ABSTRACT

The purpose of this study is to find out with what accuracy the direction of the price of Bitcoin can be predicted using machine learning methods. This is basically a time series prediction problem. While much research exists surrounding the use of different machine learning.

Techniques for time series prediction, research in this area relating specifically to Bitcoin is lacking. In addition, Bitcoin as a currency is in a transient stage and as a result is considerably more volatile than other currencies such as the USD. Interestingly, it is the top performing currency four out of the last five years. Thus, its prediction offers great potential and this provides motivation for research in the area. As evidenced by an analysis of the existing literature, running machine learning algorithms on a GPU as opposed to a CPU can offer significant performance improvements. This is explored by benchmarking the training of the RNN and LSTM network using both the GPU and CPU. This provides a solution to the sub research topic.

Finally, in analysing the chosen dependent variables, each variables importance is assessed using a random forest algorithm. In addition, the ability to predict the direction of the price of an asset such as Bitcoin offers the opportunity for profit to be made by trading the asset.

Keywords: Bitcoin Prediction, Time complexity, Machine-learning, Database architecture, RNN, LSTM.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 DOMAIN SPECIFIC INTRODUCTION

Time series prediction is not a new phenomenon. Prediction of most financial markets such as the stock market has been researched at large scale. Bitcoin presents an interesting parallel to this as it is a time series prediction problem in a market still in its beggining stage. As a result, there is high volatility in the market and this provides an opportunity in terms of prediction. In addition, Bitcoin is the leading cryptocurrency in the world with adoption growing consistently over time. Due to the open nature of Bitcoin it also poses another difficulty as opposed to traditional financial markets. It operates on a decentralised, peer-to-peer and trustless system in which all transactions are posted to an open ledger called the Blockchain. This type of transparency is not seen in other financial markets. Traditional time series prediction methods such as Holt- Winters exponential smoothing models rely on linear assumptions and require data that can be broken down into trend, seasonal and noise to be effective. This type of methodology is more suitable for a task such as predicting sales where seasonal effects are present. Due to the lack of seasonality in the Bitcoin market and it's high volatility, these methods are not very effective for this task. Given the complexity of the task, deep learning makes for an interesting technological solution based on its performance in similar areas. Tasks such as natural language processing which are also sequential in nature and have shown promising results. This type of task uses data of a sequential nature and as a result is similar to a price prediction task. The recurrent neural network (RNN) and the long short term memory (LSTM) flavour of artificial neural networks are favoured over the traditional multilayer perceptron (MLP) due to the temporal nature of the more advanced algorithms.

The aim of this research is to ascertain with what accuracy can the price of Bitcoin be predicted using machine learning.

the research variables. A brief overview of Bitcoin, machine learning and time series analysis concludes section one. Section two examines related work in the area of both Bitcoin price prediction and other financial time series prediction. Literature on using machine learning to predict Bitcoin price is limited.

Out of approximately 653 papers published on Bitcoin only 7 have related to machine learning for prediction. As a result, literature relating to other financial time series prediction using deep learning is also assessed as these tasks can be considered analogous.

1.2 PROBLEM DEFINITION

The popularity of cryptocurrencies has skyrocketed in 2017 due to several consecutive months of super exponential growth of their market capitalization, which peaked at more than $800 billions in Jan. 2018. Today, there are more than1,500 actively traded crypto currencies. Between2.9and 5.8millions of private as well as institutional investors are in the different transaction networks, according to a recent survey , and access to the market has become easier over time. Major cryptocurrencies can be bought using fiat currency in a number of online exchanges and then be used in their turn to buy less popular cryptocurrencies. The volume of daily exchanges is currently superior to $15 billions. Since 2017, over 170 hedge funds specialised in cryptocurrencies have emerged and Bitcoin futures have been launched to address institutional demand for trading and hedging Bitcoin could be effective also in predicting crypto currency prices. However, the application of machine learning algorithms to the cryptocurrency market has been limited so far to the analysis of Bitcoin prices, using random forests, Bayesian neural network , long short-term memory neural network, and other algorithms .The studies were able to anticipate, to different degrees, the price fluctuations of Bitcoin, and revealed that best results were achieved by neural network based algorithms. Deep reinforcement learning was showed to beat the uniform buy and hold strategy in predicting the prices of 12 cryptocurrencies overone-year period.

