University of Strathclyde

Department of Accounting and Finance

M.Sc. Financial Technology

Investigating the Price Dynamics of Bitcoin Through Exploratory Analysis and Deploying Machine Learning Models to Predict the Price of Bitcoin

MS996 Management Science Project

Ryan Dougan

202181897

Submitted in partial fulfilment0 of the requirements for the degree of M.Sc. Financial Technology

Word Count: 9,061

Table of Contents

Table of Figures

List of Tables

Introduction

Satoshi Nakamoto created Bitcoin as the 2008 Great Recession began to unfold, during this period banks and their role in the financial system came under increasing scrutiny. Nakamoto wanted to overcome the current system of centralised control of money. In the fiat monetary system, transactions may be undone or interfered with by third parties, and the transaction costs can be high. The idea behind Bitcoin was to enable transactions without the use of a middleman. Instead of relying on external banks and other organisations, the Bitcoin system uses cryptographic evidence to preserve the integrity of the network (Coryanne Hicks, 2020). Since its introduction in 2008, Bitcoin has reached unprecedented highs of $64,000 per coin in November 2021 (Griffith and Yaffe-Bellany, 2022). Although Bitcoin was not the first cryptocurrency, it is the most valuable coin to date, with a market capitalisation of $444 billion. Ethereum is the second most valuable cryptocurrency, with a market capitalisation of $185 billion, and Tether is third, with $66 billion (Kelly Anne Smith, 2018). The potential of cryptocurrencies and their high valuations has caught the attention of many financial institutions, who are now spending substantial amounts of resources to better understand the nature of the cryptocurrency market. The global leader in financial services, JP Morgan, now offers their wealth management clients access to six cryptocurrency funds, including the Grayscale Bitcoin Trust (Daniel, 2022). Consumers have also taken notice of the rise of Bitcoin and other cryptocurrencies, with the number of Bitcoin users now over 75 million, up from three million in 2014 (Flitter, 2021).

The burgeoning literature is evident from the volume of research published in recent years. However, there is limited research on predicting cryptocurrencies before 2015; some early studies were published by Graves and Au (2015), who developed a simple logistic regression and linear regression model. They achieved an accuracy of 55%. Another study by Shah and Zhang (2015), which was also published in 2015, deployed a Bayesian regression model to predict Bitcoin price. They experienced limited success from their model. Published research on the topic, has grown considerably since 2015 and the models being deployed are becoming more complex. Sin and Wang (2017) did exactly this when they developed an ensemble of neural networks to predict Bitcoin prices. More recently, research has involved complex models such as LSTM and ARIMA (Hua, 2020). While Dutta et al. (2019) published

a research paper where they built and deployed a Gated Recurrent Unit model, which proved to be successful in forecasting prices.

This research paper begins by discussing the background of cryptocurrencies and Bitcoin, before providing an overview of similar published literature. This is achieved by a literature review, where research papers of similar nature are highlighted and discussed. These papers include the applications of machine learning models, including alogirthms which are of interesting to this research project. Some of the most common models used were Long Short-Term Memory (LSTM), Autoregressive Integrated Moving Average (ARIMA), and Support Vector Machine (SVM). Following this, the research question of the paper is stated, and the methodology is introduced; outlining the approach for the research, discussing the data being used, and providing an overview of the models in this research. Subsequently, exploratory analysis is carried out. This includes cleaning and preparing the data to be analysed, before moving on to the descriptive analysis. Several models are deployed in the inferential analysis to predict the price of Bitcoin. These models were Long Short-Term Memory (LSTM), Autoregressive Integrated Moving Average (ARIMA), and Random Forest (RF). Lastly, the paper provides a discussion of the findings of this research.

Purpose

The purpose of this study is to investigate the dynamics of the Bitcoin market through exploratory analysis and to what extent can the price of Bitcoin be predicted through the application of machine learning techniques. Firstly, the initial exploratory analysis will aim to investigate what influences Bitcoin's price and identify influential variables from the dataset, thus will be achieved by assessing the correlation between the variables in the dataset. Furthermore, the returns and volatility of Bitcoin will be explored to gain a perspective on the nature of this cryptocurrency. Secondly, multiple machine learning models will be constructed and deployed to forecast Bitcoin prices.

Background

Cryptocurrency

The increased innovation within the Internet of Things (IoT) space has delivered unprecedented change to our daily lives. One major element of IoT, that is altering our financial landscape, is cryptocurrencies. Although 2008 is often seen as the year cryptocurrencies came to light, it can be dated back to 1990, with the introduction of DigiCash Inc.'s eCash system. DigiCash Inc. was an electronic money start-up founded by David Chaum (Frankenfield, 2021). The following year Chaum, introduced the eCash system which was based on two research papers he had written (Chaum, 1983; Chaum et al., 1992). Similar to cryptocurrencies today, DigiCash payments were sent both online and offline using cryptographic protocols to prevent double spending. The eCash system was accessible through several institutions and smartcards were available to consumers in various countries, including the United States and Finland. In the ten years that DigiCash operated, they unfortunately, failed to persuade banks to get on board with their leading edge technology. The company declared bankruptcy in 1998, ten years before the financial crisis that would lead to the creation of blockchain-based cryptocurrencies like Bitcoin (Kagan, 2021).

Interest in cryptocurrencies was rekindled in 2008 with the start of the global financial crisis. As the crisis began, Szabo (2008) stated in a blog post that cryptocurrencies could address a few issues with the current fiat currency system. The idea of bit gold was proposed. There is gold to be mined, as the name implies, and bits to be registered on a digital register. The problems of a reliable third party would be resolved by the digital record, and in his own words, Szabo highlighted "*it would be very nice if there were a protocol whereby unforgeable costly bits could be created online with minimal dependence on trusted third parties, and then securely stored, transferred, and assayed with similar minimal trust*." Despite appearing complex, Szabo's proposal is a straightforward protocol that asks users to invest money to mine digital gold or bit gold in exchange for rewards, validating the public digital register in the process. The timing of the financial crisis and the distributed nature of the protocol set his strategy apart from other failed digital currencies in the past (Lee, 2015).

To date, there are more than 18,000 different types of cryptocurrencies, with a total market cap of $2 trillion (Sofi, 2021). Although there are many definitions of what a

cryptocurrency is, they tend to fit a similar description. The Financial Conduct Authority (FCA) defines a cryptocurrency as "*a cryptographically secured digital representation of value or contractual rights that is powered by forms of distributed ledger technology and can be stored, transferred or traded electronically*" (FCA, 2019). Another definition, from the European Central Bank, is "*a type of unregulated, digital money, which is issued and usually controlled by its developers, and used and accepted among the members of a specific virtual community*" (European Central Bank, 2015).

Bitcoin

On the 3rd of October 2008, Satoshi Nakamoto uploaded the document "Bitcoin: A Peer-to-Peer Electronic Cash System", which is regarded as the birth date of Bitcoin (Marr, 2017). This paper outlined the main idea behind the virtual currency. Despite numerous attempts, Satoshi's identity is still unknown to the general public, and it's unclear whether Satoshi is a group or a person. The initial Bitcoin client was published on January 9th, 2009, and Satoshi Nakamoto and Hal Finney carried out the first Bitcoin transaction on January 12th. This transaction concerned a 50 BTC genesis block (the first block in the Bitcoin blockchain).

Bitcoin is a distributed consensus network that makes use of the proof-of-work principle to establish a reliable and secure ledger (Voigt and Rosen, 2022). It was developed to get rid of third-party middlemen (like banks), who are often engaged with digital money transfers. These third-party intermediaries incur expenses for the services they supply, which are then passed along to their customers. Additionally, the elimination of the intermediaries was a major inspiration for the idea of Bitcoin because the middlemen can exert control over the process by controlling the flow and availability of currency (Ammous, 2018). Nakamoto overcame this issue when they introduced their peer-to-peer (P2P) electronic payment system that utilises cryptography to permit transactions to happen directly between two parties without the need for a reliable intermediary. By using blockchain technology, the network maintains a public ledger where all Bitcoin transactions are recorded and copies of it are stored on servers all around the world. One of these servers, referred to as a node, can be run by anyone with a powerful computer. Instead of relying on a single point of trust (like a bank) these nodes cryptographically agree on who owns whose coins. Every transaction is shared across nodes and broadcast to the network in a public manner. About every 10 minutes miners gather these

transactions into a collection called a block, which is then permanently added to the blockchain. The process of mining is what keeps the Bitcoin network running and creates new currency. Mining involves employing advanced hardware to tackle intricate computational arithmetic problems. Cryptocurrencies are held in digital wallets and can be accessed using client software or a variety of internet and hardware solutions, like how one would maintain traditional money in a physical wallet (Sparkes, 2021). Bitcoin's market capitalisation is currently $444 billion, down from a high of $1.23 trillion on the 9th of November 2021 (CoinGecko, 2021).

Bitcoin has experienced increased use and popularity over the last decade, and this is a result of its many strengths. As Bitcoin has no single point of trust, there is no control from any central government or bank, and as a result, no middleman monitors the transactions. The algorithm requires that most trading nodes are trustworthy and employ a democratic process to settle any disputes. In addition, because only the account holder's private key can be used to sign transactions out of an account, it ensures that no one can change Bitcoin transactions made online. Furthermore, a bitcoin transaction cannot be undone, unlike conventional economic transactions, which can. To safeguard sellers from fraud and ensure buyers have mechanisms for protection, trades are mathematically irreversible (Mirzayi and Mehrzad, 2017).

There are, however, some considerations with the use of Bitcoin and its network. Firstly, Bitcoin is not widely accepted as payment and as a result users can become inconvenienced when attempting to pay for goods and services. Furthermore, given the relatively small number of transactions that can be handled in a second, the network may get quickly crowded in some circumstances, such as when demand for Bitcoin increases. In certain circumstances, Bitcoin transactions may take many days to complete or may potentially never be completed if the sender doesn't attach a large enough transaction fee. Regarding transaction fees, they may fluctuate by up to 10% on a single day. This is extremely unsustainable for businesses to take Bitcoin as a form of payment as their overhead costs become unpredictable (Zhou, 2021). Volatility is another serious concern for Bitcoin. The value of Bitcoin is prone to extreme volatility, with its value decreasing or increasing significantly on a given day. This means people may be resistant to holding Bitcoin, as if they switch their money in Bitcoin, the value may change drastically (Reiff, 2022).

Literature Review

Crypto Price Prediction

Zhesi Chen et al. (2020) published a paper in which they utilised machine learning to predict the price of Bitcoin. Although research has been undertaken in machine learning to predict Bitcoin prices with great accuracy, few have concentrated on the viability of using various modelling techniques to samples with various data structures and dimensional attributes. Several different features are used in this study, these include but are not limited to a number of transactions, ‘market capitalisation’, ‘estimated transaction value’, ‘total transactions fees’, ‘Google trends, and many more. The study discusses how to design sample dimensions for Bitcoin price prediction by utilising the right machine learning techniques. The authors approach the issue - while drawing on the Occam's razor principle and the features of the datasets – as follows: firstly the prediction sample size is divided in half into five-minute intervals with large sample sizes and daily intervals with small sample sizes. Secondly, they carry out feature engineering: they choose a small number of high-dimension characteristics for five-minute internal trading and a few features for the daily price. Thirdly, they use less difficult machine learning (ML) models like Random Forest, XGBoost, Support Vector, and Long Short-Term Memory (LSTM), as well as straightforward statistical models like Logistic Regression and Linear Discriminant Analysis. Lastly, to prevent overfitting, they use simple statistical models with high-dimension features to forecast the daily price of bitcoin and then use machine learning models to predict high-frequency prices with fewer features. For daily price prediction of Bitcoin, a collection of high-dimension features, including property and network, trading and market, attention, and gold spot price, are employed, but for 5-minute interval price prediction, basic trading features obtained from a cryptocurrency exchange are used. In comparison to more complex machine learning algorithms, statistical techniques such as Logistic Regression and Linear Discriminant Analysis for daily Bitcoin price prediction using high-dimensional features obtain an accuracy of 66%. They outperform benchmark results for daily price prediction, with the statistical methods and machine learning algorithms' greatest accuracy rates of 66% and 65.3%, respectively. For Bitcoin 5-minute interval price prediction, machine learning models such as Random Forest, XGBoost, Quadratic

Discriminant Analysis, Support Vector Machine, and Long Short-term Memory outperform statistical methods, with accuracy reaching 67.2%.

