Group 14: Ethereum Prediction

MSDS20080, MSDS20048, MSDS20060

Google Drive Link: https://drive.google.com/drive/folders/1ZHoUnJZkYXuDes-RT4xLft9kMwLWfWQj?usp=sharing

1. Objectives

Find out the pattern or useful insight from Ethereum Dataset (Google Big Query Public Ethereum Dataset). Objectives of semester project are;

1. Analysis and Visualization.
2. Future transaction prediction (Link Prediction).
3. Cluster the same user address by analyzing the Transactions of user.

2. Challenges

While doing the project, there are many hurdles we faced, which are following below;

1. Download the dataset, as one day transactions of Ethereum is approximately greater than 1 GB.
2. Analysis and apply the ML model for future prediction (link prediction and cluster), due to limited dataset consequence of under-fitting model and also visualization problem.
3. Environment Limitation (LOCAL and COLAB) - 12GB is not enough to train the model in Free tier plan on large dataset.

3. Literature Review

- Detailed analysis of Ethereum network on transaction behavior, community structure and link prediction
- Wealth Distribution and Link Predictability in Ethereum
- Ethereum transaction graph analysis
- Node2vec: Scalable Feature Learning for Networks

4. Analysis on Ethereum Dataset

We have performed the analysis on Etherum dataset. Different types of Top queries executed and visualization in graph e.g.

- Who has large amount of Ethereum amount balance.
- Blocks created over the Time i.e. day, month and year.
- Top miner, who has created the blocked.
- Transaction over the Time i.e. day, month and year.
- Who has made number of transaction.
- Who has received number of transaction.
- User transaction activity.

5. ML models and its prediction score

Use the Networkx library of python for Graph Representation & Node2Vac for Graph Embedding and different types of ML model applied

- Logistic Regression: 86% AUC
- Random Forest Classifier: 82% AUC
- Gradient Boosting Classifier: 81% AUC

While in clustering there is issue due to limited transactions data there is no similar transaction exists. Zero clusters against each address/node showing in graph node.

6. Conclusion and future work

Due to machine resources limitations we are unable to train the ML model on large data dataset however, we train the model on limited dataset as per available resource. In addition, there is still require advance DL model i.e.

- Advance Deep Learning model with memory efficient
- Custom Framework designed base on Ethereum data (Batches/parallel programming)
- Require High Computation Power & Time Require.