ML Prediction Challenge

There are two parts to this empirical challenge, but they both have the same goal: to give you experience building an ML pipeline and evaluating results. Given that this exercise is designed to be a learning experience, do not be overly concerned with how this assignment will be graded/evaluated. Feel free to experiment and take risks. If you stick to methods you already knew before this class, you likely won't get as much out of this as you could. You may work alone or in pairs, but groups larger than two will not be permitted.

Put simply, your overall goal is to build a predictor of U1031900 (# DRINKS PER DAY LAST 30 DAYS). You may use all or none of the other variables in the training data set to achieve this, but do not attempt to merge in data from other sources.

Part 1, due Jan 21 by 10 pm

Summarize the training data. The purpose of this is to explore the data and get exercise scanning through and understanding it. Please submit a pdf that contains your summary of the data. This can take any form you think is useful. The summary can be as short as one page, but it cannot be more than five pages.

Part 2, due Feb 3 by 10 pm

Build a predictor. Once you have built a predictor, you should use the test data set to submit a csv file on canvas. That csv file should only have two columns: diag_id, y_hat.

That's it. The instructions are intentionally loose. Such is the nature of starting from scratch to build an algorithm. I will likely not answer many questions like "how should we summarize the data" or "which function class should I pick". Experiment! Of course, if there are administrative questions or questions on how to do a specific task, office hours could be a good opportunity to get help, but I can guarantee the internet contains many more answers.

Good luck!