Machine learning models having the ability to outperform simple statistical methods is also echoed by Politis et al. (2021) who argue cryptocurrencies are complex and therefore require advanced analytical techniques. Politis et al highlight cryptocurrencies exhibit high volatility and steep fluctuations in comparison to fiat money, with crypto prices being dependent on a whole range of factors related to the blockchain network, social popularity, market trends, and the price of other cryptocurrencies. Because of this, straightforward statistical methods fall short of capturing the complexity of projecting cryptocurrency prices. In their paper, they develop deep learning models consisting of Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Temporal Convolutional Networks (TCN) to serve two purposes: predict the price of Ethereum and which direction the price is heading, in both the short and long term (one and seven days, respectively). The models were evaluated using the Root Mean Square Error (RMSE) and Mean Absolute Performance Error (MAPE) for the regression problem and mostly accuracy for the classification problem. From their research, the hybrid Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) model performed the best in regression, with an RMSE of 8.6% and MAPE of 3.6%. The best classification model was the Ensemble model, with an accuracy of 84.2%. The best Ensemble model, in this case, was the one that combined the predictions of the LSTM, hybrid LSTM-GRU, and hybrid LSTM-TCN models. Generally, hybrid models outperformed the individual ones and the ensemble technique led to improved results. The researchers highlighted results from predicting the weekly forecasts were worse than the daily, suggesting that it is more difficult to predict the Ether price in the long term and easier in the short term. Poongdodio et al. (2020) also conducted a study to predict the price of Ethereum, albeit they constructed different machine learning models. In this study, a Linear Regression (LR) and a Support Vector Machine (SVM) model were constructed. Their models LR and SVM obtained an accuracy of 85.46% and 96.06%, respectively. The authors advised that with the use of feature engineering a very high accuracy score of 99% was obtained.

A similar study from Jaquart et al. (2021) attempts to analyse the short-term predictability of the bitcoin market. The data within the study is from Bloomberg, Twitter, and Blockchain.com, ranging from March 4, 2019, to December 10, 2019. The dataset includes variables on minutely price data for bitcoin, gold, and oil and minutely data on the total return variants of the indices including MSCI World, S&P 500, and VIX. They use various machine

learning models, including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), to conduct their analysis. They contend that the predictions made by their model are statistically sound. The accuracy of the model's predictions of the binary market movement ranged from 50.9% to 56%, with longer forecast horizons often leading to higher accuracy. For this prediction analysis, Jaquart et al. highlight recurrent neural networks and gradient boosting classifiers are suitable. Another study predicting Bitcoin price, which achieved impressive results, is from Amjad and Shah (2017). In this paper, the authors built three models: Logistic Regression (LR), Random Forest (RF), and Linear Discriminant Analysis (LDA). From their analysis, they achieved an accuracy of greater than 70% across all three models. The authors claim, from comparing their model to classical ARIMA models, that the classification models outperform ARIMA models.

Research Question

To explore the price dynamics of Bitcoin and identify other variables which influence the price. Through this exploratory analysis, relationships amongst variables will be identified through correlations and the extreme volatility of Bitcoin will be investigated to better understand the nature of this cryptocurrency. Through the application of inferential models, several machine learning techniques will be applied to predict the price of Bitcoin.

Methodology

Approach – CRISP-DM

For this study, the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was used. The CRISP-DM model is an industry-proven guide for analytic projects and it's the analytics process standard that's used the most today (Brown, 2015). As a result of the increased use of data mining in the early 1990s, the need of standardising the lessons learned into a single methodology grew (IBM, 2021). Therefore in 1996, three early adopter companies, Daimler, NCR, and OHRA, along with two of the top tool providers of the time, SPSS and Teradata, met to form a Special Interest Group (SIG). In less than a year, the SIG was able to codify what is now known as the CRISP-DM (Sridharan, 2018). The CRISP-DM process consists of six sequential stages. These are business understanding, data understanding, data preparation, modelling, evaluation, and deployment. A CRISP-DM methodology is a suitable approach for this project as the framework can be implemented in any data science project, regardless of its domain. In addition, the methodology provides a uniform framework for managing and planning a project. The six stages of the framework are easily understandable and encourage best practices (Choudhury, 2020).

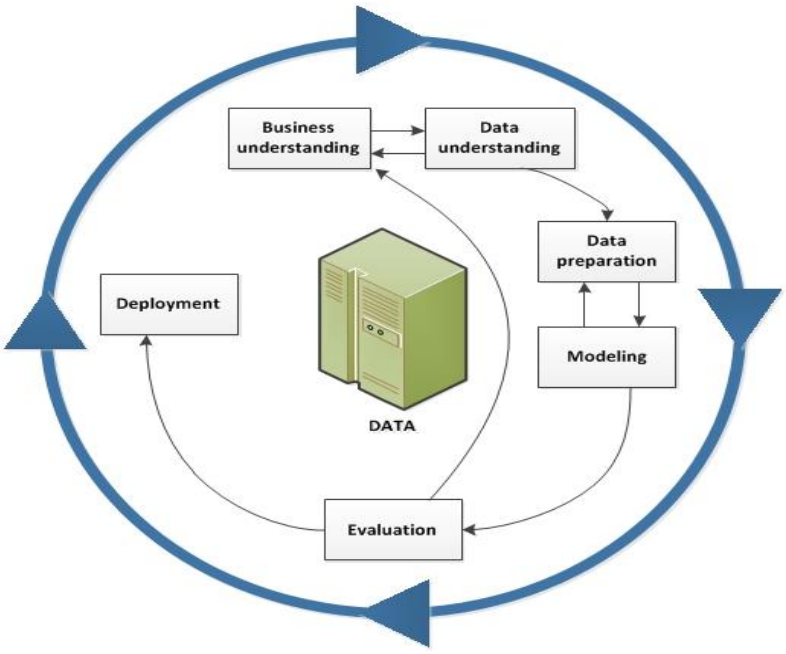

Figure 1: Crisp-DM Framework

Source: IBM, 2021

CRISP-DM Six Phases

Business Understanding – focuses on understanding the project's goals and specifications. This involves determining the project's objectives, assessing the situation, determining data mining goals, and producing a project plan.

Data Understanding – after gathering the first data, one would move on to actions to familiarise themselves with the data, detect data quality issues, and get preliminary insights from the data. Intriguing subsets may be found during this stage, to develop theories regarding hidden information.

Data Preparation – the final data set(s) are prepared for modelling during this stage, which is frequently referred to as "data munging." This stage typically consists of five tasks: select data, clean data, construct data, integrate data, and format data. This is an extensive phase of the framework, with a substantial amount of time being spent preparing the data.

Modelling – create a series of models using a variety of modelling methodologies. Typically there are four tasks in this phase: select modelling techniques, generate test design, build model, and assess the model.

Evaluation – although the previous phase evaluated the accuracy and generality of the model, this step assesses the extent to which the model satisfies the project objectives and looks for any reasons why the model may be flawed.

Deployment – the final stage involves deploying the finished model for use. This involves examining the previous evaluation results and determining a strategy for the model deployment. In addition, models should be continually monitored, to identify any new issues which may arise over time.

Data

For this project, a dataset containing fifteen variables (not including 'date') was provided in Excel format, ranging from the 7th of August 2012 to the 17th of February 2018. The data contained specific time-series information about Bitcoin and the Bitcoin network, as well as other variables on securities, commodities, and other cryptocurrencies. The wide range of variables in the dataset is beneficial to the project as this widens the scope of research. The data was pre-collected; therefore the exact source is unknown.

Table 1: Variable Descriptions

Variables	Description
BTC Price	Bitcoin price in USD
BTC Network Hashrate	Amount of computing power that is being used by the BTC network to process transactions
Average BTC Block Size	Average block size in MB
NUAU - BTC	Number of unique BTC addresses per day
Number TX – BTC	Number of BTC transactions per day
Difficulty - BTC	Refers to the level of difficulty associated with mining for new bitcoin blocks
TX Fees – BTC	Transaction fees
Estimated TX Volume USD – BTC	The total estimated value of transactions on the blockchain
Gold in USD	Gold price in USD
Ethereum Price	Ethereum price in USD
Litecoin Price	Litecoin price in USD
Bitcoin Cash price	Bitcoin cash price in USD
Cardano Price	Cardano price in USD
Nasdaq Composite Index	Nasdaq composite index
DJI	Dow Jones Industrial Average

Prediction Models

Long Short-Term Memory (LSTM)

The LSTM model, created by Juergen Schmidhuber, is a variety of recurrent neural networks (RNNs) that can learn long-term dependencies, especially in issues involving sequence prediction (Katte, 2018). In RNN models, the output of the previous step is used as the input for the following step. Schmidhuber addressed the problem of RNN long-term reliance, in which the RNN cannot predict words kept in long-term memory but can make more accurate predictions based on current data. By default, the LSTM may retain data for a very long time, therefore it is suitable for time series prediction, data processing, and classification (Saxena, 2021).

Four neural networks and a large number of memory blocks make up the LSTM model. A typical LSTM consists of a cell, an input gate, an output gate, and a forget gate (see Figure 2). Three gates regulate the information flow into and out of the cell, and the cell retains values for arbitrary periods. The LSTM method is well suited to classify, examine, and forecast time series of unknown durations (Kalita, 2022).

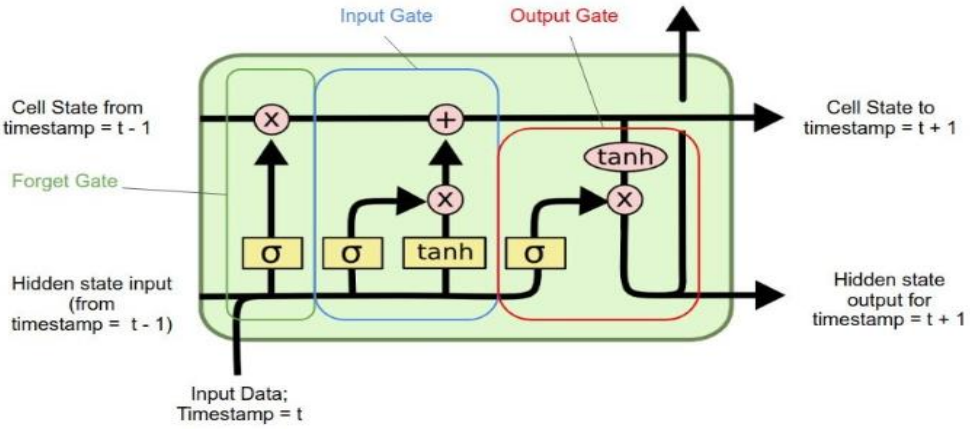

Figure 2:Long Short-Term Memory (LSTM)

Source: Turin, 2020

The LSTM model is well suited for the task of predicting Bitcoin price in this project. This is because the algorithm can capture both short and long-term seasonal trends, such as

weekly patterns, and long-term seasonal patterns like a yearly pattern. In addition, the model can manage different lengths as inputs, meaning it is easily adaptable.

Autoregressive integrated moving average (ARIMA)

Autoregressive integrated moving average, or ARIMA, is a machine learning model that uses time series data to make predictions. Instead of using actual values, the model looks at variations between values in the series to forecast future securities or financial market movements. The ARIMA model consists of three components: (1) autoregression (AR) – which refers to a model which shows a changing variable that regresses on its own prior or lagged values, and (2) integrated (I) – which represents the differencing of raw observations, to enable the time series to become stationary, (3) moving average (MA) - takes into account the relationship between an observation and a residual error from a moving average model applied to lagged observations (Hayes, 2012). The proper number of lags or amount of differencing to apply to the data will be determined by statistical software, which will also check for stationarity (Prabhakaran, 2019). The findings will then be generated, and they are frequently interpreted similarly to that of a multiple linear regression model.

The ARIMA model is a suitable machine learning approach to use for this project because to generalise the forecast, only the prior data of a time series are needed, and it performs well on short-term forecasts. Furthermore, ARIMA can model non-stationary time series data (Bora, 2021), which is suitable in this instance as the Bitcoin price statistical properties change over time, meaning it is non-stationary time series data.

Random Forest (RF)

Random Forest is an ensemble (the process of using multiple models) of decision trees. This algorithm is unique in the sense that it can be used for both classification and regression. Random Forests, which are made-up decision trees, are a straightforward model that essentially asks the question ‘yes’ or ‘no’ at each node in the tree (Beheshti, 2022). These decision trees are mostly developed using a bagging technique that employs a variety of learning models to improve the outcome. The Random Forest model is one of the most used algorithms due to its simplicity and diversity. While the trees are developing, the random forest adds more

randomness to the model. When dividing a node, it looks for the best feature from a random subset of features rather than the most crucial one. A better model is often produced as a result of the great diversity this causes (Donges, 2021).

