National College of Ireland

Prediction of Bitcoin Price using Data Mining

MSc Research Project
Data Analytics

Dharminder Singh Virk
x16102100

School of Computing
National College of Ireland

Supervisor: Dr. Eugene O'Loughlin

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Dharminder Singh Virk
Student ID:	x16102100
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Eugene O'Loughlin
Submission Due Date:	11/12/2017
Project Title:	Prediction of Bitcoin Price using Data Mining
Word Count:	XXX

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	10th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Bitcoin Price using Data Mining

Dharminder Singh Virk
x16102100
MSc Research Project in Data Analytics

10th December 2017

Abstract

Bitcoin is a computerized digital money and exchange network, represents an essential change in financial sectors, an interesting number of customers and excellent evaluation of channel inspection. In this research, dataset related to ten cryptocurrencies are used and created a new dataset by taking the closing price of each cryptocurrency for the research goal to ascertain how the direction and accuracy of price of the Bitcoin can be predicted by using data mining methods. Features engineering evaluated that all the ten cryptocurrencies are strongly correlated with each other. The task is achieved by implementation of supervised learning method in which random forest, support vector classifier, gradient boosting classifier, and neural network classifier are used under classification category and linear regression, recurrent neural network, gradient boosting regressor are used under regression category. In the classification category, support vector classifier achieved the highest accuracy of 62.31% and precision value 0.77. In regression category, gradient boosting regressor got the highest R-squared value 0.99.

1 Introduction

Machine learning algorithms are very popular in terms of solving classification and regression based problem. Although this experience is not new in the field of analytics and are generally utilized as a part of the stock market system which comes under financial market as suggested by (Kaastra and Boyd; 1996; White; 1988). Bitcoin is a problem based on classification and regression like the Stock market in its transitory stages. In this way, there is high fluctuation in the market (Brière et al.; 2015) and this create an opportunity to investigate further in this area. OkCoin, itBit, Bitfinex, CoinBase, Bitstamp are the most popular Bitcoin exchanges. Several machine learning algorithms are implemented to design models and their features changes with time as per required by the market to investigate the data more precisely. Further, Bitcoin is a digital money and its acceptance increasing constantly around the world. It works on a decentralized, distributed and trustless network in which all exchanges are presented on an open record called the Blockchain. Moreover, the price of the Ethereum was predicted based on the Ethereum and Bitcoin dataset using LSTM [1]. In this research, Bitcoin price can be predicted based on other nine cryptocurrencies. Likewise, the author used in this research

[1] Ethereum Price Prediction:`https://www.kaggle.com/phauly/get-rich-easily-by-following-these-simple-st`

the dataset of ten different cryptocurrencies. The different cryptocurrencies are Bitcoin, Litecoin, Ethereum, Nem, Neo, Dash, Monero, Ripple, Stratis, and Waves and used it for the research goal to achieve the best result.

1.1 Project Specification

1.1.1 Research Question

How the direction and accuracy of price of Bitcoin can be predicted by using data mining method?

1.1.2 Purpose

The purpose of this research is to determine how the direction and accuracy of price of Bitcoin can be predicted by using data mining methods. Study shows that research is lacking in the area of Bitcoin. Bitcoin is a problem based on time series prediction. In addition, Bitcoin is amazing unpredictable than various monetary standard which acts as a cryptocurrency lie in its transitory level. As a result, Bitcoin is amazingly unpredictable than other currencies like USD. Additionally, Bitcoin is viewed as a leading cryptocurrency, ranked as four out of the latest five years as per Malkiel and Fama (1970). Therefore, its expectation offers an opportunity, and this gives inspiration to investigate in this area. As a prove by an investigation of the existing research paper, applying machine learning methods on a Bitcoin dataset can offer outstanding performance improvements in the area. This is investigated by applying different machine algorithms such as Random Forest, Support Vector Machine, Gradient Boosting algorithm, and neural network. The result are shown in terms of accuracy. The structure of this thesis builds on existing literature which is discussed in section 2.
Further, predicting the price of the Bitcoin provide an advantage to make profit by buying or selling the asset. This paper is simply focus on the accuracy whether the price of the Bitcoin is going up or down and this facilities are provided by many exchanges [2]. Thus, the net worth which will be gained from this strategy is not only depend on accuracy but also depends on the position size.

1.1.3 Research Variables

In this research, the dataset related to ten cryptocurrencies are collected from Kaggle and the closing price of all the ten cryptocurrencies are taken as an independent variable from the datasets. Rather than focusing on one cryptocurrency, in this research nine other cryptocurrencies are utilised to determine the accuracy and direction of the Bitcoin price. If one cryptocurrency were implemented against Bitcoin, probably the accuracy and price prediction could not seem proper. As the result, nine cryptocurrencies are used for this research which will not only show the trend of each cryptocurrency but also show the correlation among each other and thus give a better result. That is why, the new dataset is created by taking the closing price from ten cryptocurrencies. In the past, study shows that McNally (2016) has used closing price average dataset which was collected from different exchanges like OkCoin, Coinbase, Bitstamp, Bitfinex and itBit to predict the price of the Bitcoin rather than any one specific exchange. Thus, the closing price of each cryptocurrency is chosen due to less noise in the dataset representing price of the

[2]Bitfinex Exchange:`https://www.bitfinex.com/`

Bitcoin going up or down. Depending on the result, the best model will be evaluated in the section 5. The more the accuracy achieved by the model the more the algorithm is best for the prediction. Moreover, performance metrics like precision, recall and f1-score, accuracy are shown in the section 5 . In addition to this, the author will discuss in this thesis, literature related to specific domain, literature related to machine learning, and supervised learning methods.

