**What factors can be used to determine whether a client will subscribe to a term deposit or not?**

Prepared by Jennifer Tatum on April 25th, 2017

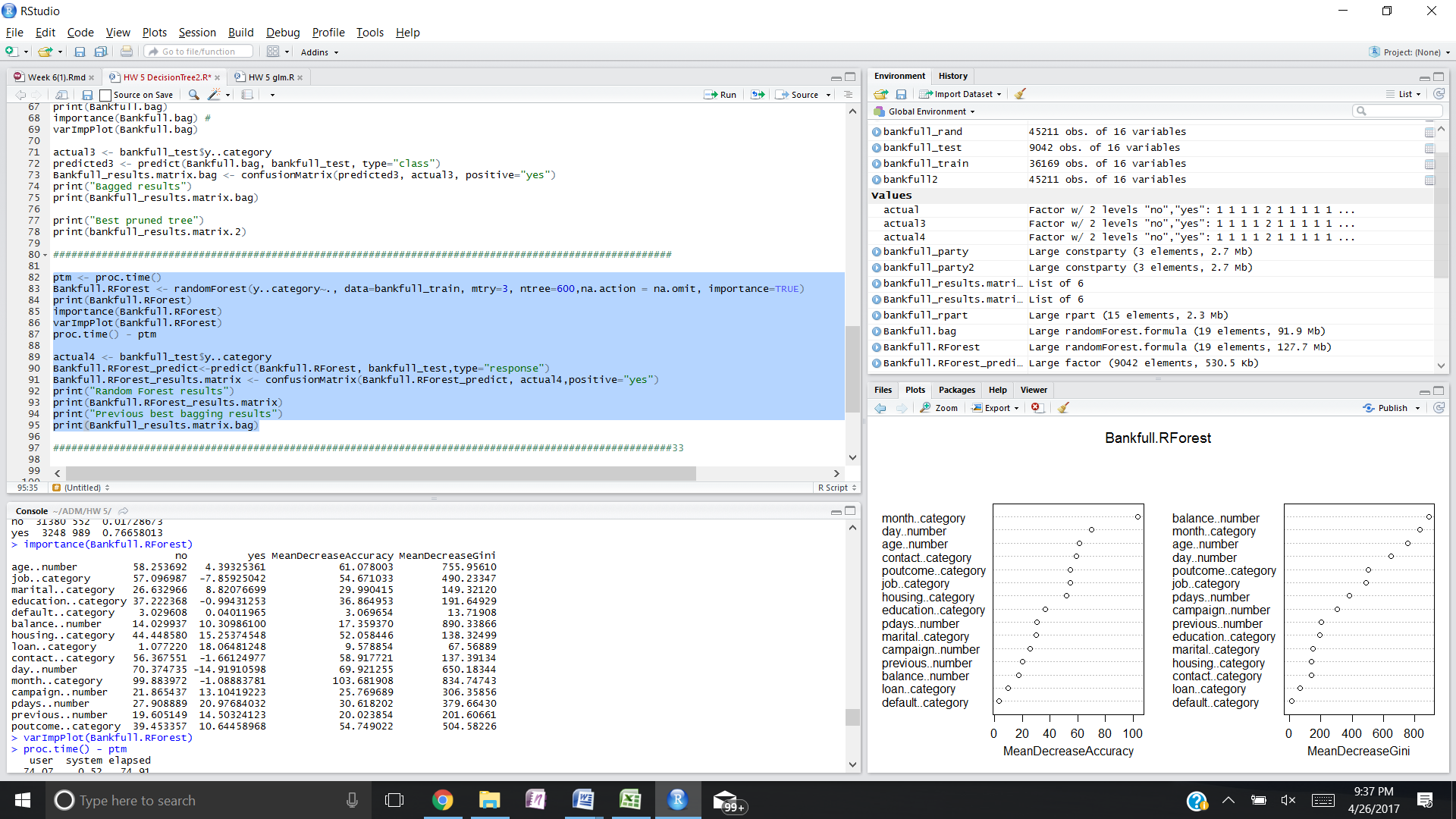
**Analysis Summary**

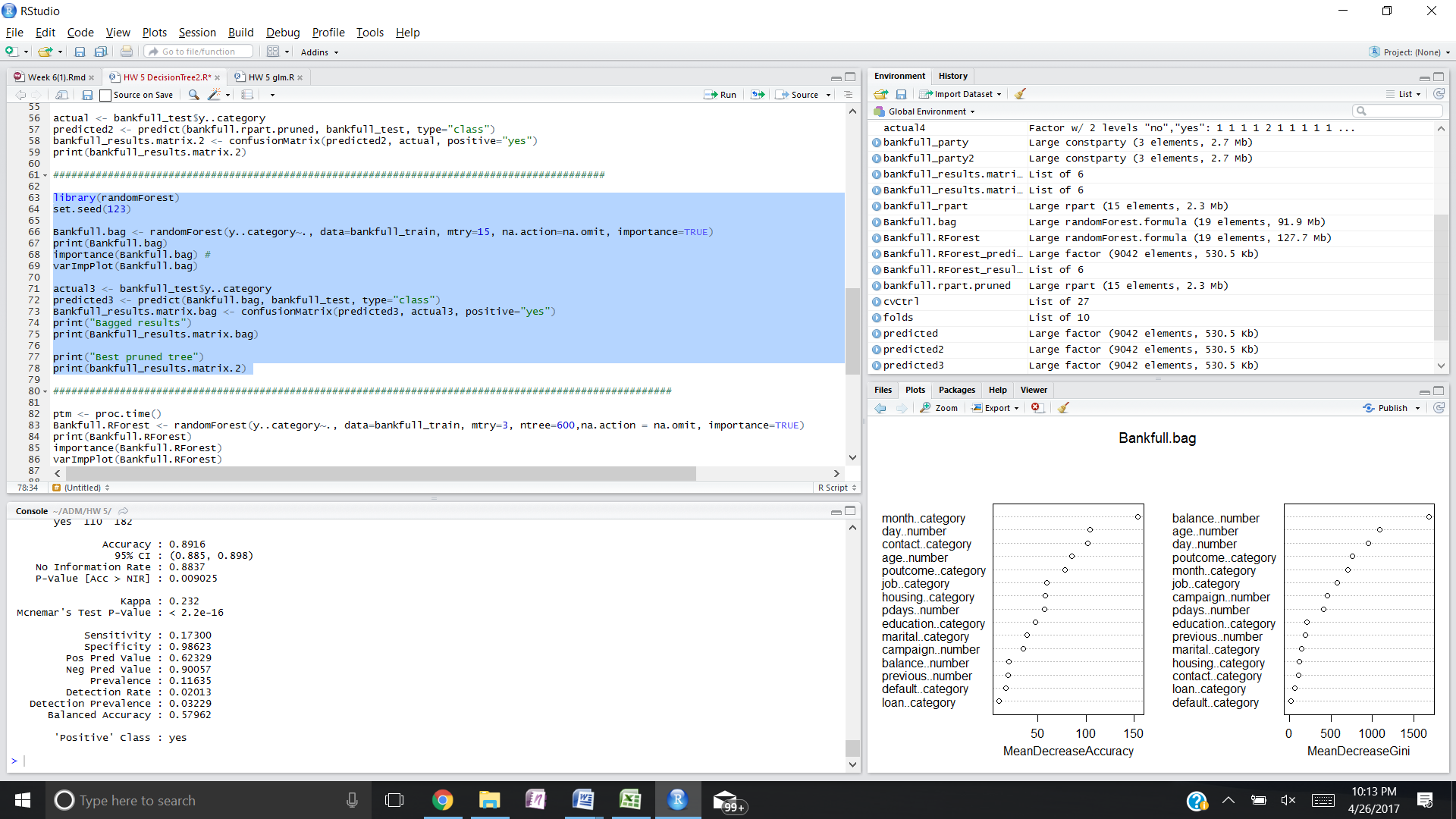
To determine what factors affect a client to subscribe to a term deposit or not, I considered a data set with 45,211 observations and 17 variables. The variable duration is dropped for this analysis because is not known before a call is performed. Also, after the end of the call whether the client subscribed or not is obviously known. Thus, this input should only be included for benchmark purposes and is discarded for predictive modeling such as this. The age of customers range from 18 to 95, with a mean and median being 41 and 39 respectively. 60% are married, 80% have at least a secondary education, and most of them have been contacted 2-3 times about the campaign.

The first model used is a decision tree model with 10 fold cross-validation. This method repeats sampling without reusing data and is the industry standard for estimating model performance. This model indicates that there is only one variable contributing to a customer subscribing to a term deposit and that is the outcome of the previous campaign. If the previous campaign was successful, around 65%, or 792 customers of 1219, will subscribe. And 90% of customers whose previous campaigns were anything but successful did not subscribe to a term deposit.

To test how well this model is at predicting subscription, there are several performance indicators to consider. The accuracy rate (how often the model predictions the correct outcome, subscribe or not subscribe) is 89%. This means that 89% of the time the model correctly predicted whether a customer in the validation data subscribed or not. The sensitivity, how well the model can predict a subscribed customer, is 17%. The specificity, how well the model predicts a not subscribed and the actual value is not subscribed, is very high at 98%.

The next model used to answer our question is random forest. Random forest shows us the variables in order of importance. Whether the random forest model splits each node considering 3 or 15 variables, the same 5 variables appear in the top 6 variables of importance: Month, day, age, previous campaign output, and job. These are the 5 most influential variables on whether a customer will subscribe to a term deposit based on this random forest model.