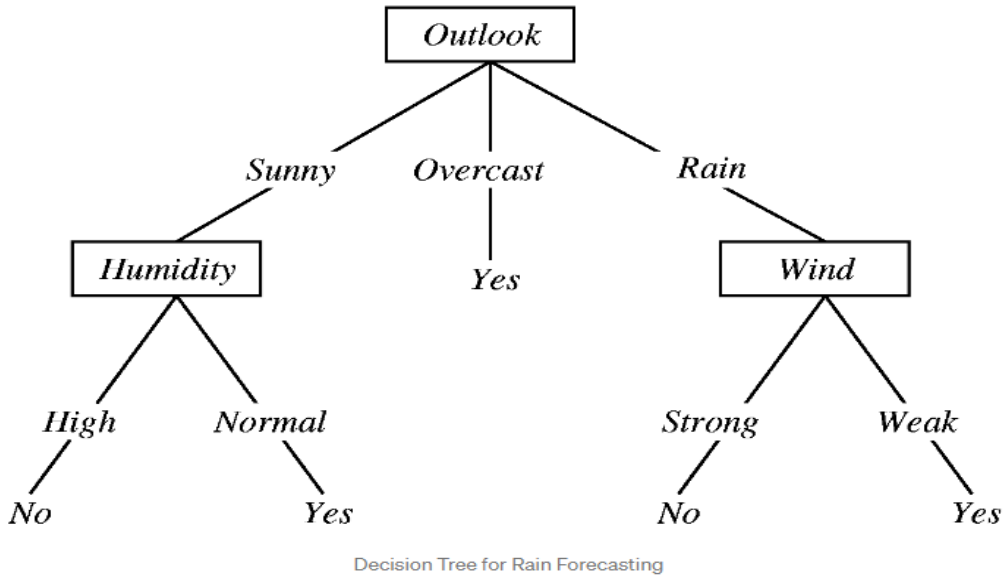

Figure 3: Decision Tree Example

Source: Yadav, 2018

Analysis and Results

Exploratory Analysis

Data Cleansing and Transformation

The dataset was originally in an xlsx Excel file before being converted into a comma-separated values (CSV) file and subsequently loaded into Jupyter Notebook, where the coding aspect of the project took place. All data points were initially examined, and the 'date' variable was converted to the index of the Pandas DataFrame. Before any exploratory analysis is conducted, data cleansing must be undertaken. This is a vital step in the process because clean data will ultimately boost overall productivity and enable the best possible information to be used in decision-making.

An extremely high number of missing values (27,543) were initially identified; however, most of these rows contained no information. After removing all rows containing no values across all columns, the number of missing values was 6,551 – this was spread over six variables. Columns '*Cardano Price', 'Bitcoin Cash Price', and 'Ethereum Price'* all contained a very high number (almost all rows for the first two) of missing values (see Figure 4). While the other three columns containing missing values were *'Gold in USD', 'Nasdaq composite index', and 'DJI'* (see Figure 4). Although there are a great number of missing values for these three variables, it is most likely not a mistake in the data as these securities are not traded on the weekend due to the stock market being closed on Saturday and Sunday.

```
date                              0
BTC Price                         0
BTC network hashrate              0
Average BTC block size            0
NUAU - BTC                        0
Number TX - BTC                   0
Difficulty - BTC                  0
TX fees - BTC                     0
Estimated TX Volume USD - BTC     0
Gold in USD                     572
Ethereum Price                 1081
Litecoin Price                    0
Bitcoin Cash Price             1790
Cardano Price                  1860
Nasdaq composite index          624
DJI                             624
dtype: int64
```

Figure 4: Missing Values for Each Column

Multiple options were considered when dealing with many missing values within this dataset. These included dropping the rows, filling with backward and forward fill, or replacing them with aggregated values e.g. mean or median. For variables *'Gold in USD', 'Nasdaq composite index', and 'DJI'*, the Pandas forward fill technique was used. This enables the missing value to be replaced by the next value (excluding missing values). Forward fill was a suitable method in this instance as the missing values for the three columns were spread sporadically throughout the dataset. As for the other three variables, '*Cardano Price', 'Bitcoin Cash Price', and 'Ethereum Price'*, they were not filled with any value (not a suitable method as the missing values were all at the beginning) or deleted, they were simply assigned to a new subset of data for further analysis. In addition, many of the variables had significant outliers, however, these were left untouched. The reasoning behind this was because of the nature of the data itself. Several of the outliers belonged to variables related to the Bitcoin network, so they may tend to spike at certain periods, e.g. TX fees may spike due to limited miners on the network or the BTC difficulty may increase due to miners solving puzzles in less than ten minutes. Therefore, as it cannot be concluded that the outliers are truly aberrant, no changes were made. Furthermore, the BTC Price value for the first 264 rows of the dataset - the dates 7th August 2012 to 27th April 2013 – was zero. As the main purpose of this project is to analysis the Bitcoin price, these rows were removed from the dataset, leaving 1,735 rows of data. Lastly, much of the data was on a different scale (see Figure 5) and this is an issue when analysing and developing machine learning models. This is a problem because when the data points are far from each other, the feature with the higher value range may dominate when calculating distances, which ultimately leads to a slow or unstable learning process. To overcome this concern, the SKLearn standard scaler technique was applied to the dataset, resulting in the data points being closer together (see Figure 5)

BTC Price	BTC network hashrate	Average BTC block size	NUAU - BTC	Number TX - BTC	Difficulty - BTC	TX fees - BTC
0.00	1.508528e+01	0.072908	32197.0	28636.0	2.036671e+06	20.927535
0.00	1.721139e+01	0.088954	35918.0	38855.0	2.036671e+06	52.180184

BTC Price	BTC network hashrate	Average BTC block size	NUAU - BTC	Number TX - BTC	Difficulty - BTC	TX fees - BTC
-0.458834	-0.514500	-1.264787	-1.212715	-1.190127	-0.527837	-0.477453
-0.458834	-0.514500	-1.216817	-1.194352	-1.085714	-0.527837	-0.245214

Figure 5: Scaled Data – Before and After

The distributions of the variables were investigated. As can be seen in Figure 6, most of the variables do not follow a normal distribution, instead, they are heavily skewed to one side. The only variable which is somewhat normally distributed is DJI. Other variables, such as the cryptocurrency prices and those directly linked to Bitcoin e.g. 'BTC network hashrate', 'Difficulty – BTC' etc are extremely right skewed. Typically, it is favoured for the variables to follow a normal distribution, however in this instance, they were not converted to a normal distribution, instead, they were left untouched. This was because of the nature of the machine learning models being constructed in this project. Some models, like logistic regression, linear regression, Gaussian Naïve Bayes, and LDA heavily favour normal distributions as they are calculated from the assumption that the distribution is bivariate or multivariate normal (Barai, 2020). However, as this project focuses on the LSTM, RF, and ARIMA models, which do not require normal distributions, the decision to leave the distributions untouched is justified.

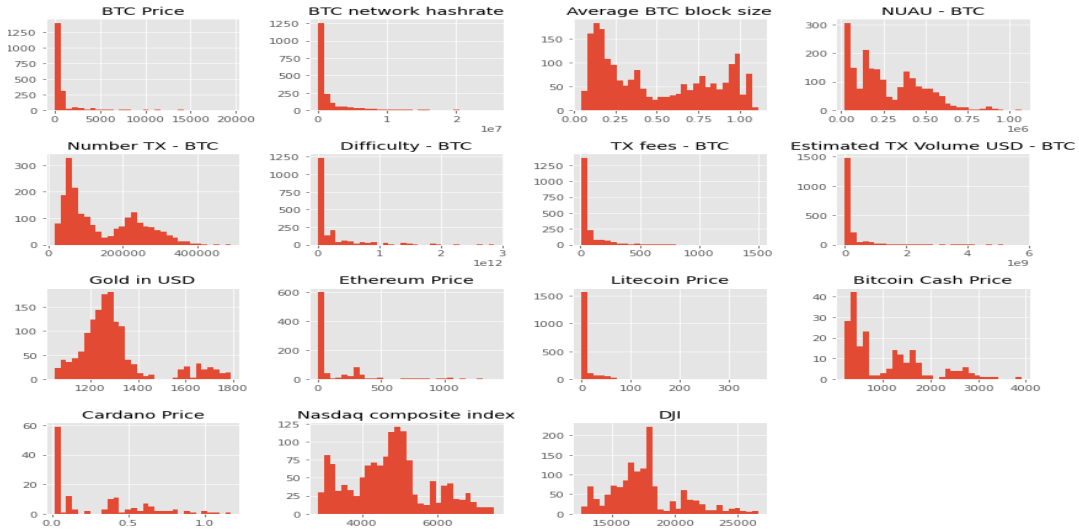

Figure 6: Distributions of Variables

Descriptive Analysis

As previously stated, the first 264 rows were dropped from the dataset, therefore Bitcoin price is analysed from 2013 to 2018 (see Figure 7). As can be seen from the figure below, there was a spike in Bitcoin value during late 2013 and early 2014 – reaching a price of $946.49 on the 7th of January 2014, before enduring a period of stability with little volatility from 2014 to 2017. 2017 was the year Bitcoin seen an unprecedented rise in value, eventually reaching a staggering $19,475.80 on 17th December 2017. Following this upward trend, the price of Bitcoin sharply declined at the beginning of 2018.

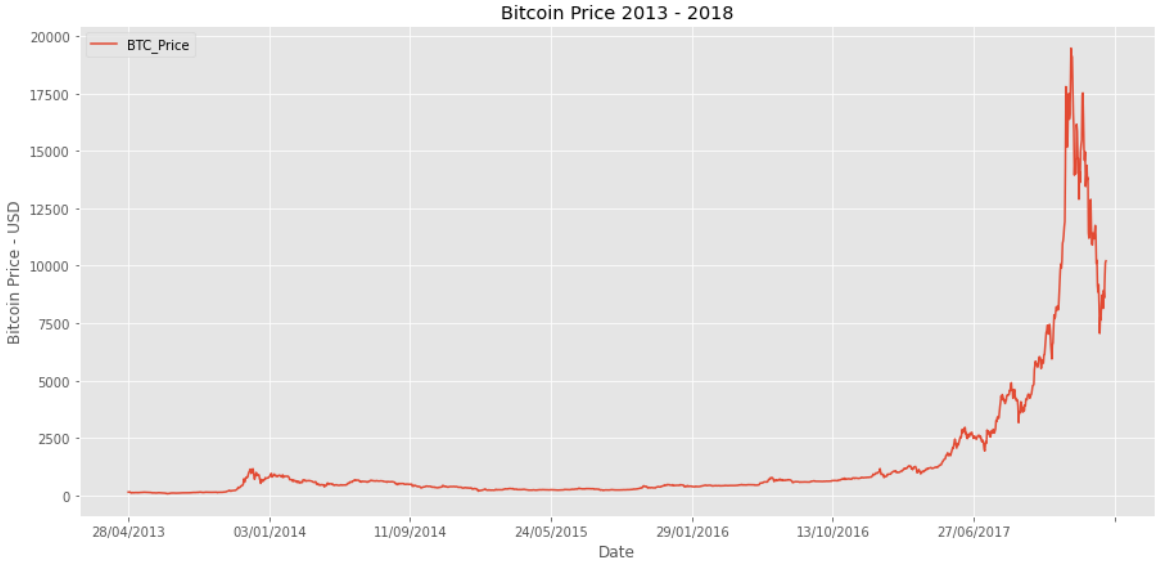

Figure 7: Bitcoin Price 2013 – 2018

Figure 8 shows a comparison of the four different cryptocurrencies that were included in the dataset. These were Bitcoin, Litecoin, Cardano, and Ethereum. It should, however, be noted that much of the data for Cardano and Ethereum was missing. Therefore, Ethereum is analysed from 2015 to 2018 and Cardano from late 2017 to 2018. Figure 8 demonstrates that Bitcoin is overwhelmingly the most valuable cryptocurrency, with the other three not coming close in value. The value of both Litecoin and Ethereum remained largely unchanged until late 2017 when they began to rise, like Bitcoin. As Bitcoin began to reach unprecedented highs from late 2017 to the beginning of 2018, the other three cryptocurrencies in question also rose, with Litecoin reaching $359.13 in December 2017, Ethereum at $1,397.48 in January 2018, and Cardano at $1.17 in January 2018 (6253.86% increase from late 2017 for Cardano).