1.2 Bitcoin

In 2008, Satoshi Nakamoto created a digital cryptocurrency and exchange system known as a Bitcoin (Madan et al.; 2015; Ron and Shamir; 2013) and came into existence and recognised for its quality in 2012 and become a popular cryptocurrency (Economist; 2013). It depends on decentralized, shared, associate with the arrangement of an assets and trade organization did by the people from the system. On November 17^{th}, 2017 one Bitcoin (Bitstamp) equals to 7831.02 US Dollar and can be gotten through exchanging for items, services, or by mining, or different monetary forms suggested by Kaminski (2014). In the Bitcoin system, cryptography is utilised to exchange the cash and to control the creation, subsequently, the Blockchain i.e., the common open system is utilised to store the computerized marked messages (Kaminski; 2014). Grocer (2013) suggested that the utilization of Bitcoin has been seen appears to be financially little, in light of dangerous cost and high volatilities. As indicated by Economist (2013), new businesses are supported by financial specialists which are related to Bitcoin and has been known as a unit of account by the German finance ministry and on 18^{th} of November, an American Senate Council informed by senior authorities that Bitcoin has legal uses. But, there may have also been many cases of Bitcoin robbery, Workplaces that changes over Bitcoin to various fiscal structures have closed. For instance, In October, FBI, America shut the Silk Street which acts as an online discussion where Bitcoin are traded for illegal products and services (Economist; 2013). Also, Bitcoin cost changes profoundly noted $230 in April 2013 and falls distinctly $70 following three months and expanded in November up to $600 as per Economist (2013).
The author has found the relationship between the Bitcoin price, the tweets and Google trends view from previous research which was done by Matta et al. (2015a) and has been utilised as an evidence that can be used as predictors. In addition to these, same methodology was implemented by Matta et al. (2015b) for another analysis in which trading volume prediction are used alternatively to predict the price of the Bitcoin and achieved solid interrelationship between Google trend view and price of the Bitcoin. The dataset utilised for this exploration was taken from one year. Furthermore, to discover same outcomes a few researcher utilized Wavelets for their examination work (Kristoufek; 2015; Delfin-Vidal and Romero-Meléndez; 2016). Likewise, the author used in this research ten different cryptocurrencies closing price to predict the accuracy and the Bitcoin's price.

Despite the fact, huge information is accessible identifying with Bitcoin and its system, the creator contends that not all researcher used this data adequately and subsequently, it may not be reflected in the cost. The goal of this paper is to gain benefit of this theory over several data mining methods.
Bitcoin is exchanged on more than 40 trades overall tolerating more than 30 distinct monetary forms and has a present market capitalization of 9 billion dollars and its interest has developed significantly with more than 250,000 exchanges now occurring every day.

Study shows that $100,000 euros had been stole from a 36-year-old person when he signed on to the unsecured system to check the Bitcoin price proposed by Nick Whigham (2017). But the situation remains unclear by the police that the Bitcoin was already hacked before the victim logged on the network. Moreover, Mt Gox, Japenese based Bitcoin exchange was hacked in January 2014 which acts as a largest Bitcoin intermediary and regarded as the leading Bitcoin exchange before hackers grabbed 85,000 BTC. Bitcoin and Ethereum largest exchange in South Korea was hacked by hackers in June 2017 as reported by the local news and the customer claimed to have lost of 1.2 billion.
According to CNBC (2017), first time Bitcoin hit new record $8000 as per the information received from industry website CoinDesk. On November 12, the price of the Bitcoin was increased more than 47 percent. Moreover, Bitcoin, world leading cryptocurrency exceeded another record that is $8200 on Monday, 20 November 2017. November has been a greatly unstable month for Bitcoin and recorded cryptocurrency price fell to $5500 by Saxena (2017b). Bitcoin was simply above $1000 when year began and achieved 850 percent overall growth. The market capitalization of digital currency has come to $137 billion. Because of this, Bitcoin become more valuable than significant organizations like McDonalds, Mastercard, British American Tobacco or Siemens.
Dong (2017) suggested that Bitcoin great worth depends on the hundreds and thousands of miners optimizing and working in distributed fashion to support its value proposition. The more inflexible and secure block chain and historical transactions directly increase the Bitcoin value in the market. After the 51% attack the total hash rate surpasses the theoretical and practical possibilities. According to Street (2017), government or other commodity does not support Bitcoin and other cryptocurrencies due to the fact of being purely digital tokens. Victims hardly have any option legally or criminally. Bitcoin transaction cannot be reversed, since criminals can rob the owner easily without being tracked. Unlike in the case of savings or checking account making it the largest limiting factor of Bitcoin.
Investopedia (2017) suggested that there are several factors which affect the prices of the cryptocurrencies from trading volumes to media and it has been in the concentration of consideration for media and merchants and the rise of the price of Ethereum is not totally surprising. According to Investopedia (2017), within 24 hours, the price of Ethereum and Bitcoin crashed by 25%. By the end of 2017, the price of the Bitcoin hit target 0f $10,000 and the price of Ethereum hit target of $500 as predicted by a billionaire Michael Novogratz, who holds about 10% of his total assets in cryptocurrencies.

2 Related Work

According to Shah and Zhang (2014), machine learning methods are not widely used for the research purposes in predicting the price of a Bitcoin. There are total 653 papers associated with Bitcoin and among them only 7 papers used machine learning algorithms for prediction of Bitcoin. Thus, this area creates an opportunity to investigate because of having limited literatures. Chen et al. (2013) designed a latent source model for the prediction of Bitcoin price which was implemented by Shah and Zhang (2014). The model got an amazing 89 percent return in 50 days with a Sharpe proportion of 4.1. The investment performance can be examined by sharp ratio whereas balancing for its risk. Therefore, buy and hold strategy was being utilised by the user when period selection is done for examining and got an amazing growth of 33%. This investigation was again at-

tempted independently but results which was received from the model were useless. This study reveals the information that in the cluster there were patterns of 20 which was used by the main author manually likewise shown in a trading books, known as the head and shoulders. Thus, risk has been seen while choosing data to acquire great outcomes.
Georgoula et al. (2015) used support vector machine by applying sentimental analysis whereas also examined the determinants of the Bitcoins price and the researcher got a positive correlation frequency in both the models. Thus, the frequency of network hash rate and Wikipedia views are strongly correlated with the Bitcoins price. In addition, Sentiment has also been utilised as a predictor of Bitcoin in other research.

Matta et al. (2015a) examined the relations between the Bitcoin views on Google trends, tweets and the price of the Bitcoin. And the output which was received from the model was not showing strong correlation, but the relationship lies from mild to moderate between the price of Bitcoin and in both positive tweets on twitter and Google Trends Views. Thus, the author used this as an evidence that they can be used as predictors. But there was only one limitation in this model that was the sample size which has been taken from 60 days. Here, in place of variable, sentiment was used. Therefore, web-based social networking channels, for example, Reddit, or twitter are used to spread the false information. As a result of pump-and-dump scheme, financier who are supposed to take advantage of such misinformation which was spread over the social media helps them to buy at a very low price or sell at an inflated price as per Gu et al. (2006). Liquidity is seen extensively constrained when it comes to Bitcoin trades. Therefore, sentiment is not used for the analysis which was collected from web-based social networking sites because of the higher fluctuation seen in the market. A similar technique was implemented by the same author in performing another research in which trading volume has been predicted instead of Bitcoin price prediction. This shows an output that Bitcoin price is strongly correlated with Google Trends views. The data sample provide periodical data of under one year, earlier the given source of data was considered for implementation. However, due to the lack in duration, the above sample data from google was found to inadequate. Other papers used wavelets to get identical results suggested by Delfin-Vidal and Romero-Meléndez (2016) and Kristoufek (2015). Kristoufek (2015) applied Wavelets Coherence analysis on a Bitcoin price and got a strong relation between the network hash rate, search engine views and mining issues with the Bitcoin price. Due to having temporal dimension, Wavelets are very good to find a strong connection between time series. Therefore, one can see relationship among factors at a particular time.

