Detecting Fradulent Ethereum Addresses: A Data Mining Approach

Anahita Hedayatifard

Professor: Dr. Arta Jamshidi
Department of Mathematics,
University Of Tehran

February, 2024

Table of Contents

What attracts frauds into blockchain and cryptocurrency transactions?

- The decentralized nature of Blockchain
- Immutable structure of blocks
- Pseudo anonymity

Examples of Types of Fraud

- **Smart-ponzi schemes**: fraudulent investment schemes that promise high returns with little or no risk.
- **Phishing**: aim to steal private keys by tricking users into entering them on fake websites.
- **Giveaway scams**: falsely promise to multiply any crypto sent. Imposter scams pose as celebrities or companies.

⋮

It was reported that the amount of money lost in crypto scam worldwide was $7.7 Billion in 2021 alone.

Table of Contents

- Dimension: 9841 rows $\times$ 51 columns
- Columns:
  - **FLAG**: whether the address is fraudulent or not
  - **Avg min between sent tnx**: Average time between sent transactions for account in minutes
  - **Avg min between received tnx**: Average time between received transactions for account in minutes
  - **Time Diff between first and last (Mins)**: Time difference between the first and last transaction
  - **MinValSent**: Minimum value of Ether ever sent
  - **MaxValSent**: Maximum value of Ether ever sent
  - **AvgValSent**: Average value of Ether ever sent

$\vdots$

```
Index(['Unnamed: 0', 'Index', 'Address', 'FLAG', 'Avg min between sent tnx',
       'Avg min between received tnx',
       'Time Diff between first and last (Mins)', 'Sent tnx', 'Received Tnx',
       'Number of Created Contracts', 'Unique Received From Addresses',
       'Unique Sent To Addresses', 'min value received', 'max value received ',
       'avg val received', 'min val sent', 'max val sent', 'avg val sent',
       'min value sent to contract', 'max val sent to contract',
       'avg value sent to contract',
       'total transactions (including tnx to create contract',
       'total Ether sent', 'total ether received',
       'total ether sent contracts', 'total ether balance',
       ' Total ERC20 tnxs', ' ERC20 total Ether received',
       ' ERC20 total ether sent', ' ERC20 total Ether sent contract',
       ' ERC20 uniq sent addr', ' ERC20 uniq rec addr',
       ' ERC20 uniq sent addr.1', ' ERC20 uniq rec contract addr',
       ' ERC20 avg time between sent tnx', ' ERC20 avg time between rec tnx',
       ' ERC20 avg time between rec 2 tnx',
       ' ERC20 avg time between contract tnx', ' ERC20 min val rec',
       ' ERC20 max val rec', ' ERC20 avg val rec', ' ERC20 min val sent',
       ' ERC20 max val sent', ' ERC20 avg val sent',
       ' ERC20 min val sent contract', ' ERC20 max val sent contract',
       ' ERC20 avg val sent contract', ' ERC20 uniq sent token name',
       ' ERC20 uniq rec token name', ' ERC20 most sent token type',
       ' ERC20_most_rec_token_type'],
      dtype='object')
```

Figure: Columns of the Data

Dropping the Irrelevant Columns

What is considered irrelevant?

- Index and Address
- Columns which are nearly constants and have zero variance
  - total ether sent contracts: Total Ether sent to Contract addresses
  - avg value sent to contract
  - max val sent to contract
  - min value sent to contract
  - ERC20 avg val sent contract

⋮

- Highly Correlated Columns:
  - ERC20 avg time between sent tnx
  - ERC20 avg time between rec tnx
  - ERC20 avg time between contract tnx


Correlation Heatmap
FLAG
Avg min between sent tnx
Avg min between received tnx
Time Diff between first and last (Mins)
Sent tnx
Received Tnx
Number of Created Contracts
Unique Received From Addresses
Unique Sent To Addresses
min value received
max value received
avg val received
min val sent
max val sent
avg val sent
total transactions (including tnx to create contract
total Ether sent
total ether received
total ether balance
Total ERC20 tnxs
ERC20 total Ether received
ERC20 total ether sent
ERC20 total Ether sent contract
ERC20 uniq sent addr
ERC20 uniq rec addr
ERC20 uniq sent addr.1
ERC20 uniq rec contract addr
ERC20 min val rec
ERC20 max val rec
ERC20 avg val rec
ERC20 min val sent
ERC20 max val sent
ERC20 avg val sent
ERC20 uniq sent token name
ERC20 uniq rec token name
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Handling Missing Data

