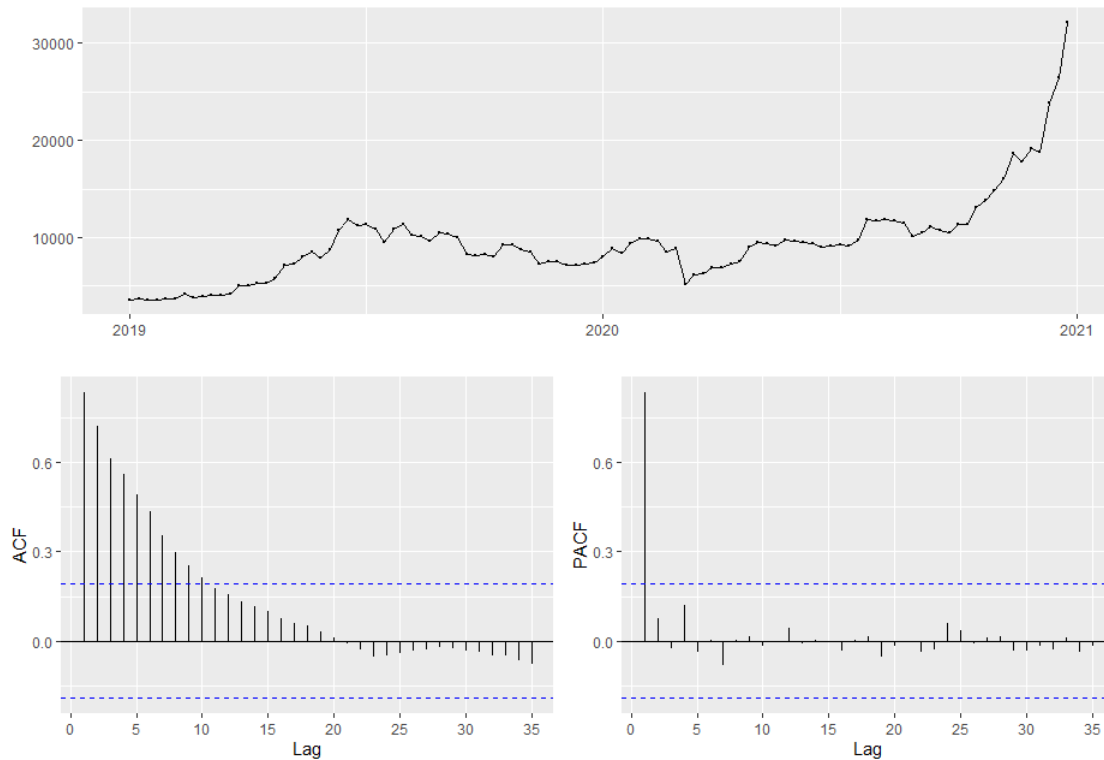
### 4.3.3 Autoregressive Integrated Moving Average (ARIMA) or Time Series

Using the data series defined in section 4.1, we will apply the Box & Jenkins approach to determine its corresponding time series ARIMA model. The figure 4.15 shows the different plots (observed, trend, seasonal and random) of the Close price of Bitcoin. An increasing trend, a seasonal fluctuating plot and the constant randomness at 1 were observed.



**Figure 4. 15 : Decomposition of multiplicative time series**

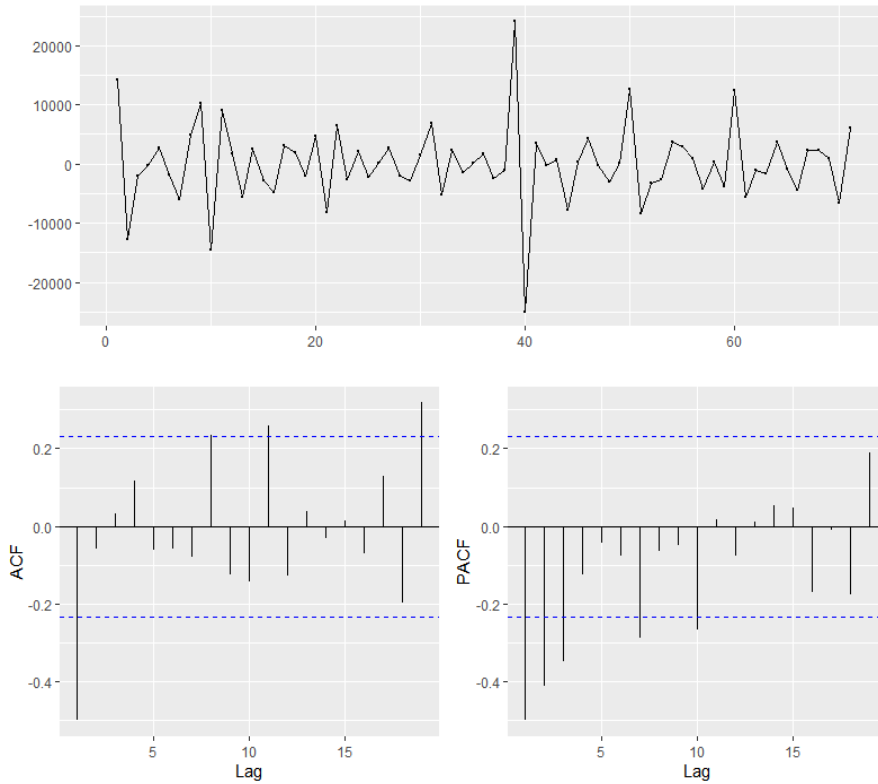
The figure 4.16 present the Close price of Bitcoin data for the years 2019 and 2020 as well as the ACF and PACF correlograms. The original data series is non-stationary and non-seasonal moreover; the corresponding sample ACF indicates as well non-stationarity because it dies down extremely slowly.



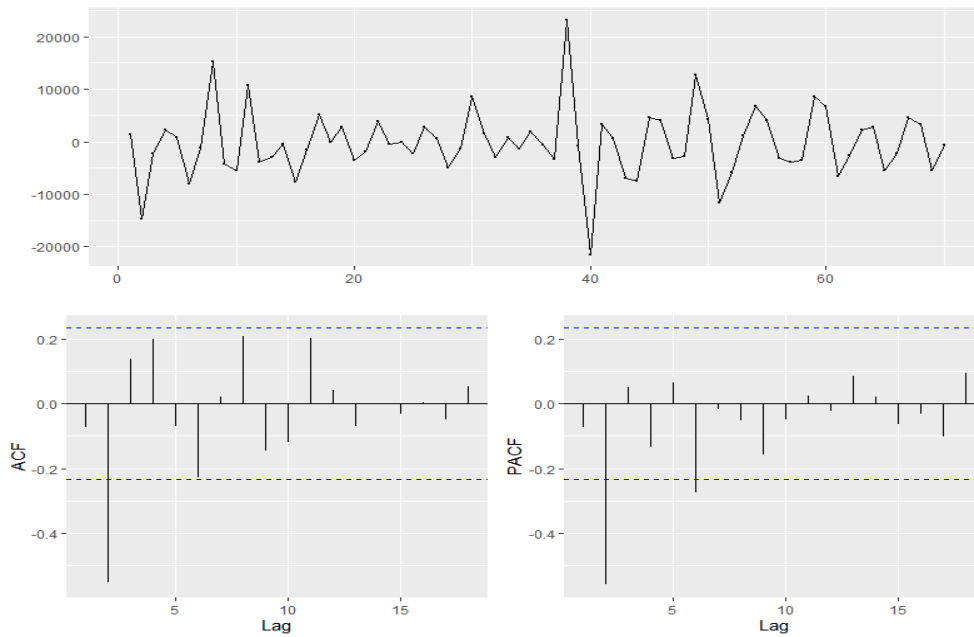
**Figure 4.16: Time series plot (top), autocorrelation (left), and partial autocorrelation function (right) plots of variables**

Thus, differencing transformation on the original data series should be performed and checked for stationarity. Figure 4.17 and 4.18 represents the first and second difference of the time series data respectively one can recognize that the data is now stationary. It shows the possible  $p$ ,  $d$  and  $q$  of ARIMA ( $p,d,q$ ) where  $d=1$  for the first difference and  $d=2$  for the second one.

The first difference of the training Close Bitcoin data indicates a fluctuation ARIMA model with  $p = 2$  and a  $q=2$  where the  $d =1$  representing the first difference. The second difference of the training Close Bitcoin data indicates ARIMA model have a  $p = 4$  and a  $q=2$  where the  $d =2$  representing the second difference.



**Figure 4. 17: Time series plot (top), autocorrelation (left), and partial autocorrelation function (right) plots of the first differenced variables**



**Figure 4. 18: Time series plot (top), autocorrelation (left), and partial autocorrelation function (right) plots of the second differenced variables**

#### 4.3.3.1. ARIMA (2, 1, 2)

The coefficients of the ARIMA (2, 1, 2) lead to a moving average as MA (2) model and an auto regression AR (2) of the first difference. The estimated parameters of AR (2) are  $\beta_1 = -0.23$  and  $\beta_2 = -0.088$  and of MA (2) with  $\phi_1 = -0.84$  and  $\phi_2 = -0.16$  thus the equation:

$$Y_t = -0.23 * Y_{t-1} - 0.088 * Y_{t-2} - 0.84 * \epsilon_{t-1} - 0.16 * \epsilon_{t-2} + \epsilon_t \quad (45)$$

The table 4.19 presents the test statistics and the p-value of Box-Ljung test which are equal to 6.95 and 0.33 > 0.05 respectively. In this case, the null hypothesis of the study fails to be rejected and we can conclude that the data values are independent with each other.