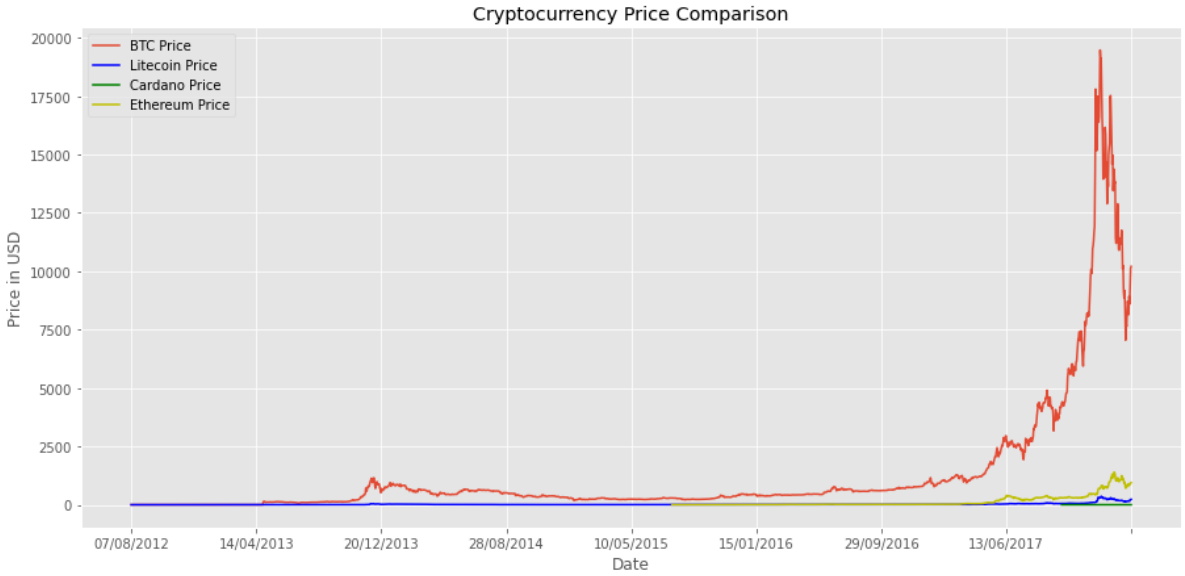

Figure 8: Cryptocurrency Price Comparison

Figure 9 provides a closer look at the rise in the value of cryptocurrencies experienced from 2017 to 2018. As previously mentioned, the price of the other three cryptocurrencies began to rise significantly in late 2017; although they did not rise at the same pace as Bitcoin. In addition, the Bitcoin price endured high levels of volatility during this period.

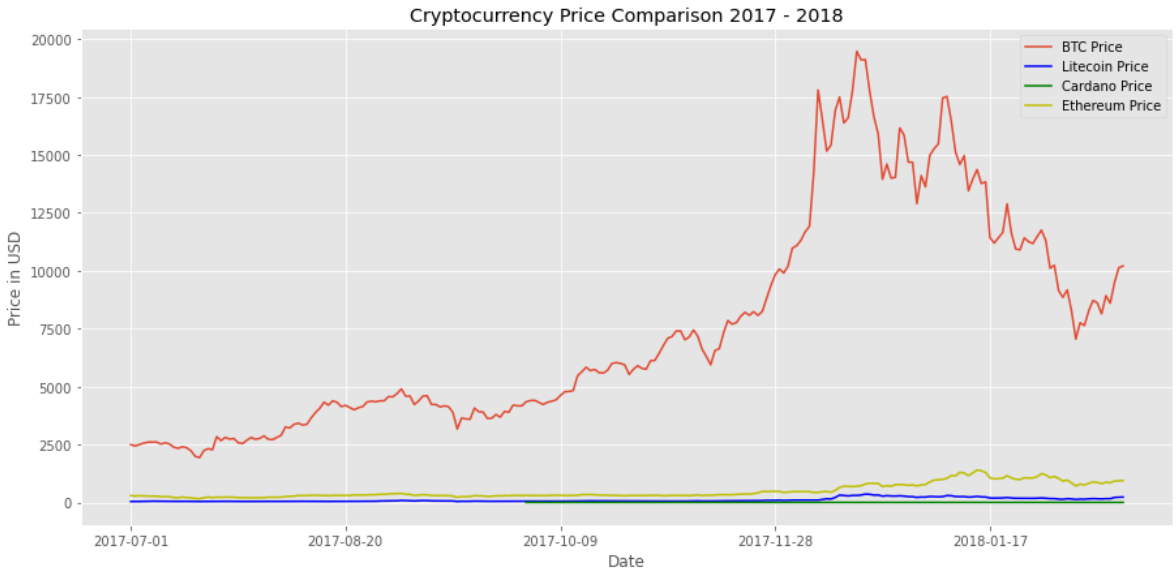

Figure 9: Cryptocurrency Price Comparison 2017 - 2019

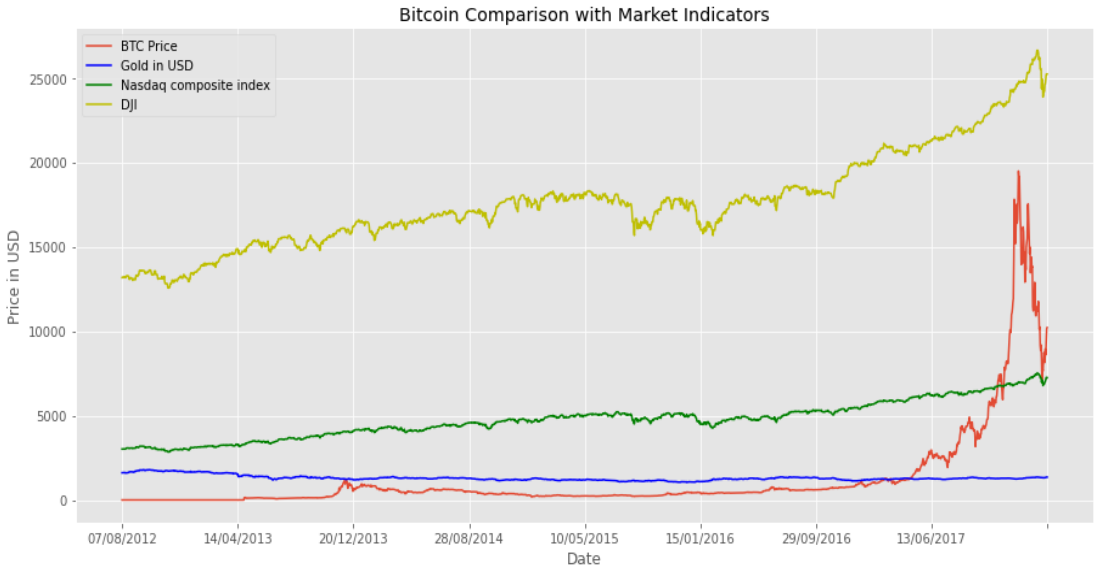

Figure 10: Bitcoin Comparison with Market Indicators

Figure 10 presents a comparison between Bitcoin and the market indicators from the dataset. These are 'Gold in USD', 'Nasdaq Composite Index', and the 'Dow Jones Industrial (DJI)'. All three indicators appear to be much less volatile than Bitcoin, which can be identified from the graph. The DJI, which consistently remains higher than Bitcoin, follows a steady increase over time, with some volatility experienced. Furthermore, the DJI begins to fall in price at the end of 2017, in line with Bitcoin which also began to fall during this period. One point, which is astonishing about Figure 10, is the sheer increase in value Bitcoin experienced in 2017 – none of the indicators experienced a sudden change in price to this degree.

```
BTC Price                       1.000000
Estimated TX Volume USD - BTC   0.963337
Litecoin Price                  0.956320
Ethereum Price                  0.904410
BTC network hashrate            0.896457
Difficulty - BTC                0.892732
Bitcoin Cash Price              0.887683
TX fees - BTC                   0.772550
DJI                             0.748465
NUAU - BTC                      0.666367
Nasdaq composite index          0.666168
Cardano Price                   0.659068
Number TX - BTC                 0.549591
Average BTC block size          0.539300
Gold in USD                    -0.072292
Name: BTC Price, dtype: float64
```

Figure 11: Bitcoin Correlation with all Variables

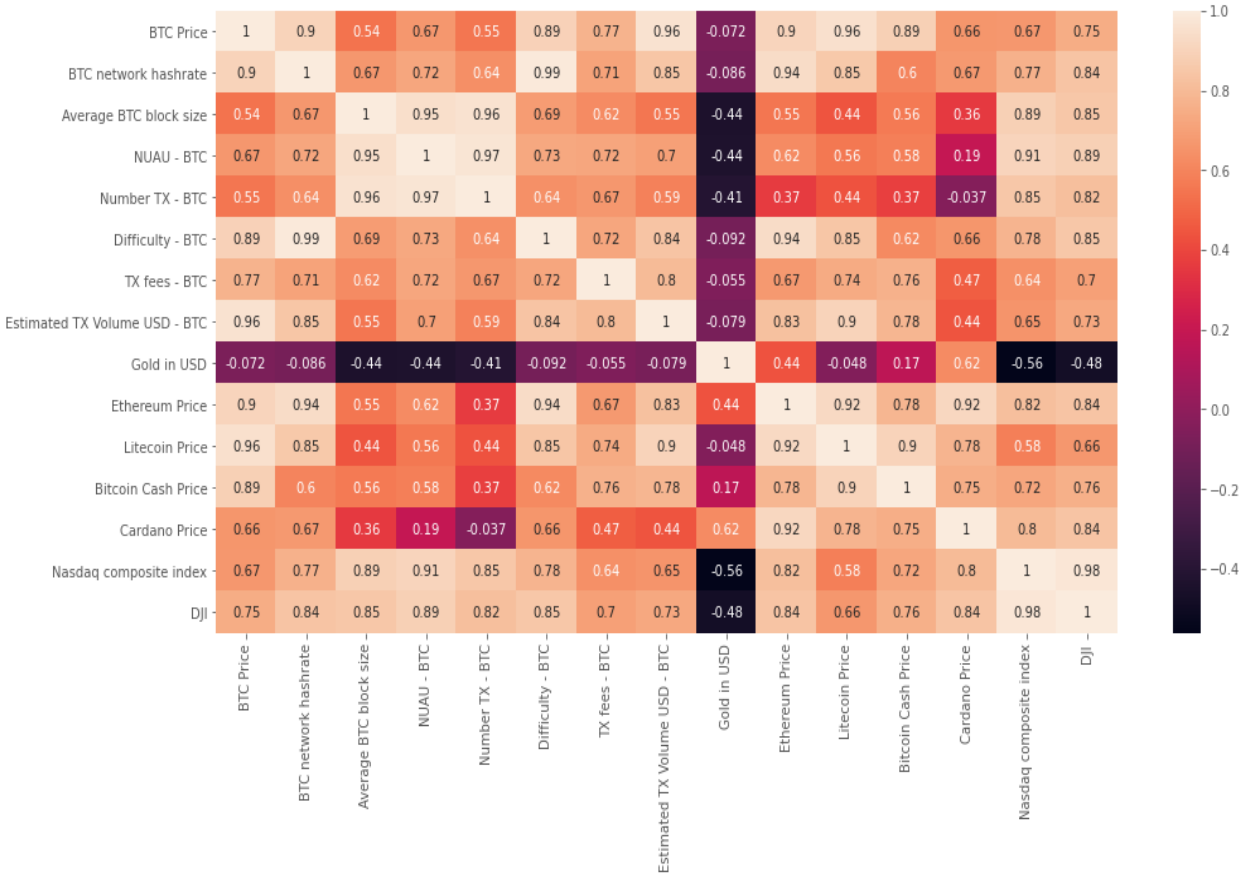

Figure 12: Heatmap of Correlations Between all Variables

Figure 11 presents the correlation between Bitcoin price and all the variables within the dataset. The correlations are displayed in ascending order. Figure 12 displays all the correlations between all variables in a heat map – lighter shades of squares represent a higher correlation value. Strong positive correlations exist heavily throughout the dataset, which is evident from the abundance of light-shaded squares. Extremely high positive correlations are present between the price of Bitcoin and the variables 'Estimated TX Volume USD (0.96), Litecoin price (0.96), and Ethereum price (0.90). Interestingly, the last two mentioned are other cryptocurrencies. TX Volume differs slightly, this is directly linked to Bitcoin itself as this is a measurement of Bitcoin transaction volume. Two of the three market indicators (Nasdaq Composite Index and DJI) have strong positive correlations with Bitcoin. Nasdaq's correlation with Bitcoin price is 0.67, while DJI's is 0.75. The remaining indicator, Gold in USD, has a correlation of -0.07 with Bitcoin, indicating there is no relationship between the two variables. Interestingly, 'Gold in USD' is the only variable that consists mainly of negative weak to

moderate correlation values with the other variables within the dataset; this is evident from the darker shaded squares in the heatmap (see Figure 12).

