EN.650.672 Security Analytics
Project Proposal
Fraud Detection In Ethereum Blockchain

Akshay Kaikottil, Liyin Li(Kenny), Pratik Kayastha, Saksham Sharma

1. Problem Definition

Over the past year, the mad rush of cryptocurrency has caught the attention of all kinds of investors. Not only that, it has also caught the attention of scammers. Crypto scams most often aim to trick a target victim into sending cryptocurrency to a compromised digital wallet. Through the use of dedicated social engineering, such as romance scams, email phishings, and even Ponzi schemes, these targets are very likely being convinced with the flourish returns promised by the attackers.

With the goal of applying our security analyst skills into solving real-world problems, we intend to use the approach of supervised learning, in building a classifier to identify fraudulent accounts among Ethereum transactions. The primary dataset that we are going to use is a labeled dataset from Kraggle, Ethereum Fraud Detection Dataset, with around 9,000 samples. In complementary to missing data or inefficient samples, we will make use of Etherscan, which is an analytic platform for querying details on any Ethereum blockchain transactions.

2. Data Introduction

The dataset contains over 9000 samples of known fraud (2179) and valid transactions (7662) made over the Ethereum platform. It contains 44 columns that include:

- Address: the address of the ethereum account
- NumberofCreatedContracts: Total Number of created contract transactions
- AvgValSent: Average value of Ether ever sent
- TotalERC20Tnxs: Total number of ERC20 token transfer transactions
- FLAG: whether the transaction is fraud or not.

2.1 What can we do with this data

- Flag address of ethereum accounts that are more probable to fraud.
- Flag transactions to be further studied by the forensics team.

2.2 Data Characteristics

Some of the important characteristics are

- Labeled
- Imbalanced
- Needs data pre-processing - Missing values through Etherscan
- Numerical
- Contains duplicate entries

3. Models & Techniques

3.1 Data Pre-processing and Features Selection

We will be making use of Exploratory Data Analysis to understand the dataset features. By reviewing some projects similar to ours, we try following different classification learning methods in our project

- Box-whisker plot
- Histogram
- Scatter plot

Since out dataset in unbalanced, we are planning to use SMOTE to balance the dataset. We are also going to try and reduce the dimensionality of our dataset usnig Principal component analysis (PCA), singular value decomposition (SVD). We then plan on choosing the best features from our dataset using Recursive Feature Elimination.

3.2 Models Selection and Evaluation

As we discussed in the earlier section that our dataset contains labels for fraud or legitimate accounts, we will be using supervised learning classification models. To determine the correct model, we need to properly understand different characteristics of the dataset. We know that our ethereum fraud detection dataset is not very large and imbalanced. We also determined that our model doesn't need to detect accounts adhoc in other words it doesn't require not fast computing power. According to our requirement, using a discriminative model would be a good choice as it would be helpful in trying to find boundaries of the classes for our problem and also it also doesn't require too much computation cost. By reviewing some projects similar to ours, we try following different classification learning methods in our project

- Random Forests
- Decision Trees
- Support Vector Machine
- Logistic Regression

For our classification problem, we will be using the Precision-Recall, ROC-AOC, Accuracy, F1 score for evaluation of different learning methods. Although, we will focus more on F1 scores as it works well with an imbalanced dataset.

To avoid the problem of overfitting, we will be applying the regularization techniques during the model evaluation. As we are also using ensemble learning methods, we will be also implementing bagging and boosting techniques and will try to evaluate them against different evaluation metrics to determine the best model.

4. Related Works

1. Detecting Fraudulent Accounts on Blockchain: A Supervised Approach
2. Data Mining based Ethereum Fraud Detection