Analysis and Prediction with Bitcoin-Blochain Dataset

Charchit Dhawan
charchitdhawan@gmail.com

1 Introduction

Bitcoin is an electronic crypto-currency created in 2008 by Satoshi Nakamoto (pseudonym). At the time the original bitcoin client was written, the idea of a purely peer-to-peer (P2P) digital currency which did not require a trusted-thirdparty to confirm transactions / prevent double spending was unique. In the bitcoin network, all transactions are public, effectively rendering double-spending impossible.Under the assumption that the majority of the network is honest, the criminal would have to have more computational power than the majority of the network in order to falsify the transaction history.
Blockchain: The bitcoin blockchain is a public ledger that records bitcoin transactions.[80] It is implemented as a chain of blocks, each block containing a hash of the previous block up to the genesis block[d] of the chain. A network of communicating nodes running bitcoin software maintains the blockchain.

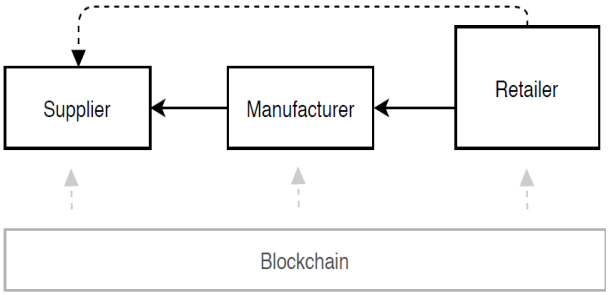
1.1 BlockChain Technology

The blockchain is an undeniably ingenious invention âĂŞ the brainchild of a person or group of people known by the pseudonym, Satoshi Nakamoto. But since then, it has evolved into something greater, and the main question every single person is asking is: What is Blockchain?

By allowing digital information to be distributed but not copied,blockchain technology created the backbone of a new type of internet. Originally devised for the digital currency, Bitcoin, (Buy Bitcoin) the tech community has now found other potential uses for the technology, **Blockchain enhances trust across a business network.**

Blockchain builts trust through the following five attributes:

- **Distributed:** The distributed ledger is shared and updated with every incoming transaction among the nodes connected to the Blockchain. All this is done in real-time as there is no central server controlling the data.
- **Secure:** There is no unauthorized access to Blockchain made possible through Permissions and Cryptography.
- **Transparent:** Because every node or participant in Blockchain has a copy of the Blockchain data, they have access to all transaction data. They themselves can verify the identities without the need for mediators.
- **Consensus-based:** All relevant network participants must agree that a transaction is valid. This is achieved through the use of consensus algorithms.
- **Flexible:** Smart Contracts which are executed based on certain conditions can be written into the platform. Blockchain Network can evolve in pace with business processes.



1.2 BITCOIN-BLOCKCHAIN DATASET

Cryptocurrencies have captured the imagination of technologists, financiers, and economists. Perhaps even more intriguing are the long-term, diverse applications of the blockchain. By increasing transparency of cryptocurrency systems, the contained data becomes more accessible and useful. The Bitcoin blockchain data are now available for exploration with BigQuery. All historical data are in the **project name:bigquery-public-data** and **dataset name:bitcoin-blockchain** are available on Google cloud, which is of 871 GB data and updates in every 10 minutes.

link of Dataset: `https://www.kaggle.com/bigquery/bitcoin-blockchain`.

a)About BigQuery: Storing and querying massive datasets can be time consuming and expensive without the right hardware and infrastructure. BigQuery is an enterprise data

warehouse that solves this problem by enabling super-fast SQL queries using the processing power of Google's infrastructure. Simply move your data into BigQuery and let us handle the hard work. You can control access to both the project and your data based on your business needs, such as giving others the ability to view or query your data.

b) How to Access Bigdata: These type of datasets can be access in BigQuery by using the GCP Console or the classic web UI, by using a command-line tool, or by making calls to the BigQuery REST API using a variety of client libraries such as Java, .NET, or Python. There are also a variety of third-party tools that you can use to interact with BigQuery, such as visualizing the data or loading the data.

BigQuery is fully-managed. To get started, you don't need to deploy any resources, such as disks and virtual machines. Get started now by running a web query or using the command-line tool.

2 Task: Machine Learning in Blockchain Technology

2.1 Problem Statement

Machine Learning processing is a logical continuity to the use of the blockchain database. You will collect data, store it on a decentralized system, and machine learning algorithms will process.

