Project Title: Rabbithole - User retention
Group Members: Raiko Marrandi, Károly Raudsepp, Andre Viibur

Github repo

Business goals:

- Background - Our client, a web3 company, seeks to onboard users to blockchain experiences and track their achievements as an on-chain resume.

- Business goals - To discover what types of on-chain activity correlate most with user retention on protocols that are sponsored on Rabbithole, model user-types and predict future usage.

- Business success criteria - We made realistic user models and divided users into groups. Discover the most important factors to increase user retention and build a prediction model for it with at least 90% accuracy.

Assessing your situation:

- Inventory of resources - Our resources include 3 team members, 3 UT IT laptops, software(jupyter notebook, google colab, google docs, google slides, dune analytics, PostgreSQL, github, python, MS Excel) on those computers and two datasets, preliminary one with a size of 290KB and a final dataset (currently being acquired).

- Requirements, assumptions, and constraints - 13 December is the final date for the project. We have no legal or security obligations, because the wallet id's aren't directly linked to any real information from the users and all of the used data is available publicly on-chain. Acceptable finished work would include user models, a machine learning model, outlines for attributes that drive retention and a poster for the presentation.

- Risks and contingencies - Dataset being heavily unbalanced and any real results that would apply in the real world would be difficult to get. Solution - Using data rebalancing methods and reselecting data that could be more relevant.

- Terminology - web3 - Idea for a version of internet that is decentralized and based on public blockchains.
  On-chain - Transactions that have been validated or authenticated and lead to an update to the overall blockchain network.
  ERC20 - Scripting standard for tokens used within the Ethereum blockchain.
  NFT - Non-fungible tokens are pieces of digital content linked to the blockchain.
  ENS - The Ethereum Name Service (ENS) is a naming system based on the Ethereum blockchain.

Ethereum - A decentralized layer 1 blockchain with smart contract functionality.

Polygon - A layer 2 sidechain network for Ethereum.

Aave - A decentralized lending system that allows users to lend, borrow and earn interest on crypto assets.

- Costs and benefits - Costs: time-costs for each member, estimated at 1800€. Benefits: up to 20 points in the course LTAT.02.002 and potential rewards from the Rabbithole team in case the project provides useful insights.

Defining you data-mining goals:

- Data-mining goals - The first goal is to understand and gather relevant data about the users from the Ethereum blockchain via the Dune Analytics platform, then merge it into one large dataset for analysis. The second goal is to have that dataset (re-)balanced and the third is to get a prediction model and patterns from it that can be used for further market research and product development.

- Data-mining success criteria - Successfully completed work will include a prediction model with at least 90% accuracy and the 5 most important patterns/factors that drive retention.

Gathering data:

- Outline data requirements - All we need to get a beautiful user retention model is necessary data. First of all we need NFT trading volume, ERC20 transfer volume owns ens and ETH balance columns. Average weekly transactions on mainnet time range is 3 months and the same goes for transactions on the polygon columns. We have to know if they have used RHQuests before and have used the platform before. Last but not least we need post quest user retention.
- Verify data availability - All the data that we are going to use is on the Dune Analytics platform. We gathered relevant data about the users from Ethereum blockchain so we have all the required data to rebalance and get patterns from it that can be used for further market research.
- Define selection criteria - We will use the Dune Analytics platform where our gathered data is. We have six csv files and one file is only for the wallets (address). All the fields that we have acquired so far are relevant to the project and its completion. In each file we have a wallet column so if we concatenate every table we get a bigger table where we have each wallet activity. The first file shows us the decentralized exchanges in each wallet (traded, class, transactions). Next table contains NFT purchase and sale transactions (sum bought, class bought, txs bought, sum sold, class sold, txs sold) and the new following table is added with a column on whether or not the person with the

given wallet owns ENS (ens). The last two tables contain average weekly transactions on mainnet (weekly_txs mainnet) and interactions (interacted).

Describing data: All data is collected from the Dune Analytics platform. We have a total of 4721 lines of wallets where we concatenated 6 files. The data types we have are numeric and textual (they come when we have mixed numbers and strings in a column). Our data contains accurate information about the activities of the wallet, for example, whether it has a lot of NFT-related purchases or sales and activity on decentralized exchanges. (you also make a brief evaluation of the suitability of the data for
your data-mining goals. For example, verify that the data includes the fields
that you expect and need to be there and sufficient cases for analysis.) SEDA MA EI TEADNUD)

Exploring data: Data types are numeric and textual (you can see it below). When looking at the data, we have to deal with undefined rows and columns, because not all wallets have all the columns when we put all the tables together. This can be solved by setting NaN to 0 instead. The data is imbalanced towards users who did not keep using the protocol.

(address object
NFT_sum_bought float64
NFT_class_bought object
NFT_txs_bought float64
NFT_sum_sold float64
NFTclass_sold object
NFT_txs_sold float64
weekly_txs mainnet float64
DEX_vol_traded float64
DEX_vol_class object
DEX_transactions float64
ens float64
interacted float64)

Verifying data quality: The quality of the data is good and suitable for achieving our end result. All undefined rows / columns should be changed to 0 before the model is created. Secondly, data should be rebalanced to maximize model accuracy.

Make a detailed plan of your project with a list of tasks. There should be at least 5 tasks. Specify how many hours each team member is going to contribute to each task.

1. Homework 10 - Käroly, Andre, Raiko - 4
2. Data gathering - Raiko -15
3. Data understanding - Käroly, Andre, Raiko - 3
4. Data preparation - Käroly, Raiko - 4
5. Modeling - Andre - 8, Käroly - 2, Raiko - 1
6. Evaluation - Käroly, Andre - 4

7. Creating a poster - Andre - 3, Károly - 7
8. Presentation - Károly, Raiko, Andre - 3
9. Other miscellaneous tasks - Károly - 3, Andre - 5

List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

1. Jupyter notebook
2. PostgreSQL (for gathering data)
3. Dune Analytics (for gathering data)
4. Python
5. Scikit-learn
6. Pandas
7. Anaconda
8. Rabbithole
9. Ethereum