Moreover, Artificial neural network and Support Vector Machine are used to predict the price of the Bitcoin by examining the Bitcoin Blockchain and the model got an accuracy of 55% and concluded that data which are collected from Blockchain has an unnatural stability where prices are controlled by exchanges whose behaviour lies outside the area of the Blockchain and therefore data are collected related to Coindesk and hash rate and further their issues are recorded through investigation by Greaves and Au (2015)

Madan et al. (2015) predicted the Bitcoin price by utilising the data from Blockchain in which Support Vector machine, Random Forest, Binomial GLM are used as machine learning algorithms in which the model got an accuracy of 98.7% with no cross-validation results and regarding the model was not sure whether it will generalize.
As far as a task related to machine learning is concerned, Bitcoin price prediction can

be viewed as like other prediction related to time series, for example stock and forex prediction which comes under financial sector. The use of ANN is not a new idea. The study shows that back propagation technique is widely implemented in this field as per Rumelhart (1986) which presumed that ANN is appropriate for prediction and modelling based on non-linear time series problem (Tang et al.; 1991; Weigend et al.; 1990). MultiLayer Perceptron was implemented by various bodies of research for predicting the stock price (Yoon and Swales; 1991; White; 1988). In the prediction of the stock price of the IBM MultiLayer Perceptron got a limited value because of the insufficient data in the model which was designed by White (1988). Moreover, the error network and trial parameters search procedure were used for this research and the limitation of this method was seen in terms of finding the global maximum value which was not guaranteed. In the study it found that random search procedure is very useful as compared with grid search by Bergstra and Bengio (2012) because random searches is used to find better models within a specific time. The point which is not recognised by the author is that to find an optimal model. Hence, in the research Bayesian optimisation is found not suitable by Ticknor (2013) whereas grid search is supported over random search. The procedure which was used here is the Bayesian regularisation to find the attributes space for an original result. The strength of each model is examined, and the best model is kept. Python library Hyperopt [3] which is an example of Bayesian optimiser which was also applied. The above approach can minimize over fitting issue, which most often problem explained in literature, majority of times it has been seen that there exists a difference in performance of model in training versus test data.

Wager et al. (2013) suggested that dropout regularisation is another way which is used to reduce the chances of overfitting the model and this technique is accepted as a noising scheme where overfitting is controlled by manipulating the data which are trained. Thus, the model increases its generalisability in terms of better performance on unseen data. Another important element which need to consider while predicting the data related to time series is that certain attributes may consists of essential information.

McNally (2016) collected dataset from Bitcoin Price Index to predict the price of the Bitcoin by using Bayesian optimised RNN, and LSTM. In the research, the author achieved 52% highest classification accuracy by using LSTM and RMSE 8% and implemented ARIMA model against deep learning models. Study shows that ARIMA model perform not better as compared to non-linear deep learning methods.

3 Methodology

In this research, CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is used which is agile and is generally utilized in Data Mining projects, since it makes the procedure simple to shift into different phases like data transformation as well as data modelling of the lifecycle, as suggested by Azevedo and Santos (2008). The datasets used for this research is collected from Kaggle. In this research ten cryptocurrencies datasets are utilized. Each dataset consists of seven features. The procedure for choosing the essential attributes for predicting the Bitcoin price is by seeing the trend of each cryptocurrency when compared with the closing price distribution of the Bitcoin.

[3]Hyperopt: `https://github.com/hyperopt/hyperopt`

3.1 Data preparation

This stage comprises of all the task that is pertinent to finalize the dataset which will be utilized for cleaning the data and for its transformation and this is not counted in a time frame because this stage is used again and again in this research until the data is fully prepared.

3.2 Feature Selection

There are total ten datasets having following features Date, Open, High, Low, Close, Volume, Market Capital. The dataset related to Bitcoin ranges from May 2013 to May 2017 and its closing price distribution is plotted in the figure 1 and observed how the price has changed over time for which generic coding was done in python. There was a spike in early 2014 and the price of the Bitcoin continued to rise. Therefore, new dataset is created by taking the closing price of each cryptocurrency which is utilized in the research goal. Now, the new dataset consists of 12 features such as Index, Delta, Bitcoin, Dash, Ethereum, Litecoin, Monero, Nem, Neo, Ripple, Stratis, Waves.

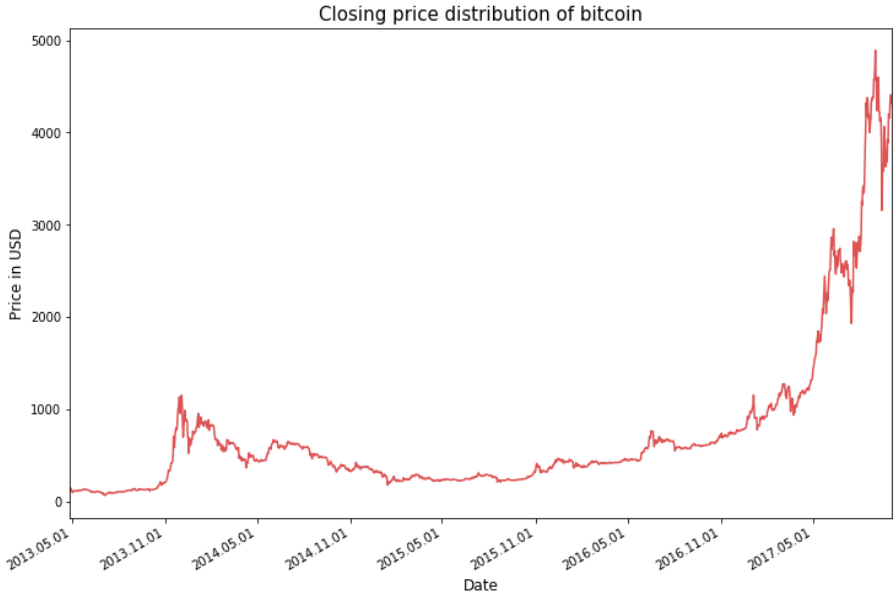

Figure 1: Bitcoin

3.3 Features Evaluation

Evaluation must be done after features selection because the features which are very large in size need more time for training. Moreover, if the optimal number is less as compared to the number of variables then the machine learning algorithm will show less accuracy. There are various techniques used for features evaluation like wrapper based selection and filter based selection. Here, in wrapper based techniques execute search result to the classifier whereas filter based methods clean the features. Thus, Bitcoin candlestick chart as shown in figure 2 reveals the information about the fluctuation in the price of the Bitcoin.