**Table 4. 19: Box-Ljung test of ARIMA (2,1,2)**

Box-Ljung test of ARIMA(2,1,2)	
X-squared	6.95
P-value	0.33

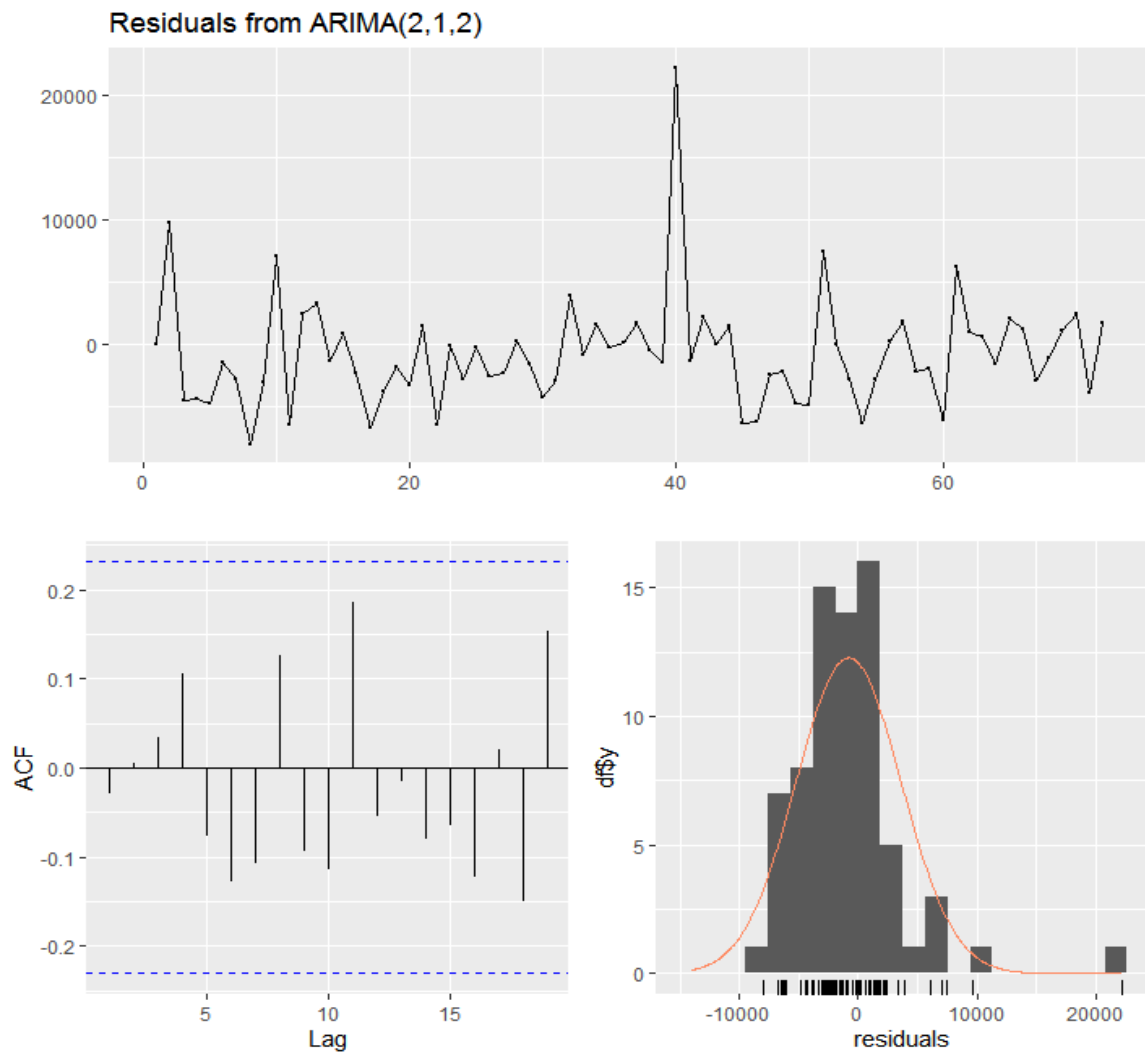
Table 4.20 shows the error measures as ME, RMSE, MAE, MPE, MAPE, and MASE. A high error is shown as MAPE = 43.47 %

**Table 4. 20: Training set error of ARIMA (2, 1, 2)**

	ME	RMSE	MAE	MPE	MAPE	MASE
<b>Training Set</b>	-804.64	4,458.00	3,133.17	-29.83	43.47	0.69

The figure 4.19 shows the residual plots of the ARIMA (2, 1, 2) where in the first figure a fluctuating trend is represented, Its corresponding ACF which indicates a randomness of the data supported by the Box-Ljung test in table 4.19 and it's corresponding histogram.





**Figure 4. 19: Time series plot Residual of the ARIMA (2, 1, 2) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables of the first difference**

The blue-grey in figure 4.20 represents short term forecasted value of the Close price of bitcoin for first 10 weeks of the year 2021. The figure 4.21 shows its ACF, PACF, histogram, and Normal Q-Q plot. The ACF plot starts at lag = 0 by default, and the autocorrelation is always 1 at lag = 0. While PACF doesn't have any first values except at lag =20. The histogram shows a positively skewed shape. The normal Q-Q plot indicates the normality of this series which in this case fall approximately along the reference line.

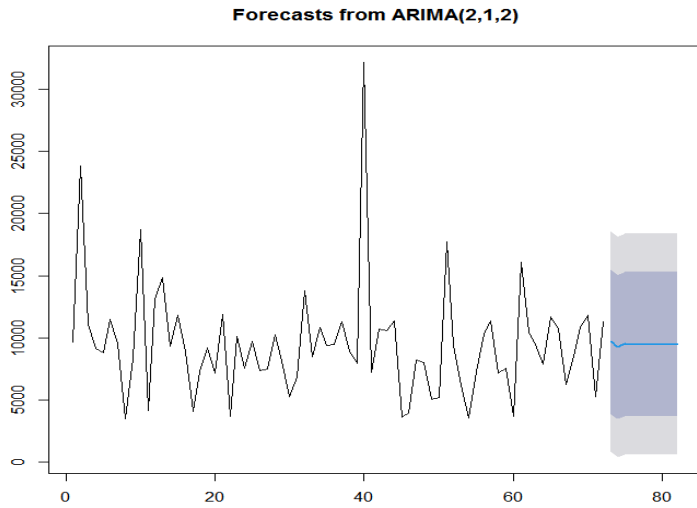


Figure 4. 20: Forecast ARIMA (2,1,2)

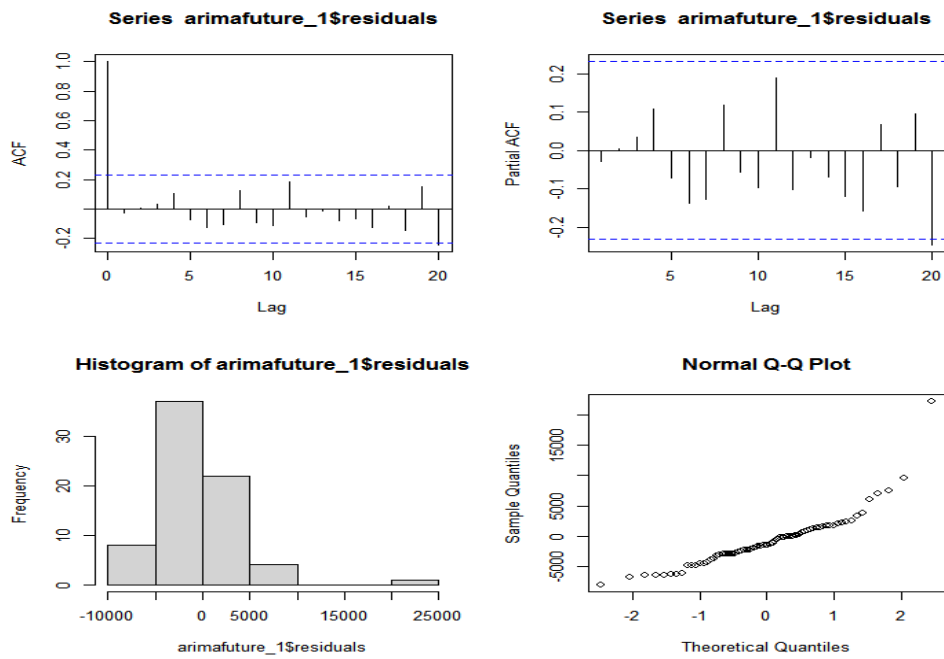


Figure 4. 21: Arima future plots of the ARIMA (2,1,2)

#### 4.3.3.2. ARIMA (4, 2, 2)

The model gives an AR (4) with a  $\beta_1 = -0.04$  ,  $\beta_2 = -0.01$  ,  $\beta_3 = 0.12$  and  $\beta_4 = 0.17$  and an MA (2) with  $\phi_1 = -2.00$  and  $\phi_2 = 1.00$  thus the equation:

(46)

$$Y_t = -0.04 * Y_{t-1} - 0.01 * Y_{t-2} + 0.12 * Y_{t-3} + 0.17 * Y_{t-4} - 2 * \epsilon_{t-1} + 1 * \epsilon_{t-2} + \epsilon_t$$

The coefficients of the ARIMA (4, 2, 2) lead to a moving average as MA (2) model and an auto regression AR (4) of the second difference.

The table 4.21 presents the test statistics and the p-value of Box-Ljung test which are equal to 6.25 and  $0.18 > 0.05$  respectively. In this case, the null hypothesis of the study fails to be rejected and we can conclude that the data values are independent with each other.

**Table 4. 21: Box-Ljung test of ARIMA(4,2,2)**

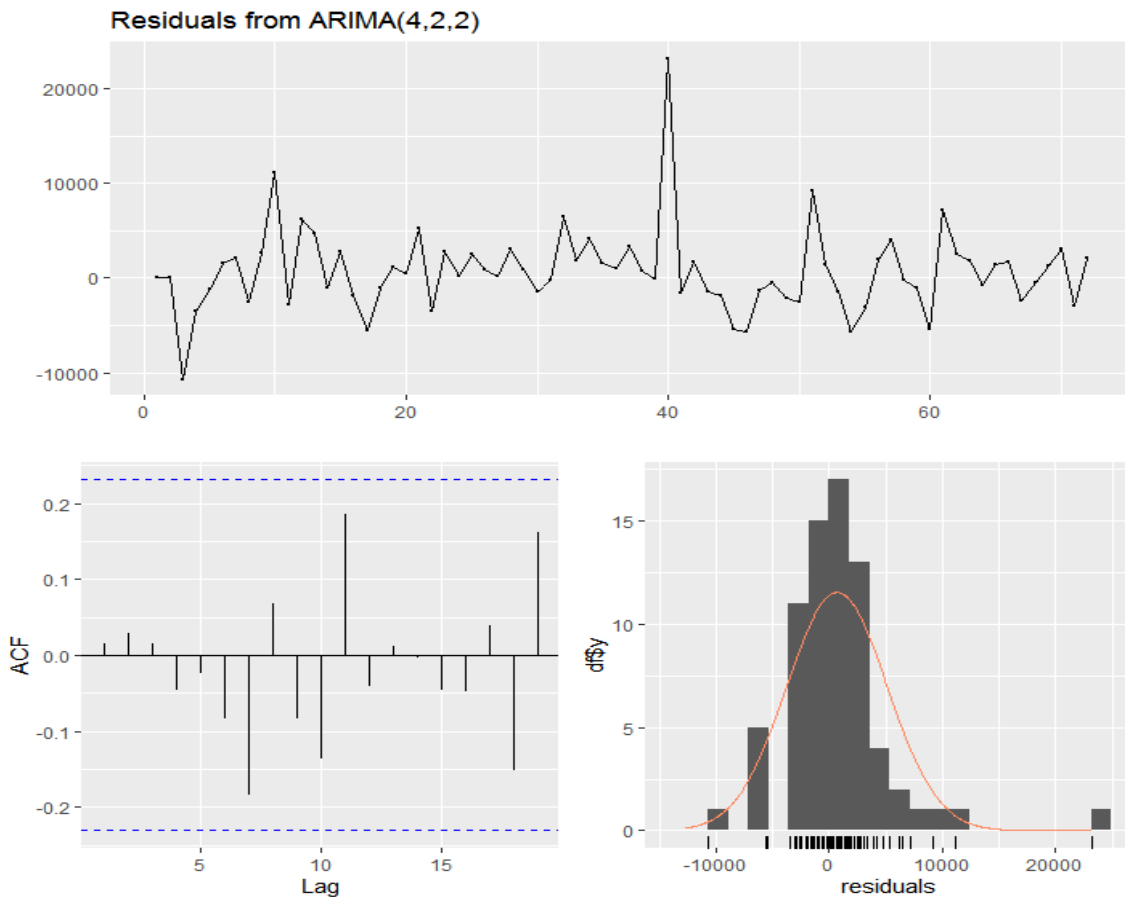
Box-Ljung test of ARIMA(4,2,2)	
X-squared	6.25
P-value	0.18

The table 4.22 shows the error measures as ME, RMSE, MAE, MPE, MAPE, and MASE with an MAPE = 34.77 % lower than the ARIMA (2,1,2) .

