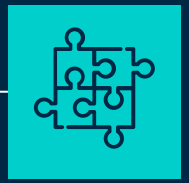
Ethereum Fraud Detection

Letian YU (FTEC)
The Chinese University of Hong Kong

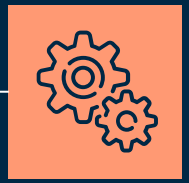
TABLE OF CONTENTS



01

PROBLEM & Data Description

- Objective
- Dataset



02

Feature Engineering

- Feature Engineering
- Data Analysis

03

Model Result & Future Improvement

- XGBoost Result
- Improvement

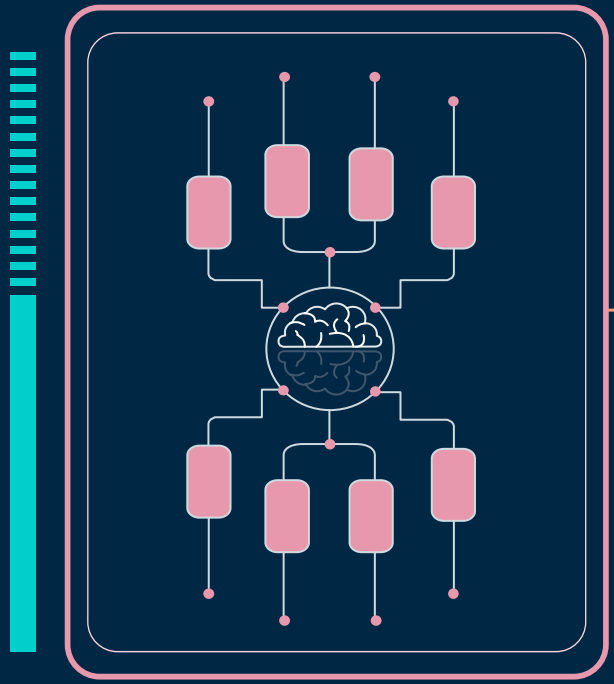
PROBLEM & Dataset

01

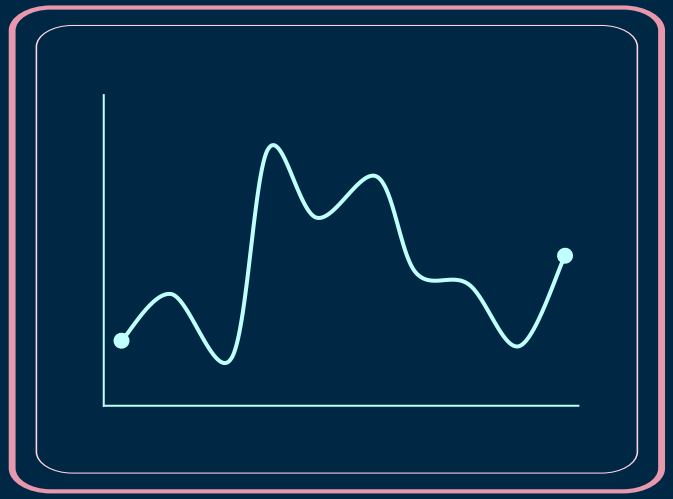
Objective & Dataset

Goal:
Detect fraudulent Ethereum address

Dataset:
train_accounts.csv
test_accounts.csv
transactions.csv



Data Description



Train Accounts:

Fraud	Non-fraud	Total
2455	22743	25198

Test Accounts:

Total	6300

Transactions:

Total Records	Unique Sender Account	Unique Receiver Account
5826604	604847	419535

Transaction Data

Feature Datatype:

#	Column	Dtype
0	from_account	object
1	to_account	object
2	transaction_time_utc	object
3	value	object
4	gas	int64
5	gas_price	int64

Dataset Preprocessing:

- Convert transaction time to date, year, year_month
- Convert value (large number) to the magnitude of digit, add flag to indicate whether the transaction is token or not
- Convert unit of gas price from GWEI to ETH (to reduce the scale of data)
- Calculate gas fee based on gas and gas price