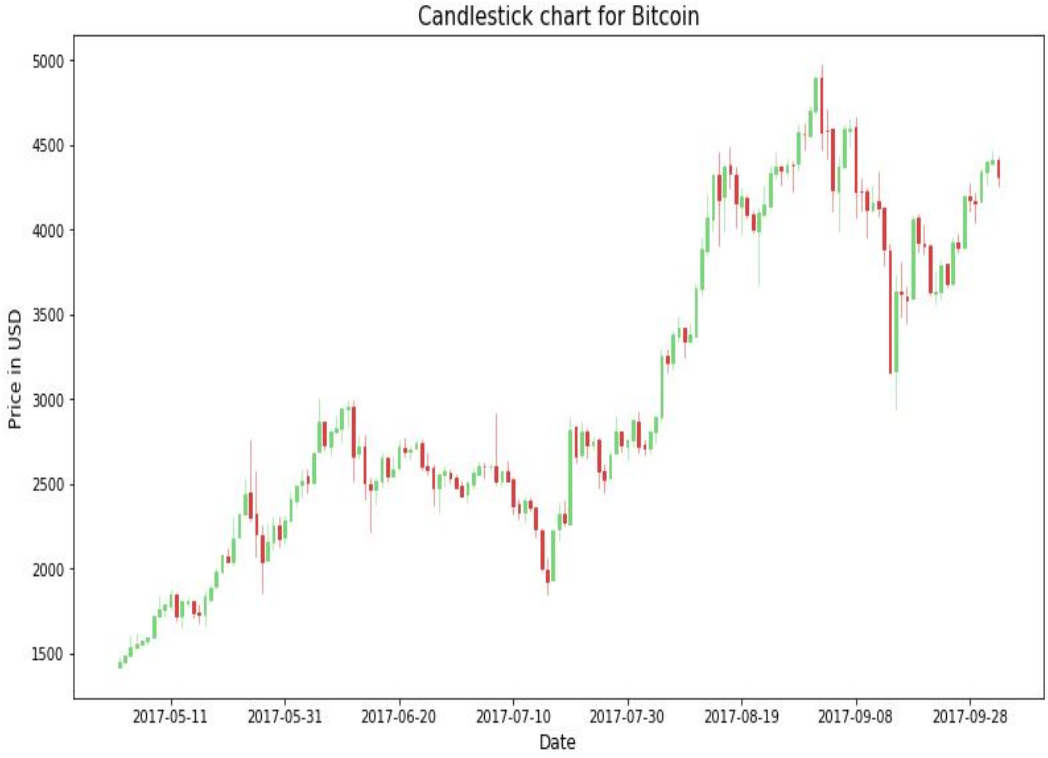

Figure 2: Candle stick chart for Bitcoin price

In the chart green candles indicates the price are going up and red candles indicates that the prices of the Bitcoin going down in US dollars from 11-May-2017 to 28-September-2017. It is clearly vivid that the Bitcoin price was approximately 1700 USD in 11-May-2017 followed by 2500 USD in 31-May-2017. Likewise, the trend was inclining and observed about 3000 USD in 20-June-2017 while 2000 USD in 10-July-2017 respectively. Moving further, the trend of Bitcoin will peak in 8-September-2017 with around 5000 US dollars whereas it will start decreasing and reached up to 3200 USD in the mid of September and it again showing inclining trend.

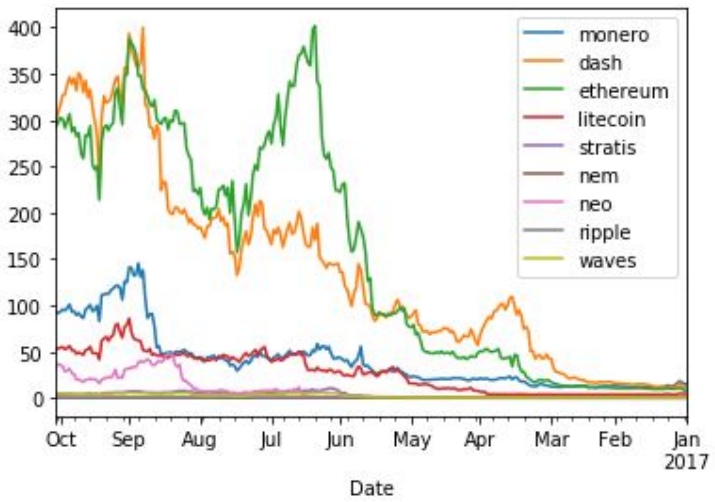

Figure 3: Line graph of nine different cryptocurrencies

All in all, the price distribution of Bitcoin was showing raising trend with coming years. This was used as a proof to further investigate in this area. Now, to see the performance of the other nine cryptocurrencies, line graph is plotted as shown in the figure 3 which provides the information from January 2017 to October 2017 in relation to one another. It is clear from the image that in Jan 2017, all nine cryptocurrency was approximately equal but from March, the prices start varying like dash assessed more with 50 followed by Ethereum with

40. The Monero, Litecoin and Neo has some fluctuation and showing trend between 50 while Waves prices seems rises from zero (0) to little higher with other cryptocurrency. Moving further, the Ethereum price has increased in the mid of June and September by 400 whereas Dash price has also observed same performance with Ethereum in the month of September with 400. Whereas Ethereum price seems all time higher in June as compared with other cryptocurrencies. Eventually, it can be said that Ethereum and Dash showing highest correlation than other cryptocurrencies such as Nem, Neo, Waves.

In this research, features evaluation play an important role in which the author got very strong correlation between Bitcoin and other nine cryptocurrencies as shown in the figure 4 which depicts the relationship of Bitcoin with other different cryptocurrencies. In the correlation map value scale lies from 0.72 to 0.96. And this shows that how much they are correlated with Bitcoin. Moreover, the figure 4 depicts the information that Neo and Ripple are moderately correlated with Bitcoin whereas Monero and Dash are the most correlated, so the author hypothesis is that they will generate the best prediction. Thus, to predict the accuracy and price of a Bitcoin classification and regression method is implemented. In the classification category, random forest, support vector classifier, gradient boosting classifier and neural network classifier are used to find the accuracy and in regression category, recurrent neural network, linear regression and gradient boosting regressor are used for the price prediction.
According to Saxena (2017a), there are many packages hold by Pythons Sklearn library which helps to develop a model for prediction and consists of tools for feature selection, data splitting, pre-processing, tuning and supervised and unsupervised methods.