**Table 4. 22: Training set error of ARIMA (4, 2, 2)**

	ME	RMSE	MAE	MPE	MAPE	MASE
<b>Training Set</b>	684.70	4,457.04	2,929.87	-8.22	34.77	0.65

The figure 4.22 shows the residual plots of the ARIMA (4, 2, 2) where in the first figure a fluctuating trend is represented. Its corresponding ACF which indicates a randomness of the data supported by the Box-Ljung test in table 4.21 and it's corresponding histogram.



**Figure 4. 22: Time series plot Residual of the ARIMA (2, 1, 2) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables of the second difference**

The blue-grey in figure 4.23 represents short term forecasted value of the close price of bitcoin for first 10 weeks of the year 2021. The figure 4.24 shows its ACF, PACF, histogram, and Normal Q-Q plot. The ACF plot starts at lag = 0 by default, and the autocorrelation is always 1 at lag = 0. While PACF doesn't have any first values except at lag =20. The histogram shows a negatively skewed shape. The normal Q-Q plot indicates the normality of this series which in this case fall approximately along the reference line.

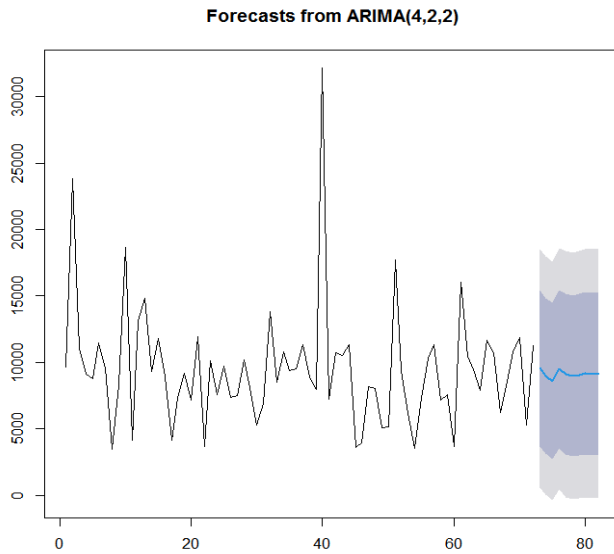


Figure 4. 23: Forecasting the ARIMA (4, 2, 2).

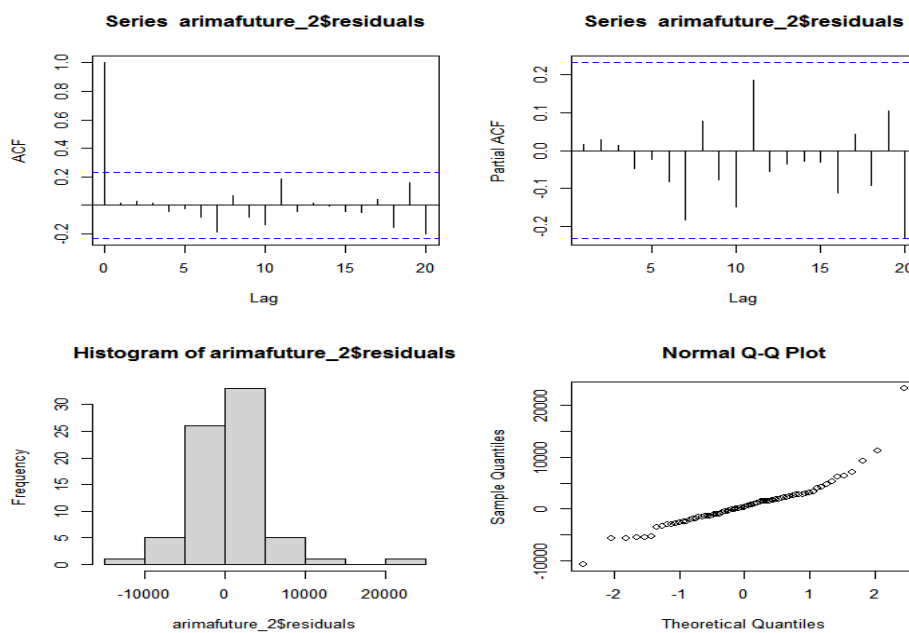


Figure 4. 24: Plots of the ARIMA (4,2,2) Future after forecasting the variables.

By applying the time series we can notice that ARIMA is not the best accuracy to forecast Bitcoin. The ARIMA (2,1,2) and (4,2,2) has an MAPE=43.47% and 34.77% respectively.

#### 4.3.4 Regression Time Series

ARIMA (2,1,2) and the ARIMA (4,2,2) are applied by combining the regression i.e. adding explanatory variable in the purpose to enhance the accuracy of prediction. The selection of explanatory variables is the results of the decision tree model and the multiple regression models i.e. only the important and significant variables are taken into consideration. The explanatory variables that are added the ARIMA model are Close CII, Close AI1, Close AI2, Blockchain addresses, Blockchain transaction, Mining commission and Unspent transaction.

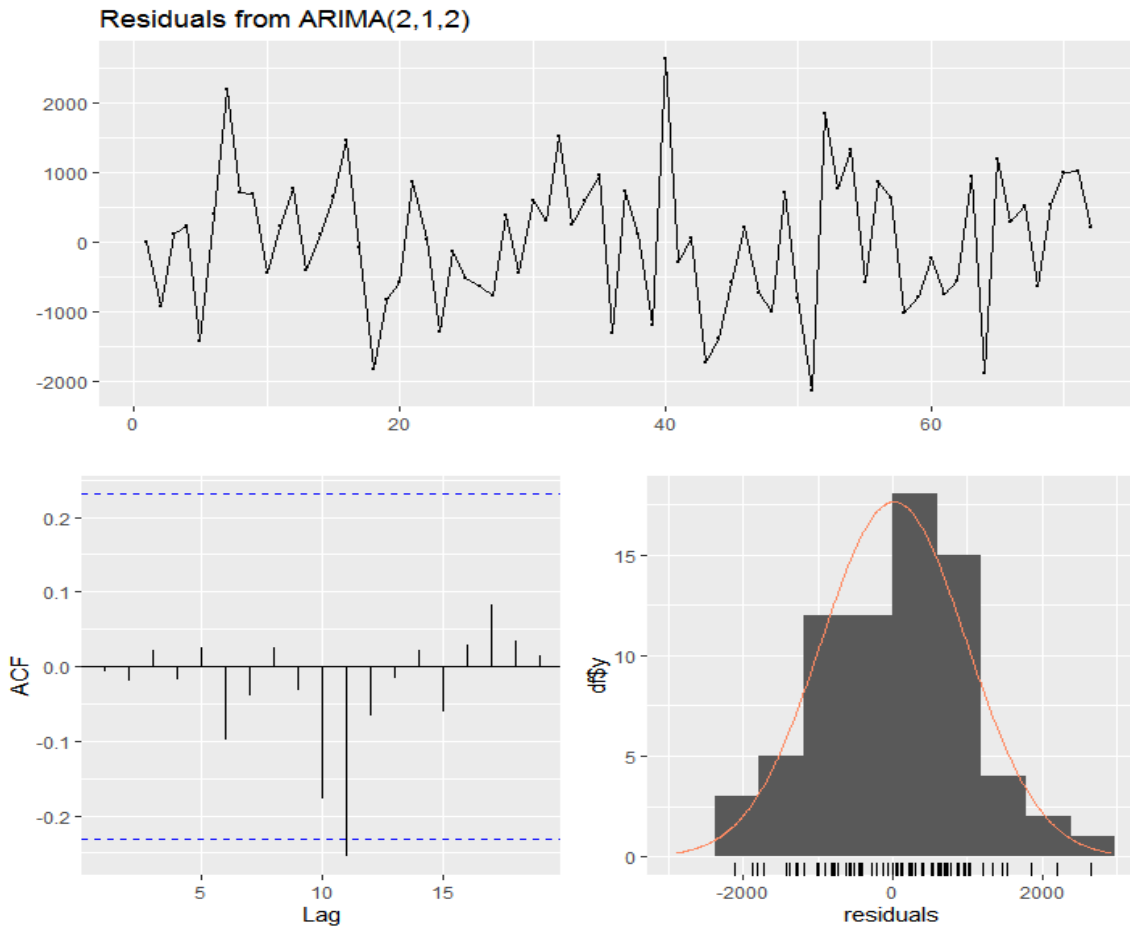
##### 4.3.4.1. Regression ARIMA (2, 1, 2)

The table 4.23 presents the test statistic and the p-value of Box-Ljung test which are equal to 10.05 and  $0.02 < 0.05$ . In this case, the null hypothesis of the study is rejected and we can tell that the data values are dependent with each other.

**Table 4. 23: Box-Ljung test of Regression ARIMA (2,1,2)**

Box-Ljung test of ARIMA(2,1,2)	
X-squared	10.05
P-value	0.02

The figure 4.25 shows the residual plots of the Regression ARIMA (2,1,2) where in the first figure a fluctuating trend is represented. Its corresponding ACF indicates that the errors are random and lies in the 95% confidence interval but not supported by the Box-Ljung test in table 4.23 and its corresponding histogram.



**Figure 4. 25: Time series plot Residual of the Regression ARIMA (2, 1, 2) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables for the first difference**

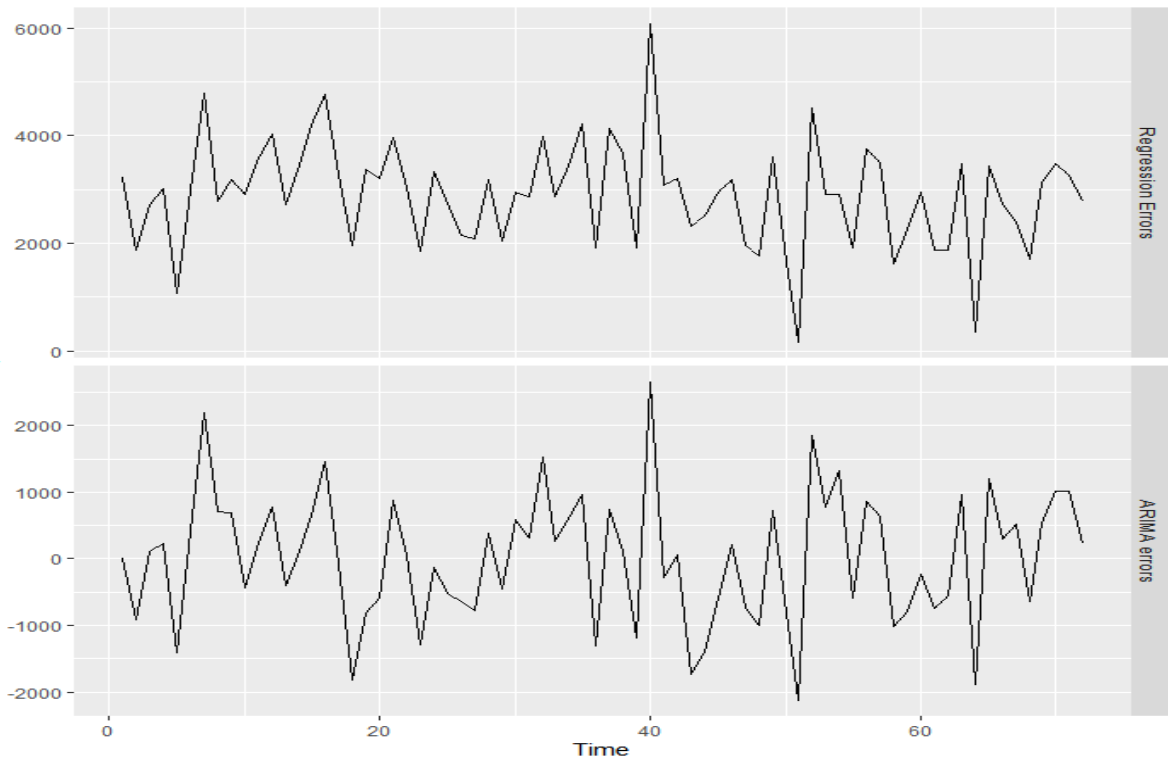
The coefficients of the Regression ARIMA (2,1,2) lead to a moving average as MA (2) model and an auto regression AR (2) of the first difference.

