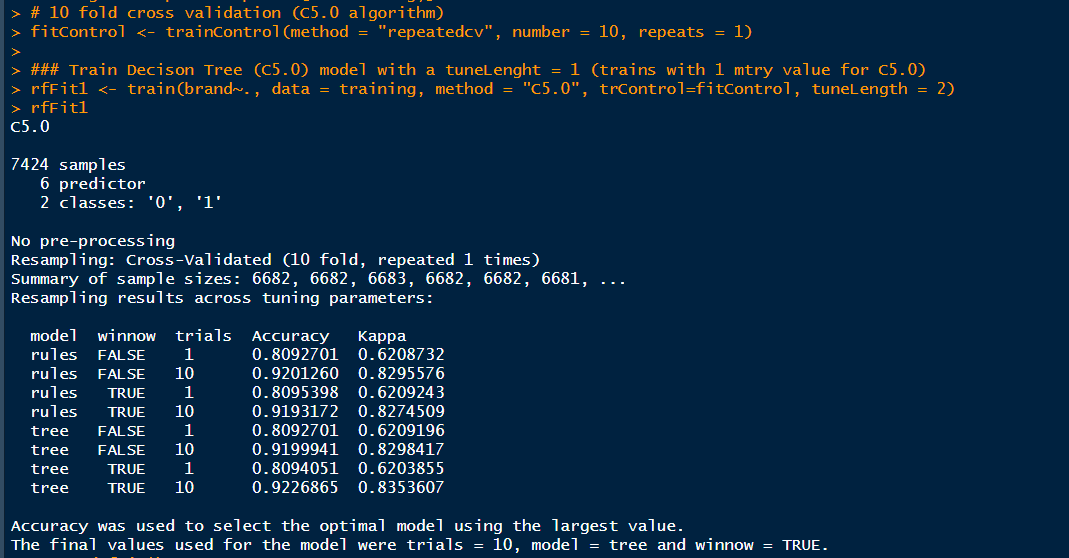
Customer Brand Preferences Report

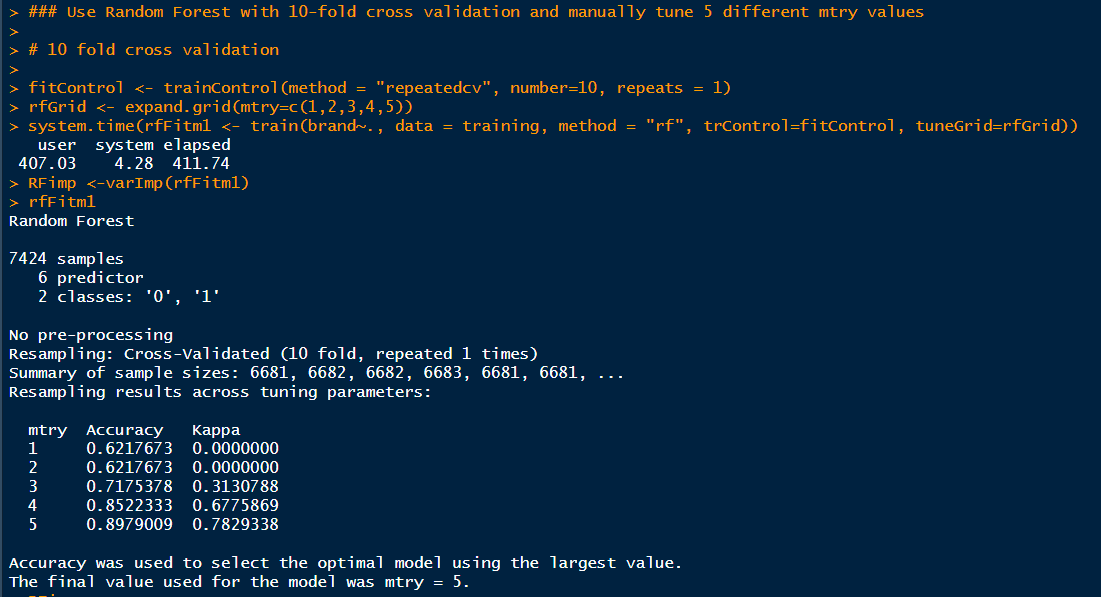
### Overall Impression

In this classification task, I first downloaded the most powerful caret package into RStudio, then I followed the plan of attack to develop a ‘pipeline’ for various models’ training and testing. After comparing the Accuracy and Kappa values and other testing results for each model, I chose C 5.0 model for my final prediction. I used serval commonly used R commands to train and test my two models including createDataPartition(), trainControl(), train(), predict(), and postResample().

### The classifiers you tried and the results of each classifier you ran exported from R

I used two classifiers in this task: C5.0 model with 10-fold cross validation, and Random Forest with 10-fold cross validation and manually tuned with 5 different mtry values.