3.3.1 Random Forest

As suggested by Polamuri (2017b), this algorithm has a distinct advantage as it can be used both for classification as well as for regression type of problems. As this algorithm is based on supervised classification algorithm and it generate forest with the number of trees. As a result, high accuracy will be recorded if there will be a higher number of trees. For example, if we implement decision tree in this place then to predict the accuracy of the price of the Bitcoin whether going up or down. To model the decision tree, training dataset will be used in which if the prices are going up, the decision tree start building the protocols with the sequence of the rise price as nodes and whether prices are going up or not as the leaf nodes. This can be done in decision

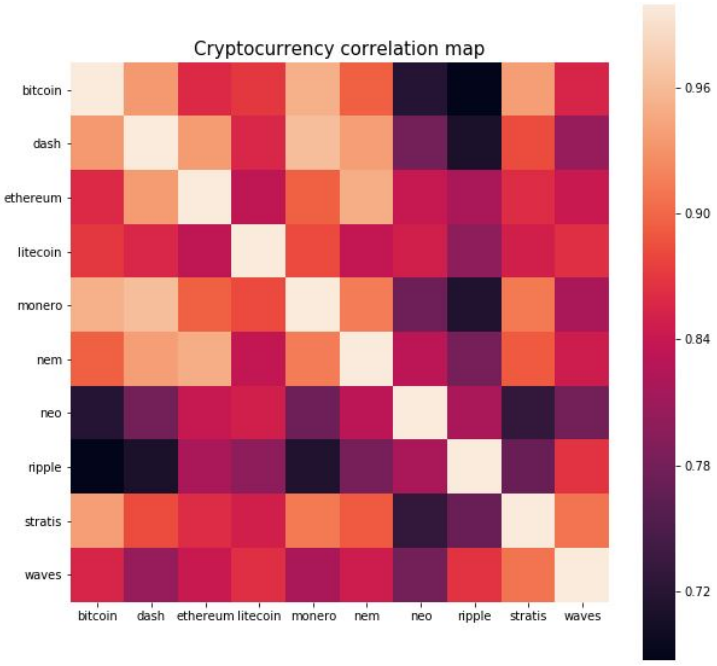

Figure 4: Cryptocurrency Correlation map

tree by using gini index calculations and the information. But in Random forest, the process is randomly selected in terms of detecting the root node and dividing the attribute nodes.

3.3.2 Support Vector Classifier (SVC)

According to Saxena (2017b), Vapnik and Chervonenkis designed Support Vector Machine (SVM). Linear classifier was the only choice to draw hyperplanes. Vapnik, Boser and Guyon in 1992 recommended kernel technique in SVM. It is popularly used as a supervised learning method but if the dataset do not contain any class labels then it will be used as a Support vector clustering which comes under unsupervised learning algorithm. According to Saxena (2017b), SVM works when dataset consist of labels set and features set and SVC builds a model to predict classes for new data and assigns new data points to one of the classes. If there are two classes, then it can be called a Binary SVM classifier. There are two kinds of SVM Classifier. they are Linear and Non-Linear SVM Classifier. Moreover, SVMs are effective when the feature is quite large. It works effectively even if the number of features are greater than the number of samples. Non-linear data can also be classified using customized hyperplanes built by using kernel trick. It is a robust model to solve prediction problem since it maximize margin. The limitation of SVM is the choice of the kernel and it lead to an increase in error percentage. With greater number of samples, it start giving poor performances.

3.3.3 Gradient Boosting Algorithm

According to Brownlee (2016b), the python library provides an implementation of gradient boosting for classification called the GradientBoostingClassifier class and regression called the GradientBoostingRegressor class. In this thesis, both gradient boosting classifier and gradient boosting regressor are used for the goal which are further discussed in the implementation section. The result are shown in the section 5.

3.3.4 Neural Network Classifier

According to Britz (2015), the neural network classifier is used for pattern recognition or classification of data by learning process.It has the ability to take out the meaningful information from the large or complicated dataset which are commonly used to detect the trend and to extract designs which are very large for any other technique to analyze. Further, how this algorithm is implemented for this research is discussed in the implementation section and result are discussed in the section 5.

3.3.5 Recurrent Neural Network

Artificial neural network consists of a class called recurrent neural network where links between neurons form a directed cycle. The internal memory of RNN is used to process input sequence in terms of arbitrary. And this makes them to recognised patterns in the dataset. Recurrent Neural Network consists of long short term memory units trained on annoted or public dataset. The result related to this method are discussed in the section 5.

3.3.6 Linear Regression

Karmanov (2017) suggested that linear regression is widely used for predictive analysis. As the research goal is to ascertain the price of the Bitcoin based on other nine cryptocurrencies trend. Thus, linear model is designed to see the prediction based on the relationship between attributes and predict the line best fit among them and based on the result whether two factors have relationship.

The Python library packages and features parameters regarding different machine learning algorithms which are used for this thesis are discussed in section 4. And Section 5 reveals the results of each methods.

4 Implementation

The author used some of the most popular classification and regression methods in this thesis. According to Saxena (2017a), there are many packages hold by Pythons Sklearn library which helps to develop a model for prediction and consists of tools for feature selection, data splitting, pre-processing, tuning and supervised and unsupervised methods.

Thus, anaconda package is installed for using this libraries. Thus, direct access is provided by Sklearn library to different modules in which different machine learning algorithms like Random Forest, Support Vector Classifier, Gradient Boosting Classifier, linear regression, RNN are implemented to train the models.

4.1 Classification

4.1.1 Random Forest

According to Polamuri (2017a), this technique is known as ensemble classification algorithm which means a group of classifiers. In this case all the default features parameters are used to implement the algorithm for the best result. Thus, random forest classifier is imported from sklearn library and the classifier are fitted to train data and then prediction is assigned to predicted variable. And then created important features and added heading of cryptocurrency to feature which are importance. Thus, the algorithm is utilized to find the accuracy and other performance measures and the result are shown in the figure 6.

4.1.2 Support Vector Classifier

According to Saxena (2017b), Support vector classifier is a supervised learning algorithm which is popular in addressing multi-classification issues. It means that the problem related to multi-classification having two target class to predict. Thus, implementing SVM classifier is done in python and used the research dataset to train a Support Vector Classifier and use the trained Support Vector Machine model to predict the accuracy of the price of the Bitcoin. As dataset is already loaded. And the dataset will use Scikit-learn package.
In this case, support vector classifier is imported from sklearn library and the support vector classifier is assigned to variable classifier. Then classifier is fitted to divide train data and the prediction is assigned to variable svm_pred. Thus, the algorithm is utilized

to find the the accuracy and other performance measures and the result are shown in the figure 6.

4.1.3 Gradient Boosting Classifier

According to Brownlee (2016b), forward stage-wise fashion is used to make a model using GB algorithm. The default features parameters are used for the implementation. Further, gradient boosting classifier is imported from sklearn library and then the gradient boosting classifier is assigned to variable classifier. Then classifier is fitted to divide train data and prediction is assigned to variable gb_pred. Thus, the algorithm is utilized to find the the accuracy and other performance measures and the result are shown in the figure 6.

