Analysis, detection & mitigation of felonious wallet accounts over the Ethereum blockchain network using machine learning techniques

DISSERTATION

Submitted in partial fulfillment of the requirements of the

MTech Data Science and Engineering Degree programme

By

Anjali Sunder Naik
2019HC04178

Under the supervision of

Dr. Vishnu Prasad V J
Senior Technical Architect
Adjunct Faculty, Department of Engineering Design, IIT Madras

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
Pilani (Rajasthan) INDIA

January, 2022

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the Dissertation entitled "**Analysis, detection & mitigation of felonious wallet accounts over the Ethereum blockchain network using machine learning techniques"** and submitted by Ms. **Anjali Sunder Naik** ID. No. **2019HC04178** in partial fulfillment of the requirements of DSECLZG628T Dissertation, embodies the work done by her under my supervision.

Vishnu Prasad
Digitally signed by Vishnu Prasad
Date: 2022.01.13 22:16:30 +05'30'

(Signature of the Supervisor)

Place: Bengaluru

Date: January,2022

Dr. Vishnu Prasad V J

Senior Technical Architect

Adjunct Faculty, Department of Engineering Design, IIT Madras

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI
I SEMESTER 21-22
DSECLZG628T **DISSERTATION**
Dissertation Outline

BITS ID No. 2019HC4178 **Name of Student: Anjali Sunder Naik**

Name of Supervisor	: Dr Vishnu Prasad V J
Designation of Supervisor	: Senior Technical Architect, Adjunct Faculty, Department of Engineering Design, IIT Madras
Qualification and Experience	: PHD Physics, 14 Years 2 Months
E- mail ID of Supervisor	: vishnu.prasad@in.bosch.com
Topic of Dissertation	: Analysis, detection & mitigation of felonious wallet accounts over the Ethereum blockchain network using machine learning techniques

Name of First Examiner: _____

Designation of First Examiner:_____

Qualification and Experience: _____

E- mail ID of First Examiner:_____

Name of Second Examiner: _____

Designation of Second Examiner:_____

Qualification and Experience:_____

E- mail ID of Second Examiner:_____

(Signature of Student)

Vishnu Prasad
Digitally signed by Vishnu Prasad
Date: 2022.01.13 22:16:56 +05'30'

(Signature of Supervisor)

Date: January,2022

Date: January,2022

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programmes Division
I SEMESTER 21-22

DSE CL ZG628T **DISSERTATION**
(EC-2 Mid-Semester Progress Evaluation Sheet)

Scheduled Month : January,2022

NAME OF THE STUDENT **: Anjali Sunder Naik**

ID NO. **: 2019HC04178**

Email Address **: sundernaik.anjali@in.bosch.com**

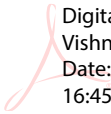
NAME OF SUPERVISOR **:Dr. Vishnu Prasad V J**

PROJECT TITLE **:Analysis, detection & mitigation of felonious wallet accounts over the Ethereum blockchain network using machine learning techniques**

EVALUATION DETAILS

EC No.	Component	Weightage	Comments (Technical Quality, Originality, Approach, Progress, Business value)	Marks Awarded
1	Dissertation Outline	10%		
2.	Mid-Sem Progress Seminar Viva Work Progress	 10% 5% 15%		

	Organizational Mentor	Additional Examiner
Name	Dr. Vishnu Prasad V J	
Qualification	PHD Physics	
Designation & Address	Senior Technical Architect, Bosch Global Software Technologies, Bengaluru Adjunct Faculty, Department of Engineering Design, IIT Madras	
Email Address	vishnu.prasad@in.bosch.com	

Signature	Vishnu Prasad  Digitally signed by Vishnu Prasad Date: 2022.01.14 16:45:29 +05'30'	
Date	January 2022	

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
SECOND SEMESTER 2020-21
DSECLZG628T **DISSERTATION**

Supervisor's Evaluation Form

Supervisor's Rating of the Technical Quality of this Dissertation Outline

EXCELLENT / GOOD / FAIR/ POOR (Please specify): **EXCELLENT**

Supervisor's suggestions and remarks about the outline (if applicable).

