Task 1:

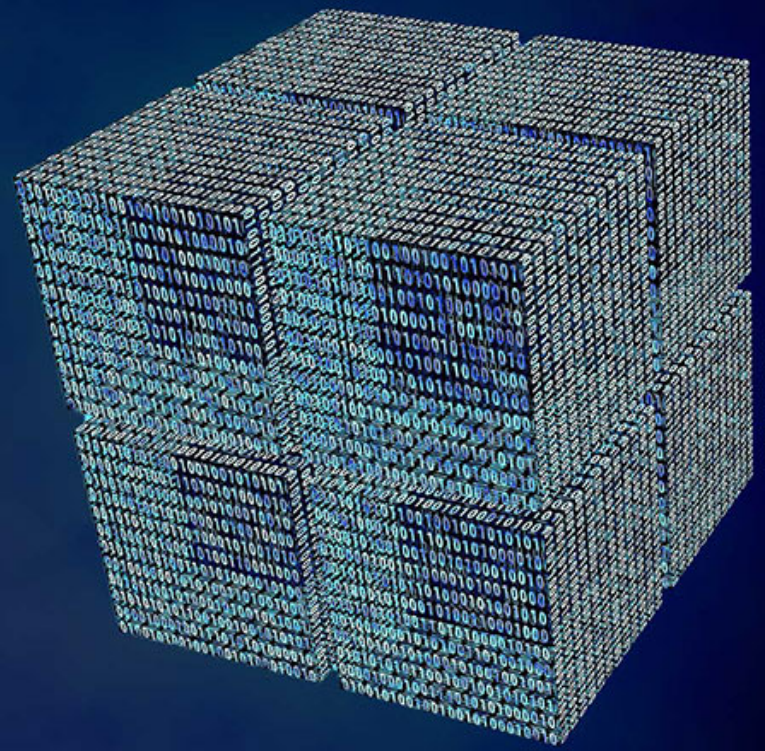
Malicious Ethereum Addresses Detection

Case Study for BMW Interview

Yichun Xie

21st Jun 2022

Content



01

Problem

Malicious Ethereum addresses Detection

02

Feature Engineering

Data Preprocessing, Feature Engineering, Exploratory Data Analysis

03

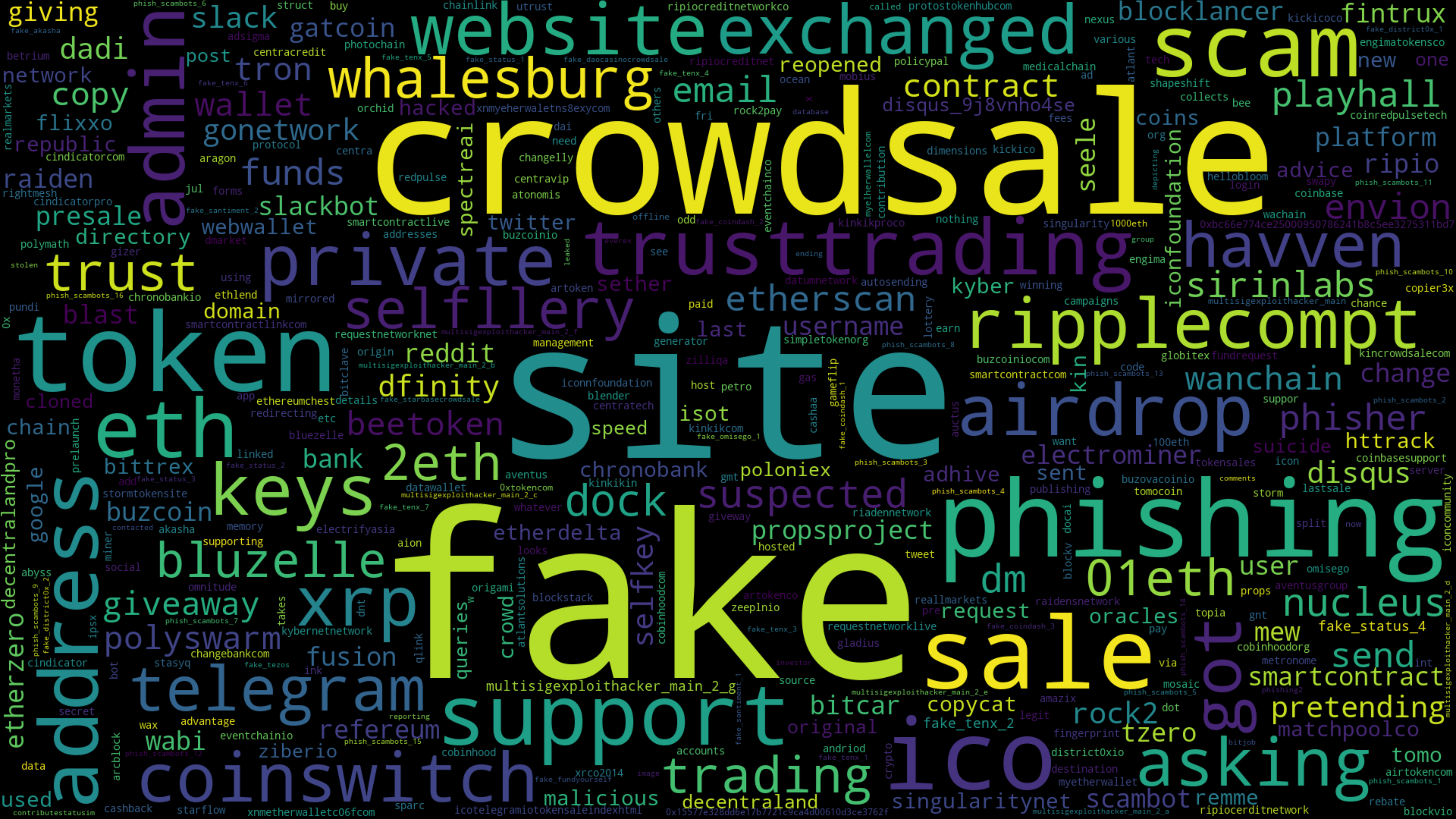
Modelling

Standard Machine Learning Models, Hyperparameter Optimization, Feature Importance