4.1.4 Neural Network Classifier

According to Britz (2015), the goal is to train a machine learning classifier that predicts the correct class 0 and 1 given the X and Y coordinates. When the data is non-linear, straight line cannot be drawn that divide the two classes. Learning the parameters for the network means finding parameters that minimize the error on the data and the function is known as loss function.
In this case, the training and test data is transformed to a matrix then feed into the input layer with a dimension of 9 and used relu as an activation function to take the features which will create the model. Relu is the default activation function and sigmod is using for binary classification. And then neural network is compiled using adam as it is optimizer and binary crossentropy for binary classification. The neural network is fitted using a batch size of 10 and number of iteration known as epoch to 100. And the output of the neural network is used to predict on the test data that has been transformed to matrix.

4.2 Regression

There are three algorithms used for regression category.

4.2.1 Recurrent Neural Network

For deep learning models, parameters are chosen with the help of three option which are available. They are heuristic search model like genetic method and grid search. In this research, the data is split between train data and test data in order to predict the price of Bitcoin for 5 days. Then Data pre-process stage is carried out to take the training value and reshape it to array. The scaling is done for the training values and changes it to X_train and y_train and reshape it into three-dimension array. Now Keras libraries and packages are imported to implement RNN with LSTM. At first, initialization of RNN is done then added the input layer and the LSTM layer. Here, the activation function sigmod is used and units is assigned by variable 4 and input shape is the shape of the data entering the model. And added the output layer. Then recurrent neural network is compiled using adam as its optimizer is selected and for loss mean squared error is used for regression. After this, the recurrent neural network is fitted using the batch size of 5 and the number of iteration known as epoch to 100. Thus, the algorithm is utilized

to predict the data. The result achieved from this method is explained in the evaluation section.

4.2.2 Linear Regression

According to Brownlee (2016c), this method is 200 years old and expect a straight line between the input and output variable which are denoted as x and y respectively. The following are the step which are used in this thesis to implement the linear regression models for prediction problem.
In this case, linear regression and train and test split are imported from sklearn library and dataset is split into train and test in which test size is taken as 0.10 and randon state as 4284 and the model is fitted and assign prediction to new_pred variable to make the prediction. Thus, the algorithm is utilized to predict the data. The result achieved from this method is shown in the evaluation section.

4.2.3 Gradient Boosting Regressor

It is used to build a model using forward stage-wise fashion which give access to optimization of random loss function. The loss function of the negative gradient is used to fit regression tree in each phase. The default features parameters are used for the better result.
Loss default value is ls which represent least square regression. There are other functions as well like lad, huber, quantile. Least absolute deviation is a full form of lad and regarded highly robust loss function.

In this case, Gradient Boosting Regressor is imported from Sklearn library. And gradient boosting is fitted with parameters in which number of estimators are 2000 and the maximum depth is 1 and R_train, Ry_train is fitted. Thus, R-squared is achieved.

5 Evaluation

The following two section shows the result which achieved by applying different categories of supervised learning method

5.1 Accuracy, Precision, Recall and F1-score

According to Joshi (2016) and Brownlee (2016a), accuracy, precision, recall, and f1-score are the important performance measures which reveals the result about the model which is regarded as best.

Accuracy is defined as the ratio of correctly predicted number to the total number. In simple word, if the model achieves high accuracy it means the model is best.

Moreover, as can be seen from figure 6, support vector classifier from classification category achieved the highest accuracy while the random forest classifier achieved the lowest accuracy. Cryptocurrencies are strongly correlated with Bitcoin as shown in the figure 4.

Moreover, the figure 4 depicts the information that Neo and Ripple are moderately correlated with Bitcoin whereas Monero and Dash are the most correlated, so the author hypothesis is that they will generate the best prediction.

The result of classification is achieved in terms of performance measures in which accuracy play an important role in terms of selecting the best model and other performance metrics are shown in figure 6. Thus, Support Vector Classifier achieved the highest accuracy of 62.31% whereas random forest achieved the lowest accuracy 52.17%. Gradient Boosting Classifier achieved 55.07% and Neural network achieved 60.86%.

Apart from this, the performance metrics consists of precision, recall and f1-score value in which Support Vector Classifier achieved the precision value 0.77. It means support vector classifier have high precision which relates to the low false positive rates. Thus, the precision result achieved from SVM classifier is good and recall value as 0.62 which is also good because the value is more than 0.5 and f1-score is 0.49. In terms of gradient Boosting classifier, precision value is 0.54 and recall value is 0.55 and f1-score is 0.54. And in neural network, precision value is 0.37, recall value is 0.61 and f1-score is 0.46.

Thus, the features which are very important in order to find the accuracy are represented in terms of score as shown in the figure 5 and it is clearly vivid that Nem denoted by 5 in the figure 5 has more score with 0.123745 followed by Litecoin denoted by 3 in the figure 5 with 0.12215 respectively. Similarly, the Neo denoted by 6 shows 0.1191 and Bitcoin denoted by 0 hold 0.111597 score. Moreover, Dash, Ethereum and Monero denoted by 1, 2, and 4 have almost similar score which is 0.110509, 0.1079 and 0.10314 respectively. The two cryptocurrency namely Stratis and Ripple denoted by 8, and 7 in the figure 5 has 0.102544 and 0.0991431 score. In conclusion, Nem has maximum score than other cryptocurrencies.

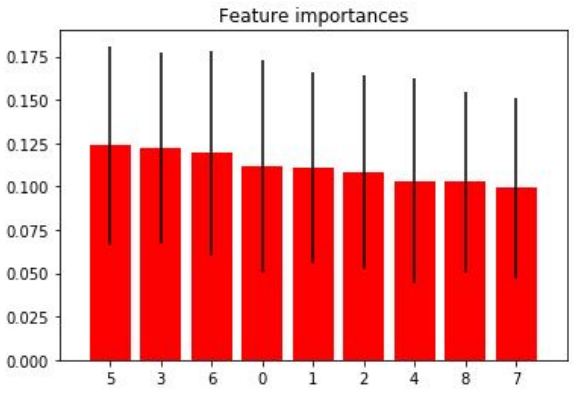

Figure 5: Important Features

	Random Forest Classifier	Support Vector Classifier	Gradient Boosting Classifier	Neural Network Classifier
Precision	0.52	0.77	0.54	0.37
Recall	0.52	0.62	0.55	0.61
f1-score	0.52	0.49	0.54	0.46
Support	69	69	69	69
Accuracy	52.17%	62.31%	55.07%	60.86%

Figure 6: Performance Metrics Table

5.2 Regression

Regression method is evaluated by seeing the Pearson's R-squared value.
According to Ng (2016), R-squared Value is defined as the proportion of variance in the dataset which is observed by the model. In other words, null model error in terms of reduction. R-squared value lies in between 0 and 1. If the model achieve higher value then it means that more variance is explained by the model. Also, it is very hard to say that the value which are achieve from the model are best because the threshold for a best R-squared value rely on the domain. Therefore, in this thesis it is utilized as a tool for comparing different models.