There is an immense probability that blockchain and machine learning could be combined, as we can use machine learning and blockchain in various fields and areas but one example can be to enhance security. as machine learning requires a lot of data and blockchain is like a ledger of data, the amalgamation of the two technologies can bring in immense opportunities in the growing technological era.Both blockchains and machine learning are new technologies that have emerged in the last decade that have far-reaching consequences for all spheres of human activity.The merger will be a game changer for the self-driving research as it can help create a marketplace and also the finance and insurance industries have a lot to gain as together, they can be used to design tools to identify and prevent fraud. Hence the amalgamation of Blockchain and machine learning is evolving and, in the future, could be the next big thing; we just have to wait and see how technologies unravel their mysteries.

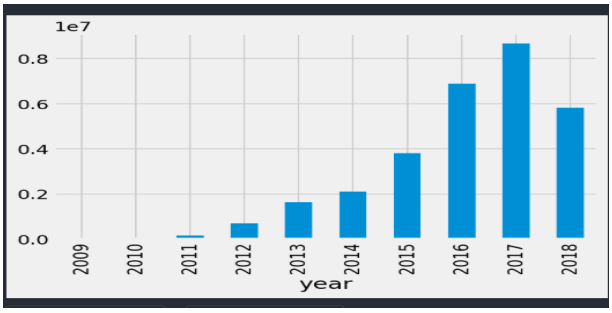
2.2 Proposed Solution

Proposed Solutions are described as follows:

- Analyse the Data-set and visualizing it
- Using different machine learning approaches for prediction using bitcoin-blockchain big-query data-set
- Comparison between different ML-approaches

2.2.1 Analyse the Bitcoin-Blockchain Data-set and Visualizing

Extracting the Data-set from the big-queries data queries and after that extraction of transaction table is been done,using that table count of transactions of bitcoins with the corresponding month and year as an attributes is shown. Using that data-set we can forecast the transactions during period of one decade as follows:



2.2.2 Apply Machine Learning to predict Bitcoin-miner

Using Random forest: Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest".
Random Forest as a Classifier: binary classification problem and we will use a random forest classifier to solve this problem. Steps followed to solve this problem will be similar to the steps performed for regression.

- **Here Accuracy achieved is 90.65 percent**

Using KNN: The first step is to import the KNeighborsClassifier class from the sklearn.neighbors library. In the second line, this class is initialized with one parameter, i.e. n-neigbours. This is basically the value for the K. There is no ideal value for K and it is selected after testing and evaluation, however to start out, 5 seems to be the most commonly used value for KNN algorithm. For evaluating an algorithm, confusion matrix, precision, recall and f1 score are the most commonly used metrics. The confusion-matrix and classification-report methods of the sklearn.metrics can be used to calculate these metrics.

- **Here Accuracy achieved is 98.4 percent**

Using LSTM-Model: Humans donâĂŹt start their thinking from scratch every second. As you read this essay, you understand each word based on your understanding of previous words.

You donâĂŹt throw everything away and start thinking from scratch again. Your thoughts have persistence.

Traditional neural networks canâĂŹt do this, and it seems like a major shortcoming. For example, imagine you want to classify what kind of event is happening at every point in a movie. ItâĂŹs unclear how a traditional neural network could use its reasoning about previous events in the film to inform later ones.

Recurrent neural networks address this issue. They are networks with loops in them, allowing information to persist.

LSTM-Network: Long Short Term Memory networks âĂŞ usually just called âĂIJLSTMsâĂİ âĂŞ are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter Schmidhuber (1997), and were refined and popularized by many people in following work.1 They work tremendously well on a large variety of problems, and are now widely used.

- **Here Accuracy achieved is 88.7 percent**

2.2.3 Comparisions

Fig: Accuracy Comparision between different Machine Learning algorithms

LSTM Model	KNN-Classification	Random Forest Classification
Accuracy: 88.625%	Accuracy: 97.7%	Accuracy: 99.05000000000001%

3 Future Scope and work

Unsupervised learning techniques revealed anomalies in a large bitcoin transaction network. We were able to identify certain users that conducted transactions in an atypical fashion, one that suggested some sort of money laundering. Unfortunately, we have no way of proving our suspicions, as we do not have labeled data that points us to cases of these hypothesized mixing services. However, our work here could help pave the way for future clustering techniques, especially by allowing one to choose features that are more revealing of patterns in the data.