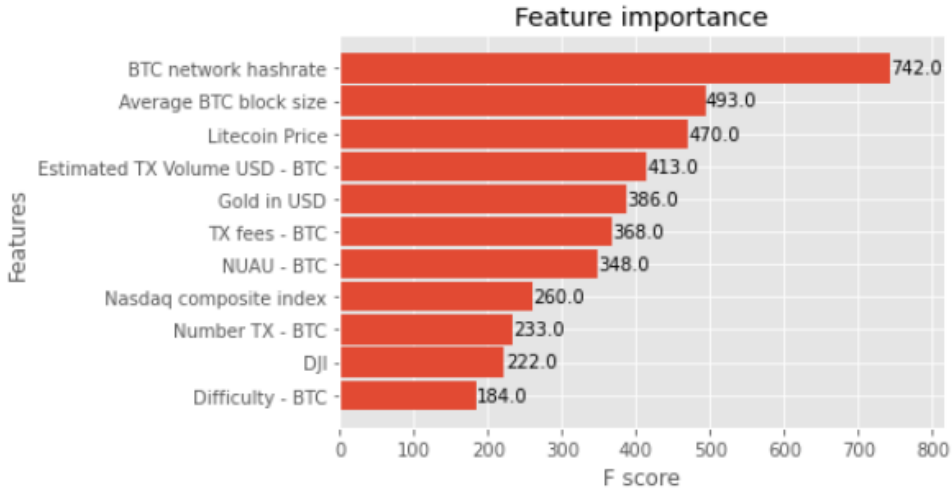

Figure 13: Feature Importance

In addition to investigating the relationships between variables from the correlation values, the feature importance technique is an insightful method for identifying variables that have a strong relationship with the target variable – Bitcoin price. As can be seen from Figure 13, many of the variables have a high level of importance. This will be useful when building models which require more than one dependent variable.

Two further columns were created in the Pandas dataframe to calculate the returns and log returns of Bitcoin price. Return and log return allows for an understanding of the return experienced on given security over time. Since it helps remove non-stationary characteristics of time series data and makes it more stable, using the log returns is often preferred for practical reasons, notably in mathematical modelling. As can be identified from Figure 14, high log return values were seen from late 2013 to early 2014, and again in 2018. During this period investors who traded frequently had the opportunity to make high returns, however, they could also have made severe losses.

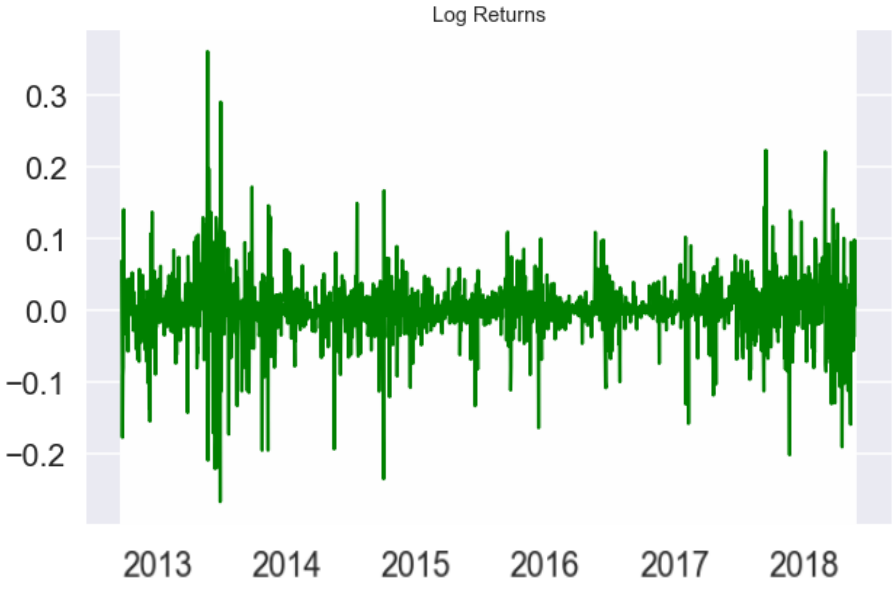

Figure 14: Log Returns

An additional column was created in the Pandas dataframe to calculate the daily volatility of Bitcoin prices. Volatility is the standard deviation of a security's annualised returns over a specific period and allows for an understanding of the potential range of price changes for the given security, or in this case, Bitcoin price. This measurement is important because it provides a sense of how risky Bitcoin is. A security, or cryptocurrency, with a high level of volatility, is inherently riskier, but this risk cuts both ways. The likelihood of success while investing in a volatile investment is raised in proportion to the risk of failure. Because of this, a lot of traders with high-risk tolerance base their trading decisions on a variety of volatility measures. Figure 15 displays the daily volatility distributions for Bitcoin. The graph shows that the majority of the volatility remained within the 0.01 to 0.05 range, however it should be noted this is a high level of volatility. At the extreme, some days experienced volatility as high as 0.14; a 14% change in price in a twenty-four-hour period is extremely risky. Bitcoin experienced huge changes in its price daily, meaning it is an extremely risky purchase for an investor.

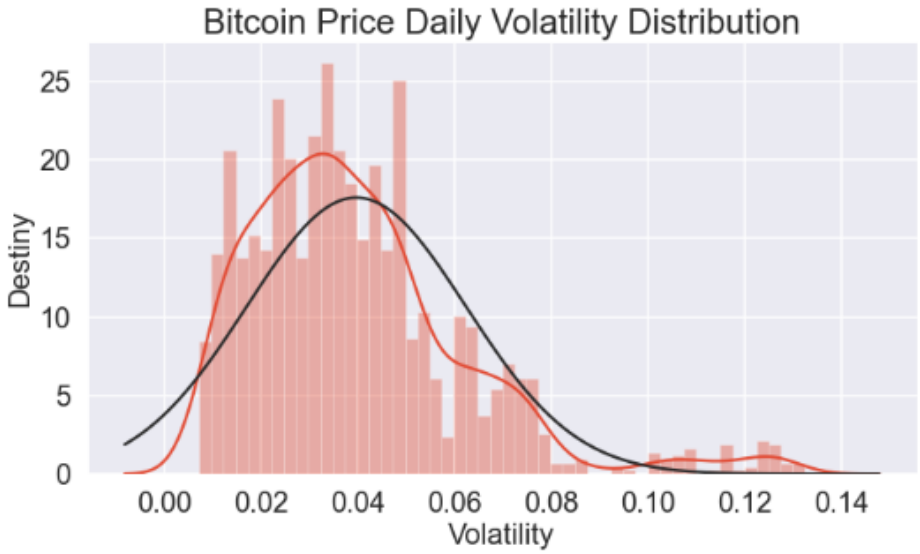

Figure 15: Daily Volatility Distribution of Bitcoin

To gain a perspective of how volatility Bitcoin may be, a comparison with the Dow Jones Industrial Average (DJI) is provided. Figure 16 displays the daily volatility distribution of the DJI.

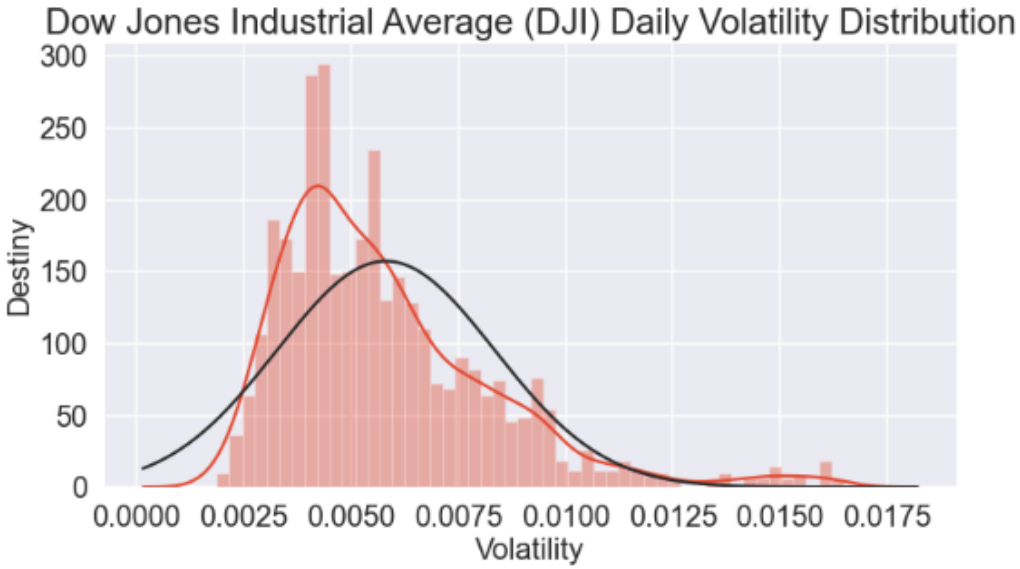

Figure 16: Daily Volatility Distribution of DJI

As can be seen from Figure 16, the daily volatility of DJI is much less than was seen for Bitcoin. Most of the daily distributions for DJI range from 0.0025 to 0.0100, while for Bitcoin the majority ranged from 0.01 to 0.05 – a huge difference when comparing both. For DJI, extreme daily volatility of 0.0175 was seen, however, this extreme for DJI was a common

daily occurrence for Bitcoin. For comparison, Bitcoin experienced volatility of 0.14, which is eight times greater than the extreme of DJI (0.0175).

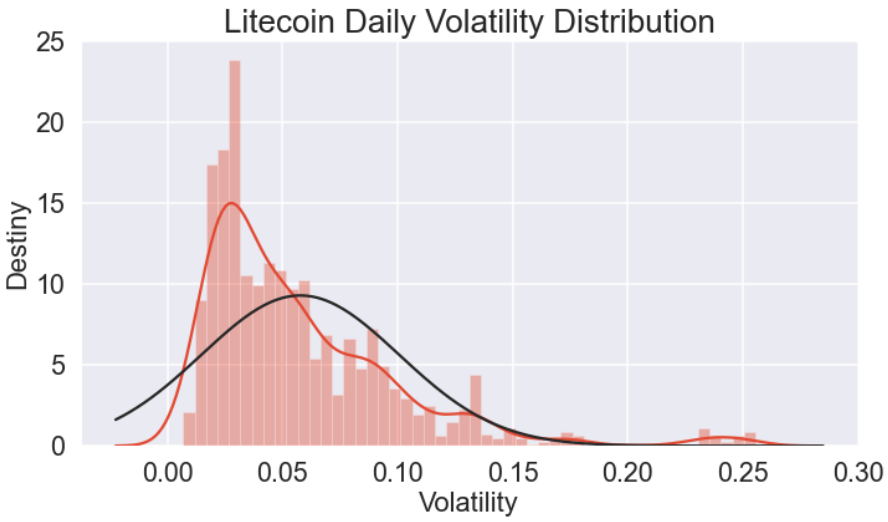

Figure 17: Litecoin Daily Volatility Distribution

The daily volatility distributions for the price of Litecoin (another cryptocurrency) can be seen in Figure 17. Similar to Bitcoin, the volatility of Litecoin is high. The majority of the daily volatility distributions ranged from 0.01 to 0.10 – a higher range than Bitcoin, and much higher than DJI. The extreme daily volatility for Litecoin reached as high as 0.25 – far greater than the extreme of Bitcoin. Interestingly, both the cryptocurrencies experienced high volatility, while the DJI experienced very little. It's evident from these comparisons that cryptocurrencies are a riskier investment, compared to typical financial securities.

To further investigate the volatility of Bitcoin, a boxplot graph grouping the daily volatility by month has been provided (see Figure 18). January and December experienced the greatest number of highly volatile days, with some reaching 0.12. All months experienced at least one or two days of extreme volatility, which can be seen from the outliers in Figure 18.