04

Solution & Discussion

Business Suggestions, Limitation

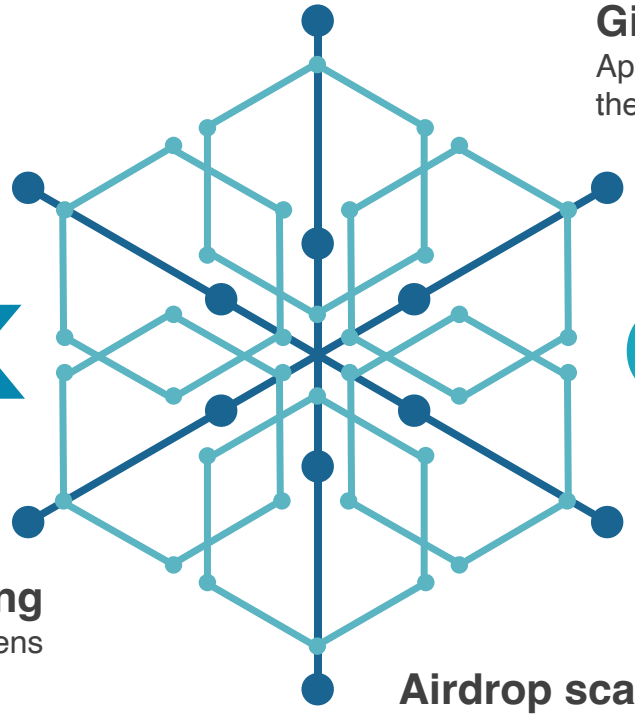

crowdsale
site
fake
phishing
token
address
admin
private
trusttrading
website
exchanged
scam
whalesburg
sale
support
ico
airdrop
ripplecompt
telegram
coinswitch
keys
eth
2eth
xrp
trust
selfllery
havven
asking
trading
bluzelle
giveaway
polyswarm
got
01eth
dm
dock
suspected
etherscan
sirinlabs
playhall
contract
email
wallet
gonetwork
funds
slackbot
presale
directory
tron
gatcoin
slack
dadi
copy
flixxo
republic
raiden
network
post
giving
blocklancer
fintrux
platform
ripio
envion
advice
seele
iconfoundation
reopened
beetoken
dfinity
reddit
domain
blast
bank
bittrex
buzcoin
chronobank
etherdelta
propsproject
wanchain
phisher
disqus
nucleus
send
smartcontract
pretending
matchpoolco
tomo
rock2
tzero
mew
request
oracles
selfkey
crowd
fusion
refereum
wabi
ziberio
malicious
decentraland
singularitynet
scambot
remme
etherzero
decentralandpro
chain
cloned
electrominer
adhive
sent
user
speed
isot
last
username
kyber
sether
twitter
webwallet
spectreai
poloniex
bitcar
copycat
original
change
httrack
suicide
kin
used

Malicious Activities

Phishing scam
Redirect users to imitation websites, ask them to reset their password or sent ETH.

Giveaway scam
Appear in many forms to ask users to send ETH to the provided wallet address, e.g., support giveaway..

Fake (crowdsale) website
Included in phishing scam and many other scams.



Social media hacks
Organizations and celebrities get hacked to post a cryptocurrency giveaway.

Fake Initial Coin Offering
Fake admin in ICOs, fake tokens

Airdrop scam
Airdrop an asset into your wallet and sending a scam website to claim the airdropped asset.

Case Study Dataset

Malicious Addresses with Comments	Malicious Transactions	Normal Transactions
663 Addresses 268 Comments From 2017-07-18 to 2020-11-17	551 Addresses 21961 Transactions From 2017-05-20 to 2022-05-05	87 Addresses 30000 Transactions From 2016-05-26 to 2022-06-08

Task: To detect malicious addresses

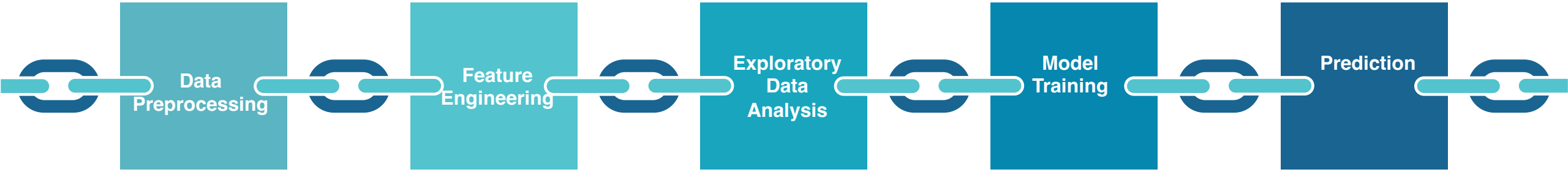
Role: member of a Crypto startup

Transactions with features:

address
from address, to address, contractAddress
input
timestamp
value
gas, gasPrice, cumulativeGasUsed
isError, txreceipt_status
blockNumber, hash, nounce, blockHash, transactionIndex
transactionIndex, confirmations

-> ML Objective: Prediction of maliciousness (Binary classification)

Machine learning Pipeline



Data Formatting, Transformation

Data types,
Lowercase addresses,
Timestamp to time,
Gas Price from Gwei to Ethr.

Select features, Create New Features

Group and aggregate features

Plots

Distributions,
Correlations

Standard machine Leaning models

Training, Evaluation Metrics,
Hyperparameter
Optimization of selected models.

Test Dataset

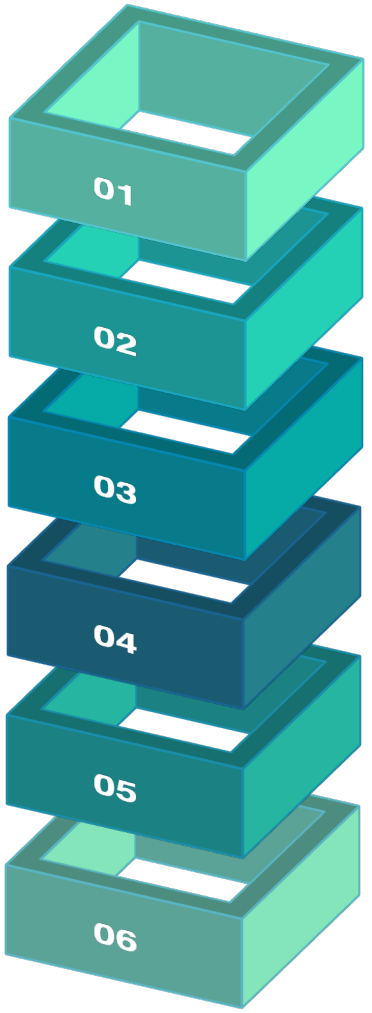
Feature importance

Feature Engineering

Smart Contract
Address types and Transaction Types.

Time
Temporal aspect.

Gas Used and Price
.



Transactions Sent and Received
Bi-directional graph.

Value
Amount of Ether in the transactions.