The table 4.24 shows the error measures as ME, RMSE, MAE, MPE, MAPE, and MASE. A low error is shown as MAPE = 9.12%

**Table 4. 24: Training set error of Regression ARIMA (2, 1, 2)**

	ME	RMSE	MAE	MPE	MAPE	MASE
<b>Training Set</b>	14.51	965.70	787.39	0.25	9.12	0.17

The figure 4.26 shows the comparison graphs between the Regression Errors and the ARIMA Error and one can concludes that both errors are random.



**Figure 4. 26: Regression Error versus the ARIMA Error (2,1,2)**

#### 4.3.4.2. Regression ARIMA (4, 2, 2)

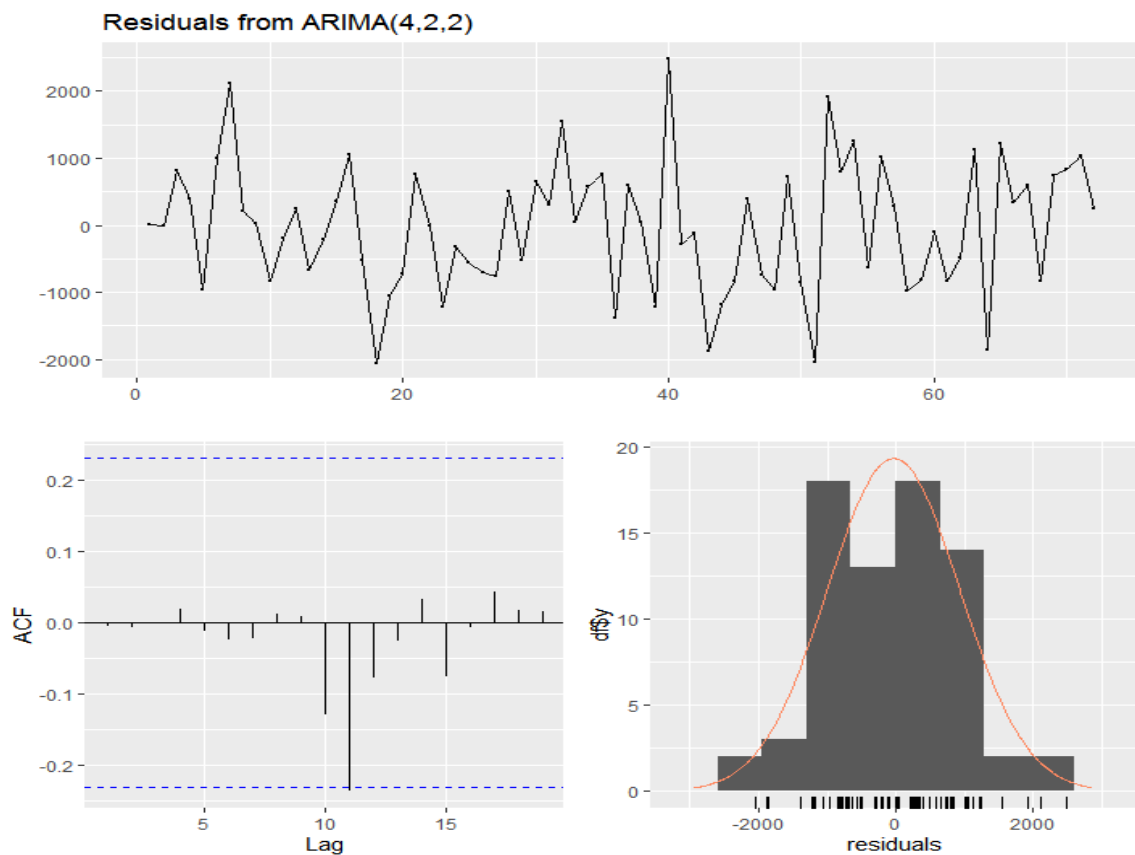
The table 4.25 presents the test statistics and the p-value of Box-Ljung test which are equal to 7.66 and  $0.053 > 0.05$  respectively. In this case, the null hypothesis of the study fails to be rejected and we can conclude that the data values are independent with each other.

**Table 4. 25: Box-Ljung test of Regression ARIMA (4,2,2)**

Box-Ljung test of ARIMA(4,2,2)	
X-squared	7.66
P-value	0.053

The figure 4.27 shows the residual plots of the Regression ARIMA (4,2,2) where in the first figure a fluctuating trend is represented. It's corresponding ACF which indicates a randomness of the error of data supported by the Box-Ljung test in table 4.27 and it's corresponding histogram.





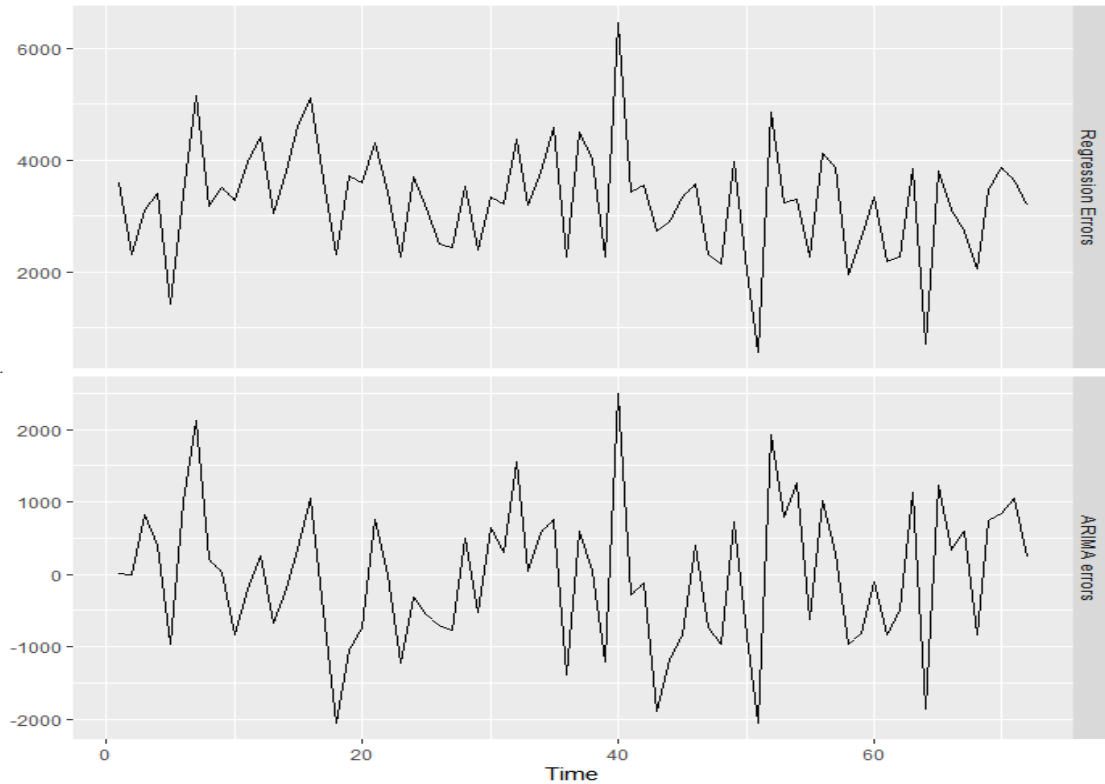
**Figure 4. 27: Time series plot Residual of the Regression ARIMA (4, 2, 2) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables for the second difference**

The table 4.26 shows the error measures as ME, RMSE, MAE, MPE, MAPE, and MASE. A low error is shown as MAPE = 9.17 %

**Table 4. 26: Training set error of Regression ARIMA (4, 2, 2)**

	ME	RMSE	MAE	MPE	MAPE	MASE
<b>Training Set</b>	-32.69	960.40	783.58	-0.61	9.17	0.17

The figure 4.28 shows the comparison graphs between the Regression Errors and the ARIMA Error and one can conclude that both errors are random.



**Figure 4. 28: Regression Error versus the ARIMA Error (4,2,2)**

By applying time series regression we can notice that Reg ARIMA can be the best used to forecast Bitcoin. The ARIMA (2,1,2) has an MAPE=9.12% and ARIMA (4,2,2) has an MAPE = 9.17% slightly higher than the previous one however the second one have random residual.

### 4.3.5 Feedforward Artificial Neural Networks

The close price of bitcoin is estimated by Feedforward artificial neural network on three steps:

- First, the input layers are the weekly Close and demand/supply variables with 1 hidden layer.
- Then, the input layers will include only the most important variable data concluded from the decision tree (Blockchain addresses, Blockchain transaction, Number of bitcoin, Mining commissions, Close AI1, Unspent transaction, Close CNYuan, Close CI1, Difficulty, and Close AI2) and assuming 7 hidden layers divided into 2 parts [4 and 3]
- Finally, the input layers include only the significant variables concluded from the multiple regression (Blockchain addresses, Blockchain transaction, Number of bitcoin and Close CNYuan) with 7 hidden layers divided into 3 parts [3,3 and 2 ]

SSE and MAPE of all three models are calculated and their graph will be presented to choose the best NN model.

#### 4.3.5.1. NN\_1 Plot

The figure 4.29 visualizes the computed neural network. Our model has 1 neuron in its hidden layer. The black lines demonstration the connections between each layer and the weights on each connection, while the blue lines show the bias term added in each step thought as the intercept of a linear model. The net is essentially a black box so we cannot say that much about the fitting, the weights and the model. The input layer has 27 inputs, the hidden layers have 1 neuron and the output layer has, with a single output since.

The MAPE of NN 1 model applied on the test data is equal 44.71 %. The SSE of the train data and the test data are 0.041 and 0.015 respectively.

