University of Reading

# Detecting Fraud on Ethereum: A Machine Learning and Blockchain Analytics Approach

Anand Patel, Supervisor: Atta Badii

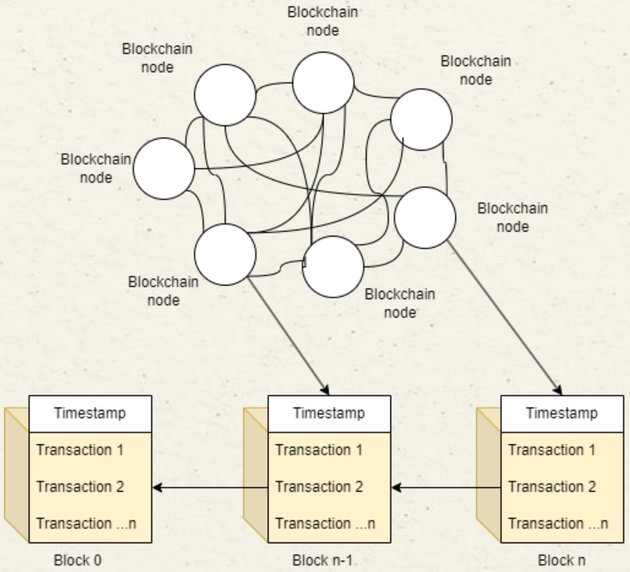
ethereum

# What is blockchain? What is Ethereum?

## Blockchain Technology

Blockchains are:

- Immutable
- Secure
- Decentralised
- Transparent
- Pseudonymous




## Ethereum

Ethereum extends on the basic blockchain:

- Self-executing
- Smart contracts
- dApps (Decentralised applications)
- Programmable
- $368 billion market capitalisation

# $14,000,000,000

## Sent to illicit addresses

CipherTrace. (2020). "Cryptocurrency Crime and Anti-Money Laundering Report," Mastercard

# Research Objectives

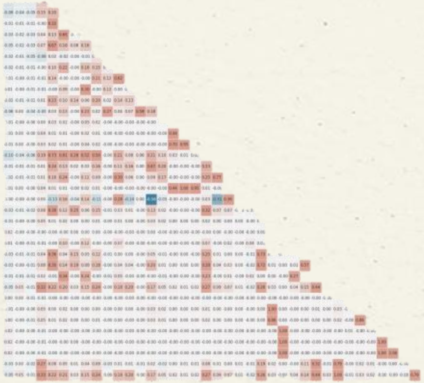
**To leverage advanced machine learning techniques to enhance the detection and prevention of fraudulent activities on the Ethereum blockchain.**