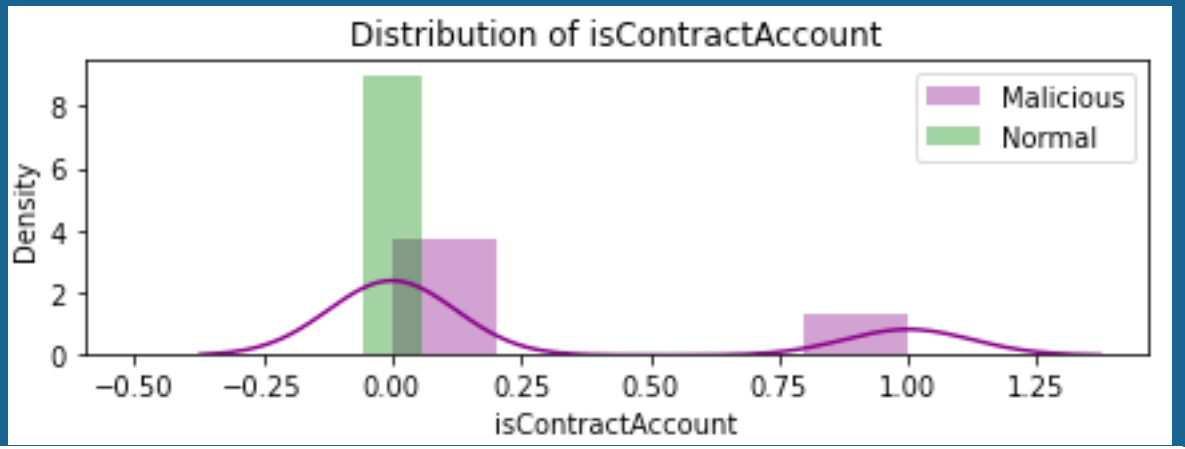
Failed and Error Transactions

Exploratory Data Analysis

Account

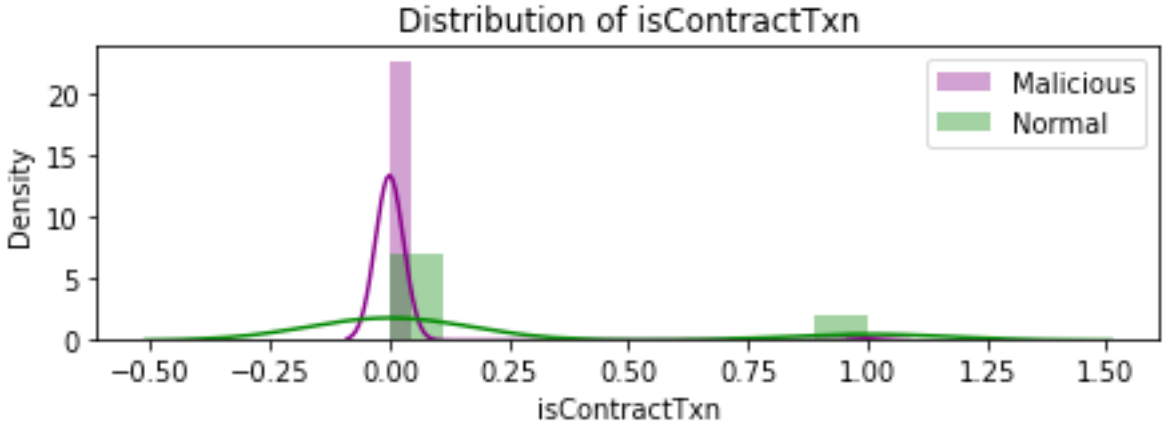
Smart Contract (SC)

Externally Owned (EOA)



Many of malicious addresses are not SC accounts, SC accounts can be malicious.

Malicious EOA accounts tend not to run on smart contract.



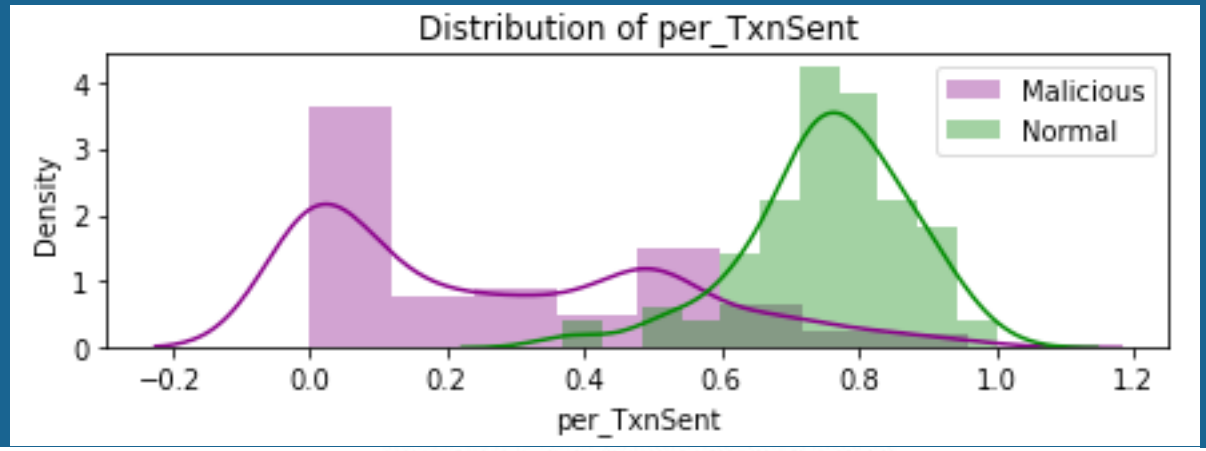
Exploratory Data Analysis

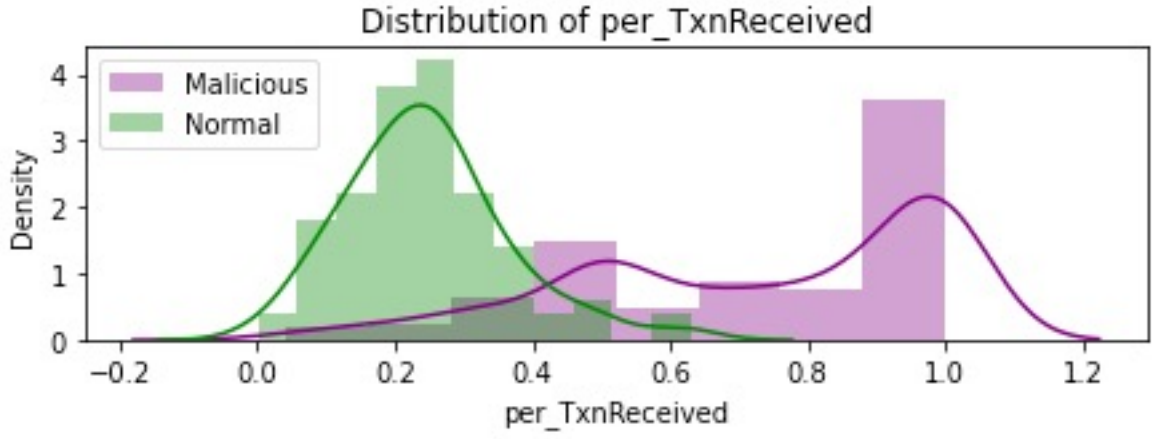
Transactions

Sent

Received

Malicious addresses have much more transactions received than sent, compared to normal addresses.