I have exported the R script and CompleteResponse.Rproj file in the attached zip file. 



### The classifier you selected to make the predictions, including a rationale for selecting the method you did and the level of confidence in the predictions.

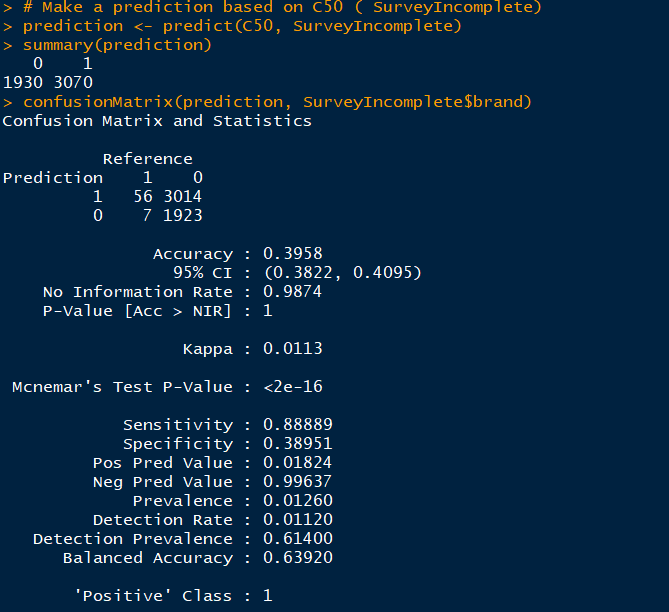
I selected the C5.0 model over Random Forest (RF) to make the predictions. Because C5.0 has a higher Accuracy and Kappa value.

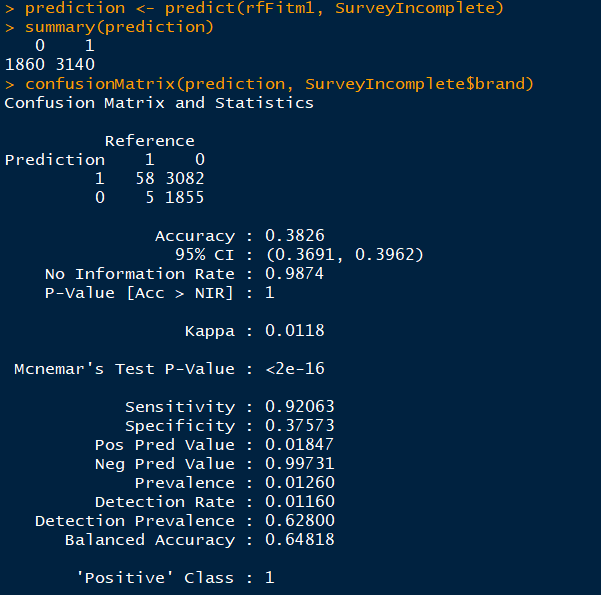
C5.0 Accuracy: 0.9226865 Kappa: 0.8353607 (trials = 10, model = tree and winnow = TRUE)

RF Accuracy: 0.8979009 Kappa: 0.7829338 (mtry = 5)

95% Level of Confidence in the Confusion Matrix: C5.0 Model has 0.3822, 0.4095 Vs. Random Forest Model has 0.3691, 0.3692 on SurveyIncomplete. Both are at 95%, and the two set of values seem close enough to each other. C5.0 model is just marginal better than RF model.

