Monika Spreitzer

C3T2

**Brand Preference Prediction using Caret**

Our objective this week is to understand computer brand preference. Our analysis will be based on 9,898 surveys which include the following data points:

* Salary ($20K – $150K)
* Age (20 – 80)
* Education Level (< High School to Doctoral/Professional Degree)
* Type of car driven (20 unique options)
* Area of residence (broken into 8 regions)
* Credit Limit ($0 - $500K)
* Computer brand preference (Sony[6154] & Acer[3744])

In addition to the ~10,000 survey responses, we received an additional 5,000 surveys which had inaccurately labeled brand preferences. Ideally, we will build a model using the complete survey data and use the model to predict brand preference for the incomplete surveys.

**High-level steps:**

1. Load complete and incomplete survey data and perform pre-processing.
2. Create models to understand consumer buying preference of complete data.
3. Evaluate the models to find the most successful.
4. Use the most successful model to predict brand preference for incomplete surveys.

**NOTE**: For this task, we will be using the Caret package, using the Random Forest and C5.0 models for classification.

Six models were created (4-RF and 2-C5.0). Here are the results of each:

1. RF Automatic Tuning Grid (length 1)
   1. Accuracy = 0.9053758
   2. Mtry 2
2. RF Automatic Tuning Grid (length 2)
   1. Accuracy = 0. 0.9133537
   2. Mtry 6
3. RF Manual Tuning Grid
   1. Accuracy = 0.9127048
   2. Mtry 3
4. RF Random Tuning Search
   1. **Accuracy = 0.9166871**
   2. Mtry 4
5. C5.0
6. Accuracy = 0.8175162
7. C5.0
8. Accuracy = 0.8993801

Based on my results, model #4 (RF Random Tuning Search) produced the best results (highest accuracy) and will be used to perform our predictions on the incomplete data set.

After predicting the brand preference for the incomplete surveys, these were the results:

summary(predictions)

Acer - 1,882

5,000

Sony - 3,118

After performing the postResample() we saw a huge decrease in accuracy (~40%); however, this is justified because the original data was extremely inaccurate (hence the reason why we had to perform these predictions).

postResample(predictions, testing$brand)

Accuracy Kappa

0.54308617 0.02732371

The following is a list of important variables ordered from greatest to least:

* salary 100.000
* age 64.733
* credit 8.925
* car 3.782
* zipcode 1.826
* elevel 0.000

Finally, please see the graphs below; they represent preferred computer brand from the 15K individuals surveyed.

In short, Sony is the preferred brand of computer with almost 2/3 of the vote.