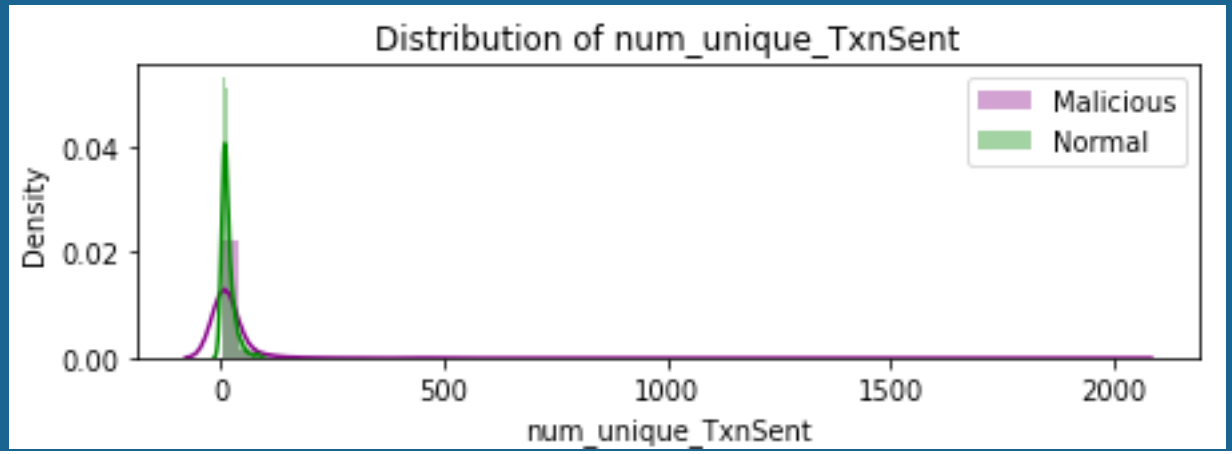


Exploratory Data Analysis

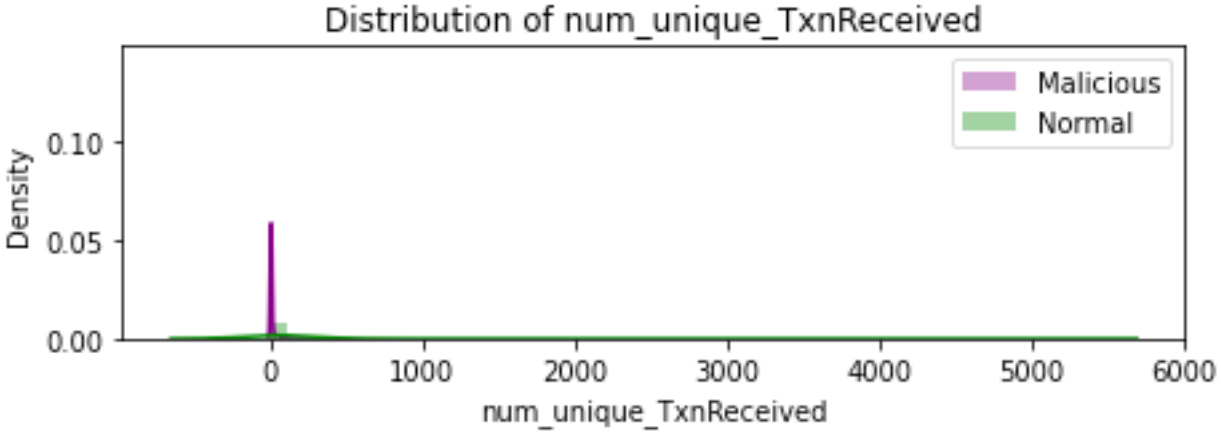
Unique Transactions

Sent

Received



Malicious addresses tend to send transactions to less unique addresses, but receive transactions from more unique addresses.