The figure 7 shows the features which are most important in terms of predicting the price of the Bitcoin. It is clearly vivid that Nem denoted by 5 in the figure has more score with 0.174 followed by Neo denoted by 6 with 0.143 respectively. Similarly, the Litecoin denoted by 3 shows 0.117 and Waves denoted by 9 and Dash denoted by 1 are almost same score which is 0.098 and 0.097 respectively. Moreover, Ripple, Delta and Stratis denoted by 7, 0, 8 have almost similar score which is 0.084, 0.0815 and 0.0765 respectively. The two cryptocurrency namely Ethereum and Monero has 0.0705 and 0.0585 score.
In conclusion, Nem has maximum score than other cryptocurrencies.
Thus, in the regression category, three algorithms were implemented and Gradient Boosting Regressor got the highest of R-squared value as compared with other two algorithms. The R-squared value of Gradient Boosting Regressor is 0.99 and linear regression got R-squared value 0.98 and RNN achieved 0.96 R-squared value.

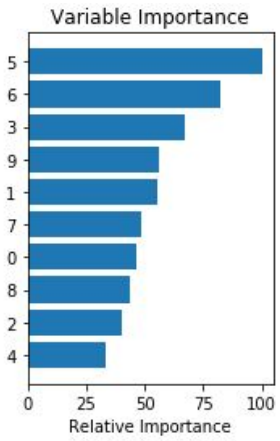

Figure 7: Important Features

6 Conclusion and Future Work

6.1 Conclusion

Machine learning algorithms such as supervised learning methods are very useful in solving real world problem. In this research, the main goal is to ascertain how the direction and accuracy of price of Bitcoin can be predicted by using data mining method. The dataset related to ten cryptocurrencies are selected to form a new dataset which is utilized to achieve the goal. The nine cryptocurrencies show strong correlation with Bitcoin as shown in the figure 4.

Thus, supervised learning method is used in which random forest classifier, support vector classifier, gradient boosting classifier, and neural network classifier are implemented under classification to find out the best accuracy from the model. Random Forest is chosen because it utilized averaging to enhance the accuracy and check over-fitting as it fits several decision tree classifier on different subsamples. Thus, random forest achieved accuracy of 52.17%. Support Vector classifier works better even when the attributes are more than the sample. Thus, Support vector classifier is very robust to predict as it

expands margin. One limitation of Support Vector Machine is the choice of kernel and with more samples it does not give good result. Thus, Support Vector Classifier achieved 62.31%. Gradient Boosting Classifier is an effective and accurate technique used for classification and as well as for regression problems. Gradient Boosting Classifier is used in this thesis because of natural handling of different cryptocurrencies data. Moreover, Gradient Boosting Regressions trees are very good in prediction. One limitation of this algorithm is scalability. Thus, Gradient Boosting classifier achieved 55.07% accuracy. Neural network classifier uses activation function like relu for the hidden layer, changes the output from the inputs of the layer. One limitation of this algorithm is hard to tune to learn well and as a result, tough to run. Here, neural network achieved 60.86% accuracy.

Moreover, the author in this research has implemented Support Vector Classifier, Random Forest, Gradient Boosting Classifier, and Neural Network Classifier and recorded accuracy of 62.31% whereas Madan et al. (2015) collected the data from Blockchain and implemented Support Vector Machine, Random Forest, Binomial GLM and recorded accuracy of 98.7% which is much more higher than the accuracy which is achieved here in this thesis. Probably, it is because of the dataset as the author used ten cryptocurrencies dataset. On the flipside, the result of this thesis is better than Greaves and Au (2015) work which shows the accuracy of 55% as discussed in the section 2.

In addition, In this thesis, random forest achieved 0.52 precision value and the value of recall is 0.52 and f1-score is 0.52. Subsequently, f1-score considers both false negatives and false positives. In addition, Support Vector Classifier achieved the precision value 0.77 and recall value as 0.62 which is good because the value is more than 0.5 and f1-score is 0.49. In terms of gradient Boosting classifier, precision value is 0.55 and recall value is 0.57 and f1-score is 0.56. And in neural network, precision value is 0.37, recall value is 0.61 and f1-score is 0.46.

As a result, Support Vector Classifier achieved the highest accuracy as compared with other algorithms and this create an opportunity for further investigation in this area by utilising different algorithms with same ideas in classification problem.

Whereas in regression methods algorithms like recurrent neural network, linear regression, and gradient boosting regressor are used. Here, using RNN algorithm, the price of a Bitcoin for five days is predicted. Moreover,RNN achieved R-squared value 0.96. The variance result which is achieved very high from regression methods make it not easy to consider it as a good model. In linear regression, important features are discussed in the regression section. Linear regressor achieved R-squared value 0.98. Gradient Boosting regressor got highest R-squared value as 0.99. Since the dataset is small that is why the value of R-squared is higher. In addition, this open the door for further investigation by using different algorithms based on regression problem to predict the price of the Bitcoin.

6.2 Future Work

There are many ways to improve the existing methodology by applying different classification methods of Deep learning. The data related to Bitcoin play an important role. Using data from different exchanges, like Blockchain, and average of closing price collected from different exchanges can provide a different insight in the model used. Unsupervised

methods can provide a different perspective about the data, and open new avenues to explore in this area. In addition, the author from his point of view recommended that one should gain knowledge regarding the technology that is behind Bitcoin for further investigation.

References

Azevedo, A. I. R. L. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview, *IADS-DM* .

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization, *Journal of Machine Learning Research* **13**(Feb): 281–305.

Brière, M., Oosterlinck, K. and Szafarz, A. (2015). Virtual currency, tangible return: Portfolio diversification with bitcoin, *Journal of Asset Management* **16**(6): 365–373.

Britz, D. (2015). Implementing a Neural Network from Scratch in Python An Introduction, `http://www.wildml.com/2015/09/implementing-a-neural-network-from-scratch/`. [Online, accessed 3-September-2015].

Brownlee, J. (2016a). Classification Accuracy is Not Enough: More Performance Measures You Can Use, `https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/`. Online, accessed 21-March-2014.

Brownlee, J. (2016b). How to Configure the Gradient Boosting Algorithm, `https://machinelearningmastery.com/configure-gradient-boosting-algorithm/`. [Online, accessed 12-September-2016].