I think prediction itself by definition come with certain degree of errors that cannot be avoided, so the slightly different values in key metrics like Kappa or Confidence is not significant.

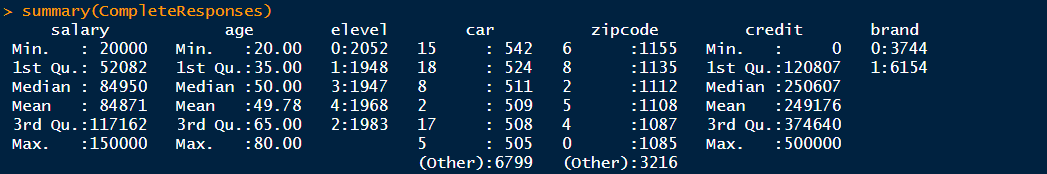


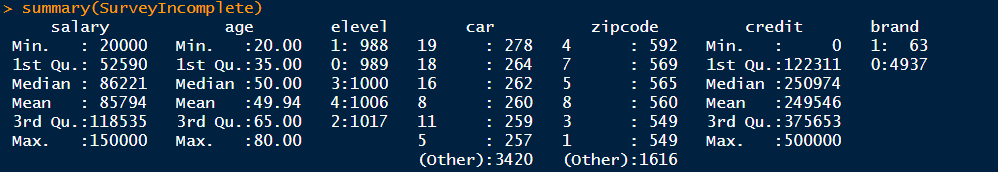


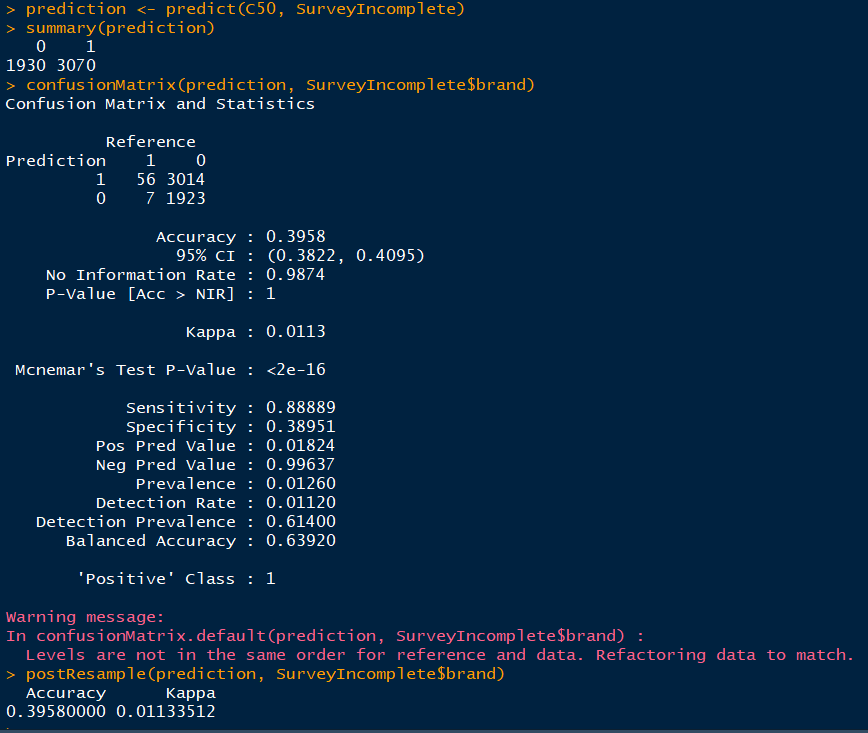
### The predicted answers to the brand preference question for the instances of survey results that are missing that answer.