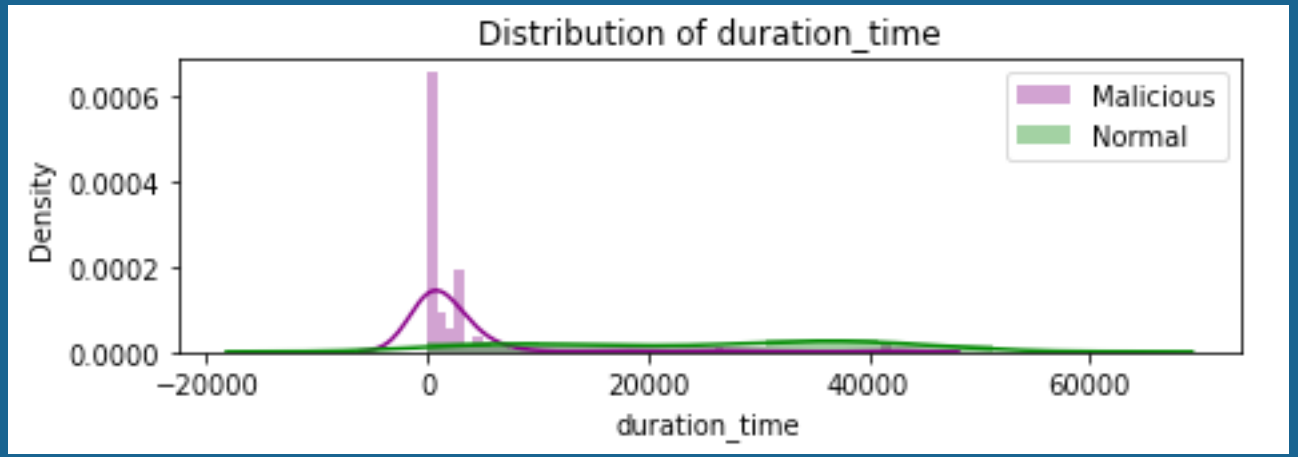


Exploratory Data Analysis

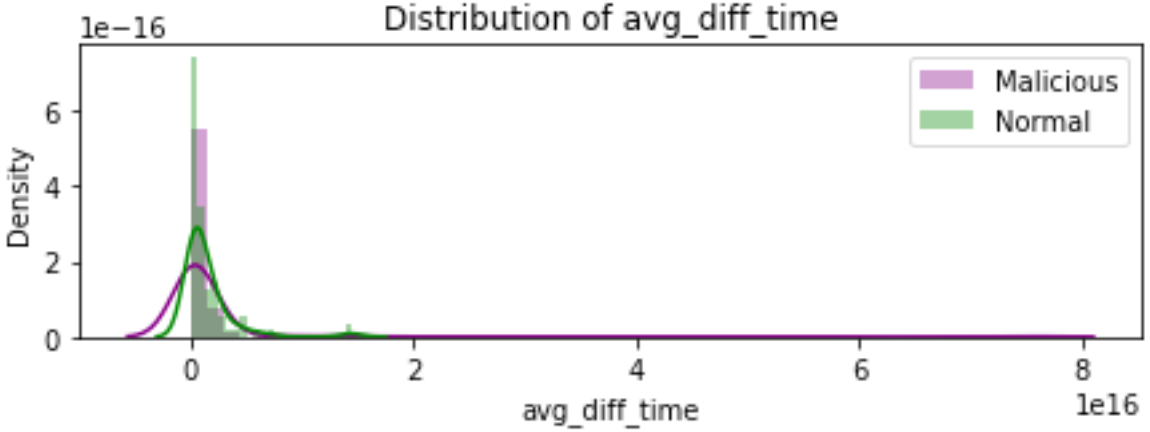
Time

Duration

Difference

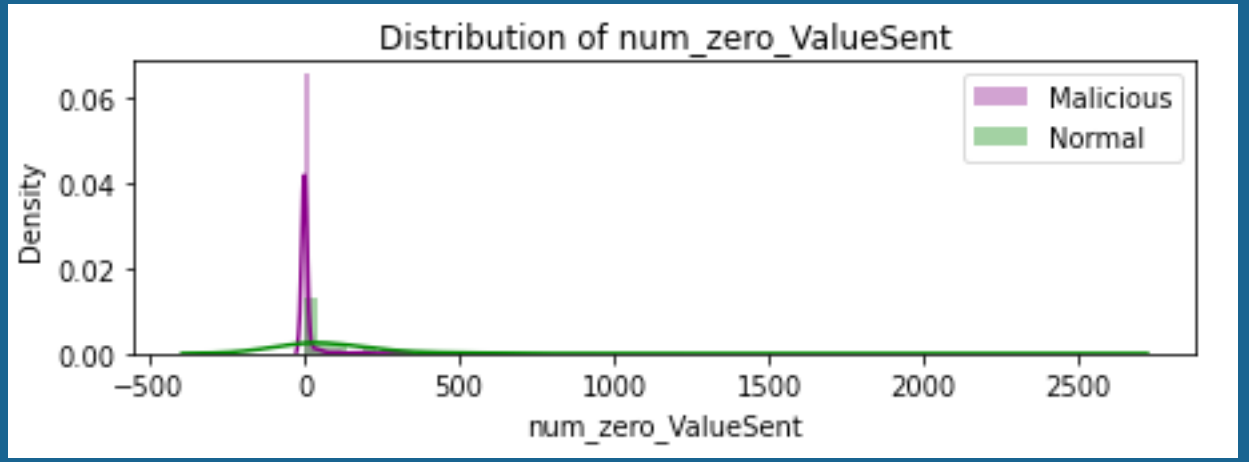


Malicious activities last shorter and with short intervals.

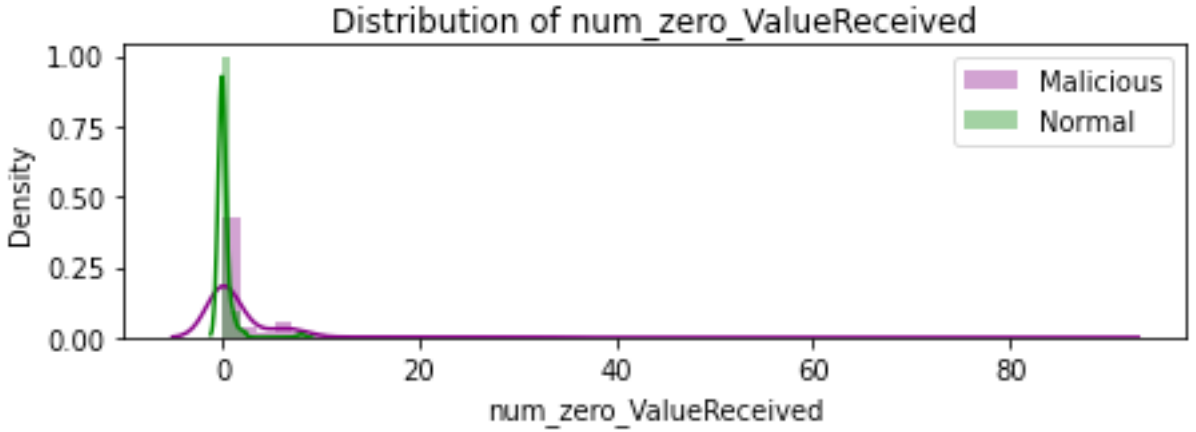


Exploratory Data Analysis

Value
Sent
Received

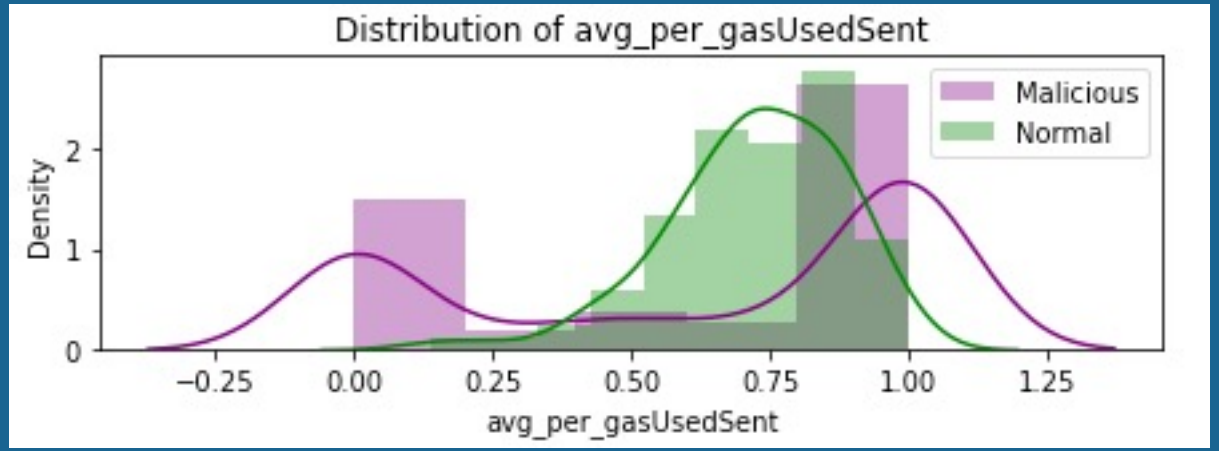


Malicious addresses send more zero value transactions, normal addresses receive more.