Vishnu Prasad
Digitally signed by Vishnu Prasad
Date: 2022.01.13 22:17:23 +05'30'

Date: January 2022

(Signature of Supervisor)

Name of the supervisor: Dr. Vishnu Prasad V J

Email Id of Supervisor:vishnu.prasad@in.bosch.com

Mob # of supervisor: +91-7598033881

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programmes Division
I SEMESTER 21-22

DSE CL ZG628T **DISSERTATION**
(Final Evaluation Sheet)

NAME OF THE STUDENT : **Anjali Sunder Naik**

ID NO. : **2019HC04178**

Email Address : **sundernaik.anjali@in.bosch.com**

NAME OF THE SUPERVISOR : **Dr. Vishnu Prasad V J**

PROJECT TITLE : **Analysis, detection & mitigation of felonious wallet accounts over the Ethereum blockchain network using machine learning techniques**

(Please put a tick (✔) mark in the appropriate box)

Sl. No.	Criteria	Excellent	Good	Fair	Poor
1	Work Progress and Achievements	✔			
2	Technical/Professional Competence	✔			
3	Documentation and expression	✔			
4	Initiative and originality	✔			
5	Punctuality	✔			
6	Reliability	✔			
	Recommended Final Grade	✔			

EVALUATION DETAILS

EC No.	Component	Weightage	Marks Awarded
1	Dissertation Outline	10%	
2	Mid-Sem Progress Seminar		

	Viva Work Progress	10% 5% 15%	
3	Final Seminar/Viva	20%	
4	Final Report	40%	
	Total out of	100%	

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
SECOND SEMESTER 2020-21

DSECLZG628T **DISSERTATION**

Dissertation Title : Analysis, detection & mitigation of felonious wallet accounts over the Ethereum blockchain network using machine learning techniques

Name of Supervisor : Dr. Vishnu Prasad V J

Name of Student : Anjali Sunder Naik

ID No. of Student : 2019HC04178

Abstract

As of 2021, a survey from Coin Market Cap indicates that there are nearly over 6,000 digital coins in the market, a severe increase from just a handful since 2013. However, a large portion of these cryptocurrencies might not be that significant. The total market cap of all the crypto assets, including stable coins and tokens has shown a significant rise from year 2020 and has hit 2.4 trillion. Cryptocurrencies has vast potential of revolutionizing and transforming compliance-free peer-to-peer transactions. However, an end user must overcome certain challenges related to privacy, security, and control. As the transactions are recorded in a publicly distributed ledger known as blockchain, hackers have a large attack surface to gain access to critical and sensitive data. In the rapidly growing crypto currency space, the technological advent of cryptocurrencies and their respective benefits has been veiled with several illicit financing activities operating over the network such as ransomware, terrorist financing, hacking, data manipulation during transaction process, phishing, fraud, money laundering, bribery etc. Chainalysis, a firm that tracks every crypto currency transaction and serves as an advisor to an array of government authorities has published a report that shows that the amount of cryptocurrency spent on dark net markets rose 60% to reach a new high of $1.15billion from July 2020 to June 2021.

In this work, the primary focus is on the Ethereum network, which has seen over 1373 million transactions since its inception. Propelled with the rise in use of machine learning techniques in the research dimensions of financial domain, this is an attempt to explore the possibility to use various machine learning algorithms to analyze and detect the illicit accounts using the transaction history. Many criminals don't typically transfer funds directly to and from their linked addresses when transacting with regulated exchanges. A vast majority of bad actors will move their funds at least one time. CipherTrace analysts found that a typical cryptocurrency exchange's dark market exposure will typically double at two hops out (transactions once removed from the exchange).