Sample:

from_account	to_account	transaction_time_utc	value	gas	gas_price
a20151	b966524	2020-05-04 13:21:32	13000000000000000	21000	1080000123
a25907	b31505	2020-05-04 13:22:10	0	1500000	1200000000
a20151	b31501	2020-05-04 13:22:10	0	60000	847000023

Transaction Data

	gas	gas_price	transaction_year	gas_fee	is_token	value_digit
count	5826604.00	5826604.00	5826604.00	5826604.00	5826604.00	5826604.00
mean	245096.38	55.44	2019.30	15262534.33	0.63	7.17
std	537469.03	208.43	1.05	96734096.46	0.48	8.11
min	21000.00	0.00	2016.00	0.00	0.00	1.00
25%	50000.00	6.00	2019.00	459000.00	0.00	1.00
50%	90000.00	20.00	2020.00	1890000.00	1.00	1.00
75%	250000.00	60.00	2020.00	9000000.00	1.00	17.00
max	12022226.00	171397.02	2021.00	69239660614.74	1.00	23.00

Feature Engineering

02

KEY IDEA:

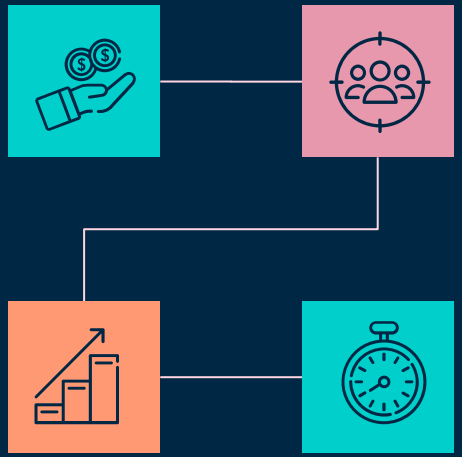
Construct feature for each account in the train-test list

SOLUTIONS

Construct features for accounts in the list (as sender & as receiver)

