# Lab 9

1. Feature engineering was necessary in this problem largely due to the prevalence of missing values. By using statistical measures, we can impute categorical and numerical data from similar instances within the dataset. While not perfect, the provides an informed means of filling in the gaps in our dataset. Further engineering steps were taken to encode categorical data so that it may be accurately interpreted by the machine learning algorithm.
2. Feature scaling is the practice of taking all of the numeric values for a column and scaling such that they fit within a set range, typically 0 to 1. When this technique is applied to all features in a dataset, it allows two features that may be in very different scales (for example .0001 and 10000) to have equal weighting when considered by the machine learning algorithm.
3. It is reasonable, but not great. Around 80% accuracy in the training and validation set means that the model is attaining a fair amount of discernability within the dataset. The high ratio of false positives to false negatives is perhaps a cause for concern. While ultimately the goal is to limit both counts, having similar amounts of false negatives and false positives is an indication of model balance. Now, if it is the circumstance where a false positive is seen as less impactful than a false negative, it may be appropriate to skew a model in such a direction.
4. When it comes to improving accuracy, there are a number of things that could be attempted. In this problem we used particular implementations when it came to feature imputation, encoding and scaling. Each one of these steps could have performed differently, potentially creating an increase in accuracy. Besides increasing the size of the dataset, performing feature ranking and dropping the lowest performers, and varying the testing/training split, Gaussian Naïve Bayes as implemented by scikit-learn has on optional parameter variance smoothing for calculation stability that could be adjusted.
5. Naïve Bayes is called naïve because it assumes an independence between variables. Naïve Bayes is called idiotic because variables in any sophisticated classification problem are almost never independent. In our problem, we are attempting to train a classifier based on a set of characteristics about a person including demographic and financial information that is very much interrelated. Surprisingly enough, despite this false intuition, the Naïve Bayes classifier still performs relatively well on the dataset, so perhaps Savant Bayes is more appropriate.