Various machine learning algorithms are evaluated on publicly available accounts flagged by the Ethereum community for their illegal activity coupled with valid accounts. A smart contract deployed on the public blockchain network is further used to track the illicit accounts, and hence proposed as a possible mitigation technique for flagging suspicious wallet addresses. External parties can query this smart contract to validate a blacklisted account and enable the law

enforcement agencies take appropriate actions on the stolen coins/Ponzi schemes. The proof of truth data on the blockchain ledger will serve as a benchmark for future analysis.

Key Words: Blockchain, Big data, Fraud-detection, Ethereum, Machine-learning

Vishnu Prasad

Digitally signed by Vishnu Prasad
Date: 2022.01.13 22:17:56 +05'30'

(Signature of the Student)

(Signature of the Supervisor)

Anjali Sunder Naik

2019HC04178

Dr Vishnu Prasad V J

Senior Technical Architect, Adjunct Faculty

Department of Engineering Design, IIT Madras

Contents

Chapters

1. Introduction & Background

The world wide web has revolutionized information, and the Web2 has revolutionized interactions. The Web3, widely known as the next age of the internet is an idea for a new iteration of the world wide web based on blockchains. The current internet system with its client-server based architecture and centralized data management has many unique points of failure in terms of privacy of personal data and inefficiencies in the backend operations. Blockchain came into existence with a need to resolve the existing trust issues in the current data-intensive transaction systems.

In general terms, a blockchain is an immutable distributed ledger, maintained within a distributed network of nodes, wherein each node maintains a copy of the ledger by applying transactions, validated by a consensus protocol. Blockchain technologies has a very interesting evolution cycle and has come quite far from cryptographically secured chain of blocks to decentralized applications.

In the early 90s, Stuart Haber and Scott Stornetta published their first work on blockchain, which served as a basis for the 2009 Satoshi Nakamoto's popular Bitcoin Whitepaper. The first bitcoin purchase took place in 2010, which further grew into a $1 billion marketplace in 2013.At the same time, Vitalik Buterin released the Ethereum whitepaper. The Ethereum genesis block was further created in 2015, which has grown to house 1.278M transactions at present.

Ethereum is a blockchain framework which lets a person send cryptocurrency to anyone for a minimal fee. It also powers applications that can be consumed across all the participants in the network. It can be simply termed as the "world's programmable blockchain". The introduction of Ethereum as a blockchain platform opened a new world to investors, developers, researchers, bankers, money launderers and hackers. The pseudo-anonymity nature of the actors in the network has paved way to felonious activities without a trace. Hence this work attempts to provide a solution to the growing need of detecting the felonious activities in the Ethereum network.

2. Problem Statement

As of 2021, a survey from Coin Market Cap indicates that there are nearly over 6,000 digital coins in the market, a severe increase from just a handful since 2013. However, a large portion of these cryptocurrencies might not be that significant. The total market cap of all the crypto assets, including stable coins and tokens has shown a significant rise from year 2020 and has hit 2.4 trillion. Cryptocurrencies has vast potential of revolutionizing and transforming compliance-free peer-to-peer transactions. However, an end user must overcome certain challenges related to privacy, security, and control. As the transactions are recorded in a publicly distributed ledger known as blockchain, hackers have a large attack surface to gain access to critical and sensitive data. In this rapidly growing crypto currency space, the technological advent of cryptocurrencies and their respective benefits has been veiled with several illicit financing activities operating over the network such as ransomware, terrorist financing, hacking, data manipulation during transaction process, phishing, fraud, money laundering, bribery etc. Chainalysis, a firm that tracks every crypto currency transaction and serves as an advisor to an array of government authorities has published a report that shows that the amount

of cryptocurrency spent on dark net markets rose 60% to reach a new high of $1.15 billion from July 2020 to June 2021.