Gas

Max, min, mean, std for gas price, gas fee, and gas



General Info

of transactions,
# of year coverage,
Sender/receiver fraud or not

Value

of token transactions,
Max, min, mean, std
for value scale

Transaction Count

Max, mean, std of transaction count by year, month, date

Feature Engineering

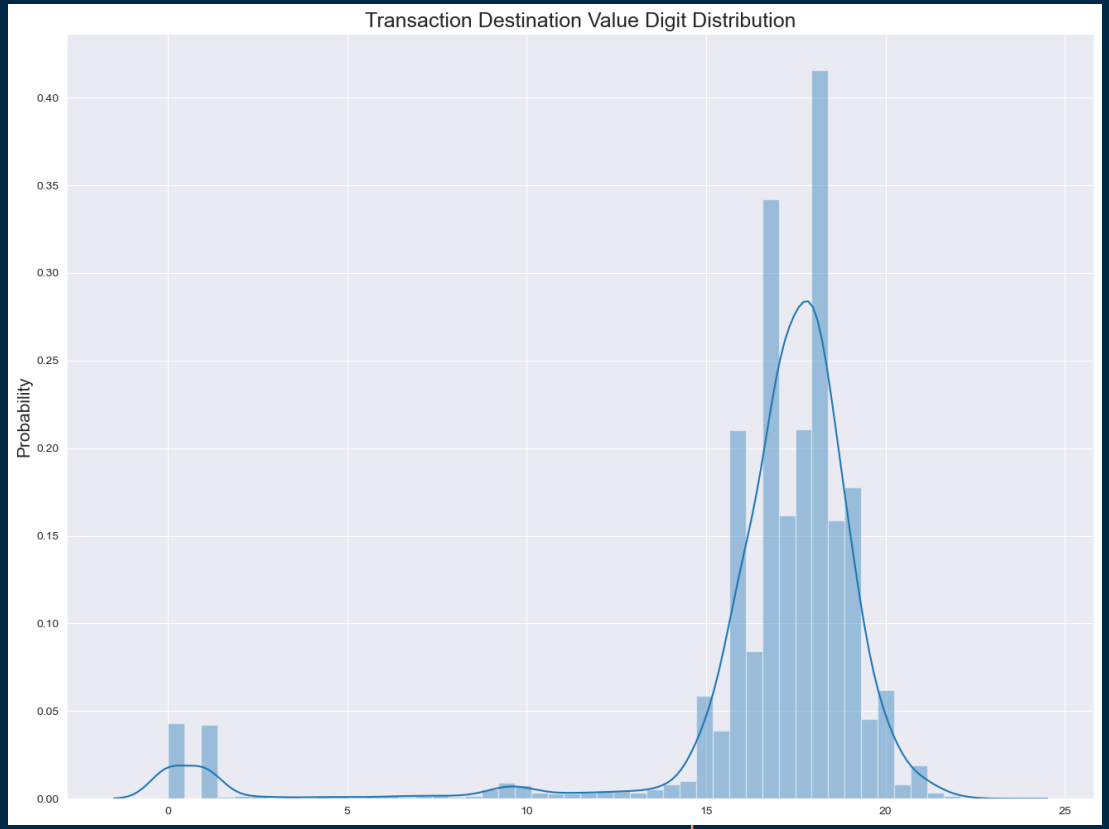
Constructed features: 57

Further processing:

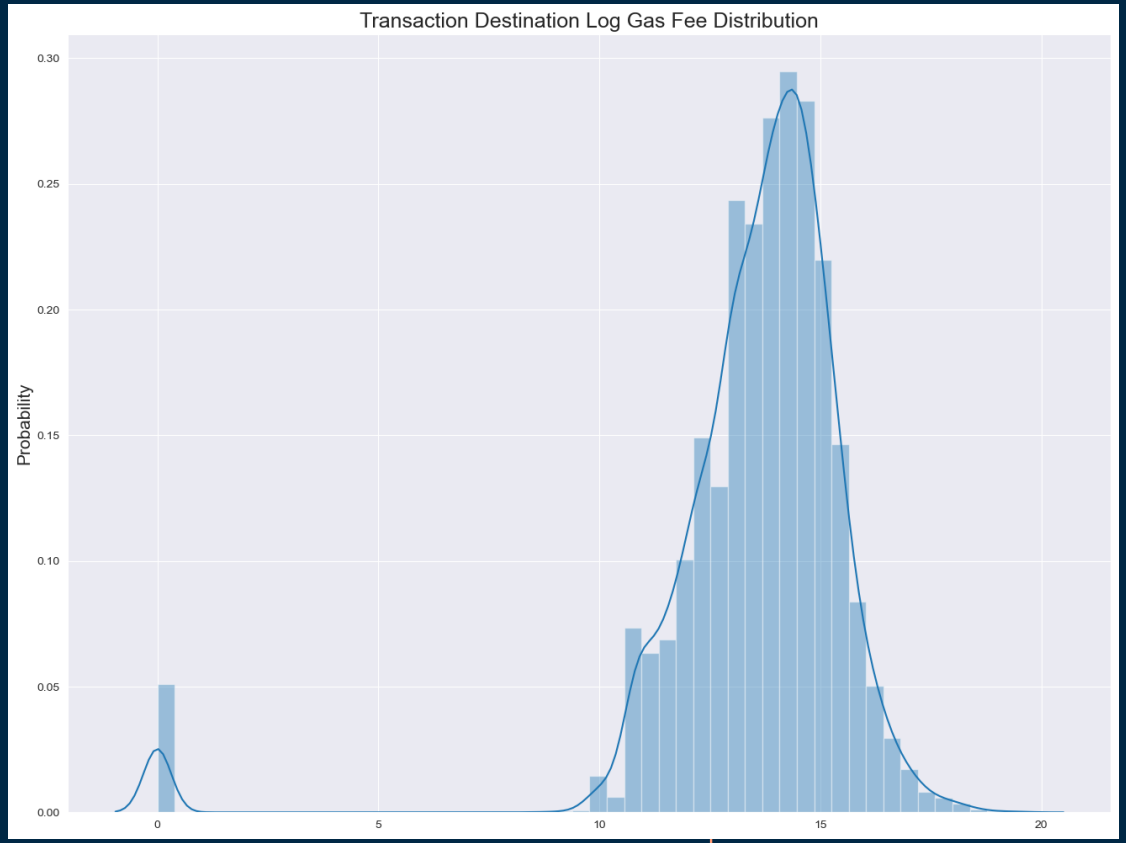
- Fill null-values
- Log-transformation for skewed distribution with long tail

