Customer Brand Preferences

From: Usman Iftikhar

Two sets of data are provided to predict the brand of computers customer prefer. The first set contains 9,898 observations with six predictors to predict the brand that is a factor with 2 levels. The second data set is the incomplete data set with the same variables of 5,000 observations. The analysis explores two classification algorithms to predict the brand customers prefer, C.50 decision tree and random forest algorithms. The complete data set is split into training set with 75% of the data and test set with 25% of the data. The optimized model from the complete data set is then applied to the incomplete data set for brand prediction.

## C5.0 Classifier

For the C5.0 model, a 10-fold cross validation and an Automatic Tuning Grid with a tuneLength of 2 is used. Performance output from this trained model is provided below in Table 1. The final values used for the model are highlighted (trials = 10, model = tree, winnow = TRUE) because it has the highest accuracy and kappa. The importance given to each of the predictor variables is given in Table 2. C5.0 model gave highest importance to age and salary, very little to education level, and no importance to zip code, car, nor credit.



Table 1. Performance output of C5.0



Table 2. Variable importance with C5.0 decision tree

## Random Forest Classifier

For the Random Forest (RF) mode, a 10-fold cross validation and manual grid with 5 different mtry values (mtry – 1, 2, 3, 4, 5) is used. Performance output from this model are provided below in Tablle 3. The value used for this model is mtry = 3 because it has the highest accuracy and kappa. RF gave some importance to all the predictor variables except education level, but the highest importance was given to salary. The importance given to each of the predictor variables is given in Table 4.



Table 3. Performance output of RF



Table 4. Variable importance with RF

## Product Prediction

C5.0 was chosen as the best model for this task due its higher accuracy and kappa. The model was then applied to the 25% of the test set data from the complete responses data set. The model predicted the brand in the test data set with an accuracy of 0.918 and a kappa of 0.826. The same model was then applied to the incomplete data set.

From the entire data (incomplete and complete data sets) the brand preferred by customers is Sony with a total count of 9,284 (Figure 1). The total count for the Acer brand was 5,614.