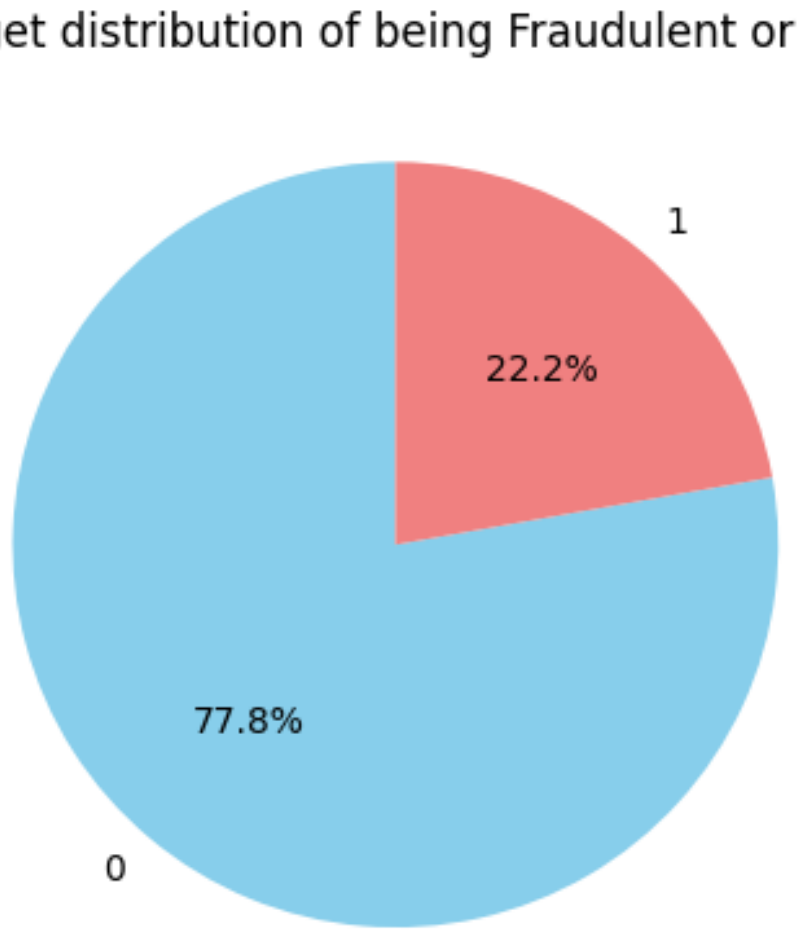
As most addresses that had missing data were due to very little or even no transactions being made, the missing data was imputed by zeros.