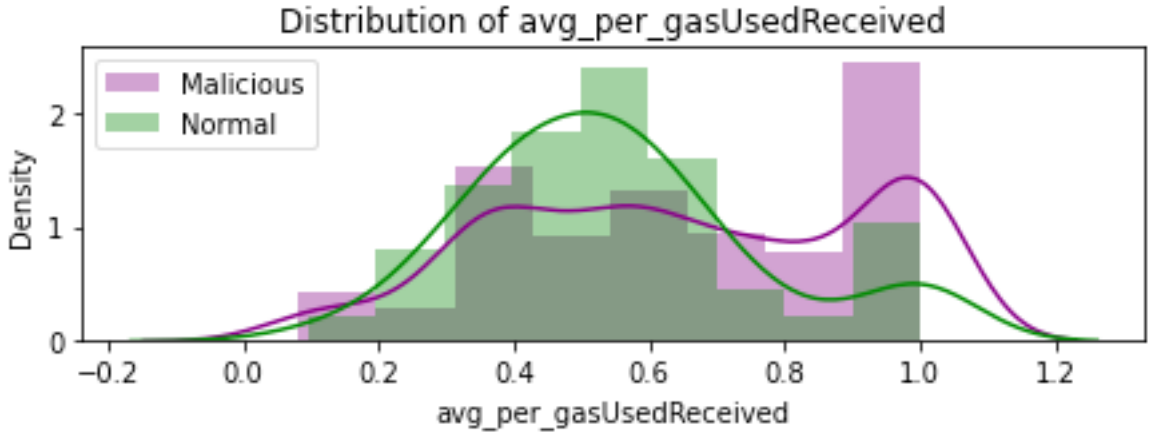


Exploratory Data Analysis

Average Percentage Gas Sent Received

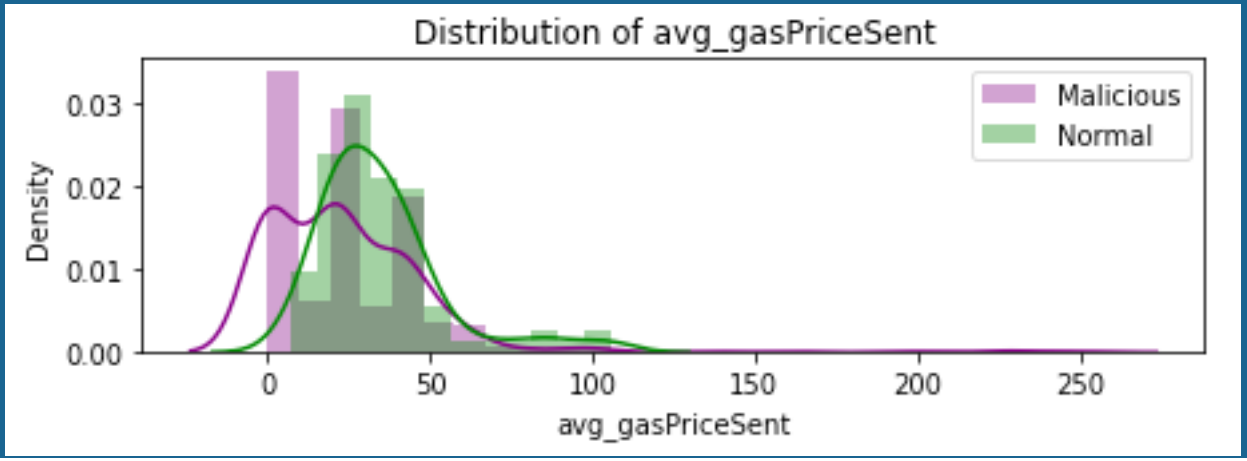


Malicious addresses tend to use the upper limit of the gas.