## Comprehensive Analysis of Ethereum

- Examine Ethereum Architecture
- Catalogue Ethereum Fraud Types

## Evaluation of Machine Learning Models:

- Implement a variety Models
- Feature Engineering
- Compare Models



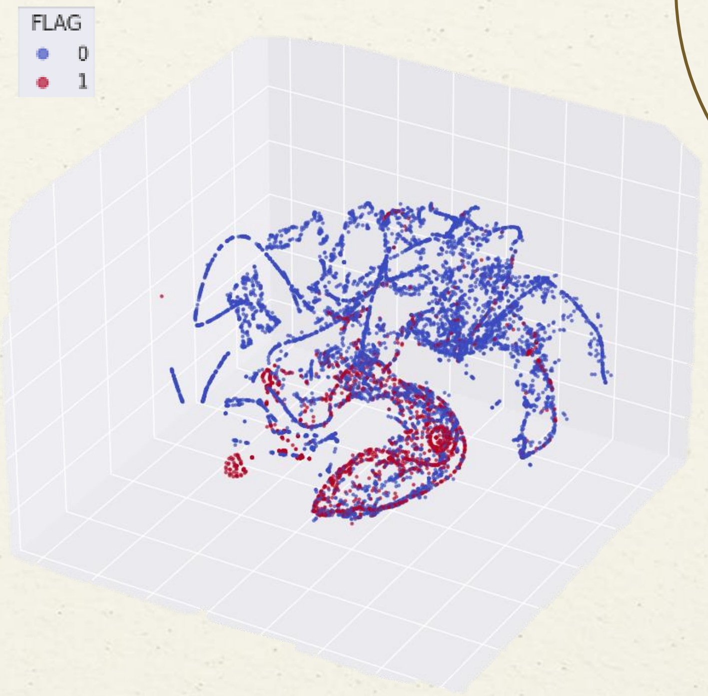
Correlation matrix of the dataset

# 9841

The Kaggle dataset contains Ethereum accounts of which 22% are labelled fraudulent

# Dataset shortcomings

- Imbalance
- Bias
- Poor documentation

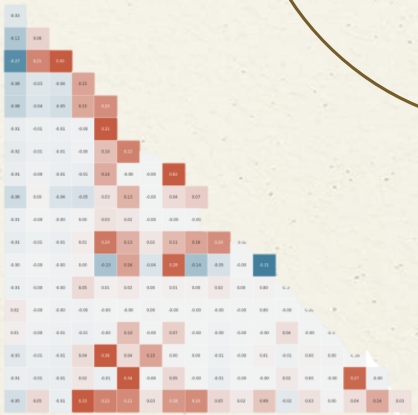



3D t-SNE visualisation of the dataset, (Blue: normal account, Red: Illicit account)

# Data Pre-processing

To address some of the dataset issues, the following steps were taken:

- Drop Categorical Variables
- Handle missing values
- Remove features with zero variance
- Dimensionality reduction via correlation analysis
- Dataset Preparation and Split
- T-SNE visualisation
- Address data class imbalance



Correlation matrix of the dataset after pre-processing

# Methodology

## Theoretical framework

Models are proposed based on the nature of the dataset and the proven efficacy in similar tasks the from existing literature