3. Objective of the project

In this work, the primary focus is on the Ethereum network, which has seen over 1373 million transactions since its inception. Propelled with the rise in use of machine learning techniques in the research dimensions of financial domain, this is an attempt to explore the possibility to use various machine learning algorithms to analyze and detect the felonious accounts using the transaction history. Many criminals don't typically transfer funds directly to and from their linked addresses when transacting with regulated exchanges. A vast majority of bad actors will move their funds at least one time. CipherTrace analysts found that a typical cryptocurrency exchange's dark market exposure will typically double at two hops out (transactions once removed from the exchange).

Various machine learning algorithms are evaluated on publicly available accounts flagged by the Ethereum community for their illegal activity coupled with valid accounts.

4. Uniqueness of the project

This project will provide a compact comparison of various machine learning techniques to identify fraudulent activities in the Ethereum network.

5. Benefit to the organization

The analysis would benefit the organization in below ways:

- There would be more trust and worldwide adoption of the distributed ledger technology and possible regulation can be achieved in its usage.
- The felonious wallet accounts can be tracked and validated which would internally enable the law enforcement agencies take appropriate actions on the fraudulent activities
- The analysis data would serve as a benchmark for future analysis.

6. Scope of work

- Collection of blacklisted Ethereum wallet accounts data and its relevant historical transactions.
- Cleansing of data and identification of relevant attributes
- Data preprocessing & raw feature extraction
- Identification of Machine Learning algorithms for the obtained data set
- Data Analysis using various algorithms and evaluation of key metrics
- Hyperparameter tuning to identify ideal model algorithm
- Visualizations of results

7. Solution architecture

MapReduce is central to Apache Hadoop and distributed data processing. By leveraging data locality, MapReduce allows functions to run in parallel across multiple server nodes in a cluster. We use this feature of MapReduce algorithm to efficiently process the input data.

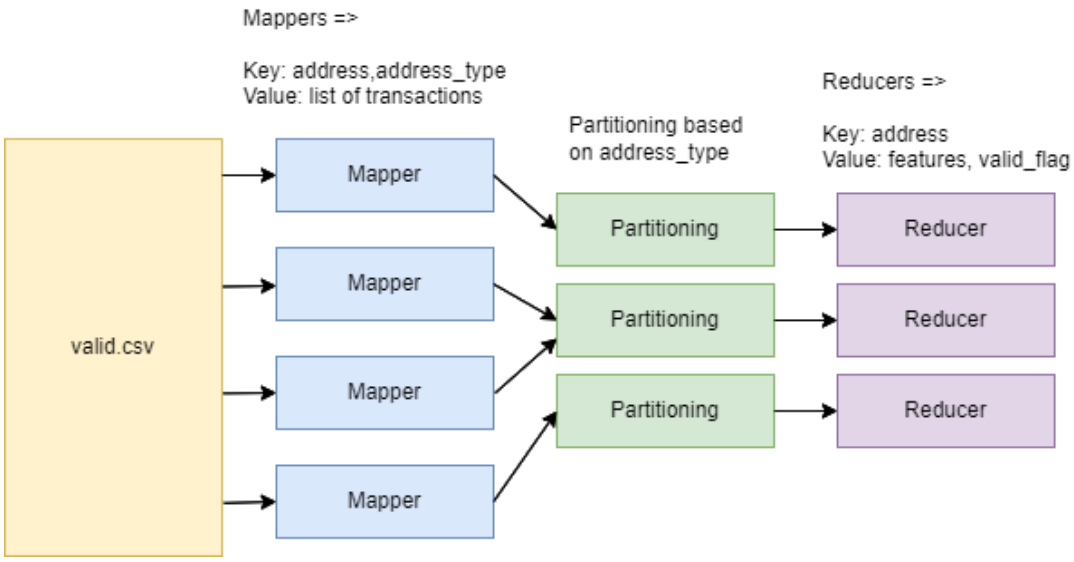


Figure 1: Map Reduce Pipeline for Felonious Data

```
Map-Reduce Framework
        Map input records=8477
        Map output records=6924
        Map output bytes=3118225
        Map output materialized bytes=3144968
        Input split bytes=139
        Combine input records=0
        Combine output records=0
        Reduce input groups=5678
        Reduce shuffle bytes=3144968
        Reduce input records=6924
        Reduce output records=5678
        Spilled Records=13848
        Shuffled Maps =3
        Failed Shuffles=0
        Merged Map outputs=3
        GC time elapsed (ms)=317
        Total committed heap usage (bytes)=4127195136
```

