

COURSE

**CSYE7374 54713 Parallel Machine Learning & AI SEC 02 Summer 2020**

FINAL PROJECT REPORT

**Detecting malicious mining pool behavior within Nakamoto consensus using parallelized ML**

INSTRUCTOR

Dr. Handan Liu

PRESENTERS

Srinjoy Chakravarty

(NUID 001086186)

Srishti Ashok Mishra

(NUID 001305178)

Contents

[Introduction 4](#_Toc47535167)

[Objective 4](#_Toc47535170)

[Motivation 4](#_Toc47535171)

[Methodologies 5](#_Toc47535172)

[Problem Analysis and Algorithms 5](#_Toc47535173)

[• Confusion Matrix 5](#_Toc47535174)

[• F1 Score 5](#_Toc47535175)

[• F-beta score 5](#_Toc47535176)

[• Precision-recall Curve 5](#_Toc47535177)

[• Precision Score 5](#_Toc47535178)

[• Random Forest Classifier 5](#_Toc47535179)

[• Recall Score 5](#_Toc47535180)

[Description of Dataset 5](#_Toc47535181)

[Result and Analysis 5](#_Toc47535189)

[Conclusion 5](#_Toc47535190)

# Introduction

#### What is the Blockchain System?[¶](https://www.kaggle.com/saketc/blockchain-system-anomaly-detection#What-is-the-Blockchain-System?)

“A blockchain is a time-stamped series of an immutable record of data that is managed by a cluster of computers not owned by any single entity. Each of these blocks of data is secured and bound to each other using cryptographic principles”. In simple terms, blockchain is a chain of blocks; the words “block” and “chain” are the digital information and public database, respectively. The core components of blockchain architecture:

1. Node — user or computer within the blockchain
2. Transaction — smallest building block of a blockchain system
3. Block — a data structure used for keeping a set of transactions which is distributed to all nodes in the network
4. Chain — a sequence of blocks in a specific order
5. Miners — specific nodes which perform the block verification process
6. Consensus— a set of rules and arrangements to carry out blockchain operation

Machine learning algorithms when applied on big data, allow us to analyze the actual extent of ‘cybersecurity’ behind the ‘proof-of-work’ generated by various entities, powering the distributed ledger technologies, that claim to be pseudonymous and decentralized in nature

# Background

# Contentious hard forks such as the one between Bitcoin ABC and Bitcoin SV (Satoshi’s Vision) result in ‘hashwars’. In such a climate, malicious mining pools often emerge to abuse rules allowed within Nakamoto consensus, such as the one proposed by SharkPool in 2017. Selfish game theoretical interests cause malicious behaviors like mining empty blocks, that often go unpunished.

# Objective

By classifying address signatures, we can identify mining entities and their potentially malicious behavior on the network to help the community make more responsible mining decisions.

# Motivation

Smaller altcoins which have hard forked from Bitcoin, are particularly vulnerable to selfish mining attacks such as the one proposed by SharkPool. Our ML model aims to detect and coalesce various miners within these smaller hard forks (Litecoin, Bitcoin Gold, Bitcoin Diamond, Bitcoin Cash) into possible mining pools, and analyse their empty block production patterns.

# Methodologies

# Problem Analysis and Algorithms

# Confusion Matrix

# F1 Score

# F-beta score

# Precision-recall Curve

# Precision Score

# Random Forest Classifier

# Recall Score

# Naïve Bayes

# Artificial Neural Network

# Description of Dataset

# The source of dataset is **Google BigQuery Bitcoin Blockchain Dataset**. It consists of two projects: Transactions and Blocks containing features like block\_id, is\_miner, address, transaction\_id, timestamp, input, output, etc., which updates every 10 minutes.

# Data is extracted by connecting to Google Datastore. A service account key for BigQuery is a JSON file which is obtained from GCP console and will be used further.

# The google BigQuery api client python libraries includes the functions you need to connect your Jupyter Notebook to the BigQuery : pip install google-cloud-bigquery

# Set the environment variable called **“GOOGLE\_APPLICATION\_CREDENTIALS”** for pointing your Notebook to your service account key

# SQL queries are executed to obtain the required fields

# Extraction repositories were bigquery-public-data.crypto\_bitcoin.transactions and bigquery-public-data.bitcoin\_blockchain.blocks

# To begin the pre-processing of data, drop the columns with null values, split the data into training and testing sets to train the model

# Concepts:

# Pool - A pool is a platform with specialized software where miners combine the computational power of their equipment for more efficient mining of a certain cryptocurrency.

# Mining Pool - A pool for mining is an association of miners, who are seeking for best reward. The computing power of each participant’ device makes up the total hashrate of the pool. Such cooperation gives much more chances to find a block and get a reward, which is distributed among the participants according to the system accepted by the operator of the particular pool for mining.

# In simpler terms - A pool for mining can be compared to a lottery pool. Your chances of winning the popular lottery are very low, but when you team up with a lot of other people and agree to share the money you win if you succeed, your chances are multiplied by the number of participants.

# Result and Analysis

# Bitcoin Mining Pool Classifier

# The data is extracted using SQL Queries from Google’s Big Query

# There are two main tables to examine here in the schema: a “blocks” table and a “transactions” table, which are joinable.

# #insert df.info() image to get the details of the columns from the dataset

# The question we are trying to answer is, to classify whether or not a given transaction was generated by a mining pool

# To train the model, Random Forest Classifier is used.

# To check how good our model has been trained, we use metrics like – confusion matrix, recall-score, precision\_score, precision\_recall\_curve, f1\_score, fbeta\_score.

# #image of Precision/Recall Curve

# #image of Confusion Matrix

# To understand which features, provide the most signals about the mining pool, we plot the top 12 features that are important according to the analysis

# #image of the features

# To check if there are any false positive associated with dark mining pools, we check if y\_test is false and y\_pred is true

# The list of false positive addresses is then used for network analysis

# Network Analysis: Bitcoin blockchain data

# Consider one of the false positive address from the previous code, using sql, query all the transactions that are related to the false positive address

# Data cleaning and processing: making datetime type transformation, converting satoshis to bitcoin, converting sending of bitcoin to negative value

# EDA includes, unique addresses that are included in the transaction history, top addresses, transaction activity across time and plotting the Date vs Transaction graph to see the volume of transaction per day

# 

# 

# Furthermore, difference between input transactions and output transactions, total amount of bitcoins sent each month,

# 

# Conclusion