A Portuguese banking institution used data obtained through its marketing campaign to predict whether clients are likely to subscribe to a term deposit. The bank’s marketing campaigns were centered on making direct phone calls. The data acquired shows a variety of information pertaining to the potential clients, including such metrics as age, job, marital status, education, and loan history. They also recorded information regarding the specifics of each phone call and marketing representative, including the duration of the call, the day/time of the call, the number of times the client has been called and whether their previous call was successful.

We must begin by conducting further analysis to accurately predict whether each client will subscribe to their term deposit. Our project will be considered successful if we can accurately predict which set of conditions will lead to a client subscribing to a term deposit. We must construct a model that uncovers the precise groups of individuals and the times they are likely to subscribe. In doing so, we can present these findings to the Portuguese banking institution so they can focus their marketing efforts in a more efficient manner.

In terms of cleanliness, our main issue was the ‘unknown’ values were not read as null and required correction. Several of the columns delivered as ‘object types’ with values of either yes or no, so we decided to change these datatypes to Boolean. We replaced missing values on several of the columns with the most common value, and all instances of yes and no with 0 and 1. Numerous column headers were problematic and were updated for clarity. We our grouped a lot of our categorical data together in preparation for creating dummy variables, which we plan on using to create either a decision tree or random forest. We decided to drop the “default” column because only 3 of the 40,000 values were yes. We also dropped the “pdays” column because they decided to put the value ‘999’ for every new customer (which skewed the data) and it already felt redundant due to the similar ‘campaign’ and ‘previous’ columns.

We assumed that single people, people with higher incomes, people with personal loans and people with home loans were more likely to say yes. After analyzing the ‘campaign’, ‘previous’ and ‘poutcome’ columns we assume that clients will be more likely to say yes if they did so the first time and will be less likely to respond or say yes if they’ve already been contacted several times beforehand. Duration factors in too heavily because longer phone calls will almost always lead to yes, so we will neglect this column during our analysis.

After creating the train\_test\_split and running a random forest, we struggled to get an R-Squared value above 0.2. After some discussion, we decided to go back and create grouped values for duration and age so that we could have more predictors. After creating roughly five bins for each duration and age and recreating our dummy variables we brought our predictor total up to 38. We also decided to normalize our economic indicators to see if it would have any impact on the model.

We then created our first random forest using 200 estimators and a max depth of 7 which resulted in an R-Squared value of 0.37. While a significant increase from our first attempt, we still would have liked to seen a stronger model.

We then created a bar graph and list of all important features using the feature\_importances\_ function and learned that the nr\_employed (number of employees quarterly indicator) and bin\_duration\_LongestCall were by far the most important predictors at 0.36 and 0.26. The third most important predictor only had an importance of 0.069.

We then attempted to fine tune the model using the cross validation method with the hyperparameters bootstrap, max\_depth, min\_samples\_split, min\_samples\_leaf, max\_features and n\_estimators. The resulting parameters that the cross-validation method recommended actually gave us a lower R-Squared value than the parameters we had come up with on our own. We also attempted to run a RandomizedSearch and ran into similar issues.

We were hoping several features would prove to be better predictors than the rest, but this turned out not to be the case. None of our variables turned out to be great predictors and only a small amount of people from the original dataset said yes to subscribing to a term deposit in the first place. These two factors hold our model back from reaching its full potential, but I believe that the model is strong given the information we had to work with.