And finally we will take these 5 variables and run a logistic regression to show if these variables have a greater influence on subscription than just random chance. When we do we see that again the previous campaign outcome has the most influence on whether a customer will subscribe to a term deposit. If the previous campaign was successful, the customer is 12 times more likely to subscribe than if the campaign was anything other than successful. Student, retired and unemployed customers are 2.3, 1.9, and 1.4 times more likely to subscribe respectively. And the months of March, October and September are 3.6, 2.2 and 2 times more likely to subscribe than the other months.

These figures are valid when you look at them independently. Meaning all of the other variables have to be constant to say that a customer is 12 times more likely to subscribe if their previous campaign was successful. This is true when we look at probabilities as well. Holding everything else constant, a customer whose previous campaign was successful has a 92% probability of subscribing to a term deposit. And student, retired and unemployed customers have probabilities of 70%, 65% and 59% of subscribing respectively. And the months of March, October and September have probabilities of 78%, 69% and 67% respectively.

Using the validation set to determine how well the logistical regression is performing, I predicted a subscription if the regression model’s probability for subscription for a customer was over 50%. The model’s accuracy rate was 89%, specificity 99%, sensitivity 16% and precision 62%. In generally the model performed the same as the first two models when focusing on accuracy.

**Commentary**

I believe that the Decision tree model with 10 fold cross-validation, Random forest and Logistic Regression methods all provided strong analysis of the transactions. The previous campaign outcome was by far the strongest indicator of subscription and I therefore recommend that customers with previously successful campaigns should be contacted about a term deposit. Next, students should be contacted as they have a high probability of subscription. And finally customers that were last contacted in March, September or October should be contacted as they also have a high probability of term deposit subscription.