CRYPTOCURRENCY (ETHEREUM) FRAUD DETECTION

GA DSI 33 - Jimmy Ong

AGENDA

01 ABOUT

- Introduction
- Problem Statement

02 PREPARATION

- Data Cleaning
- EDA
- Feature Selection

03 MODEL

- Modeling
- Model Evaluation

04 SUMMARY

- Conclusion
- Recommendation

01

ABOUT

- Introduction
- Problem Statement

01 INTRODUCTION

What is Cryptocurrency?

- Digital coins and tokens
- Real-world value
- Value has been increasing over the years

01 INTRODUCTION

Total Value of Cryptocurrency

- 3 Trillion USD (2021)
- 7x of Singapore GDP

01 INTRODUCTION

Fraud Cases

- 10 Billion USD lost
- 15x most expensive private jet

01 PROBLEM STATEMENT

Many Fraud cases from ethereum.

As a investor myself, I want to :

- Know insights on Fraud
- Main Features of Fraud

To reduce chances of getting scam by frauds

02

PREPARATION

- Data cleaning
- EDA
- Feature Selection

02 PREPARATION

DATA CLEANING

- Removed duplicates
- Tidying up data for EDA

FEATURE SELECTION

- Filter Method
- Embedded Method

EDA

- Heatmap
- Feature Importance
- Bar plot
- Feature Impact

02 FEATURE SELECTION

FILTER

- Uses correlations between variables
- Eg. Heatmap

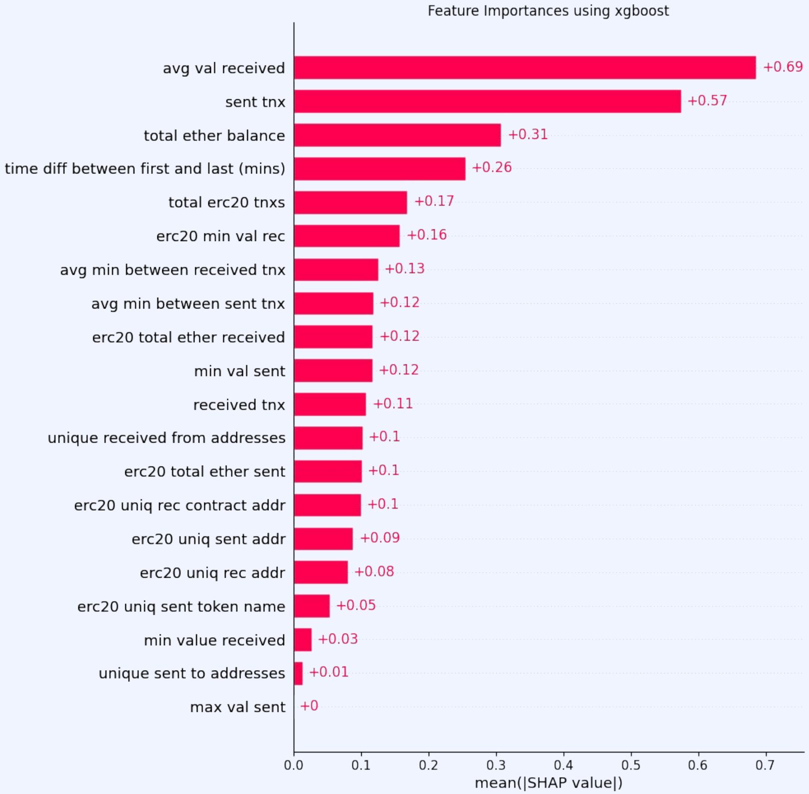
EMBEDDED

- Uses machine model to select features
- Eg. Feature Importance

02 FEATURE SELECTION

Selected 19 features for modeling

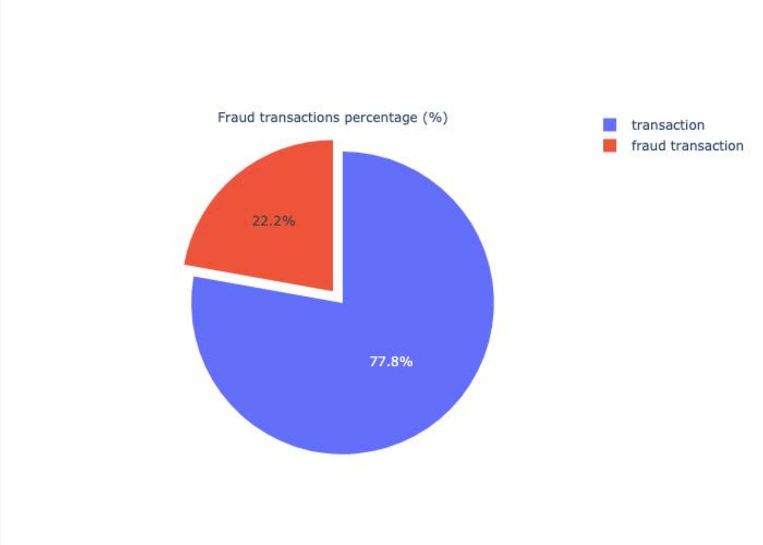
- Reduction of features by 62%
- Improve computation efficiency