Explorary Data Analysis (Sample)

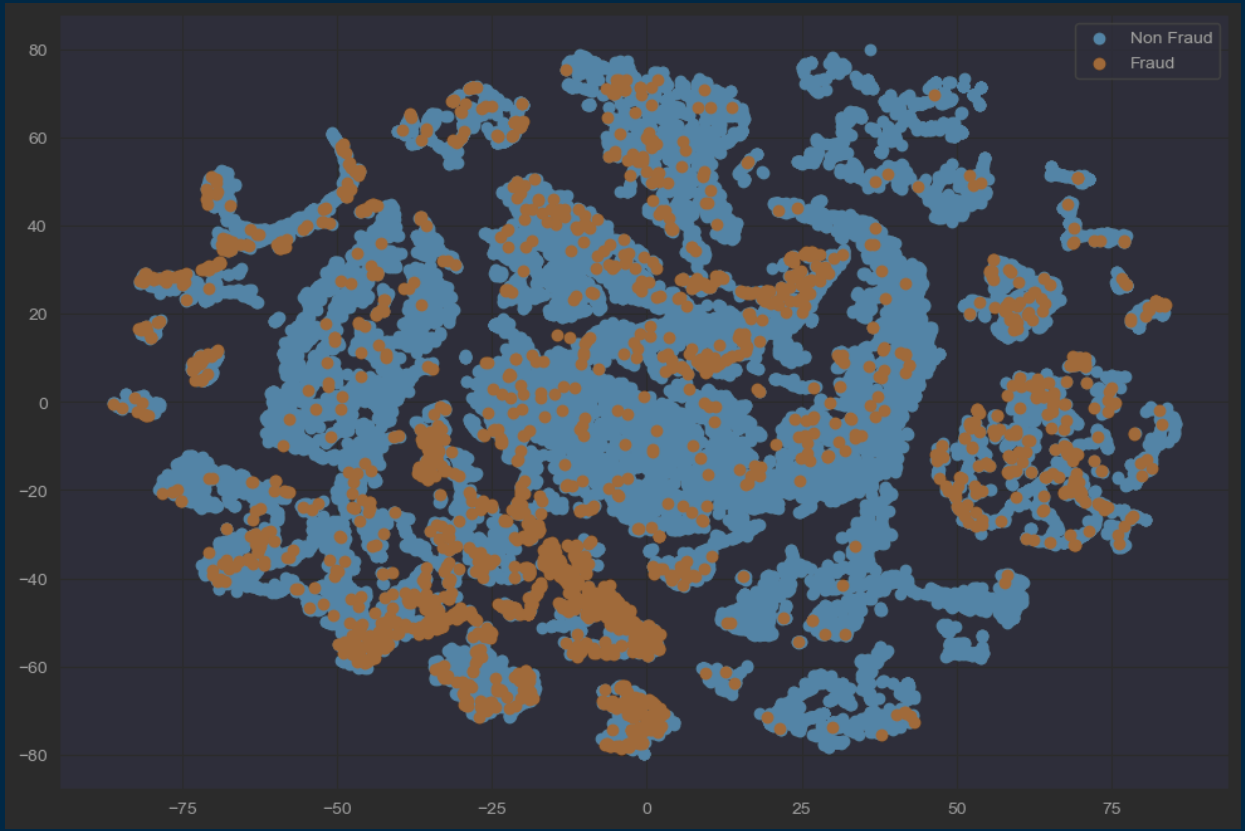
Receiver Value Log Sacle Distribution



Receiver Log Gas Fee Mean Distribution



Sampled T-SNE Visualization (2D)