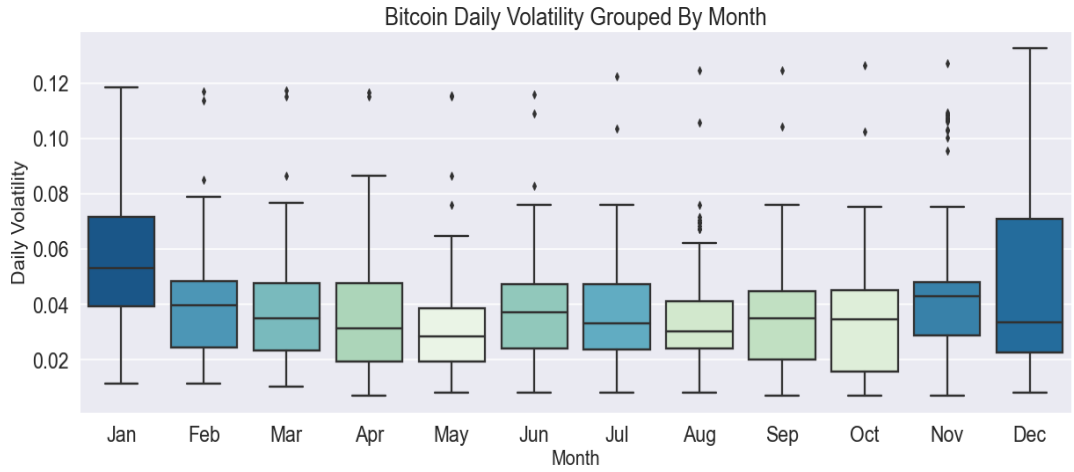

Figure 18: Bitcoin Daily Volatility Grouped by Month

Figure 19 groups the daily volatility by year. As can be seen from the graph, the price of Bitcoin experienced high volatility in 2013, before enduring a period of greater price stability from 2014 to 2017. The least amount of volatility was during 2016, however from then onwards the daily volatility began to increase each year. In 2018 the volatility returned to levels seen before throughout 2013, although the volatility remained around 0.08, unlike in 2013 when there were lows of 0.02 and highs of 0.13.

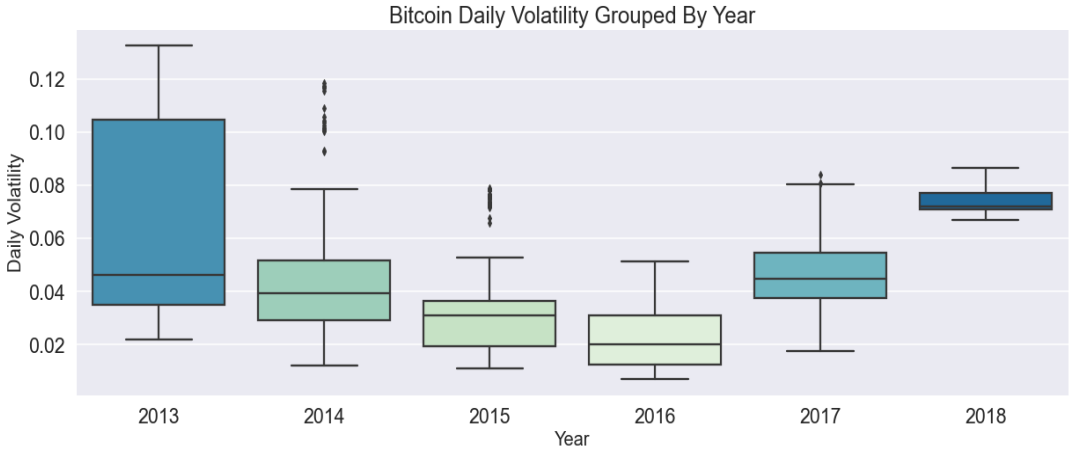

Figure 19: Bitcoin Daily Volatility Grouped by Year

Inferential Analysis

For the inferential analysis, several machine learning models were attempted. However, in the end, the final three models were LSTM, ARIMA, and Random Forest.

Long Short-Term Memory (LSTM)

The validation loss for the LSTM model is shown in Figure 20. For this model, the epoch value was set to 250. As expected, the validation loss for the training set is very low. For the test set, the validation begins high, before quickly falling to 0.025. Multiple epoch values were tested, with 250 producing the best results.

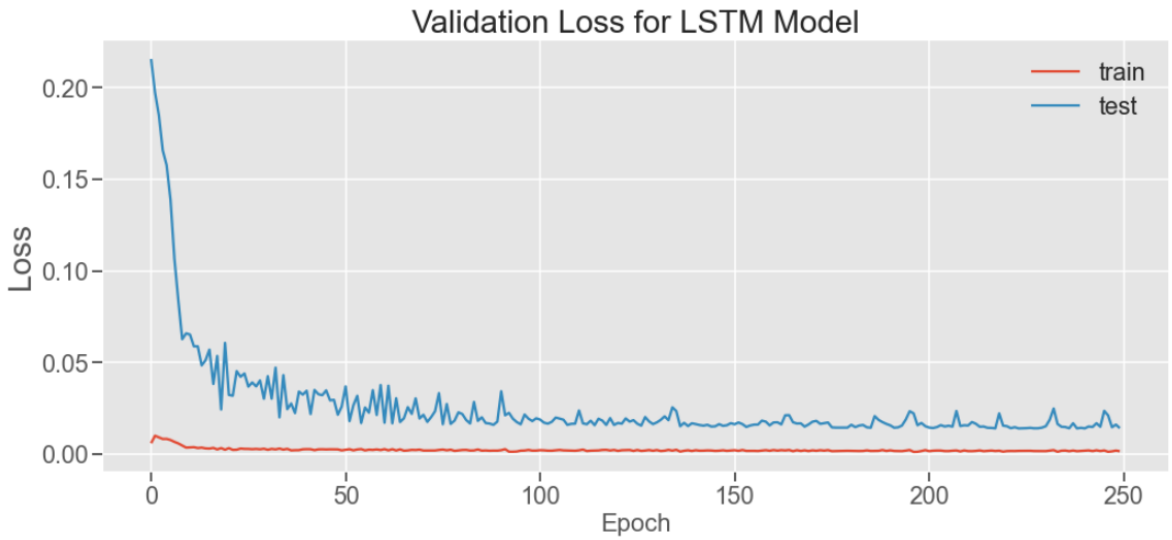

Figure 20: Validation Loss for LSTM Model

Figure 21 displays the predicted output values for the LSTM model. The green line (prediction values) follows the same trend as the red line (true price) successfully. As observed in Figure 21, the predictions seem to be a little off at the beginning, as can be seen from the green line being positioned slightly above the red. As the days increase, the prediction line progressively improves. The LSTM produced an RMSE of 147.05 for the training data and an RMSE of 516.52 for the testing data.

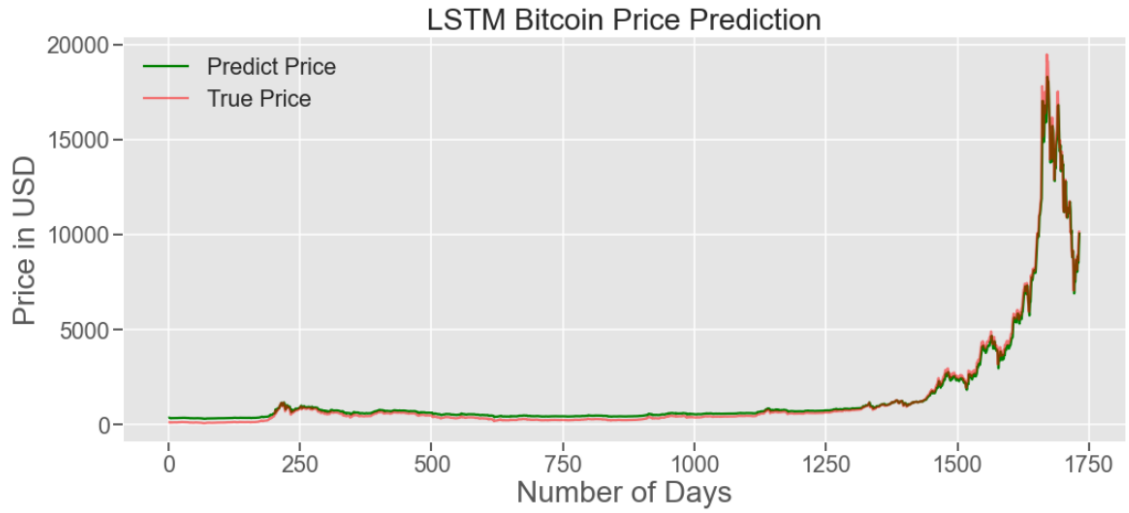

Figure 21: LSTM Bitcoin Price Prediction

Autoregressive Integrated Moving Average (ARIMA)

Figure 22 presents the predicted output from the ARIMA model for the test data, which contains the data from 2017 onwards (hence why the prediction is over 350 days). Similar to LSTM, the green line (predicted values) follows the red line (real values) closely. This is evident from 0 to 275 days, however, where there is more volatility in the price, the algorithm does not seem to predict as accurately, which can be observed from 275 days onwards. The RMSE for the ARIMA test set 543.67.

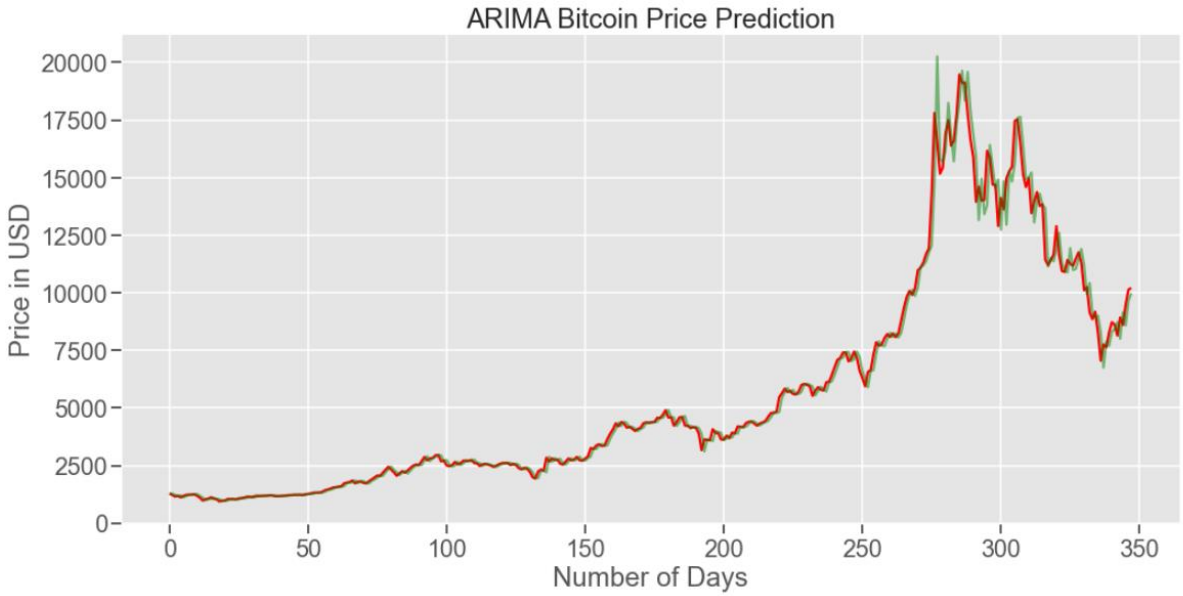

Figure 22: Arima Bitcoin Price Prediction

Random Forest Regressor (RF)

The final model constructed for this analysis was a Random Forest model. Figure 23 displays how well the prediction price followed the actual prices. As evident from the graph, the green line (prediction values) follows a very similar trend to the red line (true price). Unlike the LSTM and ARIMA models, the Random Forest algorithm includes more than one dependent variable. In this instance, a total of seven dependent variables were used to predict the Bitcoin price. These variables were 'BTC network hashrate', 'Average BTC block size', 'NUAU – BTC', 'TX fees – BTC', 'Estimated TX Volume USD – BTC', 'Litecoin Price', and 'DJI'. These variables were selected because they have the strongest relationship, which was identified through the feature importance method applied in the exploratory analysis. This Random Forest model produced an RMSE of 279.59 for the training data and an RMSE of 124.79 for the testing data. In addition, as this algorithm is a meta estimator that fits several classifying decision trees to the data, the accuracy scores were able to be obtained from the classifying. An accuracy score of 98.24% was achieved for the training data and 94.6% for the testing data.

