Viome internship data challenge

This assignment approximates a potential situation that you might be tasked with as a data scientist at Viome. A common problem we face is trying to uncover the associations between our customers' microbiome and some health or wellness outcome, such as a specific illness of interest. We have provided you with a dataset consisting of simulated microbiome data, along with a target outcome. Think of this data as the output of a next-generation RNA-sequencing pipeline (the measured activity of the microbes found in our customers' stool samples), and a binary indicator giving the customer's response to the question *"do you suffer from disease X?"*

Dataset description

- Each row represents a distinct (stool) sample, indexed by the customer's ID.
- The columns of the dataset represent distinct microbes (e.g. bacterial species).
- The values in the data give the count of RNA reads that were mapped to the microbe's genome: more reads means that microbe is more active in the sample.
- The total number of reads in each row depends primarily on the duration the sequencer was run, *not the absolute activity of the microbes.* Running the sequencer for longer always results in more reads, which will come from microbes in proportion to their true activity in the sample. Therefore we suggest that you divide each read count by the total number of reads in the sample, and work with proportions rather than absolute counts.

Accessing the data

In the provided archive, you'll find the following CSV files:

1. data.csv
   The microbiome data, with each row corresponding to one sample and each column corresponding to one organism
2. labels.csv
   The disease labels in the same order as the data, with 1 meaning the sample is from a customer with the disease
3. user_ids.csv
   The customer IDs in the same order as the data, indicating which sample comes from which customer

Project instructions

You should load the data from the supplied CSV files, join the labels and user IDs appropriately, and provide some kind of descriptive summary or visualization of any aspects of the data you feel are informative.

Then, try to build an appropriate predictive model that we could use to determine whether a new sample is taken from someone with the disease or someone without the disease.

Make sure you follow standard best practices to fit your model, and report some standard measurement of your model's performance that you believe will generalize to new data. Following best practices and describing the performance of your model correctly is more important than building the best-performing model possible.

You should submit a notebook (python or R) that shows the steps you went through in detail. We should be able to reconstruct exactly what you did from the submission. You can include comments if you think it will help us understand your thought process. If you run out of time, you can also describe anything else you would have liked to try in comments.

We expect the task to take no more than 1-2 hours. Please do not spend more than 3 hours maximum on this.

Feel free to use all the typical tools and resources you would use for such a task, except for consulting other people!

Let us know if you have any questions.

Good luck!
- Viome data science team