Model Result

03

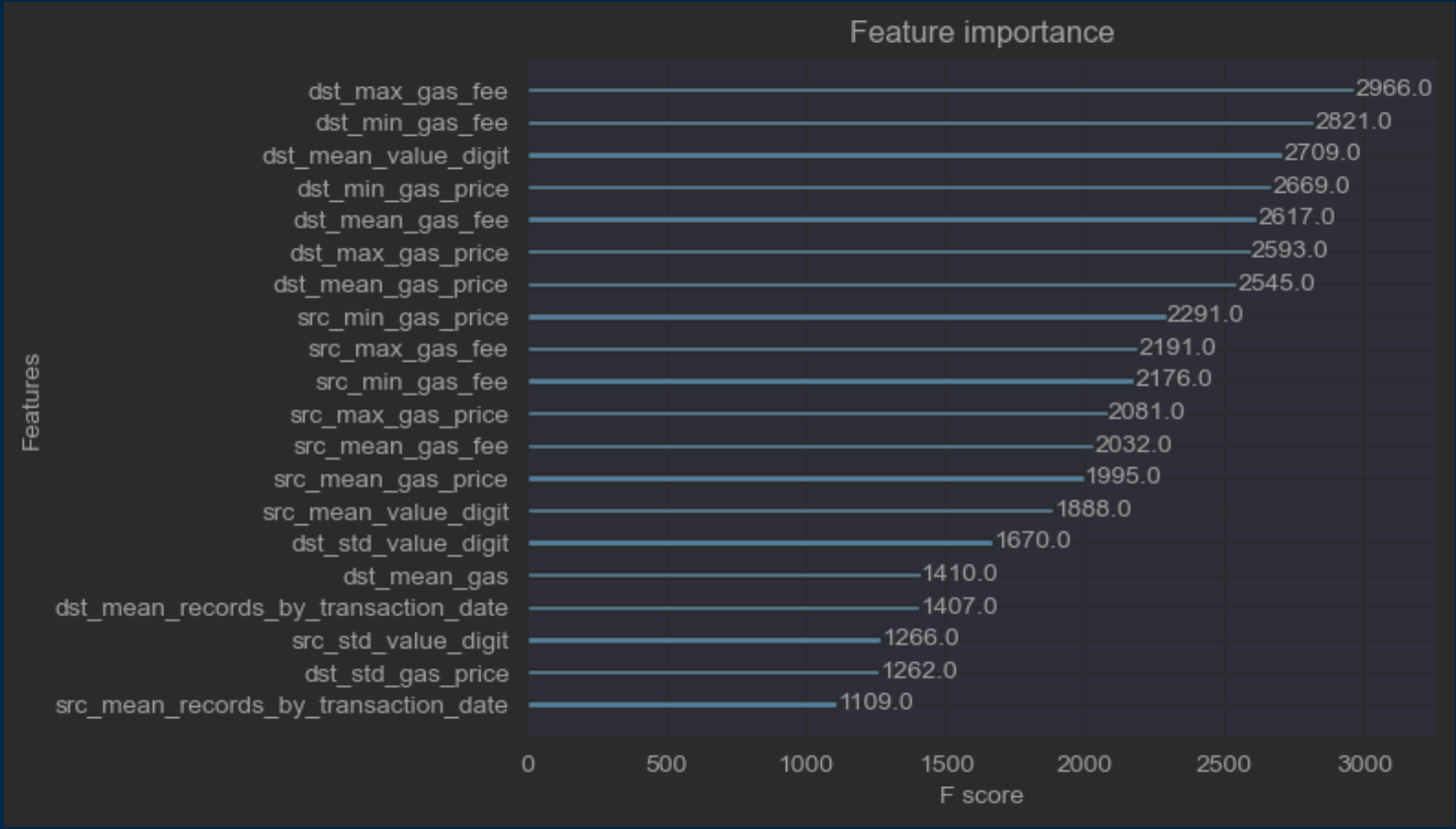
Model: XGBoost

- Grid-search for parameter tuning
- Stratified-kfold cross validation to avoid overfitting
- Evaluation Metrics: f1-score (Best: 0.7531 on training data)

- Tuned parameters: learning rate, n estimators, max depth, subsample, gamma

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$$

Feature Importance (by f-score)