## 1 Modelling

Propose potential models for the dataset

## 2 Training

Train the models to attain the best performance metrics

## 3 Evaluate

Compare the models to identify the best performer

# Methodology

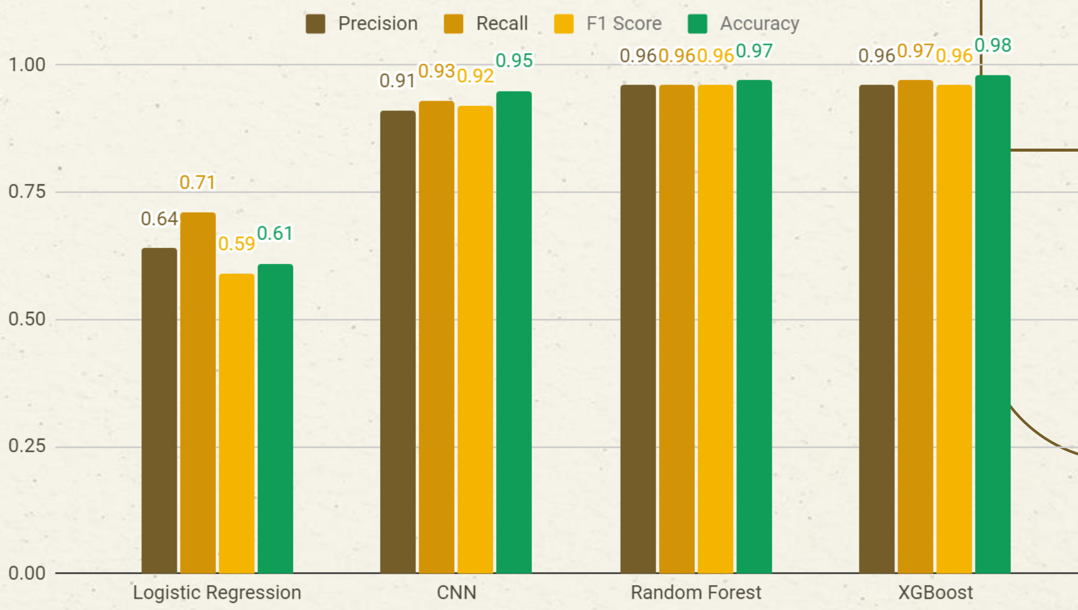
Models

- **Logistic Regression** $P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}$
- **Random Forest** $Y = \text{Majority vote}\{f_1(X,\Theta_1), f_2(X,\Theta_2), \ldots, f_k(X,\Theta_k)\}$
- **Neural Network (CNN)** $z = b + \sum_{i=1}^{n} w_i \cdot x_i$ $\quad a = \sigma(z)$
- **XGBoost** $\hat{y}_i = \sum_{k=1}^{N} f_k(x_i)$ $\quad \text{Objective} = \sum_i L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$

# Initial results

XGBoost, an advanced implementation of gradient boosting, has shown the best performance in this comparative analysis of machine learning models for detecting fraud on the Ethereum blockchain.

- High Precision, 0.99 for classifying non-fraudulent accounts and 0.93 for fraudulent transactions
- F1-Score weighted average = 0.98
- Overall accuracy of 98%
- This model will now go on to be hyperparameter tuned to optimise the model further




## Logistic Regression

Performed significantly worse than other models tested

## XGBoost

Had the best performance scores

# Hyperparameter Tuning XGboost

This is done to improve model performance as different hyperparameters can significantly impact the predictive accuracy of the model. GridSearchCV was the method used to do an exhaustive search over specified parameter values.

## Parameters explored:

LearningRate: [0.01, 0.1, 0.5, 0.75]
N Estimators: [50, 100, 200, 250]
Subsample: [0.2, 0.5, 0.9, 1.2]
Max Depth: [3, 4, 5, 6]
Colsample by Tree: [0.3, 0.7, 1.2]

## Optimal Parameters found:

ColSample by Tree: 0.7
Learning Rate: 0.1
Max Depth: 6
Number of Estimators: 250
Subsample: 0.5

## Achieved Recall: 98.85%

# Conclusions

## XGBoost

Demonstrated that XGBoost is the best model for detecting fraud on the Ethereum dataset

## Majority of fraud found

The model's high recall ensures that nearly all fraudulent transactions are detected, minimising the risk of fraud slipping through the system.

## Impact

Improved fraud detection capabilties can improve the integrity and trustworthiness of the Ethereum network

## Future Research

Future work could include real time detection by connecting an ethereum node to the machine learning model

# Thanks!

## Reflections

- Sourcing a high-quality labelled dataset was difficult
- The dataset needed a lot of pre-processing, e.g. dealing with the data imbalance
- I highly enjoyed this challenging project as it united my two admired fields of Computer Science, Machine learning and Blockchain computing and gave me a deeper understanding of those fields respectively

# References


- Steven Farrugia, Joshua Ellul, George Azzopardi, Detection of illicit accounts over the Ethereum blockchain, Available at: https://doi.org/10.1016/j.eswa.2020.113318
- Chen, T. and Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. arXiv. Available at: https://arxiv.org/abs/1603.02754
- LaValley, M.P., 2008. Logistic Regression. Circulation, 117(18), pp.2395-2399. Available at: https://doi.org/10.1161/CIRCULATIONAHA.106.682658
- Schratz, P., Muenchow, J., Iturritxa, E., Richter, J. and Brenning, A., 2019. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling, 406, pp.109-120. Available at: https://doi.org/10.1016/j.ecolmodel.2019.06.002
- Kaggle, 2021. Ethereum Fraud Detection Dataset. Available at: https://www.kaggle.com/datasets/vagifa/ethereum-frauddetection-dataset/discussion