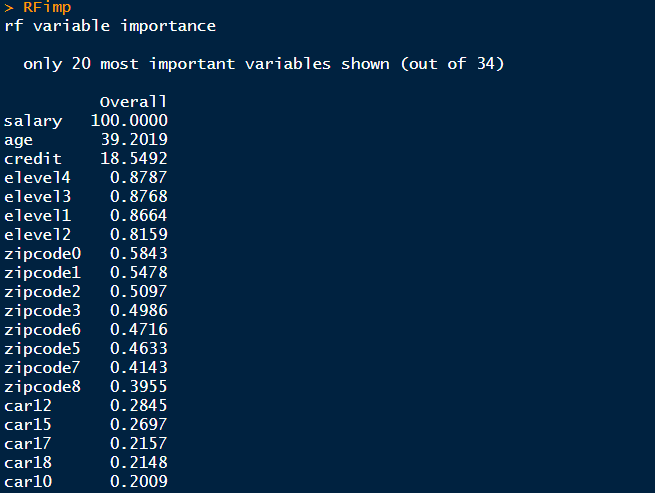
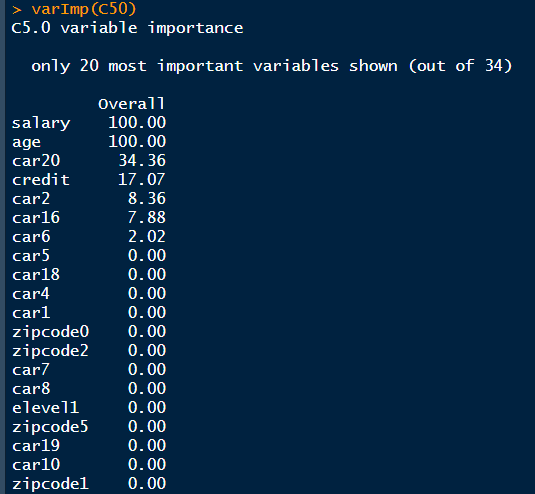
After checking the summary() function, I noticed the make-up and distribution for the brand attribute in SurveyIncomplete are heavily skewed (63 vs 4937). This is expected as the Surveyincomplete does not have the correct brand data. Therefore, we should simply ignore the brand values in this summary() call.

I also ran VarImp function to find out the variable importance for each attribute. Obviously, both salary and age have significant influence among all variables for both modellings, which helps us to further predict the answer to the brand preference. I further filled the prediction values for the survey results in SurveyIncomplete file. i.e. replace its value with what’s the model predicts (Acer, 0=1930 and Sony, 1=3070).









### A chart that displays the customer preference for each brand based on the combination of the actual answers and the predicted answers to the brand preference survey question.

The aggregated counts for both datasets are 15,000. As we can see, people tend to purchase Sony more than Acer. As I mentioned earlier, I filled the predicted values for the survey results in SurveyIncomplete file.

### After making the predictions using the test set use postResample() to assess the metrics of the new predictions compared to the ***Ground Truth.*** Did something interesting happen here? If so, be prepared to explain why!

I first used postResample() command with SurveyIncomplete, the Accuracy rate was only 0.3958, I started to question myself, is it a good model? Why the Accuracy rate is so low given that the original Accuracy rate with CompleteSurvey is about 0.9226865 (page 1). I further applied postResample() command with the original CompleteSurvey data on 25% of the testing partition. I got a very good Accuracy rate of 0.9155 and 95% of confidence interval is 0.9039, 0.9262.

The underlying reason was “brand” attribute in the surveyIncomplete contains no "truth". Brand in SurveyIncomplete is corrupt data. There is no ground truth for comparison, therefore the low accuracy rate should be thrown away.

Accuracy and Kappa can only be found via cross validation and predict with a test set for this task. The test set, on the contrary, DOES have ground truth for brand which explains the high Accuracy rate in the 25% testing partition.