Feature Importance (built-in function)

feature	importance
src_dst_account_fraud_or_not	0.140521
dst_mean_records_by_transaction_date	0.075922
dst_year_count	0.060107
src_min_value_digit	0.046148
src_mean_gas	0.041872
src_std_records_by_transaction_year	0.035083
dst_std_records_by_transaction_month_year	0.027341
dst_std_records_by_transaction_year	0.025232
dst_min_gas	0.023041
src_max_gas	0.020630
dst_total_transactions	0.020091
src_total_transactions	0.017963
src_max_value_digit	0.015503
dst_std_value_digit	0.014914
dst_std_gas_price	0.014847

Model Comparison: Logistic Regression

	CV	Tuning	F1-score
XGBoost	Yes	Yes	Around 0.75 on training set
Logistic Regression	Yes	Only iter_num	Aroud 0.59 on training set

Future Improvement

Features:

- More features using transaction time data
- Construct features with more information (e.g., # of token to pct of token)

Dataset:

- Data scale not balanced
- Need to carefully check data quality and feature meaning

Model:

- Try deep learning model as alternative comparison

Task:

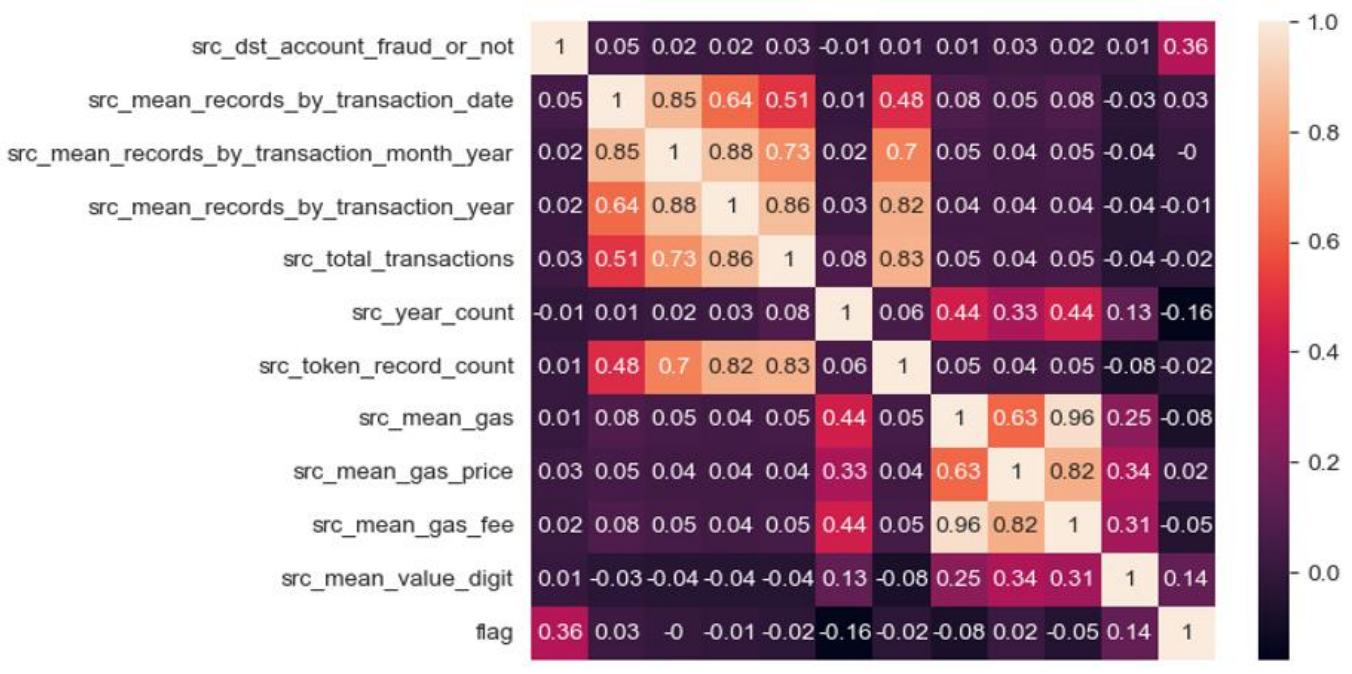
- Change to classification task on transaction record (need to design the metrics more carefully)

THANKS!

Appendix

Feature Correlation

Correlation Analysis (Pearson coefficient)



Correlation Analysis (Pearson coefficient)