Aline Barbosa Alves 2019-08-27

Customer Brand Preferences Report

This report covers the data analytics activities of classification analysis exploring the data to predict customer brand preferences. Its objective is to help Blackwell to decide with which manufacturer we should pursue a deeper strategic relationship. To do that the sales team engaged a market survey with our existing customers but the brand preference was not adequately captured for all of the respondents. So to predict the incomplete answers, we received 3 documents that contain information about the complete responses, the survey incomplete and survey key to help us to understand the data.

First, we imported the complete responses file into RStudio and start to take some general insights. It contains 9898 answers with information about salary, age, level of education, car, zipcode, credit and brand and no missing values. Salary goes from 20k to 150k and the mean is 84871. The customers age is from 20 to 80 and the mean is 50. The level of education, type of car and zipcode are almost equally divided. While credit is from 0 to 500k and the mean is 249176. And the preferred brands are Sony and Acer with 6154 and 3744 customers respectively. After familiarizing with the data and before split the data, we converted the level of education, car, zipcode and brand from numeric to factor since this make more sense because the information is in levels.

So we were able to split the data in training and testing set. We choose to divide in 75% to train and 25% to test. Then we built a C5.0 model with 10-fold cross-validation, automatic tune grid and tune length of 2. The results are the following:

```
                                                C5.0

7424 samples
   6 predictors
   2 classes: 'Acer', 'Sony'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 6681, 6682, 6681, 6682, 6682, 6682, ...
Resampling results across tuning parameters:

  model  winnow  trials  Accuracy   Kappa
  rules  FALSE    1      0.8312188  0.6647246
  rules  FALSE   10      0.9186452  0.8268407
  rules   TRUE    1      0.8283761  0.6583844
  rules   TRUE   10      0.9207982  0.8309462
  tree   FALSE    1      0.8285232  0.6522414
  tree   FALSE   10      0.9205276  0.8317923
  tree    TRUE    1      0.8256807  0.6457380
  tree    TRUE   10      0.9183733  0.8268375

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were trials = 10, model = rules and winnow = TRUE.
```

We also built a Random Forest model with 10-fold cross-validation and manually tuned with 5 values of mtry. We tried some different mtry to see which ones to choose in our model and the range between 11 and 15 was the best one. These were the results:

```
                                             Random Forest

7424 samples
   6 predictors
   2 classes: 'Acer', 'Sony'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 6682, 6681, 6682, 6682, 6681, 6682, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  11    0.9210640  0.8324529
  12    0.9202552  0.8306920
  13    0.9201204  0.8303659
  14    0.9201209  0.8303470
  15    0.9199864  0.8300207

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 11.
```

It's important to notice that both of the models were built with all the features. So we used a variable importance function to see which features were more important to the models. Both of the models had salary and age as the most important features. Then we built the same models but with only these features and the results were:

```
                                                 C5.0

7424 samples
   2 predictors
   2 classes: 'Acer', 'Sony'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 6681, 6683, 6681, 6681, 6682, 6681, ...
Resampling results across tuning parameters:

  model  winnow  trials  Accuracy   Kappa
  rules  FALSE    1      0.8273077  0.6507850
  rules  FALSE   10      0.9179667  0.8250582
  rules   TRUE    1      0.8273077  0.6507850
  rules   TRUE   10      0.9179667  0.8250582
  tree   FALSE    1      0.8231351  0.6359999
  tree   FALSE   10      0.9203908  0.8314307
  tree    TRUE    1      0.8231351  0.6359999
  tree    TRUE   10      0.9203908  0.8314307
```

```
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were trials = 10, model = tree and winnow = TRUE.
```

```
                                  Random Forest

7424 samples
   2 predictors
   2 classes: 'Acer', 'Sony'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 6683, 6681, 6681, 6681, 6682, 6681, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  11    0.9045031  0.7969954
  12    0.9046396  0.7973949
  13    0.9039665  0.7959553
  14    0.9045045  0.7969812
  15    0.9054475  0.7990297

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 15.
```

As we can see, the models with the feature selection got a little worse than with all the features. So we decided to continue with the Random Forest with all the features as it was the model with best accuracy and kappa and the time of processing was not a real problem to this project. After choosing a model, we used it in the test set. Then we used a function that calculate the performance across resamples and the result was pretty good:

```
 Accuracy     Kappa
0.9260307 0.8426520
```

This result means that our model is choosing the right brand in 92,60% of the cases. As it was a good result we could apply the model to the survey incomplete data. But first, we had to import the data into RStudio and convert the same features as we did on the other set. After running the model we had the following result:

```
Acer Sony
1881 3119
```

We can conclude that Sony is the favorite brand from the customers. The figure 1 as the following numbers shows us the preferences for Sony and Acer for the entire existing customers survey:

```
Acer:5652
Sony:9246
```

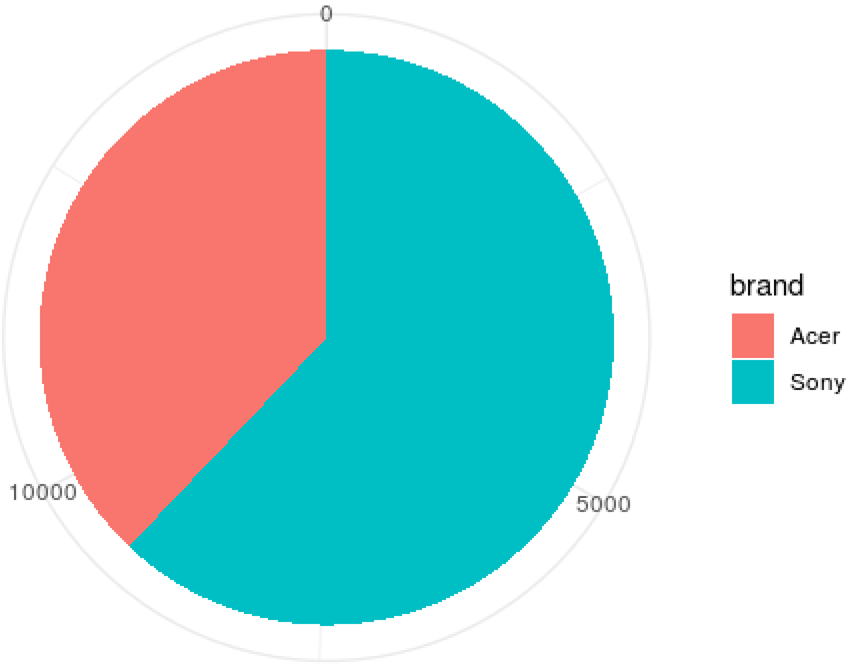

Figure 1: Preferences for brands for the entire existing customers survey

So we recommend that the company should pursue a deeper strategic relationship with Sony because 62% of our actual customers prefer this brand. We can do that with more than 92% of accuracy for these clients. But it’s essencial to notice that the features from this data are almost equally divided so it may not work for any type of data. This means that if we want to use this information to other clients we should collect more diverse data.