02 EDA : FRAUD TRANSACTIONS

Fraud Transactions

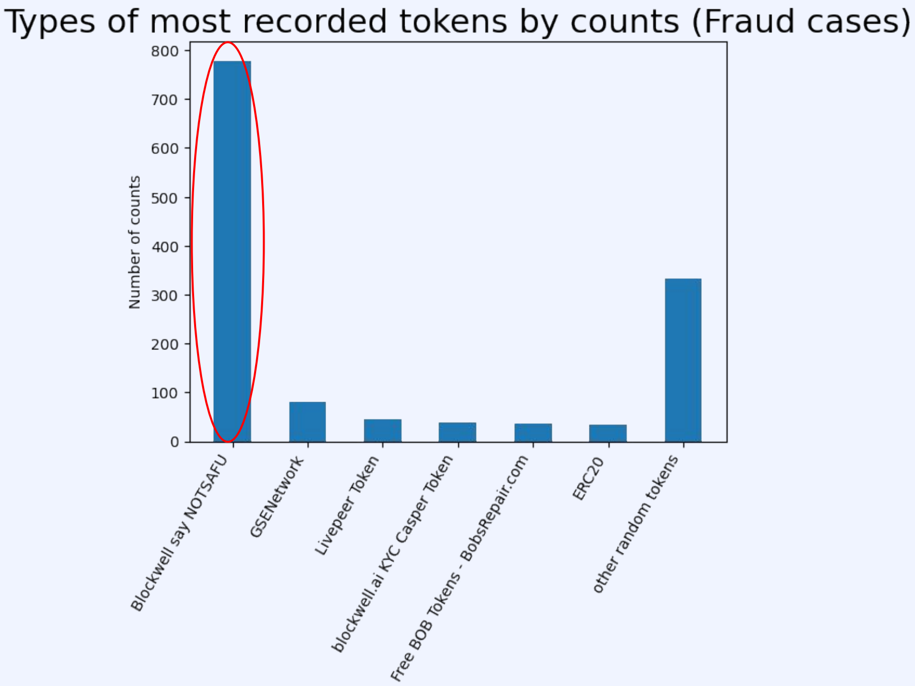
- 22% of total transactions



02 EDA : FRAUD TOKENS

Common traits Fraud transactions in tokens

- Use of famous crypto names

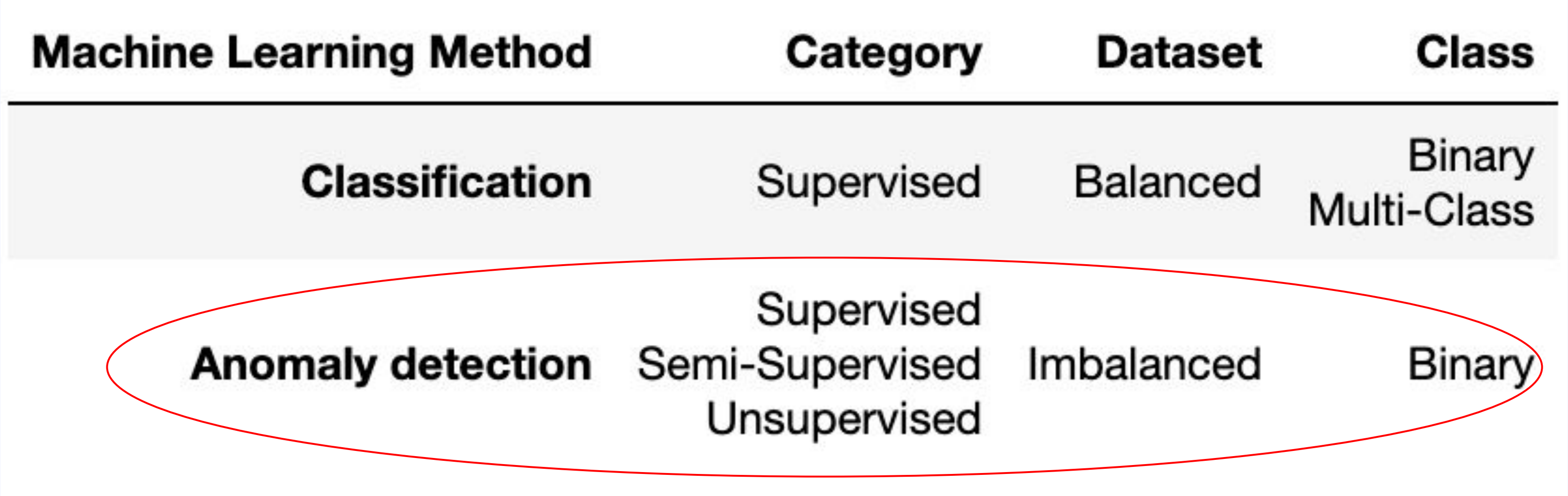


03

MODEL

- Modeling
- Model Evaluation

03 MODELING : MACHINE LEARNING METHODS



Machine Learning Method	Category	Dataset	Class
Classification	Supervised	Balanced	Binary Multi-Class
Anomaly detection	Supervised Semi-Supervised Unsupervised	Imbalanced	Binary

03 MODELING : RECALL VS ACCURACY SCORE

Recall Score : How many times the model correctly identify True Positive(Fraud)

Accuracy Score : How many times the model made correct predictions

Precision Score : How many times the model correctly predict positive class

03 MODELING : PR AUC VS ROC AUC CURVE

PR AUC curve : Focus on precise and recall scores

ROC AUC curve : Focus on accuracy scores

03 MODELING : OTHER STEPS

Smote - for oversampling training dataset

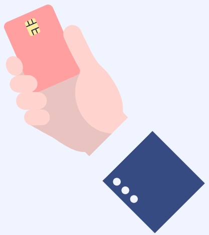
Stratified K-fold - for cross validation due to imbalance dataset

