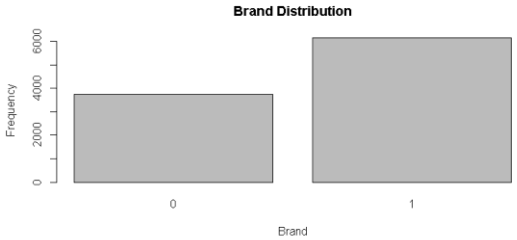
CUSTOMER PREFERENCE PREDICTION

Report

Abstract

Blackwell Electronics' sales team engaged its customers in a survey to find out their brand preferences. Not all the surveys were completed so we had to predict the missing responses using data from the completed responses in other to enable us to make informed decision about the company's strategic partnerships.

Data Exploration and Pre-Processing



From the brand distribution histogram on our completed survey dataset above, we can see that Sony is preferred about 20% more than Acer.

Updating data types

I updated the datatype of Brand to *Factor*, to avoid my model seeing it as a numeric value.
CompleteResponses$brand<-as.factor(CompleteResponses$brand)

C5.0 Model Analysis

Variable Importance report

Looking at our quantitative evidence below for variable importance, it's clear that Salary has the highest importance and is the only feature that significantly affects a customer's brand preference. Removing all other features would have been ideal if we were certain we would always get surveys with salary data. However, keeping the other features, ensures the model is still usable even if the salary attribute isn't available in future datasets.

Variable	Importance
Brand	100.000
Salary	10.9646
Age	1.803
Car	0.4613
Credit	0.2900
Zipcode	0.1534
Elevel	0.000

C5.0 Model Evaluation

model	winnow	trials	Accuracy	Kappa
tree	**FALSE**	**10**	**0.922145**	**0.834707**
rules	TRUE	10	0.9213373	0.831257
tree	TRUE	1	0.9187788	0.827744
rules	FALSE	10	0.9185104	0.826324
rules	TRUE	1	0.9185098	0.827156
rules	FALSE	1	0.9182412	0.826466
tree	TRUE	10	0.9181061	0.826389
tree	FALSE	1	0.9175677	0.82504

C5.0 Post Resample Results

Accuracy	Kappa
0.919159	0.8288586

Random Forest Model Analysis

Random Forest Model Cross Validation Results

To avoid bias in my parameter tuning, I used the *Random Search* approach to generate the best mtry value after several iterations.

mtry	Accuracy	Kappa
3	**0.922277**	**0.834968**
2	0.921873	0.834397
4	0.920257	0.830572
5	0.918236	0.826221
1	0.863146	0.699627

Random Forest Post Resample Result

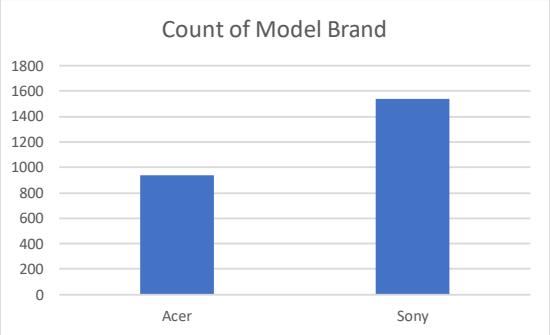
Accuracy	Kappa
0.919968	0.8301456

Confusion Matrix Stats

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  841  103
         1   95 1435
Accuracy : 0.92
                 95% CI : (0.9086,
0.9304)
    No Information Rate : 0.6217
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8301
 Mcnemar's Test P-Value : 0.6189

            Sensitivity : 0.8985
            Specificity : 0.9330
         Pos Pred Value : 0.8909
         Neg Pred Value : 0.9379
             Prevalence : 0.3783
         Detection Rate : 0.3399
   Detection Prevalence : 0.3816
      Balanced Accuracy : 0.915
```

Selected Model and Predicted Results.

Based on the reports above, the Random Forest Classifier had a bit better Accuracy & Kappa than C5.0 and that was the method I used to make my predictions. Our results show that most users still preferred Sony laptops even in the incomplete data set, and the major influencing factor is the user's salary. The higher the salary, the more likely they are to prefer a Sony device.



Merging Tables Together

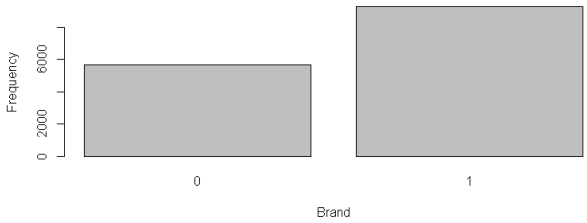
I added the results from the prediction run to the *SurveyIncomplete* dataset. This updated our values and ensured that every entry had a brand preference.

```
SurveyIncomplete$brand<-prediction
_results
```

I merged the CompletedSurvey and the updated SurveyIncomplete tables into one data frame, using *rbind* to get a complete view.

```
full_survey_view <-rbind(CompleteR
esponses,SurveyIncomplete)
```

Summary



Looking at the reports from our data set, it's clear that there is a high preference for Sony devices from our customers.

Based on this and the confidence we have in the Machine Learning results, I highly recommend that we pursue a deeper strategic relationship with Sony to ensure we provide the best products and experience to our customers.

SONY®