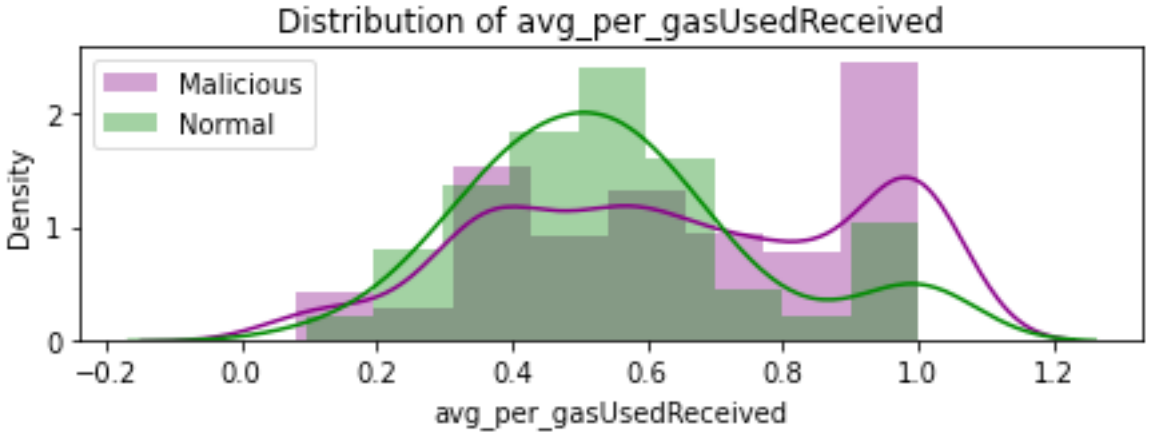


Exploratory Data Analysis

Average Gas Price Sent Received

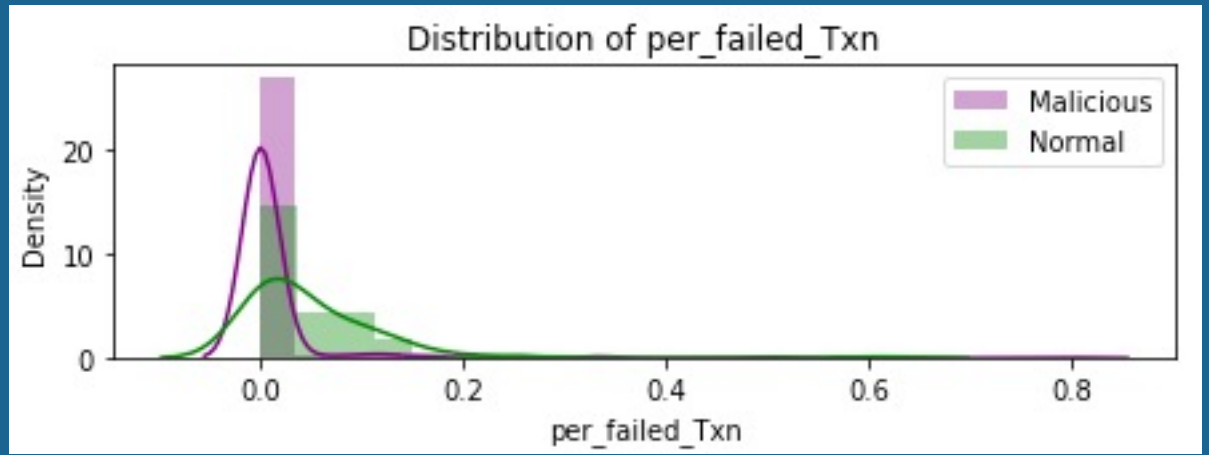


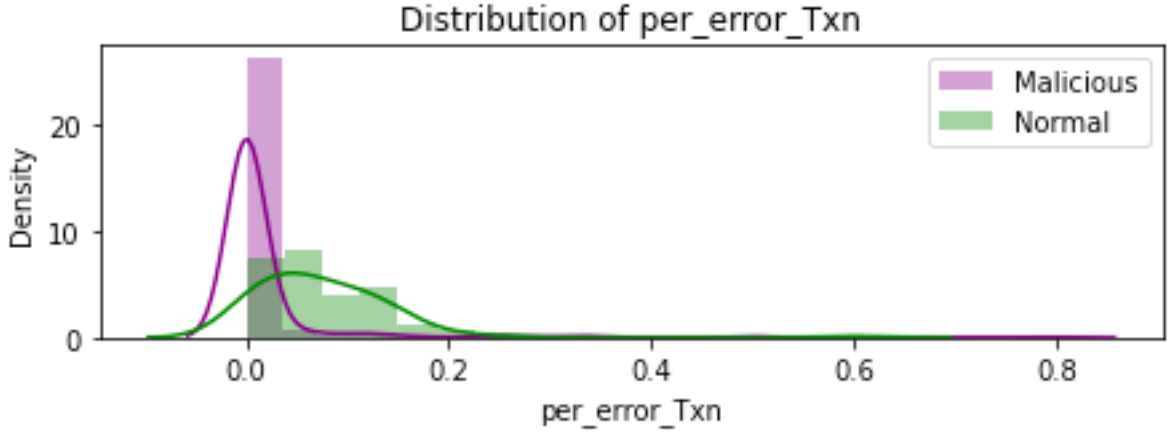
Malicious addresses set the gas price to be lower when sending the transactions, but when they receive transactions, it's much higher.



Exploratory Data Analysis

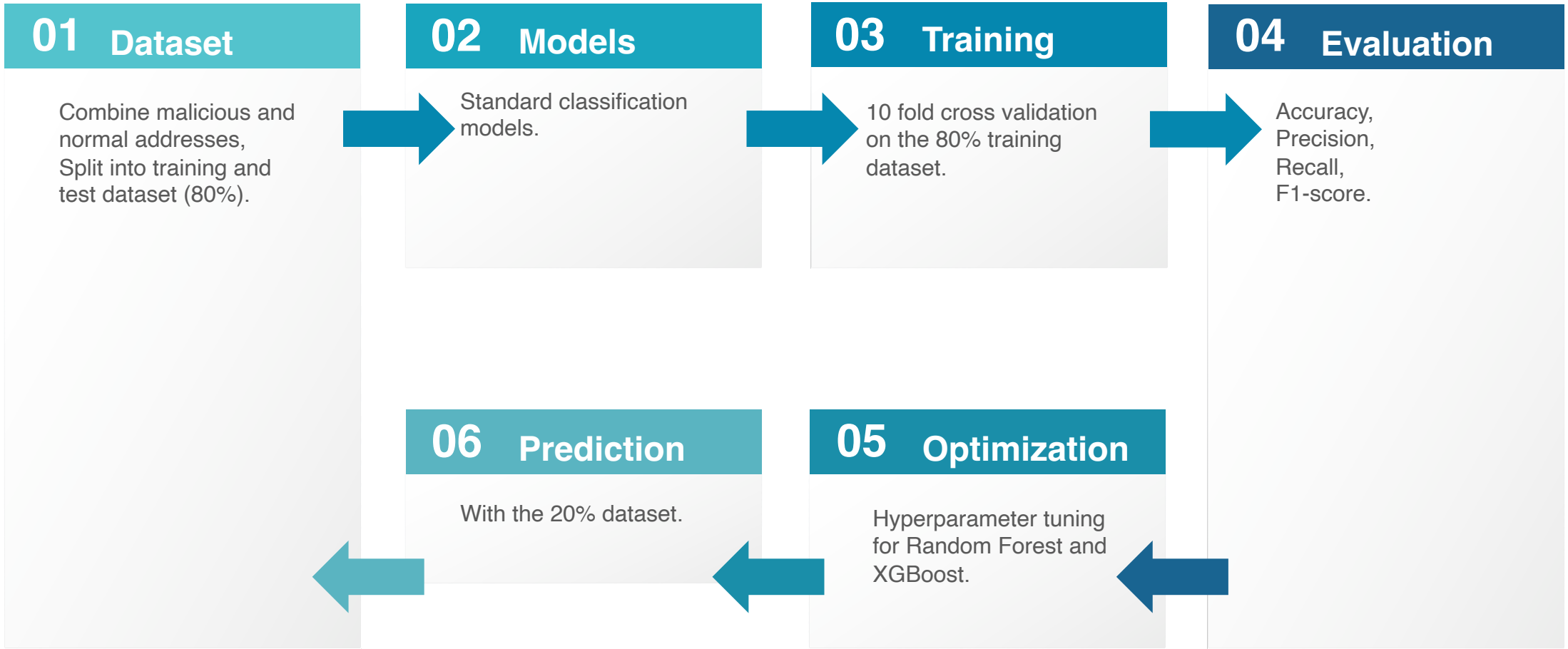
Transactions
Failure
Error





Malicious transactions are less likely to have error or fail.

Modelling


01 Dataset
Combine malicious and normal addresses, Split into training and test dataset (80%).
02 Models
Standard classification models.
03 Training
10 fold cross validation on the 80% training dataset.
04 Evaluation
Accuracy, Precision, Recall, F1-score.
05 Optimization
Hyperparameter tuning for Random Forest and XGBoost.
06 Prediction
With the 20% dataset.