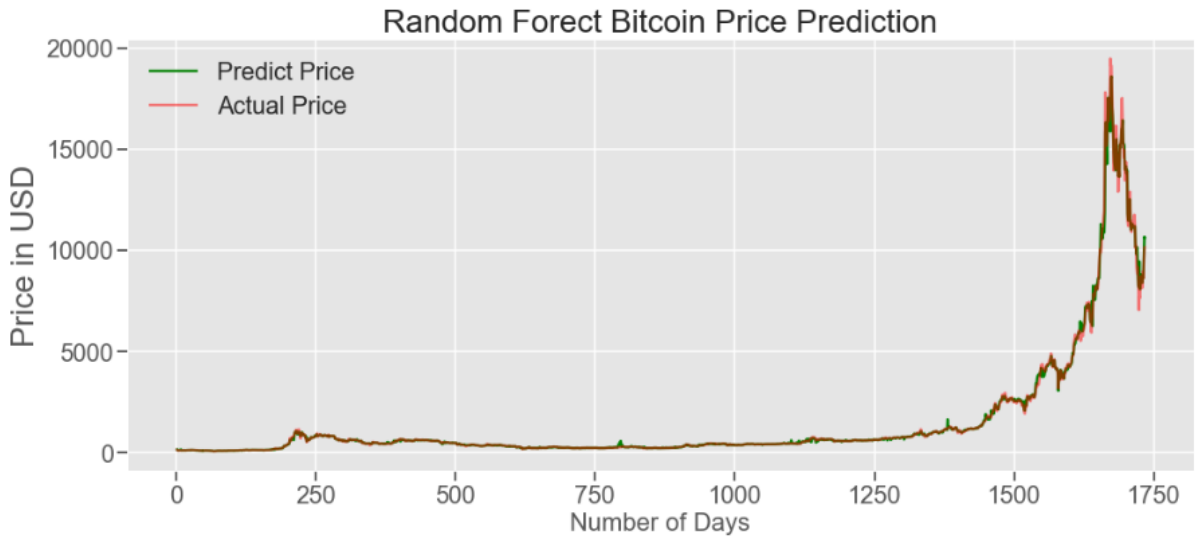

Figure 23: Random Forest Bitcoin Price Prediction

Discussion

The purpose of this research was to investigate the dynamic of the Bitcoin market through exploratory analysis and to what extent can the price of Bitcoin be predicted through the application of machine learning techniques. Through the exploratory analysis, it was discovered that Bitcoin experienced an unprecedented rise in price from 2013 to 2018 (see Figure 7). During this period, there were times of high volatility, most notably from late 2013 to early 2014, and from 2017 to 2018; Bitcoin experienced higher levels of stability from the period of 2014 to 2017 (see Figure 7). It was revealed that the other three cryptocurrencies in this analysis (Ethereum, Litecoin, and Cardano) followed a similar rise to that of Bitcoin in 2017 (see Figures 8 and 9). This suggests that cryptocurrencies are closely related. A comparison of Bitcoin to other market indicators, such as Gold, Nasdaq Composite Index, and DJI, revealed at first glance that Bitcoin is much more volatile than the market indicators. Interestingly, when Bitcoin began to rise in value in late 2017, as did the DJI; similarly, when Bitcoin started to dip in 2018, the DJI followed a similar trend (see Figure 9). Similar to Ethereum, Litecoin, and Cardano, both DJI and Nasdaq Composite Index followed a similar trend to Bitcoin, thus suggesting there is a relationship between the two market indicators and Bitcoin.

Upon further investigation of the relationships that exist within the dataset, it was discovered Bitcoin has a strong positive correlation with Ethereum (0.90), Litecoin (0.96), and Cardano (0.66) (see Figure 11). This suggests when one cryptocurrency begins to increase in value, the others in question also do so; when the cryptocurrency market does well, all cryptocurrencies do well, and on the contrary, when the market does poorly, all cryptocurrencies do poorly. This finding is in line with the work of Gkillas et al. (2018), who found distinct patterns of highly significant bivariate dependence in the distribution tails of several of the most fundamental and popular cryptocurrencies. Therefore, Gkillas et al. claim the cryptocurrency market is highly correlated, resulting in cryptos following similar trends. Similarly, strong positive correlations exist between Bitcoin and two of the market indicators: DJI (0.75) and Nasdaq Composite Index (0.67) (see Figure 11). Again, like the cryptocurrencies, this strong positive relationship suggests when Bitcoin performs well, both the DJI and Nasdaq Composite Index follow a similar trend, albeit not as significant as Ethereum (0.90) and Litecoin (0.96). These findings are echoed by the work of Zhang et al. (2018) who argue, from their analysis, that the cryptocurrency market is highly correlated with

the DJI. Interestingly, there was very little correlation between Bitcoin and gold (-0.07). This is neither a strong positive correlation, nor a strong negative correlation This means the price of gold is not influenced by Bitcoin, and vice versa. This is a very intriguing finding, as just recently there have been articles published claiming Bitcoin is the new gold (Duke, 2022; Alin, 2022), yet there is next to no correlation between the two. This finding is similar to the work of Baur and Hoang (2021), who suggests there is no correlation between them both. They argue this could be a result of the substitution effect, where investors sell gold and buy Bitcoin, and the catching up effect, where investors buy Bitcoin to catch up with the market value of gold. Another strong positive correlation exists between Bitcoin and the variables directly linked to Bitcoin and the Bitcoin network, these are 'Estimated TX Volume' (0.96), 'Network Hashrate' (0.89), 'Difficulty' (0.89), 'Bitcoin Cash Price' (0.89), 'TX fees' (0.77), 'NUAU – BTC' (0.67), 'Number TX' (0.55), and 'Average block size' (0.54) (see Figure 11).

As previously stated, Bitcoin experienced significant volatility from the period of 2013 to 2018, which was observed in Figure 7 and Figure 8. However, by plotting the daily volatility distributions into a histogram graph, the severity of the volatility can be identified (see Figure 15). The daily volatility for Bitcoin typically stayed within the 0.01 to 0.05 range, with extreme volatility reaching as high as 0.14 on some days. This is an extreme swing in price. To appreciate how significant this swing is, we can consider that the average annual return of the stock BP has been 2.91% (Macrotrends, 2022), yet Bitcoin can fluctuate by 14% in value in one day. For comparison, the daily volatility of the DJI mainly ranged from 0.0025 to 0.0100, while days of extreme volatility reached 0.0175 – this is notably lower than the volatility observed for Bitcoin. Similarly to Bitcoin, the daily volatility of Litecoin mainly ranged from 0.01 to 0.10, and extreme days see volatility as high as 0.25. Both the cryptocurrencies experienced extremely high volatility, with Litecoin experiencing more extreme days. However, in comparison, the volatility of the DJI was very low. This suggests that cryptocurrencies are typically more volatile than traditional securities. This is confirmed by Vejačka (2014), who published a research paper stating that cryptocurrencies are more volatile in comparison to basic investment instruments. Baur and Dimpfl (2018) further this point, by stating that cryptocurrencies are indeed volatile, however, the volatility increased more to positive shocks than negative shocks. Baur and Dimpfl argue positive shocks can be explained by the herding behaviour of uneducated investors, buying out of concern for missing out on rising bitcoin prices, and pump-and-dump schemes. The contrarian actions of knowledgeable

investors can be used to explain the lesser and hence asymmetric volatility response to negative shocks.

To further understand the volatility of Bitcoin, Figure 18 grouped the daily volatility observations by month. January and December contained the days with the highest volatility, with some days reaching 0.12 in January and 0.13 in December. Although not as frequently as January and December, all months featured at least one or two days with extreme volatility, with most months having a day with at least 0.10. This suggests Bitcoin is generally volatile across all months, but some months experience more than others. The cause of January and December experiencing extreme volatility may be linked to the 2018 'Crypto Winter' crash. During this period, Bitcoin, and the crypto market in general, lost huge amounts in value (Weissman, 2018). Furthermore, daily volatility observations were grouped by year, as can be seen in Figure 19. High levels of price volatility were experienced in 2013, 2014, and 2018, with the year 2015 and 2016 being more stable. The year 2013 was particularly fascinating as volatility ranged from 0.02 to 0.13 and in 2014, volatility reached 0.12 on several days throughout the year, while 2018 saw less of a range, with lows of 0.07 and highs of 0.09. The extreme volatility of 2013 and 2014 suggests this was a rocky period for Bitcoin, however, there are multiply explanations for the levels of volatility experienced. In 2013, the Bank of China and the Ministry of Industry in China issued a warning that banks are not allowed to handle transactions related to Bitcoin. This led to a dramatic fall in the value of Bitcoin (Sergeenkov, 2021). In addition, 2014 was also highly volatile for Bitcoin, as the popular Japanese crypto exchange, Mt. Gox was hacked, resulting in the exchange, which handled 70% of all Bitcoin transactions, declaring they had lost 650,000 to 850,000 coins. This eventually led to Mt. Gox filing for bankruptcy on February 28th, 2014 (Frankenfield, 2022). As expected, the log returns of Bitcoin, which was presented in Figure 14, followed a similar pattern to the volatility levels. High log returns were evident in periods of high volatility.

For this analysis, three models were utilised for Bitcoin prediction. The RMSE scores for all three models can be observed in Table 2. The best performing model was the Random Forest. This finding was surprising as this was the least complicated model of the three, however, it was the only model which was multivariate; there were seven independent variables, and therefore the algorithm had many features to learn from. The RMSE score for the training data was 279.59 and 124.79 for the test data. As this was a classification algorithm, the accuracy was measured. The accuracy achieved for the training set was 98.24%, and for the testing set, it was 94.6% The high levels of accuracy from the Random Forest model are in

line with the work of Amjad and Shah (2017), who achieved high accuracy in their random forest model for predicting time series Bitcoin price data.

Table 2: RMSE Scores for Models

	Model		
RMSE	LSTM	ARIMA	Random Forest
TEST	516.52	543.67	279.59
TRAIN	147.05		124.79

The other two models, LSTM and ARIMA, performed worse than the random forest model. LSTM achieved an RMSE score of 147.05 for the training data. This is low because it is the data the algorithm is learning from; however, it may have been potentially overfitting slightly. The RMSE score for the test set was 516.52. The ARIM model performed the worst of the three models, with an RMSE score of 543.67 for the test data. Interestingly, these two algorithms are known for their ability to accurately predict time series data, yet a less complicated model, random forest, managed to achieve better results. Both the LSTM and ARIMA algorithms were trained on only the Bitcoin price data. As the random forest was trained on seven features, this suggests multivariate algorithms may be better suited in this instance to predict the Bitcoin price. The better performance of the random forest algorithm is also highlighted by Amjad and Shah (2017), who claim classification models can outperform LSTM and ARIMA when predicting Bitcoin.

Limitations

This literature review had some limitations. First, time constraints were an issue as this project was researched and written within a short period. This meant not as much time was dedicated to the initial research step as one would have liked. However, the paper does include a wide range of research on multiple topics from different journals, but with more time these topics could have been researched in greater detail. Furthermore, if time allowed more models would have been built and tested, resulting in a deeper understanding of the complexities of each model. The second limitation was the data itself. The Bitcoin information within the dataset ranged from 2013 – 2018, there the models built were only trained on this data. Therefore, if the models were to be deployed on present-day Bitcoin data, they may struggle to predict the price. In addition, the exploratory analysis could have been more in-depth if there were other cryptocurrency data, this would mean the comparison between Bitcoin and other cryptocurrencies was more diverse.

Future Work

Improvements to the findings of this work could be achieved by considering and addressing the limitations. As the ARIMA was the worst performing algorithm, improvements could be made by reassessing the data inputted and applying different parameters. Furthermore, the model could be extended to the SARIMA or SARIMAX model, which may produce better results. In addition, data could extend to include more variables that are influential to Bitcoin, this would ultimately increase the accuracy of the models. Sentiment analysis data could also be included, as previous work has suggested sentiment plays a crucial role in the price of Bitcoin (Gontyala, 2021). This sentiment data could be obtained from social media platforms such as Reddit and Twitter, as well as Google Trends.

Conclusion

The purpose of this study was to investigate the dynamic of the Bitcoin market through exploratory analysis and investigate to what extent can the price of Bitcoin be predicted through the application of machine learning techniques. From the exploratory analysis, correlation values between Bitcoin price and other variables were identified. Bitcoin has a strong positive correlation with Ethereum (0.90), Litecoin (0.96), and Cardano (0.66) (see Figure 11). This means that cryptocurrencies tend to move in the same direction, when Bitcoin does well, all cryptocurrencies do well, and the opposite applies when Bitcoin is doing poorly. In addition, strong positive correlations exist between Bitcoin and market indicators DJI (0.75) and Nasdaq Composite Index (0.67). However, there was no correlation between Bitcoin and gold. Furthermore, extreme volatility for Bitcoin was experienced from 2013 to 2014, and again in late 2017 to early 2018. Some days saw volatility as high as 0.14 for Bitcoin, while Litecoin, another cryptocurrency, reached 0.25. For comparison, the volatility of the DJI reached a mere 0.0175.

Three models were built for this analysis, these were Random Forest, ARIMA, and LSTM. The ensemble method, Random Forest, achieved the best results. The RMSE score for the training data was 124.79 and 279.59 for the test data. ARIMA achieved an RMSE score of 543.67, while the score of LSTMS for the test data was 516.52. A slight difference between them both.

