DS5559 Recommendations for Final Project

-if your dataset is > 3GB, consider taking a meaningful sample to get within size requirement.

-early on, drop fields and records that are not needed.

-categorical variables with many levels can often be bucketed effectively. Conduct EDA to understand how best to bucket.

-be sure to scale features in regression models

-when testing if the pipeline works properly, try on sample of data to save runtime

-when building tree models like GBT, start small to insure things work. This means small number of trees, shallow depth (e.g., 3). GBT and RF can take a long time to train, and are more likely to use up memory.

-when the labels are imbalanced, first split the data into train/test. Next, can change the label proportions in the train set ONLY.