Performance

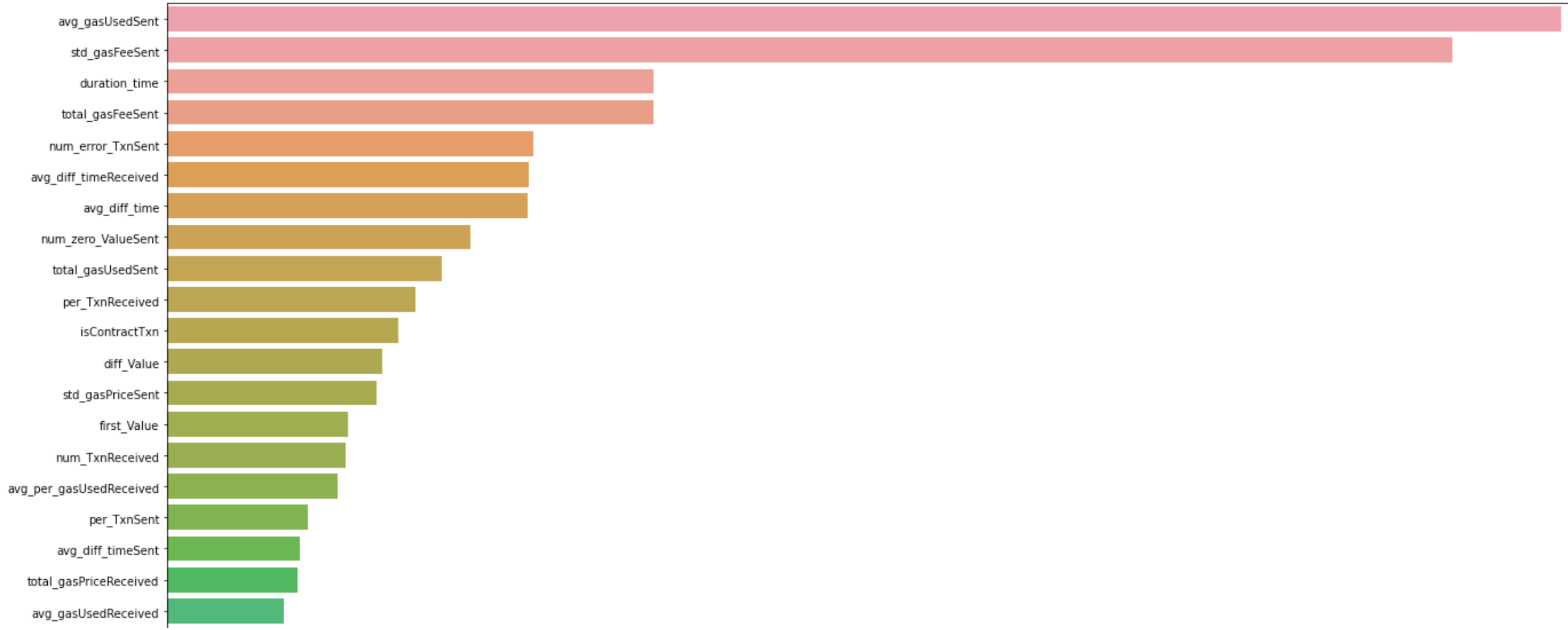
	val_accuracy	val_precision	val_recall	val_f1	test_accuracy	test_precision	test_recall	test_f1
Logistic Regression	0.858824	0.955820	0.880354	**0.914815**	0.875000	0.942857	0.908257	**0.925234**
SVM	0.872549	0.873333	0.997727	**0.931354**	0.851562	0.851562	1.000000	**0.919831**
KNN	0.888235	0.923547	0.950253	**0.936462**	0.859375	0.902655	0.935780	**0.918919**
Random Forest	0.968627	0.975648	0.988737	**0.982070**	0.945312	0.955357	0.981651	**0.968326**
XGBoost	0.966667	0.973478	0.988687	**0.980974**	0.945312	0.963636	0.972477	**0.968037**

Random Forrest and XGBoost outperform other models.

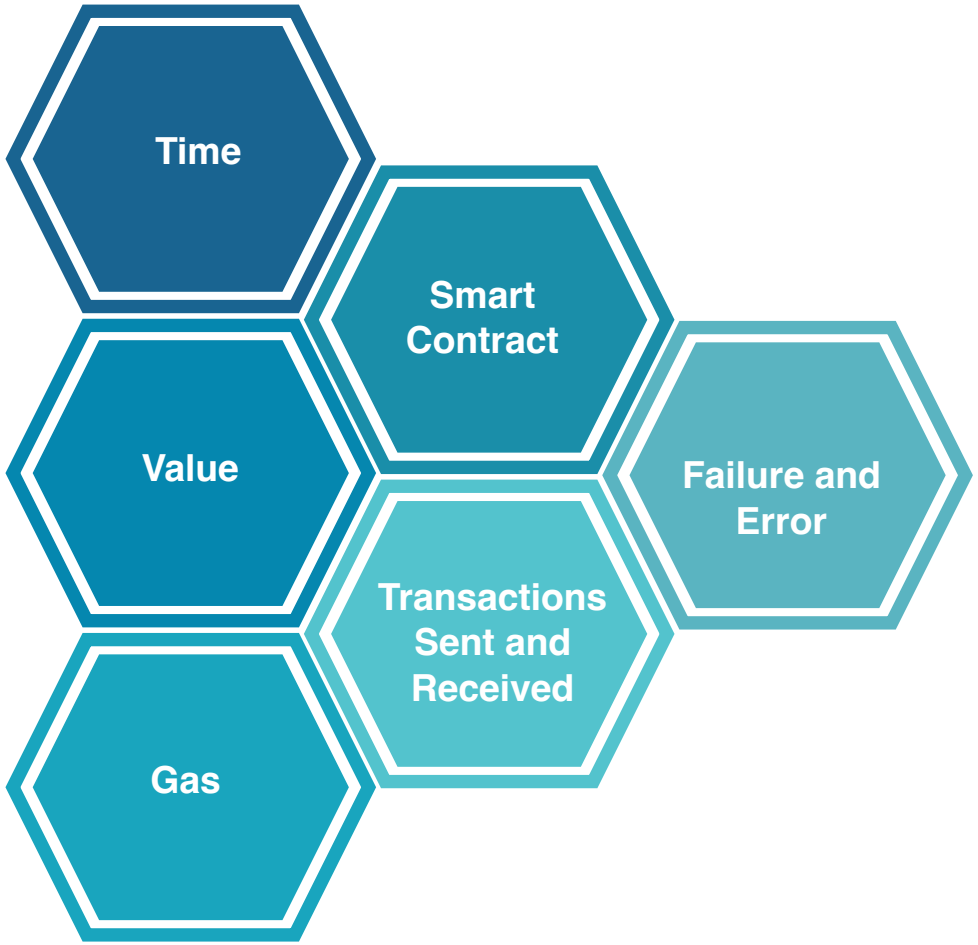
-> Hyperparameter tuning

-> Confusion Matrix: TP=107, TN=15, FP=4, FN=2

Feature Importance



Solution & Discussion



- Vulnerabilities of Smart Contract
- Phishing Scams
- Detection of Abnormal Value

Task 2

Bad Suppliers Detection

Bad Suppliers Detection

Supply Chain for a Food Company

Time
Amount
Quantity

Construct features of good suppliers, manifest larger errors for bad suppliers.

Supplier
Location
Time
Item
Amount of items
Item Price
Returned amount

Class imbalance, majority of suppliers are good, abnormal behaviors are unpredictable.

Weather, Transport, Global Pandemic…


THANK YOU