Predict Which Brand of Products Customers Prefer

***Classifiers Used & Parameters***

*Train, Test & Sample Sets*

* Sample Set: 20% of all *Complete Responses*
  + Train Set: 75% of Sample Set
  + Test Set: 25% of Sample Set

*C5.0*

* Parameters
  + Cross Validation Folds: 10
  + Repetitions: 1
  + Automatic Tuning Grid
    - tuneLength = 2
  + Optimized Parameters
    - Trials: 10
    - Model: Rules
    - Winnow: TRUE
      * Accuracy: 0.8807594
      * Kappa: 0.7387474
* Important Features
  + Mazda: 100
  + Salary: 100
  + Chevrolet: 99.47
  + Age: 95.40
  + East North Central: 93.94

*Random Forest*

* Parameters
  + Cross Validation Folds: 10
  + Repetitions: 1
  + Manual Grid
    - mtry Values Used: 2, 4, 8, 16, 32
    - At first, I used 1,2,3,4,5 but noticed they returned very poor performance. I brought them up to the ones shown above. Once I reached 16, I noticed that if I made the mtry value much higher than that, the accuracy suffered.
  + Optimized mtry Value: 16
    - Accuracy: 0.9053907
    - Kappa: 0.79647652
* Important Features
  + Salary: 100
  + Age: 70.349

***Selected Model***

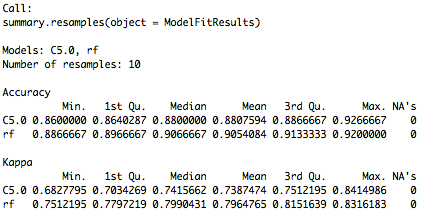


Figure 1: C5 & Random Forest Result Comparison

*Random Forest Classifier*

* I chose the Random Forest classifier over the C5.0 classifier, because even though the C5.0 classifier has the max accuracy and kappa values, the RF classifier constantly outperforms the C5.0 classifier. It can be seen that 25% of all RF classifiers reach at least a 91.33% accuracy. On the other hand, 25% of C5.0 classifiers ran are only able to reach an accuracy of 88.67%. The average accuracy of the RF classifier is 90.54% and 88.08% for the C5.0 classifier.

***Prediction***

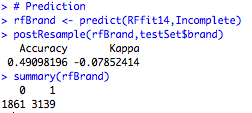
******

Figure 2: Predicted Brands Compared with Test Set

*Prediction vs. Test Set*

Using the postResample() function, I compared the brand prediction I made of the incomplete surveys against the test set I used to measure the performance metric of the classifiers. I found it interesting that, even though there is not supposed to be any correlation between the predicted values and the test set, there was a 50% accuracy. Even though, I did not expect this, it makes sense; the brand value is either a 0 for the Acer brand or a 1 for Sony. This means that there is a 50% chance that the predicted values compared against another set will match. The kappa value is negative indicating that much of the accuracy gotten is due to simple probability, even if there is no correlation between the data sets.

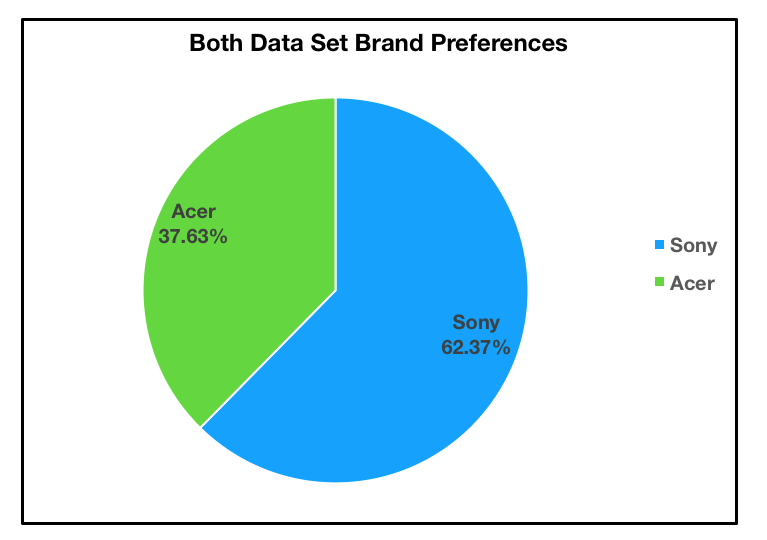
***Results***

Figure 3: Predicted & Given Customer Brand Preference (15,000 Responses)

Figure 3 shows the computer brand preference of the customers; it includes the predicted brand preferences, as well as the known brand customer preferences. It can be concluded that the majority of customers, exactly 62.37%, prefer Sony as their computer brand. Out of the 15,000 customers, 9,356 prefer the Sony brand.

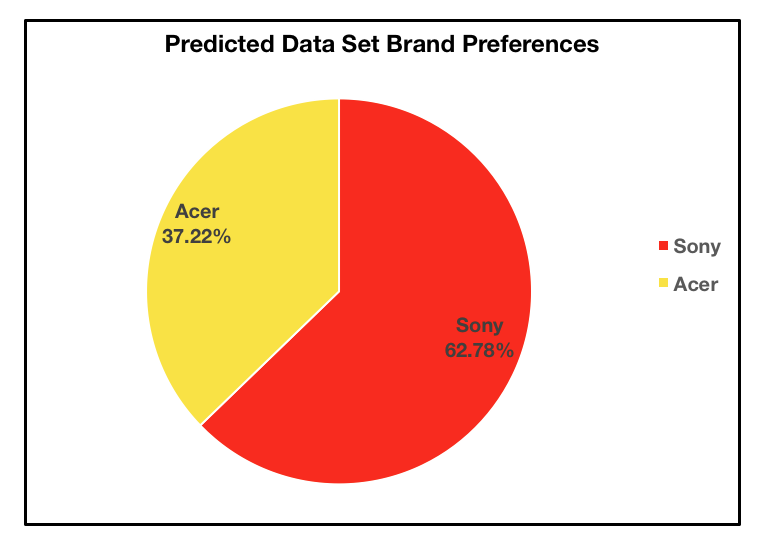


Figure 4: Predicted Computer Brand Preference (5,000 Responses)

The predicted brand preference values agree with the overall brand preference; the Random Forest predicts that 62.78% of all customers will prefer the Sony computers. Compared to the overall brand preference percentage, the Random Forest classifiers predicts a negligible 0.41% percentage difference between the two data sets. Out of the 5,000 predicted customer preferences, the Random Tree predicted 3,139 of the customers would prefer Sony over Acer.