Figure 2: Map Reduce Framework for Felonious Data

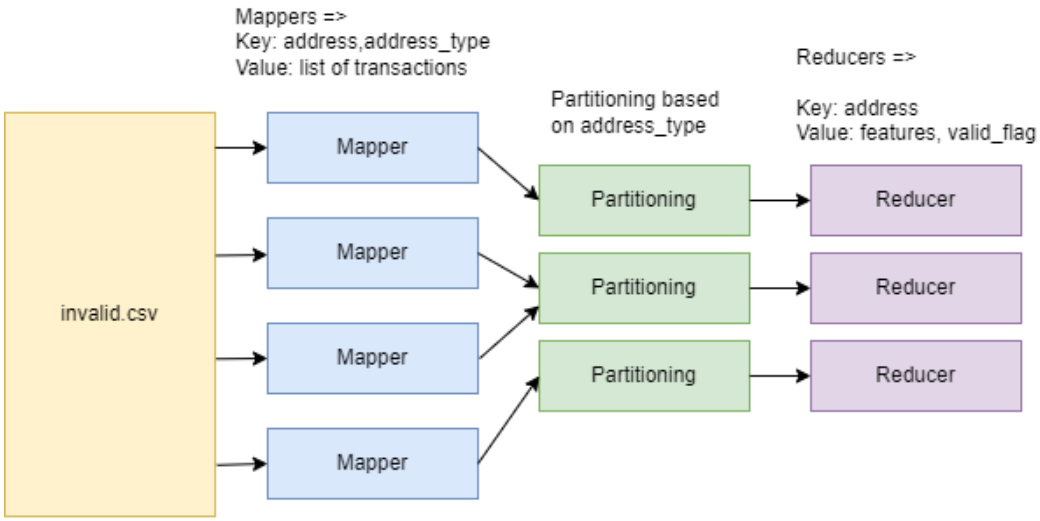


Figure 3: Map Reduce Pipeline for Non-Felonious Data

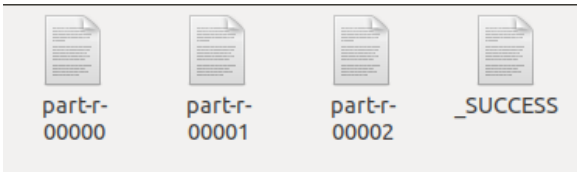


Figure 4: Snapshot of Map-reduce execution

8. Resources needed for the project

Programming Language	Python, Java
Tools	Eclipse, Jupyter-Lab
Hardware+ Cloud Platform	Google Collab+ GPU

9. Project plan & Deliverables

S No.	Task	Expected date of completion	Names of Deliverables
1	Research and data collection from various sources	3 weeks	Research & Data
2	Data cleansing & identification of data attributes	2 weeks	Feature extraction
3	Identification & Analysis of various Machine Learning algorithms for the obtained data set	6 weeks	Models
4	Hyperparameter tunning to identify the best model architecture	3 weeks	Optimal Model
5	Visualization of results and Documentation	2 weeks	Metrics UI, Project Report

10. Work accomplished so far

10.1 Understanding the Requirements

In this work, we are attempting to investigate the Ethereum accounts transaction data and to detect felonious activities.

10.2 Architecture and Design

10.2.1 Data Collection

We collect the data from various open sources such as Ethereum foundation backlisted data, Ether scan, which is a block explorer and analytics platform for Ethereum. The Ether scan provides a

segregated data based on word cloud, and we filter the felonious data based on key words such as phishing, hack, etc. We also collect data from Harvard verse and GitHub.