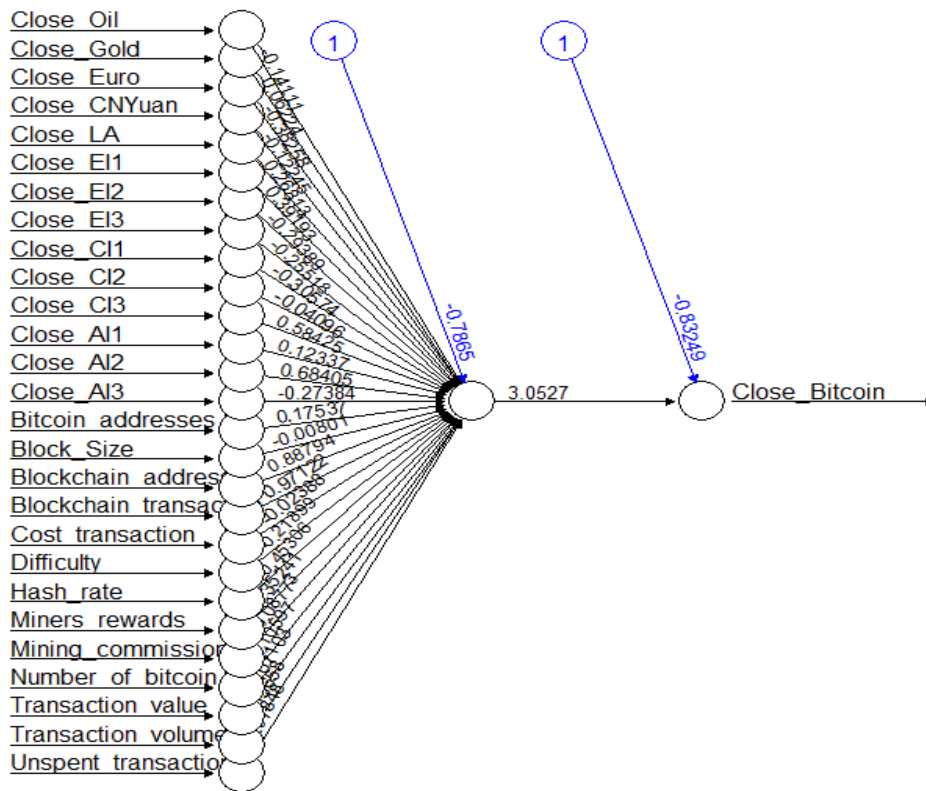


Figure 4. 29: Neural Network of the weekly Close and demand/supply variables

#### 4.3.5.2. NN\_2 Plot

The figure 4.30 visualizes the computed neural network. Our model has 7 neurons in its hidden layer. The black lines demonstration the connections between each layer and the weights on each connection, while the blue lines show the bias term added in each step thought as the intercept of a linear model. The net is essentially a black box so we cannot say that much about the fitting, the weights and the model. The input layer has 10 inputs (Blockchain addresses, Blockchain transaction, Number of bitcoin, Mining commissions, Close AI1, Unspent transaction, Close CNYuan, Close CI1, Difficulty, and Close AI2), the hidden layers have 7 neurons and the output layer has a single output the Close Bitcoin since we are doing regression.

The MAPE of NN 2 model applied on the test data is equal 81.3 %. The SSE of the train data and the test data are 0.045 and 0.014 respectively.

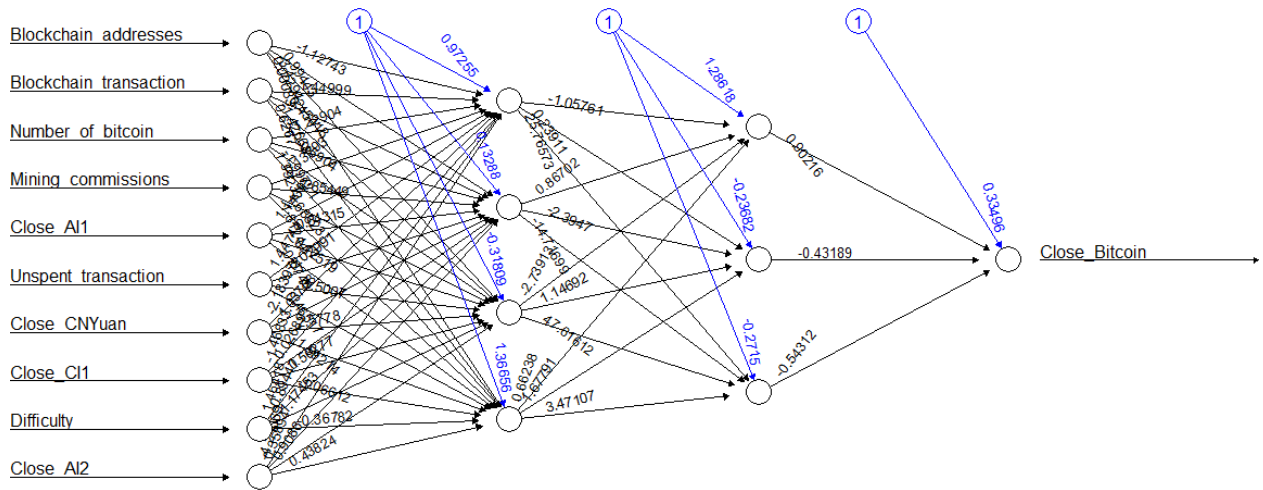


Figure 4. 30: Neural Network of the most important variables of decision tree

### 4.3.5.3. NN\_3 Plot

The figure 4.31 visualizes the computed neural network. Our model has 7 neurons in its hidden layer. The black lines demonstration the connections between each layer and the weights on each connection, while the blue lines show the bias term added in each step thought as the intercept of a linear model. The net is essentially a black box so we cannot say that much about the fitting, the weights and the model. The input layer has the most significant 4 inputs (Blockchain addresses, Blockchain transaction, Number of bitcoin, and Close CNYuan) in the multiple regression, the hidden layers have 7 neurons and the output layer has a single output the Close Bitcoin since we are doing regression.

The MAPE of NN 3 model applied on the test data is equal 24.7 %. The SSE of the train data and the test data are 0.06 and 0.012 respectively.

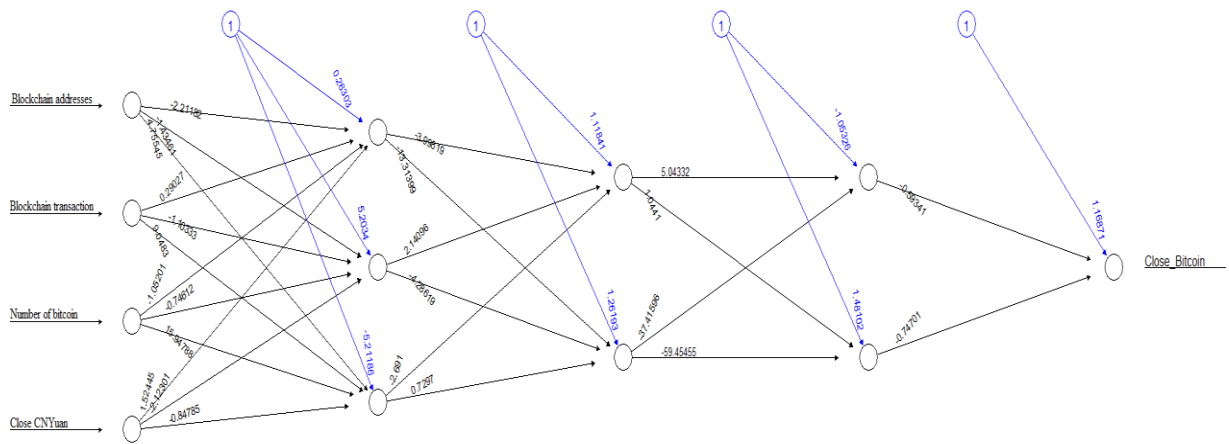


Figure 4. 31: Neural Network of the significant variables of the multiple regression

The figure 4.32 sums up the SSE and the MAPE values for the different NN hidden layers where the most suitable one is the NN 3 which has the lowest MAPE = 24.7 % with the highest train = 0.06.

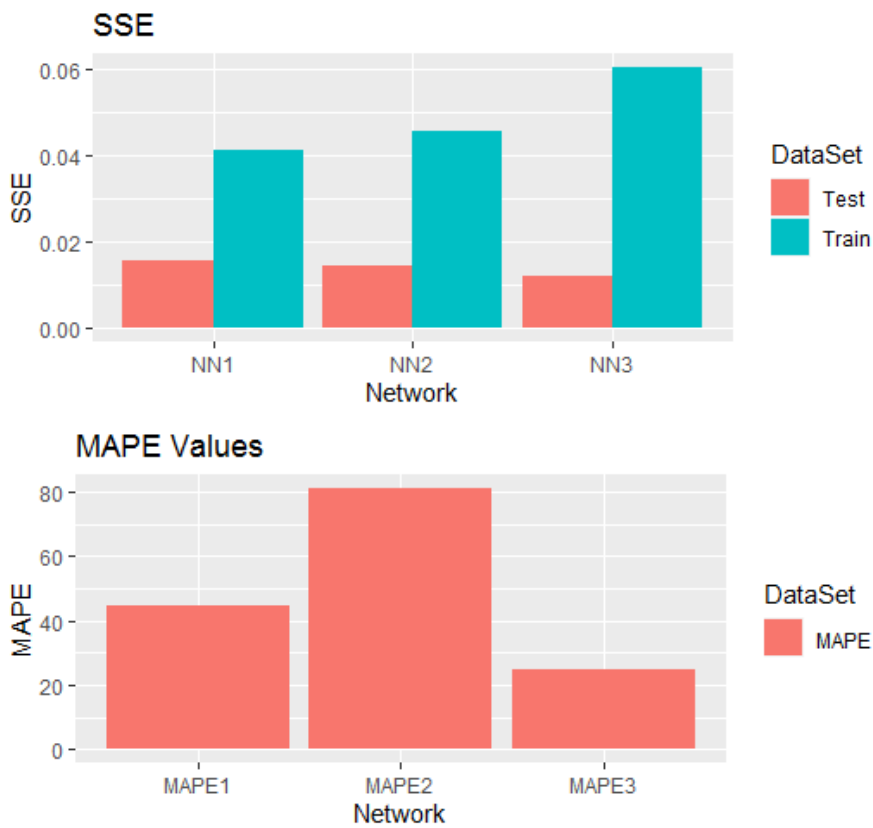


Figure 4. 32: Histograms of the SSE and MAPE value for different NN

### 4.3.6 Regression Time Series Neural Network

The Regression Time Series Neural Network (NNAR) model is a forecasting model that combines the multiple regression, time series ARIMA model and Neural Network. Forecasting is done in three steps:

1. The order of auto regression determined by time series which indicates the number of previous values.
2. The NN is trained by taking the order of auto regression.
3. Adding independent variables to predict the dependent variables.

