Explanation on Classification Tree

My main dependent variable is whether the customer will have a savings account with Santander. I would like to look at savings account exclusively because Santander’s business is more focused on retail and consumer banking, which means that consumers are more likely to be attracted to Santander’s savings program that usually offers a higher interest rate than other mainstream retail banks like JPMorgan, Bank of America, and Citi Bank. Therefore, examining whether consumers open savings account will be extremely important for Santander.

I used three methods of classification trees: single decision tree, prune decision tree, and random forest decision trees. Single classification tree was used to examine pictures. Prune tree was to decide the optimal cost complexity of the classification tree. Random Forest trees are used to enhance the strength of the model through averaging the results over many classification tree.

I used cross-validation on the training data by partitioning the data into different months. I trained the data on one month and used the next month as the test data. This makes sense because we would like to use current consumer data to predict whether they will purchase and enroll in a savings account next month.

The decision tree models do really good job in terms of predicting whether consumers will have a savings account. However, one caveat is that most consumers do not have savings account to begin with. Hence we have to examine our models more carefully.