10.2.2 Data Cleansing

In the first phase of data cleansing, we filter the Ethereum addresses with null transaction. We also convert the address to lower case and remove the duplicated records. Once that is achieved, we segregate the accounts as:

1. Externally Owned Address (EOA)
   The end users create EOAs to be a part of the Ethereum network. Participants get the private key for each account to perform transactions.

2. Smart Contract
   These are the self-executing code which can be invoked by EOAs or another contract as an internal transaction.

   The result of this process is labelled data for both the categories of the Ethereum accounts.

10.2.3 Feature Engineering

We extract a total of 44 features from the Externally Owned Address and 18 features from the Smart Contracts. We can treat Ethereum transaction data as big data, and hence map reduce is the optimal programming model to efficiently extract the features in parallel over the large dataset in a distributed manner.

Once the features are extracted, we use information gain as a parameter to obtain the top 10 features relevant for the data analysis. Information gain is a feature selection mechanism, which evaluates the gain of each feature variable against the context of the target variable.

We use the ether scan API to extract the transaction list for each address in the dataset. The sample response for an address looks like below:


```
{
    "status": "1",
    "message": "OK",
    "result": [
        {
            "blockNumber": "6127109",
            "timeStamp": "1533974291",
            "hash": "0x1890d018b54fc773ca153701f64b0668d278e15ee9f99abad11635d24ec0babe",
            "nonce": "0",
            "blockHash": "0xa4a5635e484879021678290a785d8ef245c959d6f1613e9ec0f94ce13c088c8c",
            "transactionIndex": "105",
            "from": "0x8ddfdf60aaffe05c623ba193a186abd1f8024946",
            "to": "0xbceaa0040764009fdcff407e82ad1f06465fd2c4",
            "value": "25533614328758401081460",
            "gas": "21000",
            "gasPrice": "9000000000",
            "isError": "0",
            "txreceipt_status": "1",
            "input": "0x",
            "contractAddress": "",
            "cumulativeGasUsed": "7124989",
            "gasUsed": "21000",
            "confirmations": "7863508"
        },
```


Figure 5: Sample API response for an address

Externally Owned Accounts		
Feature	**Description**	**Unit**
f1_total_transactions_sent	The total number of transactions sent from the given address	Integer
f2_total_transactions_received	The total number of transactions received from the given address	Integer
f3_value_out	The total ether sent from the given address	Big Integer
f4_value_in	The total ether received from the given address	Big Integer
f5_value_difference	Absolute difference [(f3_value_out) - (f4_value_in)]	Big Integer
f6_number_of_distinct_address_contacted	The number of distinct addresses contacted	Integer
f7_total_transactions_sent_received	The total number of transactions performed by the address	Integer
f8_total_transactions_sent_to_unique_address	The total number of transactions sent to a unique address	Integer
f9_total_transactions_received_from_unique_address	The total number of transactions received from a unique address	Integer
f10_first_transaction_time	The block timestamp wherein the first transaction was performed	Long
f11_last_transaction_time	The block timestamp wherein the last transaction was performed	Long
f12_transaction_active_duration (seconds)	[(f11_last_transaction_time) - (f10_first_transaction_time)]	Long