The node input is defined by the auto regression and the NN inputs are deduced from the time series forecasting. The predicted values are the output of NN. To avoid over fitting, the hidden nodes should be selected properly by trial and error or experimentation as no theoretical basis is found (Nagwani, 2016)

NNAR (p,P,k) is the general denotation of the model where p = lagged inputs number, P= seasonal lags and k = number of nodes in hidden layers where the model is:

$$Y_t = f(Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}, Y_{t-m}, Y_{t-2m}, \dots, Y_{t-Pm}) + \epsilon_t \quad (47)$$

Which f represents the neural network with k hidden number, m= the length of the seasonal and  $\epsilon_t$  the residual series.

Three models are represented NNAR (1, 1, 2), NNAR (1, 1, 4) and NNAR (1, 1, 6).

#### 4.3.6.1. NNAR (1,1,2)

NNAR (1,1,2) is forecasted with 1 input lag, 1 seasonal lag and 2 hidden layers with all the variables.

The table 4.27 presents the test statistic and the p-value of Box-Ljung test which are equal to 2.2333 and 0.13  $>0.05$ . In this case, the null hypothesis of the study fails to be rejected and we can tell that the data values are independent with each other.

**Table 4. 27: Box-Ljung test of Regression NNAR(1,1,2)**

Box-Ljung test of NNAR (1,1,2)	
X-squared	2.233
P-value	0.13

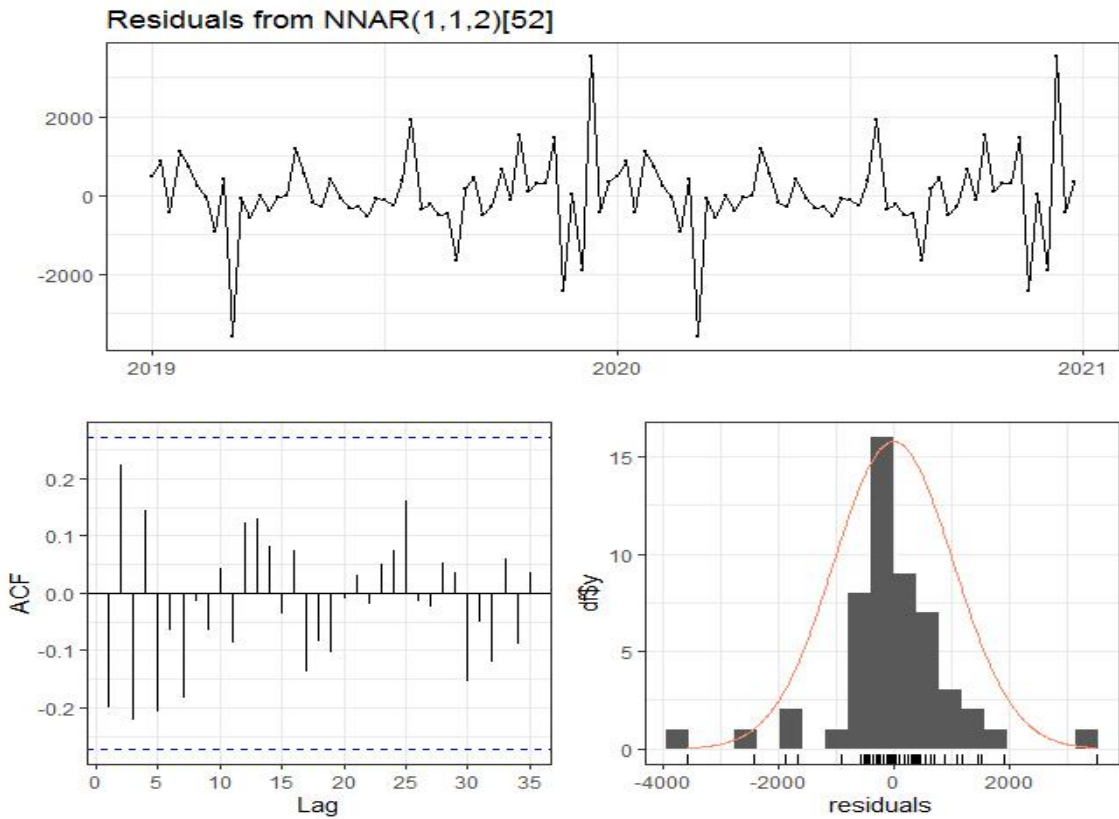
The table 4.28 shows the error measures as ME, RMSE, MAE, MPE, MAPE, and MASE. A low error is shown as MAPE = 6.27%

**Table 4. 28: Training set error of NNAR (1, 1, 2)**

	ME	RMSE	MAE	MPE	MAPE	MASE
Training set	-1.47	1,026.86	40	-1.03	6.27	0.14

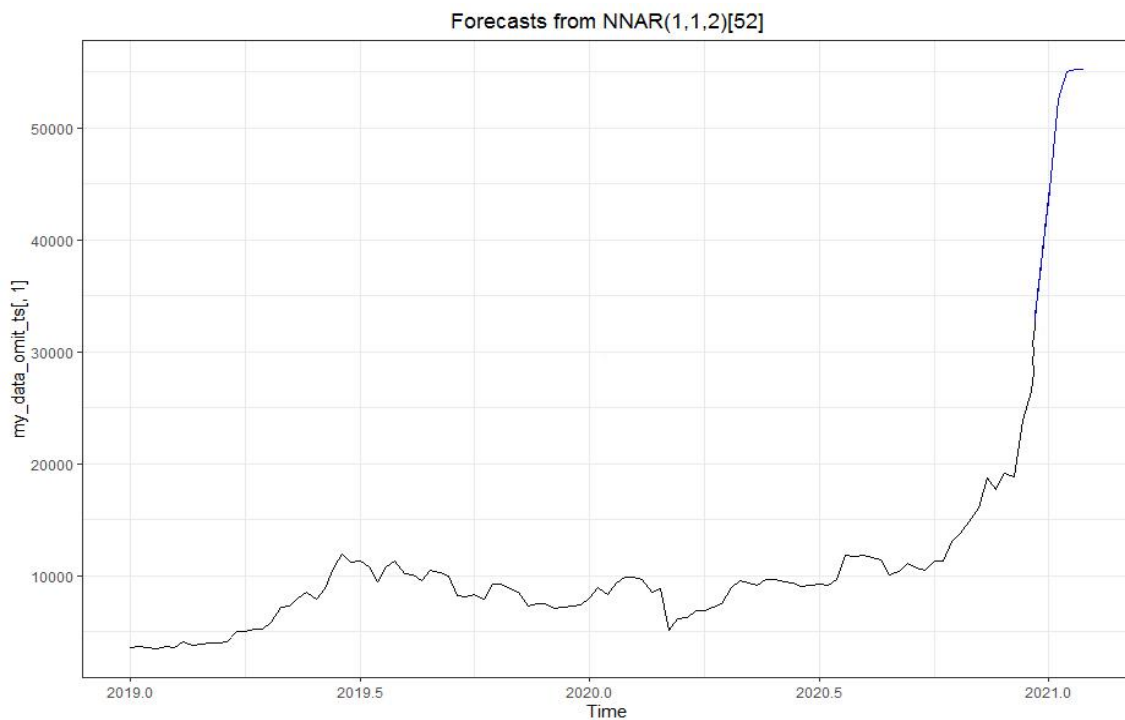
The figure 4.33 shows the residual plots of the Regression NNAR (1,1,2) where in the first figure a fluctuating trend is represented. Its corresponding ACF indicates that the errors are random and lies in the 95% confidence interval but not supported by the Box-Ljung test in table 4.27 and its corresponding histogram.





**Figure 4. 33: Residual of the NNAR (1,1,2) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables**

The figure 4.34 shows the forecast plot of time series in NNAR(1,1,2) plotted as a blue line.



**Figure 4. 34: NNAR (1,1,2)**

#### 4.3.6.2. NNAR (1,1,4)

NNAR(1,1,4) is forecasted with 1 input lag, 1 seasonal lag and 4 hidden layers with the Close CNYuan, Blockchain addresses, Blockchain transaction and number of bitcoin.

The table 4.29 presents the test statistic and the p-value of Box-Ljung test which are equal to 0.36 and  $0.55 > 0.05$ . In this case, the null hypothesis of the study fails to be rejected and we can tell that the data values are independent with each other.

**Table 4. 29: Box-Ljung test of Regression NNAR(1,1,4)**

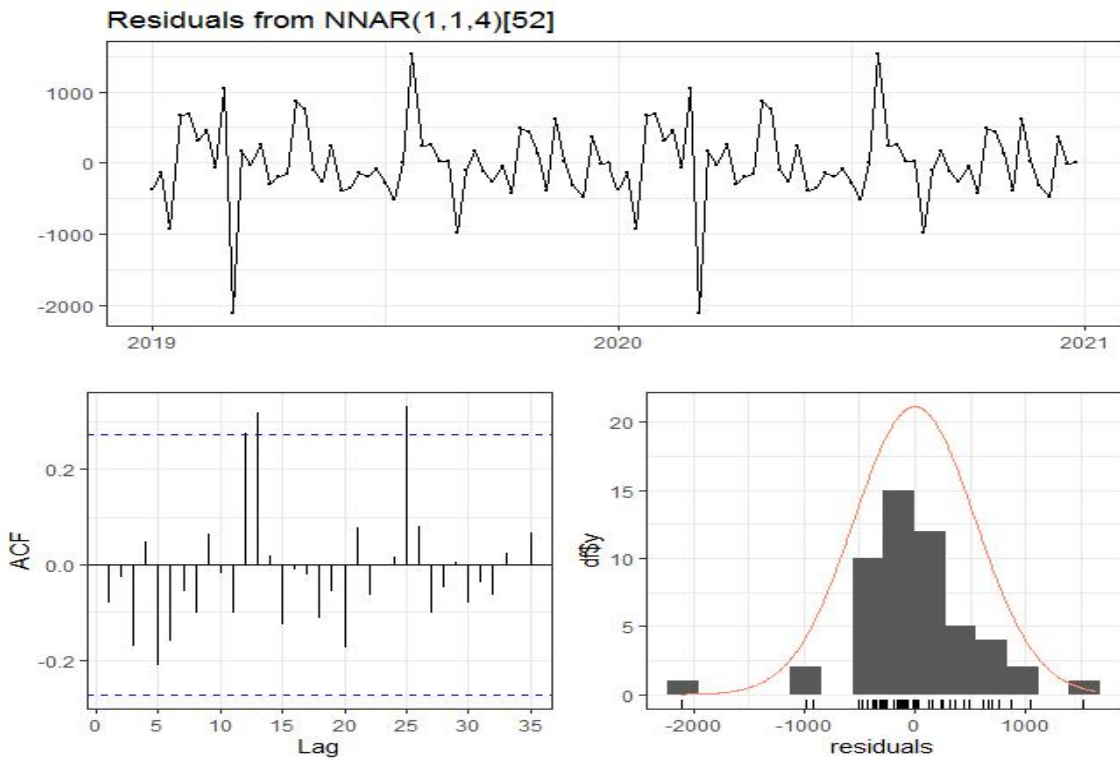
Box-Ljung test of NNAR (1,1,4)	
X-squared	0.36
P-value	0.55

The table 4.30 shows the error measures as ME, RMSE, MAE, MPE, MAPE, and MASE. A low error is shown as MAPE = 4.05%