Recall Score - for showing the proportion of true anomalies identified

PR AUC - Better representation for imbalance dataset

ROC - Better for balanced dataset

03 MODELING



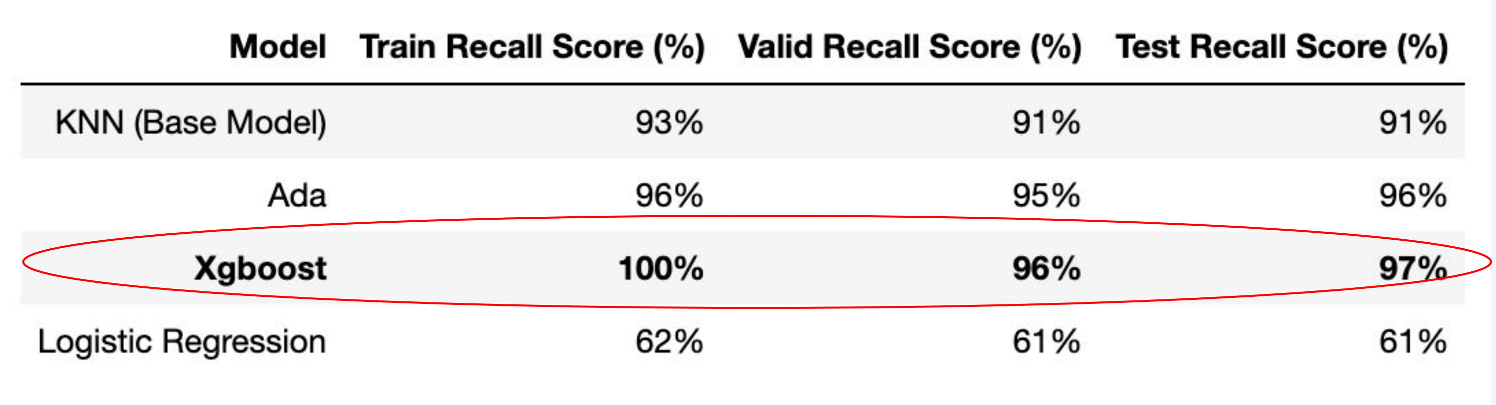
Logistic Regression

KNN

ADA

Xgboost

03 MODEL EVALUATION SCORE

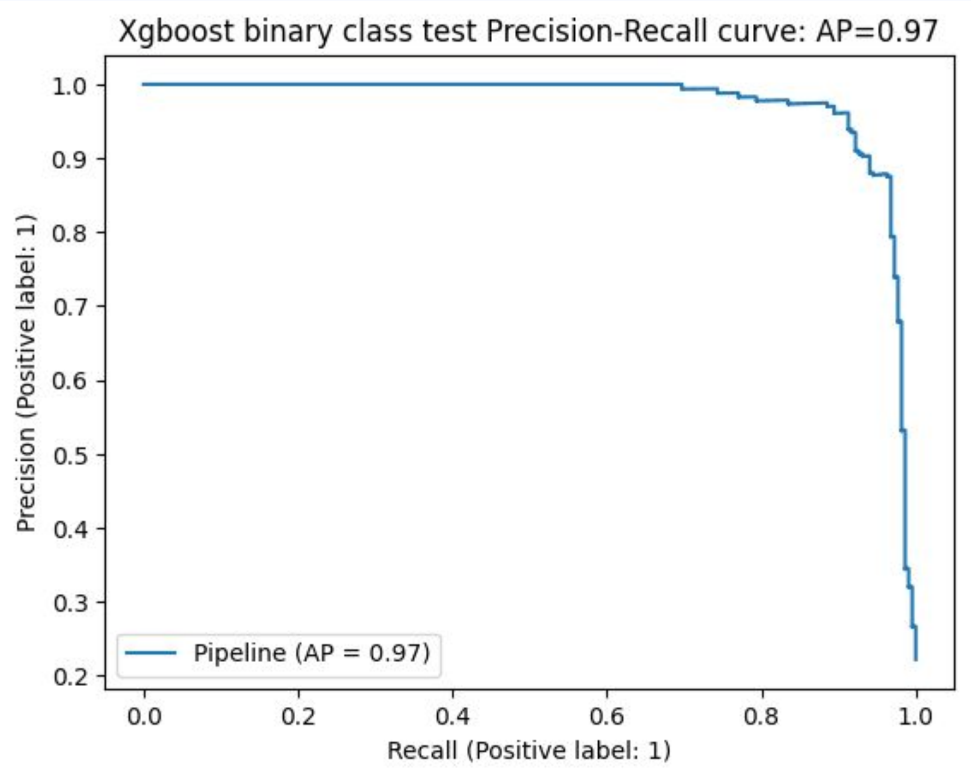


Model	Train Recall Score (%)	Valid Recall Score (%)	Test Recall Score (%)
KNN (Base Model)	93%	91%	91%
Ada	96%	95%	96%
Xgboost	**100%**	**96%**	**97%**
Logistic Regression	62%	61%	61%

For every 100 fraud transactions, around 97 of frauds are detected.

03 MODEL EVALUATION : BEST PR CURVE

- Fill most of the area under curve
- Lesser mistakes made for identifying non-fraud as fraud



04

SUMMARY

- Conclusion
- Limitations
- Recommendations

04 CONCLUSION

97%

MODEL SCORE

22%

OF TRANSACTIONS ARE FRAUD

63%

Reduction of features

FAMOUS CRYPTO NAMES

ARE USED IN TOKEN FOR FRAUD CASES

04 LIMITATIONS

OUTDATED DATASET

Data might not be relevant as crypto industry changes at a fast pace

INSUFFICIENT DATA

Dataset has around 9000 transactions compared to millions of transactions per day

04 RECOMMENDATIONS

BLACKLIST

TOKENS THAT USES
FAMOUS CRYPTO NAMES

THE END