CS-GY 6923 - Project Proposal

Ethereum Fraud Detection

Motivation/Aim of the project:

Ethereum is a decentralized blockchain platform that establishes a peer-to-peer network that securely executes and verifies application code, called smart contracts. A transaction once done can't be undone in blockchain network. So it is very important to identify any illegal transactions in the network and prevent them from corrupting the entire blockchain network. Ethereum fraud detection is of paramount importance in the world of blockchain and cryptocurrencies. This is crucial for maintaining the integrity and security of the platform, protection of users, and preserving Ethereum ecosystem.

My project is to build a machine learning model which classifies a given transaction as fraud or valid transaction.

Dataset:

I'm planning to use the Ethereum Fraud Detection dataset from kaggle (link). This dataset contains 9840 transactions and 50 features for each transaction and each transaction is labeled as fraud or valid transactions. This data set is imbalanced and contains lot of feature which are not relevant. Preprocessing needs to be done on this dataset before use.

Dataset: https://www.kaggle.com/datasets/vagifa/ethereum-frauddetection-dataset/data

ML model:

Since this is a classification task. I'm planning to use

1) logistic regression
2) Decision trees (XGBClassifier)
3) Support Vector Machines (with different kernels)

to model this dataset. If possible, I intend to build a Neural Network model on this dataset.

Accuracy/Error measures:

Since this is a classification task I'm going to be using

1. Confusion Matrix (precision, recall, f1-score)
2. AUC-ROC (Area under the ROC Curve)
3. Log-loss for logistic regression

Since cryptocurrency and blockchain technology has become increasingly popular for the past few years, there is some work done on this problem. There are some results you can find on the internet, but I didn't find any well documented results for this problem. I've have some domain knowledge on blockchain technology, so I am planning to do data preprocession to eliminate the features which are irrelevant to improve results of this problem.

Name: Tekuri Sai Akhil

NetId: st5050