**Table 4. 30: Training set error of NNAR (1, 1, 4)**

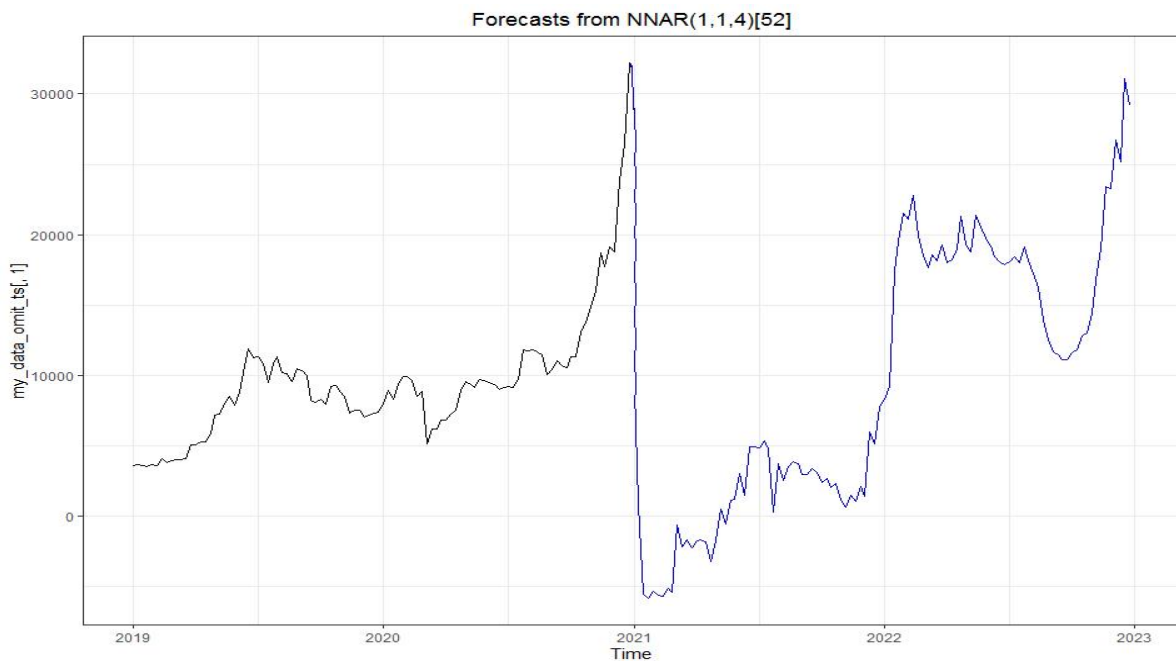
	ME	RMSE	MAE	MPE	MAPE	MASE
<b>Training set</b>	-0.42	543.67	375.59	-0.51	4.05	0.08

The figure 4.35 shows the residual plots of the Regression NNAR (1,1,4) where in the first figure a fluctuating trend is represented. Its corresponding ACF indicates that the errors are random and lies in the 95% confidence interval but not supported by the Box-Ljung test in table 4.29 and its corresponding histogram.



**Figure 4. 35: Residual of the NNAR (1,1,4) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables**

The figure 4.36 shows the forecast plot of time series in NNAR (1,1,4) as the fluctuating where the last part is increasing designed as a black line of the model and the fluctuating blue line shows the forecasted part of the model.



**Figure 4. 36: NNAR (1,1,4)**

**4.3.6.3. NNAR (1,1,6)**

NNAR (1,1,6) is forecasted on a 20 networks average where each of which is a 12-6-1 network with 85 weights options were linear output units with the Close Bitcoin, Blockchain addresses, Blockchain transaction, Number of bitcoin, Mining commissions, Close AI1, Unspent transaction, Close CNYuan, Close CI1, Difficulty, and Close AI2.

The table 4.31 presents the test statistic and the p-value of Box-Ljung test which are equal to 1.37 and 0.24 >0.05. In this case, the null hypothesis of the study fails to be rejected and we can tell that the data values are independent with each other.

**Table 4. 31: Box-Ljung test of Regression ARIMA (2,1,2)**

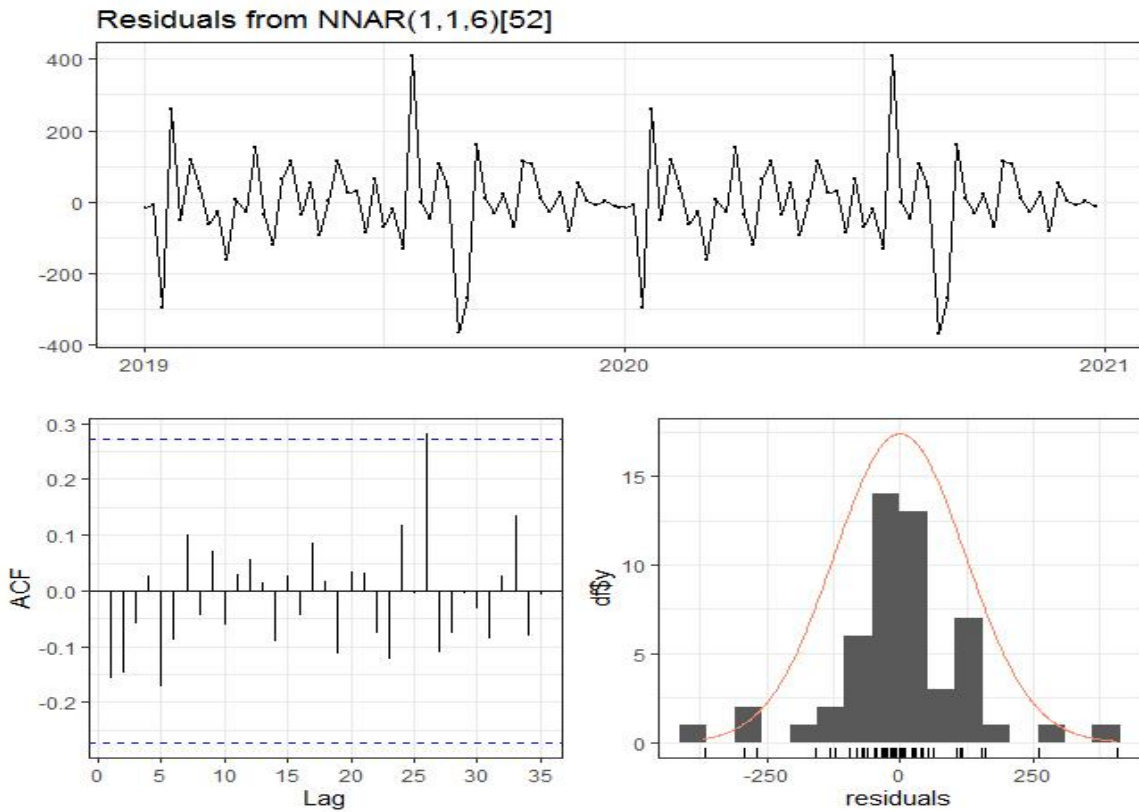
Box-Ljung test of NNAR (1,1,2)	
X-squared	1.37
P-value	0.24

The table 4.32 shows the error measures as ME, RMSE, MAE, MPE, MAPE, and MASE. A low error is shown as MAPE = 0.85%

**Table 4. 32: Training set error of NNAR (1, 1, 6)**

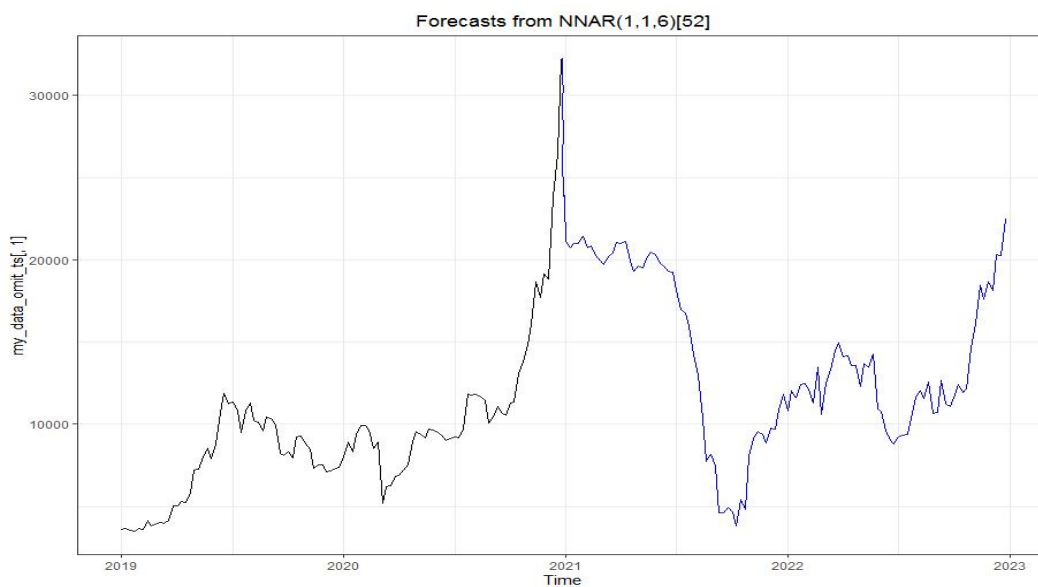
	ME	RMSE	MAE	MPE	MAPE	MASE
<b>Training set</b>	-0.69	122.07	81.59	-0.07	0.85	0.02

The figure 4.37 shows the residual plots of the Regression NNAR (1,1,6) where in the first figure a fluctuating trend is represented. Its corresponding ACF indicates that the errors are random and lies in the 95% confidence interval but not supported by the Box-Ljung test in table 4.31 and its corresponding histogram.



**Figure 4. 37: Residual of the NNAR (1,1,6) (top), autocorrelation (left), and residual histogram (right) plots of the weekly Close variables and the demand supply variables**

The figure 4.38 shows the forecast plot of time series in NNAR (1,1,6) data fluctuating where the last part is increasing designed as a black line of the model and the fluctuating blue line shows the forecasted part of the model.



**Figure 4. 38: NNAR (1,1,6)**

The table 4.33 concludes all the machine learning types MAPE's of this study:

**Table 4. 33: MAPE of the models**

<b>Machine Learning types</b>	<b>MAPE</b>
<b>Regression Decision Tree</b>	10.26%
	9.10%
	6.92%
<b>Multiple Regression</b>	8.80%
	9.15%
	9.67%
	11.12%
	10.76%
<b>ARIMA(2,1,2)</b>	43.47%
<b>ARIMA(4,2,2)</b>	34.77%
<b>Reg ARIMA(2,1,2)</b>	9.12%
<b>Reg ARIMA(4,2,2)</b>	9.17%
<b>Artificial Neural Networks</b>	44.71%
	81.30%
	24.70%
<b>NNAR(1,1,2)</b>	6.27%
<b>NNAR(1,1,4)</b>	4.05%
<b>NNAR(1,1,6)</b>	0.85%

## Chapter 5

### Conclusion and Future Work

A cryptocurrency is a medium of exchange, such as the US dollar where it uses encryption techniques to control the creation of monetary units and the transfers of funds. Bitcoin is the most traded cryptocurrency which is founded during the financial crisis in 2009, the one for which the blockchain technology was invented. It is one of the interesting decentralized forms that appeared to address economic problems related to centralized currency. Bitcoin is bank-free internet money that gives you a complete control over your finances.