The Bitcoin's value varies just like any other stock . There are many algorithms used on stock market data for price forecast. However, the parameters affecting Bitcoin are different. Therefore it is necessary to foretelling the value of Bitcoin so that correct investment decisions can be made. The price of Bitcoin does not depend on the business events or intervening government authorities, unlike the stock market. Thus, to forecast the value we feel it is necessary to leverage machine learning technology to predict the price of Bitcoin. So the project aim is to predict the price of bitcoin and help investor's make better investments. This research is concerned with predicting the price of Bitcoin using machine learning. The goal is to ascertain with what accuracy can the direction of Bitcoin price in USD can be predicted.

The price data is sourced from the Bitcoin Price index. The task is achieved with varying degrees of success through the implementation of a Bayesian optimized recurrent neural network (RNN) and Long Short-Term Memory (LSTM) network.

1.3 PROJECT PURPOSE

The purpose of this study is to find out with what accuracy the direction of the price of Bitcoin can be predicted using machine learning methods. This is basically a time series prediction problem. While much research exists surrounding the use of different machine learning techniques for time series prediction, research in this area relating specifically to Bitcoin is lacking. In addition, Bitcoin as a currency is in a transient stage and as a result is considerably more volatile than other currencies such as the USD. Interestingly, it is the top performing currency four out of the last five years1. Thus, its prediction offers great potential and this provides motivation for research in the area. As evidenced by an analysis of the existing literature, running machine learning algorithms on a GPU as opposed to a CPU can offer significant performance improvements. This is explored by benchmarking the training of the RNN and LSTM network using both the GPU and CPU. This provides a solution to the sub research topic.

Finally, in analysing the chosen dependent variables, each variables importance is assessed using a random forest algorithm. In addition, the ability to predict the direction of the price of an asset such as Bitcoin offers the opportunity for profit to be made by trading the asset. To implement a full trading strategy based on the results of the models is worthy of a dissertation in itself and as a result this paper will focus solely on the accuracy in which direction the price can be predicted. In basic terms, the model would initiate a short position if the price was predicted to go up and a long position if the price was predicted to go down. Several Bitcoin exchanges offer margin trading accounts to facilitate this too. The profitability of this strategy would be based not only on the accuracy of the model, but also on the size of the positions taken. This is outside the scope of this research but could be addressed in future work.

While we will try to build a predictive model for the Bitcoin price value calculator,we are aware in advance that price may differ greatly because of internal and external factors to Bitcoin. By internal factors we are presuming factors inside the Bitcoin security.

By external we are referring to agents which influence indirectly the price of Bitcoin(exchange closures, replacing cryptocurrencies, speculation markets, the fact that as its believed widely over 80% of Bitcoins in circulation is concentrated in a limited number of investors etc.) Anyway, we shall compare our results to other models built for cryptocurrency prediction. Let's not forget that in the first month of 2018 there were models which predicted that Bitcoin would surpass the 100,000.00 USD per Bitcoin till the end of the year, while we are barely reaching the 7,000.00 USD value just 2 months before the end of the year.

1.4 PROJECT FEATURES

The main feature of this system is to propose a general and effective approach to predict the bitcoin price using data mining techniques. The main goal of the proposed system is to analyze and study the hidden patterns and relationships between the data present in the bitcoin dataset. The solution to the bitcoin analysis problem can provide extremely useful information to prevent investors from loosing money which is being invested on bitcoin. Most of the existing work solves these problems separately by different models. so dealing with this becomes more important. The analysis and prediction plays an important role in the problem definition.

The constant increase in bitcoin usage has become an extremely serious problem, with the development of technology and hi-tech tools having a significantly greater impact on the bitcoin price. The large amounts of information also poses a challenge to analyze such data and identify similarities or relations between the data. Also there is a challenge of inconsistency that can occur in the data due to incompleteness in the dataset. Therefore, there is an urging need of proper techniques to analyze large volumes of data to get some useful results out of it. So the main aim of this project is to propose a general and effective approach to predict the bitcoin price using data mining techniques.

The main features of the proposed system are:

- More efficient.
- Better bitcoin price monitoring systems.
- Reduces the costs of storage, maintenance and personnel.
- It reduces the time complexity of the system.
- System that has a simpler architecture to understand.
- Processing of large amount of data becomes easier.