References

Alin, V. (2022). *Why Deutsche Bank Thinks Bitcoin Will Be the New Gold*. [online] Coinmonks. Available at: https://medium.com/coinmonks/why-deutsche-bank-thinks-bitcoin-will-be-the-new-gold-d8edfc235da1 [Accessed 1 Aug. 2022].

Ammous, S. (2018). *The Bitcoin standard : the decentralized alternative to central banking*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Barai, A. (2020). *Normal Distribution and Machine Learning*. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/normal-distribution-and-machine-learning-ec9d3ca05070#:~:text=Note%3A%20Normality%20is%20an%20assumption [Accessed 25 Jul. 2022].

Baur, D.G. and Dimpfl, T. (2018). Asymmetric volatility in cryptocurrencies. *Economics Letters*, 173, pp.148–151. doi:10.1016/j.econlet.2018.10.008.

Baur, D.G. and Hoang, L. (2021). The Bitcoin gold correlation puzzle. *Journal of Behavioral and Experimental Finance*, 32(2214-6350), p.100561. doi:10.1016/j.jbef.2021.100561.

Beheshti, N. (2022). *Random Forest Regression*. [online] Medium. Available at: https://towardsdatascience.com/random-forest-regression-5f605132d19d [Accessed 5 Jul. 2022].

Bora, N. (2021). *Understanding ARIMA Models for Machine Learning*. [online] Capital One. Available at: https://www.capitalone.com/tech/machine-learning/understanding-arima-models/ [Accessed 29 Jul. 2022].

Brown, M.S. (2015). *What IT Needs To Know About The Data Mining Process*. [online] Forbes. Available at: https://www.forbes.com/sites/metabrown/2015/07/29/what-it-needs-to-know-about-the-data-mining-process/?sh=177aafa6515f [Accessed 28 Jul. 2022].

Chen, Z., Li, C. and Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, 365, p.112395. doi:10.1016/j.cam.2019.112395.

Choudhury, A. (2020). *Why Is CRISP-DM Gaining Grounds*. [online] Analytics India Magazine. Available at: https://analyticsindiamag.com/why-is-crisp-dm-gaining-grounds/#:~:text=Benefits%20of%20Using%20CRISP%2DDM&text=CRISP%2DDM%20encourages%20best%20practices [Accessed 8 Jul. 2022].

CoinGecko (2021). *Cryptocurrency Prices, Charts, and Crypto Market Cap*. [online] CoinGecko. Available at: https://www.coingecko.com/.

Coryanne Hicks (2020). *The history of Bitcoin*. [online] US News & World Report. Available at: https://money.usnews.com/investing/articles/the-history-of-bitcoin [Accessed 2 Jul. 2022].

Daniel, W. (2022). *Bitcoin has 'significant upside' and could rise to $38,000, JPMorgan says*. [online] Fortune. Available at: https://fortune.com/2022/05/25/what-is-bitcoin-worth-price-outlook-jpmorgan-significant-upside-jamie-dimon-crypto/ [Accessed 3 Jul. 2022].

Donges, N. (2021). *A complete guide to the random forest algorithm*. [online] Built In. Available at: https://builtin.com/data-science/random-forest-algorithm#how [Accessed 24 Jul. 2022].

Duke, S. (2022). Bitcoin could be the new gold, says Goldman Sachs. *www.thetimes.co.uk*. [online] 6 Jan. Available at: https://www.thetimes.co.uk/article/bitcoin-is-the-new-gold-says-goldman-sachs-5sgb82bp3 [Accessed 2 Aug. 2022].

Dutta, A., Kumar, S. and Basu, M. (2019). A Gated Recurrent Unit Approach to Bitcoin Price Prediction. *SSRN Electronic Journal*. doi:10.2139/ssrn.3514069.

European Central Bank (2015). Virtual currency schemes – a further analysis. [online] doi:10.2866/662172.

FCA (2019). *Cryptoassets: our work*. [online] FCA. Available at: https://www.fca.org.uk/firms/cryptoassets#:~:text=Cryptoassets%20are%20cryptographically%20secured%20digital [Accessed 12 Jul. 2022].

Flitter, E. (2021). Banks Tried to Kill Crypto and Failed. Now They're Embracing It (Slowly). *The New York Times*. [online] 1 Nov. Available at: https://www.nytimes.com/2021/11/01/business/banks-crypto-bitcoin.html [Accessed 4 Jul. 2022].

Floyd, D. (2022). *How Bitcoin Works*. [online] Investopedia. Available at: https://www.investopedia.com/news/how-bitcoin-works/.

FRANKENFIELD, J. (2021). *What is DigiCash?* [online] Investopedia. Available at: https://www.investopedia.com/terms/d/digicash.asp#:~:text=DigiCash%20was%20a%20company%20founded.

Frankenfield, J. (2022). *Mt. Gox History*. [online] Investopedia. Available at: https://www.investopedia.com/terms/m/mt-gox.asp#citation-17 [Accessed 2 Aug. 2022].

Gkillas, K., Bekiros, S. and Siriopoulos, C. (2018). Extreme Correlation in Cryptocurrency Markets. *SSRN Electronic Journal*. doi:10.2139/ssrn.3180934.

Gontyala, S.P. (2021). Prediction of Cryptocurrency Price based on Sentiment Analysis and Machine Learning Approach.

Greaves, A. and Au, B. (2015). *Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin*. [online] Available at: http://snap.stanford.edu/class/cs224w-2015/projects_2015/Using_the_Bitcoin_Transaction_Graph_to_Predict_the_Price_of_Bitcoin.pdf.

Griffith, E. and Yaffe-Bellany, D. (2022). Bitcoin Plummets Below $20,000 for First Time Since Late 2020. *The New York Times*. [online] 18 Jun. Available at: https://www.nytimes.com/2022/06/18/technology/bitcoin-20000.html#:~:text=Bitcoin%20has%20erased%20some%20%24900 [Accessed 1 Jul. 2022].

Hayes, A. (2012). *Autoregressive Integrated Moving Average (ARIMA)*. [online] Investopedia. Available at: https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp#:~:text=An%20autoregressive%20integrated%20moving%20average%2C%20or%20ARIMA%2C%20is%20a%20statistical.

Hua, Y. (2020). Bitcoin price prediction using ARIMA and LSTM. *E3S Web of Conferences*, 218, p.01050. doi:10.1051/e3sconf/202021801050.

IBM (2021). *CRISP-DM Help Overview*. [online] www.ibm.com. Available at: https://www.ibm.com/docs/en/spss-modeler/18.2.0?topic=dm-crisp-help-overview.

Jaquart, P., Dann, D. and Weinhardt, C. (2021). Short-term bitcoin market prediction via machine learning. *The Journal of Finance and Data Science*, 7, pp.45–66. doi:10.1016/j.jfds.2021.03.001.

Kagan, J. (2021). *What is DigiCash?* [online] Investopedia. Available at: https://www.investopedia.com/terms/d/digicash.asp#:~:text=DigiCash%20was%20a%20company%20founded.

Kalita, D. (2022). *An Overview on Long Short Term Memory (LSTM)*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/03/an-overview-on-long-short-term-memory-lstm/ [Accessed 26 Jul. 2022].

Katte, A. (2018). *Meet Juergen Schmidhuber — The Father Of LSTM, An Outsider In The World Of Deep Learning*. [online] Analytics India Magazine. Available at: https://analyticsindiamag.com/meet-juergen-schmidhuber-the-father-of-lstm-an-outsider-in-the-world-of-deep-learning/#:~:text=Meet%20Juergen%20Schmidhuber%20%E2%80%94%20The%20Father [Accessed 26 Jul. 2022].

Kelly Anne Smith (2018). *13 types of cryptocurrency that aren't bitcoin*. [online] Bankrate. Available at: https://www.bankrate.com/investing/types-of-cryptocurrency/ [Accessed 28 Jun. 2022].

Lee, D. (2015). *Handbook of digital currency : bitcoin, innovation, financial instruments, and big data*. Amsterdam: Elsevier/ Ap.

M., P., Sharma, A., V., V., Bhardwaj, V., Sharma, A.P., Iqbal, R. and Kumar, R. (2020). Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system. *Computers & Electrical Engineering*, 81, p.106527. doi:10.1016/j.compeleceng.2019.106527.

Macrotrends (2022). *BP Return on Investment 2006-2021 | BP*. [online] www.macrotrends.net. Available at: https://www.macrotrends.net/stocks/charts/BP/bp/roi.

Marr, B. (2017). *A Short History Of Bitcoin And Crypto Currency Everyone Should Read.* [online] Forbes. Available at: https://www.forbes.com/sites/bernardmarr/2017/12/06/a-short-history-of-bitcoin-and-crypto-currency-everyone-should-read/?sh=1cfbbe5b3f27 [Accessed 1 Jul. 2022].

Maverick, J.B. (2022). *What Is the Average Annual Return for the S&P 500?* [online] Investopedia. Available at: https://www.investopedia.com/ask/answers/042415/what-average-annual-return-sp-500.asp#:~:text=The%20index%20has%20returned%20a [Accessed 2 Aug. 2022].

Mirzayi, S. and Mehrzad, M. (2017). *Bitcoin, an SWOT analysis.* [online] IEEE Xplore. doi:10.1109/ICCKE.2017.8167876.

Politis, A., Doka, K. and Koziris, N. (2021). Ether Price Prediction Using Advanced Deep Learning Models.

Prabhakaran, S. (2019). *ARIMA Model - Complete Guide to Time Series Forecasting in Python | ML+*. [online] Machine Learning Plus. Available at: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/ [Accessed 24 Jul. 2022].

Reiff, N. (2022). *Why Bitcoin Has a Volatile Value*. [online] Investopedia. Available at: https://www.investopedia.com/articles/investing/052014/why-bitcoins-value-so-volatile.asp [Accessed 10 Jul. 2022].

saxena, S. (2021). *LSTM | Introduction to LSTM | Long Short Term Memor*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/ [Accessed 28 Jul. 2022].

Sergeenkov, A. (2021). *China Crypto Bans: A Complete History*. [online] www.coindesk.com. Available at: https://www.coindesk.com/learn/china-crypto-bans-a-complete-history/#:~:text=2013%3A%20China%20bans%20banks%20from [Accessed 2 Aug. 2022].

Shah, D. and Zhang, K. (2015). Bayesian regression and Bitcoin.

Sin, E. and Wang, L. (2017). Bitcoin Price Prediction Using Ensembles of Neural Networks. *International Conference on Blockchain and Cryptocurrency (ICBC)*.

Sofi (2021). *Understanding The Different Types of Cryptocurrency*. [online] SoFi. Available at: https://www.sofi.com/learn/content/understanding-the-different-types-of-cryptocurrency/.

Sparkes, M. (2021). *What is bitcoin and how does it work?* [online] New Scientist. Available at: https://www.newscientist.com/definition/bitcoin/.

Sridharan, M. (2018). *CRISP-DM - A Framework For Data Mining & Analysis*. [online] thinkinsights.net. Available at: https://thinkinsights.net/data-literacy/crisp-dm/#Characteristics-of-CRISP-DM [Accessed 26 Jul. 2022].

Szabo, N. (2008). *Bit gold*. [online] Blogspot.com. Available at: http://unenumerated.blogspot.com/2005/12/bit-gold.html [Accessed Jul. 18AD].

Turing (2020). *Comprehensive guide to LSTM & RNNs.* [online] www.turing.com. Available at: https://www.turing.com/kb/comprehensive-guide-to-lstm-rnn.

Vejačka, M. (2014). *Basic Aspects of Cryptocurrencies Basic Aspects of Cryptocurrencies*.

Voigt, K. and Rosen, A. (2022). *What Is Bitcoin, and How Does It Work?* [online] NerdWallet. Available at: https://www.nerdwallet.com/article/investing/what-is-bitcoin [Accessed 2 Jul. 2022].

Weissman, C.G. (2018). *What the hell happened to crypto this year?* [online] Fast Company. Available at: https://www.fastcompany.com/90285052/beyond-the-bubble-what-happened-to-bitcoin-in-2018 [Accessed 5 Jul. 2022].

Yadav, P. (2018). *Decision Tree in Machine Learning*. [online] Medium. Available at: https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96.

Zhang, W., Wang, P., Li, X. and Shen, D. (2018). The inefficiency of cryptocurrency and its cross-correlation with Dow Jones Industrial Average. *Physica A: Statistical Mechanics and its Applications*, 510(0378-4371), pp.658–670. doi:10.1016/j.physa.2018.07.032.

Zhou, J. (2021). *Bitcoin Transactions Are Slow and Costly. Let's Explain Why.* [online] Medium. Available at: https://medium.com/geekculture/bitcoin-transactions-are-slow-and-costly-lets-explain-why-a3f6f2e326db [Accessed 12 Jul. 2022].