Gaining popularity leads eventually that Bitcoin might replace official currencies in the future stages of life. Meanwhile, many countries started to accept Bitcoin as a way of payment with few restrictions where others consider it illegal and criminalized.

Since estimating and predicting its value is a necessary step, thus in this study, the weekly bitcoin price was forecasted by including many variables types as commodities like crude oil or gold, indexes as European, Asian or Chinese, other cash currencies like euro or yuan and demand/ supply variables live blockchain addresses, miners rewards, transaction value and of course other kinds of variables using several machine learning and deep learning:

- First, a Regression Decision Tree is considered to sort the variables included in this study by their importance criteria with an MAPE of 6.92%
- Then, multiple regression was used to fit the data having the independent variable as the important variable concluded from the decision tree. After removing all the outlier and the non-significant variables the MAPE was equal 10.76%.
- The lowest MAPE among ARIMA model was reached by having the autoregressive of order 4 and a moving average of order 2 applied on the second difference of the close price of bitcoin and it was equal to 7.85%.
- The MAPE of the artificial neural network by having 3 or more column of hidden was 24.7%.
- Finally, the MAPE of the combination of multiple Regression, Time Series and Neural Network NNAR (1,1,6) was 0.85% .

In conclusion, Decision tree, multiple regression, ARIMA, regression time series and Feedforward artificial neural network standalone were not sufficient to fit, estimate and predict close price of bitcoin. However the accuracy of their combination was greater than 99%. This study reflected the short term prediction subject. Therefore, future studies can focus on studying the long term Bitcoin forecasting.



## References

- Anupriya and S. Garg, "Autoregressive Integrated Moving Average Model based Prediction of Bitcoin Close Price," 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2018, pp. 473-478, doi: 10.1109/ICSSIT.2018.8748423. <https://ieeexplore.ieee.org/document/8748423>
- Artificial Neural Networks for Machine Learning - Every aspect you need to know about. (2017, July 15). DataFlair. <https://data-flair.training/blogs/artificial-neural-networks-for-machine-learning/>
- Ayaz, Z., Fiaidhi, J., Sabah, A., & Anwer Ansari, M. (2020, April 9). Bitcoin Price Prediction using ARIMA Model. TechRxiv.. [https://www.researchgate.net/publication/340567768\\_Bitcoin\\_Price\\_Prediction\\_using\\_ARIMA\\_Model](https://www.researchgate.net/publication/340567768_Bitcoin_Price_Prediction_using_ARIMA_Model)
- Bellis, M. (2019, January 27). The History of Money. ThoughtCo; ThoughtCo. <https://www.thoughtco.com/history-of-money-1992150>
- Brownlee, J. (2016, April 5). Linear Discriminant Analysis for Machine Learning. Machine Learning Mastery. <https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/>
- CFI (2020) Autoregressive Integrated Moving Average (ARIMA) - ApplicationsCorporateFinanceInstitute. <https://corporatefinanceinstitute.com/resources/knowledge/other/autoregressive-integrated-moving-average-arima/>
- Chen, J. (2019). Autoregressive Integrated Moving Average (ARIMA). Investopedia. <https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp>
- Chen, J. (2019). Neural Network Definition. Investopedia. <https://www.investopedia.com/terms/n/neuralnetwork.asp>
- Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, 365, 112395. <https://www.sciencedirect.com/science/article/abs/pii/S037704271930398X>
- Cointelegraph. (2017). How to Buy Ethereum. Cointelegraph. <https://cointelegraph.com/bitcoin-for-beginners/what-is-bitcoin>

- Crossland, T. (2020, July 28). The History and Future of Neural Networks. Retrieved from The AI Journal website: <https://aijournal.com/the-history-and-future-of-neural-networks/>
- Felizardo L., Oliveira R., Del-Moral-Hernandez E. and Cozman F., "Comparative study of Bitcoin price prediction using WaveNets, Recurrent Neural Networks and other Machine Learning Methods," 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), Beijing, China, 2019, pp. 1-6, doi: 10.1109/BESC48373.2019.8963009.  
<https://ieeexplore-ieee-org.neptune.ndu.edu.lb:9443/document/8963009>
- Fork, A. (2017, April 5). A brief history of Money. Medium. <https://blog.humaniq.co/a-brief-history-of-money-66e076f70652>
- Georgios Drakos. (2019, May 23). Decision Tree Regressor explained in depth. Retrieved from GDCoder website: <https://gdcoder.com/decision-tree-regressor-explained-in-depth/>
- Gregersen E. (2020). Bitcoin. In *Encyclopædia Britannica*. Retrieved from <https://academic-eb-com.neptune.ndu.edu.lb:9443/levels/collegiate/article/Bitcoin/574115>
- Kumar et al., "Empirical Analysis of Bitcoin network (2016-2020)," 2020 IEEE/CIC International Conference on Communications in China (ICCC Workshops), Chongqing, China, 2020, pp. 96-101, doi: 10.1109/ICCCWorkshops49972.2020.9209945.  
<https://ieeexplore-ieee-org.neptune.ndu.edu.lb:9443/document/9209945>
- Loh, Eng & Ismail, Shuhaida & Khamis, Azme & Mustapha, Aida. (2020). Comparison of Feedforward Neural Network with Different Training Algorithms for Bitcoin Price Forecasting. *ASM Science Journal*. 1-7. 10.32802/asmscj.2020.sm26(1.5).[https://www.researchgate.net/publication/340605517\\_Comparison\\_of\\_Feedforward\\_Neural\\_Network\\_with\\_Different\\_Training\\_Algorithms\\_for\\_Bitcoin\\_Price\\_Forecasting](https://www.researchgate.net/publication/340605517_Comparison_of_Feedforward_Neural_Network_with_Different_Training_Algorithms_for_Bitcoin_Price_Forecasting)
- Mangla, N Bhat A., Avabratha G., Bhat N. (2019). Bitcoin Price Prediction Using Machine Learning. *INTERNATIONAL JOURNAL OF INFORMATION AND COMPUTING SCIENCE*  
[https://www.researchgate.net/publication/333162007\\_Bitcoin\\_Price\\_Prediction\\_Using\\_Machine\\_Learning](https://www.researchgate.net/publication/333162007_Bitcoin_Price_Prediction_Using_Machine_Learning)

- Midrack, R. L. (2019). *How Neural Networks and Artificial Intelligence Impact Your Life*. Lifewire. <https://www.lifewire.com/neural-networks-4155332>
- Mirzayi S. and Mehrzad M., "Bitcoin, an SWOT analysis," 2017 7th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, 2017, pp. 205-210, doi: 10.1109/ICCKE.2017.8167876.  
<https://ieeexplore.ieee.org/abstract/document/8167876>
- Mudassir, M., Bennbaia, S., Unal, D., & Hammoudeh, M. (2020). *Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach*. *Neural Computing and Applications*.  
<https://link.springer.com/article/10.1007/s00521-020-05129-6>
- Murray, C. (2019). *Investing In Bitcoin: Everything You Need To Know Before You Buy*. *Money Under 30*. <https://www.moneyunder30.com/everything-you-need-to-know-about-bitcoins#:~:text=Bitcoin%20is%20a%20currency%20designed,bank%20controls%20the%20currency%20supply>.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied Linear Statistical Models*. Irwin Chicago.
- Papadopoulos, P. & Vassiliadis, S. & Rangoussi, M. & Konieczny, T. & Gralewski, J. (2017). *BITCOIN VALUE ANALYSIS BASED ON CROSS-CORRELATIONS*. *Journal of Internet Banking and Commerce*. 22.  
[https://www.researchgate.net/publication/323993232\\_BITCOIN\\_VALUE\\_ANALYSIS\\_BASED\\_ON\\_CROSS-CORRELATIONS](https://www.researchgate.net/publication/323993232_BITCOIN_VALUE_ANALYSIS_BASED_ON_CROSS-CORRELATIONS)
- Prabhakaran S. (2019, February 18). *ARIMA Model - Complete Guide to Time Series Forecasting in Python | ML+*. *Machine Learning Plus*.  
<https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>
- Rokach, Lior & Maimon, Oded. (2005). *Decision Trees*. 10.1007/0-387-25465-X\_9.  
[https://www.researchgate.net/publication/225237661\\_Decision\\_Trees/](https://www.researchgate.net/publication/225237661_Decision_Trees/)
- Sas Intitute (2019) *Machine Learning: What it is and why it matters*. Sas.Com.  
[https://www.sas.com/en\\_us/insights/analytics/machine-learning.html](https://www.sas.com/en_us/insights/analytics/machine-learning.html)
- Sayad (2020) <https://www.saedsayad.com/lda.htm>

- Schabenberger, O. (2016). *The difference between Statistical Modeling and Machine Learning, as I see it.* . <https://www.linkedin.com/pulse/difference-between-statistical-modeling-machine-i-see-schabenberger>
- Sena, D., & Nagwani, N. K. (2015). A NEURAL NETWORK AUTOREGRESSION MODEL TO FORECAST PER CAPITA DISPOSABLE INCOME. Retrieved from: <https://www.semanticscholar.org/paper/A-NEURAL-NETWORK-AUTOREGRESSION-MODEL-TO-FORECAST-Sena-Nagwani/93b79de5d49e26e933b5a731318390fe907d4957#references>
- Sin E. and Wang L., "Bitcoin price prediction using ensembles of neural networks," 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, 2017, pp. 666-671, doi: 10.1109/FSKD.2017.8393351. <https://ieeexplore-ieee-org.neptune.ndu.edu.lb:9443/document/8393351>
- T. I. Adegboruwa, S. A. Adeshina and M. M. Boukar, "Time Series Analysis and prediction of bitcoin using Long Short Term Memory Neural Network," 2019 15th International Conference on Electronics, Computer and Computation (ICECCO), Abuja, Nigeria, 2019, pp. 1-5, doi: 10.1109/ICECCO48375.2019.9043229. <https://ieeexplore-ieee-org.neptune.ndu.edu.lb:9443/document/9043229>
- Wakefield,K(2019) A guide to machine learning algorithms and their applications. [https://www.sas.com/en\\_ie/insights/articles/analytics/machine-learning-algorithms.html](https://www.sas.com/en_ie/insights/articles/analytics/machine-learning-algorithms.html)
- What is a Neural Network? (2018, August 14). Forcepoint. <https://www.forcepoint.com/cyber-edu/neural-network>
- Wonderopolis(2014),Who Invented Money? | Wonderopolis. Wwww.Wonderopolis.Org. <https://www.wonderopolis.org/wonder/who-invented-money#:~:text=No%20one%20knows%20for%20sure>
- Yellin, T., Aratari, D., & Pagliery, J. (2018). What is bitcoin? - CNNMoney. CNN Money. <https://money.cnn.com/infographic/technology/what-is-bitcoin/index.html>
- Yildirim, S. (2020, July 27). 11 Most Common Machine Learning Algorithms Explained in a Nutshell. Medium. <https://towardsdatascience.com/11-most-common-machine-learning-algorithms-explained-in-a-nutshell-cc6e98df93be>