Ethereum investor profiler

Capstone Proposal

Domain background

This project explores the domain of crypto investor profiling. For every transaction that is executed on the public blockchains, there is a trace that we can analyze. As opposed to other asset classes, we have available data into the behavior of individual investors. How can we derive valuable insight into their behavior? Or to go even further, can we use the principle of BYOD (bring your own data) and cluster a user based only on a provided Ethereum address?
Personally, I am curious about crypto-economic systems and machine learning. I would like to build a data product that combines both.

Problem statement

While crypto data on public blockchains are open, there are not many ML applications yet that leverage the available data. One of the reasons is that the underlying data structures are based on OLTP systems which are difficult to analyze (as opposed to OLAP). Recently, the data has been made available on Bigquery and is ready to be used.
The challenge is in modeling the data and using it to feed the ML models. Also, the data is not labeled which is also why the problem is formulated as an unsupervised machine learning task.
As pricing data is the most accessible, most of the crypto machine learning projects are focused on predicting future prices using that. On the other hand, this project is focused on getting a deeper insight into investor behavior.
The problem can be formulated as follows: Given a set of Ethereum investor addresses, how much variance can be explained by splitting them into N buckets?

The datasets and inputs

The data used in Ethereum's blockchain data which is available as a public BigQuery dataset. This is an input from which we can build relevant features to differentiate between different investor behavior. Individual Ethereum addresses' features are fed into the ML model. The dataset can be accessed here.

A solution statement

The final result is a web app that returns an investor profile based on the user's provided Ethereum address. The user doesn't need to provide any of their own personal data as all data that is needed can be extracted from the publicly-available Ethereum data. Besides that, we will evaluate the explained variance for different algorithms and choose the one that performs the best.

A benchmark model

With a given number of clusters, the benchmark model will be to classify every Ethereum address into the most common cluster. Then to improve on that more clustering algorithms can be used such as K-means, Density-based clustering, and Hierarchical clustering.

A set of evaluation metrics

The challenge with the task is that there are no ground truth labels to use for objective evaluation.
There are two possible paths to take (or both):

- Use percentage explained variance (PVE) metric on the chosen converged clustering algorithms
- Use an evaluation metric such as Silhouette Coefficient which also doesn't demand known ground truth labels

An outline of the project design

- Feature engineering (data modeling to build features for each of the sampled Ethereum addresses)
- Clustering algorithms (fit the data on the chosen algorithms)
- Evaluation (for each of the algorithms and choose the best performing)
- Deploy the endpoint, test on individual instances
- Use a Lambda function to access the deployed endpoint
- Use the API gateway to trigger the Lambda function
- Build the web UI
- Provide the investor profile inference via the web UI