Brownlee, J. (2016c). How To Implement Simple Linear Regression From Scratch With Python, `https://machinelearningmastery.com/implement-simple-linear-regression-scratch-python/`. [Online, accessed 26-October-2016].

Chen, G. H., Nikolov, S. and Shah, D. (2013). A latent source model for nonparametric time series classification, *Advances in Neural Information Processing Systems*, pp. 1088–1096.

CNBC (2017). Bitcoin hits new record high, breaking $8,000 for the first time, `http://cnb.cx/2AeG9Yy/`. [Online, accessed 20-November-2017].

Delfin-Vidal, R. and Romero-Meléndez, G. (2016). The fractal nature of bitcoin: Evidence from wavelet power spectra, *Trends in Mathematical Economics*, Springer, pp. 73–98.

Dong, R. (2017). Cash is King, Is Bitcoin Cash king?, `https://www.linkedin.com/pulse/bitcoin-vs-cash-rui-dong/`. [Online, accessed 20-November-2017].

Economist, T. (2013). Bitcoin under pressure, `https://www.economist.com/news/technology-quarterly/21590766-virtual-currency-it-mathematically-elegant-increasing` [Online, accessed 30-November-2013].

Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D. N. and Giaglis, G. M. (2015). Using time-series and sentiment analysis to detect the determinants of bitcoin prices.

Greaves, A. and Au, B. (2015). Using the bitcoin transaction graph to predict the price of bitcoin.

Grocer, S. (2013). The Wall Street Journal. 2013. Beware the Risks of the Bitcoin: Winklevii Outline the Downside, `https://blogs.wsj.com/moneybeat/2013/07/02/beware-the-risks-of-the-bitcoin-winklevii-outline-the-downside/`. [Online, accessed 2-July-2013].

Gu, B., Konana, P., Liu, A., Rajagopalan, B. and Ghosh, J. (2006). Identifying information in stock message boards and its implications for stock market efficiency, *Workshop on Information Systems and Economics, Los Angeles, CA*.

Investopedia (2017). Ethereum Reaches Record High And Bitcoin Price Follows Suit, `http://www.investopedia.com/news/bitcoin-price-sets-new-record-and-ethereum-follows-suit?utm_source=twitter&utm_medium=social&utm_campaign=shareurlbuttons/`. [Online, accessed 25-November-2017].

Joshi, R. (2016). Accuracy, Precision, Recall F1 Score: Interpretation of Performance Measures, `http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/`. Online, accessed 9-September-2016.

Kaastra, I. and Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series, *Neurocomputing* **10**(3): 215–236.

Kaminski, J. (2014). Nowcasting the bitcoin market with twitter signals, *arXiv preprint arXiv:1406.7577* .

Karmanov, F. (2017). Linear Regression in Python: A Tutorial - Springboard Blog , `https://www.springboard.com/blog/linear-regression-in-python-a-tutorial/`. [Online, accessed 8-August-2017].

Kristoufek, L. (2015). What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis, *PloS one* **10**(4): e0123923.

Madan, I., Saluja, S. and Zhao, A. (2015). Automated bitcoin trading via machine learning algorithms.

Malkiel, B. G. and Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work, *The journal of Finance* **25**(2): 383–417.

Matta, M., Lunesu, I. and Marchesi, M. (2015a). Bitcoin spread prediction using social and web search media., *UMAP Workshops*.

Matta, M., Lunesu, I. and Marchesi, M. (2015b). The predictor impact of web search media on bitcoin trading volumes, *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on*, Vol. 1, IEEE, pp. 620–626.

McNally, S. (2016). *Predicting the price of Bitcoin using Machine Learning*, PhD thesis, Dublin, National College of Ireland.

Ng, R. (2016). Evaluating a Linear Regression Model — Machine Learning, Data Science, Algorithmic Trading Public Policy, http://disq.us/t/2byn8ws/.

Nick Whigham, n. (2017). The $155K heist in plain sight and the problem with Bitcoin, http://www.news.com.au/technology/online/hacking/restaurantgoer-has-bitcoins-stolen-over-unsecured-public-wireless-network/news-story/a82d75eb85763ee3d38678b430df3bf0/. [Online, accessed 23-November-2017].

Polamuri, S. (2017a). Building Random Forest Classifier with Python Scikit learn , http://dataaspirant.com/2017/06/26/random-forest-classifier-python-scikit-learn/. [Online, accessed 26-June-2017].

Polamuri, S. (2017b). How the Random Forest algorithm works in Machine Learning, http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/. [Online, accessed 22-May-2017].

Ron, D. and Shamir, A. (2013). Quantitative analysis of the full bitcoin transaction graph, *International Conference on Financial Cryptography and Data Security*, Springer, pp. 6–24.

Rumelhart, D. E. (1986). Learning internal representation by back propagation., *Parallel distributed processing: exploration in the microstructure of cognition* **1**.

Saxena, R. (2017a). Building Decision Tree Algorithm in Python with SCIKIT LEARN, https://dataaspirant.com/2017/02/01/decision-tree-algorithm-python-with-scikit-learn/. [Online, accessed 01-February-2017].

Saxena, R. (2017b). Svm classifier, Introduction to support vector machine algorithm, https://dataaspirant.com/2017/01/13/support-vector-machine-algorithm/. [Online, accessed 13-January-2017].

Shah, D. and Zhang, K. (2014). Bayesian regression and bitcoin, *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, IEEE, pp. 409–414.

Street, T. (2017). Cryptocurrencies Are at Greater Risk of Being Hacked, https://www.thestreet.com/story/14397153/1/cryptocurrencies-are-at-greater-risk-of-being-hacked.html/. [Online, accessed 21-November-2017].

Tang, Z., de Almeida, C. and Fishwick, P. A. (1991). Time series forecasting using neural networks vs. box-jenkins methodology, *Simulation* **57**(5): 303–310.

Ticknor, J. L. (2013). A bayesian regularized artificial neural network for stock market forecasting, *Expert Systems with Applications* **40**(14): 5501–5506.

Wager, S., Wang, S. and Liang, P. S. (2013). Dropout training as adaptive regularization, *Advances in neural information processing systems*, pp. 351–359.

Weigend, A. S., Huberman, B. A. and Rumelhart, D. E. (1990). Predicting the future: A connectionist approach, *International journal of neural systems* **1**(03): 193–209.

White, H. (1988). Economic prediction using neural networks: The case of ibm daily stock returns.

Yoon, Y. and Swales, G. (1991). Predicting stock price performance: A neural network approach, *System Sciences, 1991. Proceedings of the Twenty-Fourth Annual Hawaii International Conference on*, Vol. 4, IEEE, pp. 156–162.