f13_last_txn_bit	0 if last transaction is incoming transaction else 1	Integer
f14_last_transaction_value	Total ether transferred in last transaction	Big Integer
f15_average_incoming_ether	Average value of ether in the incoming transactions	Big Integer
f16_average_outgoing_ether	Average value of ether in the outgoing transactions	Big Integer
f17_average_percentage_gas_incoming	Average % of gas used in the incoming transactions	Double
f18_average_percentage_gas_outgoing	Average % of gas used in the outgoing transactions	Double
f19_outgoing_gas_price	Total gas price in the outgoing transaction	Long
f20_incoming_gas_price	Total gas price in the incoming transaction	Long
f21_average_incoming_gas_price	Average gas price in the outgoing transaction	Double
f22_average_outgoing_gas_price	Average gas price in the incoming transaction	Double
f23_total_failed_transactions_incoming	Total failed transactions in the incoming transaction	Integer
f24_total_failed_transactions_outgoing	Total failed transactions in the outgoing transaction	Integer
f25_total_failed_transactions	Total failed transactions	Integer
f26_total_success_transactions_incoming	Total successful transactions in the incoming transaction	Integer
f27_total_success_transactions_outgoing	Total successful transactions in the outgoing transaction	Integer
f28_total_success_transactions	Total successful transactions	Integer
f29_gas_used_incoming_transaction	Total gas used in the incoming transaction	Long

f30_gas_used_outgoing_transaction	Total gas used in the outgoing transaction	Long
f31_percentage_transaction_sent	Percentage of transactions sent	Double
f32_percentage_transaction_received	Percentage of transactions received	Double
f33_standard_deviation_ether_incoming	Standard deviation of ether in incoming transaction	Double
f34_standard_deviation_ether_outgoing	Standard deviation of ether in outgoing transaction	Double
f35_standard_deviation_gas_price_incoming	Standard deviation of gas price in incoming transaction	Double
f36_standard_deviation_gas_price_outgoing	Standard deviation of gas price in outgoing transaction	Double
f37_first_transaction_bit	0/1 (0 if last transaction is incoming transaction else 1)	Bit
f38_first_transaction_value	Ether value in first transaction	Long
f39_mean_in_time (seconds)	Mean time difference between incoming transaction	Double
f40_mean_out_time	Mean time difference between outgoing transaction	Double
f41_mean_time	Mean time difference between all transaction	Double
f42_transaction_fee_spent_incoming	Total fee spent in incoming transaction	Long
f43_transaction_fee_spent_outgoing	Total fee spent in outgoing transaction	Long
f44_transaction_fee_spent	Total fee spent	Long

Table I: Features extracted from Externally owned accounts

Smart Contract		
Feature	**Description**	**Unit**
f1_contract_creation_time	Contract creation time	Long
f2_transaction_fee_spent_contract_creation	Transaction fee spent in contract creation	Long
f3_percentage_gas_used_contract_creation	Gas price used in contract creation	Double

f4_gas_price_contract_creation	Block timestamp for contract creation	Long
f5_first_contract_invoke_time	First contract invoke timestamp	Long
f6_last_contract_invoke_time	Last contract invoke timestamp	Long
f7_active_duration (seconds)	Active duration of the contract address	Long
f8_total_invocations	Total invocations from the contract address	Integer
f9_total_unique_invocations	Total unique invocations from the contract address	Integer
f10_avg_gas_used_contract_invocations	Average percentage of gas used in contract invocations	Double
f11_total_gas_price_contract_invocations	Total gas price used in contract invocations	Long
f12_avg_gas_price_contract_invocations	Average gas price used in contract invocations	Double
f13_total_tx_fee_contract_invocations	Total transaction fee used in contract invocations	Long
f14_avg_tx_fee_contract_invocations	Average transaction fee used in contract invocations	Double
f15_total_ether_contract_invocations	Total ether in contract invocations	Big Integer
f16_average_ether_contract_invocations	Average ether in contract invocations	Big Integer
f17_total_gas_used_contract_invocations	Total gas used in contract invocations	Long
f18_avg_gas_used_contract_invocations	Average gas used in contract invocations	Double

Table II: Features extracted from Smart Contract Address

11. Key challenges faced during the project

- Map-reduce execution failure due to irregular data in the dataset.
- Accounts will null transactions/lack of account information
- Feature extraction for data which has less than 2 transactions

12. Potential Risk and Mitigation plan

- Collection of meaningful data and the amount of data collected
- Scaling data