1.5 MODULES DESCRIPTION

1.5.1 DATA GATHERING

The first step in this project or in any data mining project is the collection of data to be studied or examined to find the hidden relationships between the data members. The important concern while choosing a dataset is that the data which we are gathering should be relevant to the problem statement and it must be large enough so that the inference derived from the data is useful to extract some important patterns between the data such that they can be used to predict the future events or can be studied for further analysis. The result of the process of gathering and creating a collection of data results into what we call as a Dataset. The dataset contains large volume of data that can be analyzed to get some knowledge from the databases. This is an important step in the process because choosing the inappropriate dataset can lead us to incorrect results.

1.5.2 DATA PREPROCESSING

The primary data collected from the internet resources remains in the raw form of statements, digits and qualitative terms. The raw data contains error, omissions and inconsistencies. It requires corrections after careful scrutinizing the completed questionnaires. The following steps are involved in the processing of primary data. A huge volume of raw data collected through field survey needs to be grouped for similar details of individual responses.

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Therefore, certain steps are executed to convert the data into a small clean data set. This technique is performed before the execution of Iterative Analysis. The set of steps is known as data preprocessing.

The process comprises:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

Data Preprocessing is necessary because of the presence of unformatted real world data. Mostly real world data is composed of:

- **Inaccurate data (missing data) -** There are many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with biometrics and much more.
- **The presence of noisy data (erroneous data and outliers) -** The reasons for the existence of noisy data could be a technological problem of gadget that gathers data, a human mistake during data entry and much more.
- **Inconsistent data -** The presence of inconsistencies are due to the reasons such that existence of duplication within data, human data entry, containing mistakes in codes or names, i.e., violation of data constraints and much more.

1.5.3 CLASSIFICATION

This technique is used to divide various data into different classes. This process is also similar to clustering. It segments data records into various segments which are known as classes. Unlike clustering, here we have knowledge of different clusters. Ex: Outlook email, they have an algorithm to categorize an email as legitimate or spam.

CHAPTER 2

LITERATURE SURVEY

2.1 DATA MINING

Literature survey is that the most vital step in code development method. Before developing the tool it's necessary to see the time issue, economy and company strength. Once these things are satisfied, then next steps is to determine which operating system and language can be used for developing the tool Once the programmers begin building the tool the programmers would like heap of external support. This support is obtained from senior programmers, from book or from websites Before building the system the on top of thought area unit taken under consideration for developing the projected system.. We have to analyze the Data mining Outline Survey:

2.1.1 Data Mining Survey

Data mining is a data analysis technique which allows us to study and identify different patterns and relationships between the data. In other words, data mining is a technique which can be employed to extract information from large and extensive datasets and convert the information into a prominent structure so that it can be used further for gaining inference and knowledge on the data so as to prevent the crimes.

Data mining contains techniques for analysis which involve various domains. For instance, some of the domains involved in data mining are Statistics, Machine Learning and Database systems. Data mining is additionally spoken as “Knowledge discovery in databases (KDD)”.

The real task of data mining systems is the semi-automatic or automatic analysis of large volumes of data to extract previously unknown relationships such as groups of data members(clustering analysis),unusual records(outlier or anomaly detection),and dependencies. Normally, this includes database techniques like spatial indices.

2.1.3 Techniques in Data Mining:

1. Classification: This technique is used to divide various data into different classes. This process is also similar to clustering. It segments data records into various segments which are known as classes. Unlike clustering, here we have knowledge of different clusters. Ex: Outlook email, they have an algorithm to categorize an email as legitimate or spam.

2. Association: This technique is used to discover hidden patterns in the data and also for identifying interesting relations between the variables in a database. Ex: It is used in retail industry.

3. Prediction: This technique is used only for particular uses. It is used extract relationships between independent and dependent variables in the dataset. Ex: We use this technique to predict profit obtained from sales for the future.

4. Clustering: A cluster is referred to as a group of data objects. The data objects that are similar in properties are kept in the same cluster. In other words we can tell that clustering is a process of discovering groups or clusters.

5. Here we do not have prior knowledge of the clusters. Ex: It can be used in consumer profiling.

6. Sequential Patterns: This is an essential aspect of data mining techniques its main aim is to discover similar patterns in the dataset. Ex: E-commerce websites suggestions are based on what we have bought previously.

7. Decision Trees: This technique is a vital role in data mining because it is easier to understand for the users. The decision tree begins with a root which is a simple question. As they can have multiple answers we get our nodes of the decision tree also the questions in the root node might lead to another set of questions. Thus, the nodes keep adding in the decision tree. At last, we are allowed to make a final decision on it. Apart from these techniques there are certain other techniques which allow us to remove noisy data and also clean the dataset. This helps us to get accurate analysis and prediction results.