Min-Max Normalization

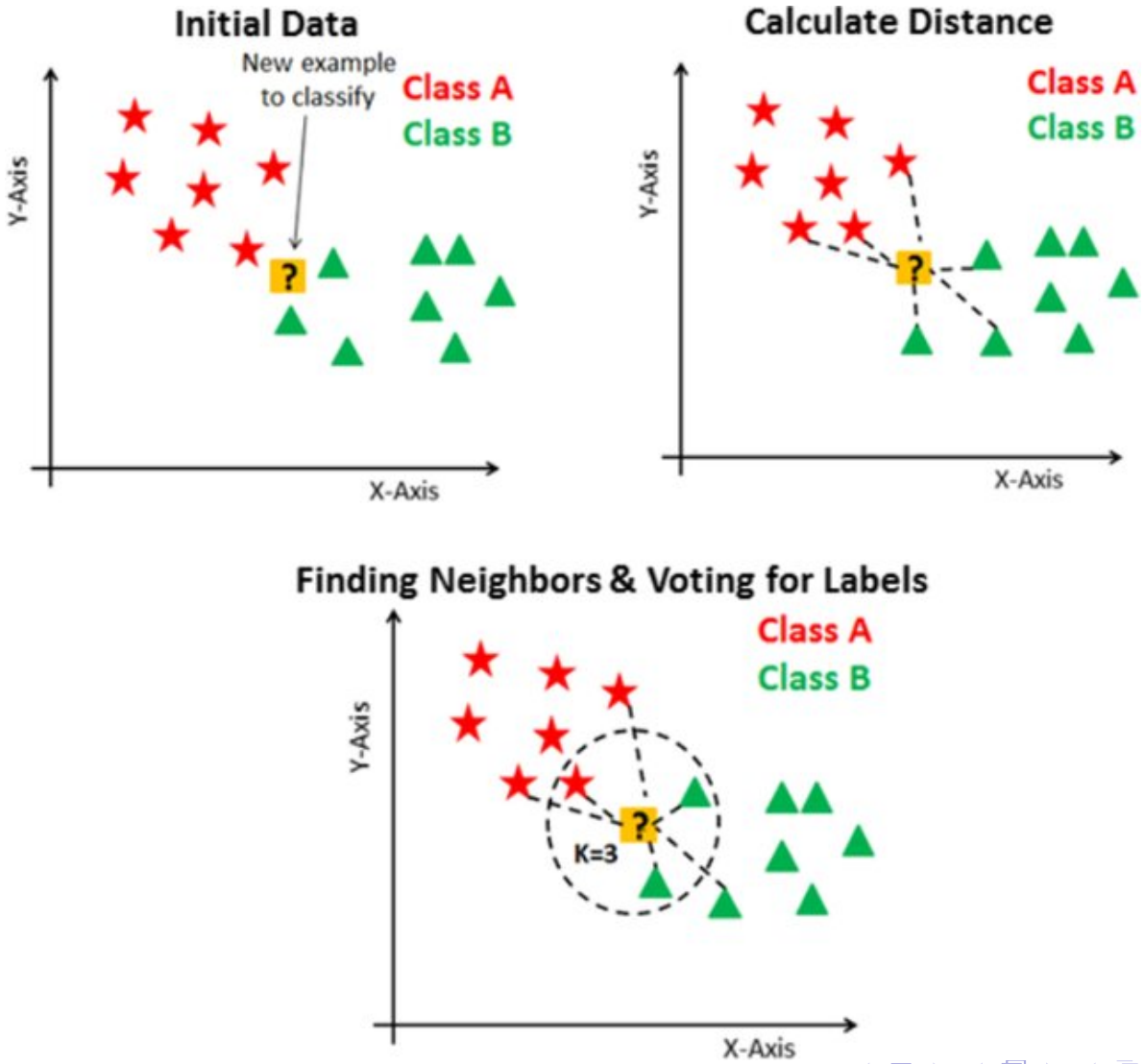
$$x^*_{mm} = \frac{x - min}{max - min}$$

- **Equal Contribution**: By scaling features to a common range, each feature contributes equally to the analysis and modeling process. This is important when using distance-based algorithms.
- **Handling Outliers**: Min-max scaling can make models more robust to outliers by constraining the range of values.

A Word of Caution: Unbalanced Data



How does KNN Work?



- **K N**earest **N**eighbors
- a Supervised Algorithm
- Used for Classification and Regression Tasks
- Voting Method:

$$\text{vote of each data point} := \frac{1}{d^2}$$

- Splitting the data set into training and test sets
- Balancing the training data set using the method of oversampling the minority class:

$$x = \frac{p(records) - rare}{1 - p}$$

  - **x** := the required number of resampled records
  - **p** := the desired proportion of rare values in the balanced data set
  - **records** := the number of records in the unbalanced data set
  - **rare** := the current number of rare target values

Classification Report

- **Precision**: What percent of your predictions were correct? the ability of a classifier not to label an instance positive that is actually negative.
- **Recall**: What percent of the positive cases did the model catch? the ability of a classifier to find all positive instances
- **F1 measure**: What percent of positive predictions were correct? As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

Precision	Recall	F1 measure	Support
$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$2\frac{Precision \times Recall}{Precision+Recall}$	$TP + FN$

Model Results

-	Precision	Recall	F1 measure	Support
0	0.96	0.90	0.93	765
1	0.70	0.86	0.77	218

Thank You