**Customer Brand Preferences Report**

**Introduction**

This report predicts the customers' brand preferences that were missing from the incomplete surveys and suggests an area for further investigation. *Appendix: The Process* outlines the analysis and the results from the different models that were run are in the file, *Results for all model run to predict brand preference.*

**Recommendation**

***Issues with the predictions***

The data used to make predictions of brand preference does not represent the customer population. It is recommended that the survey is repeated with a randomly selected sample of customers.

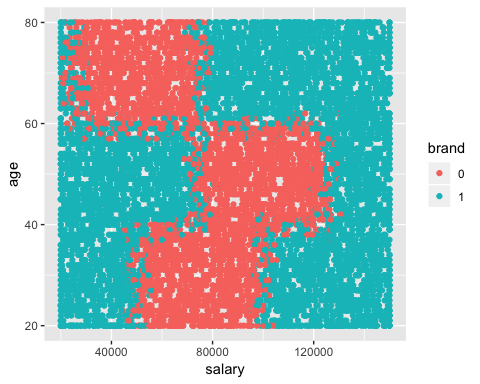
In the data of complete and incomplete surveys, all the variables are uniformly distributed with pretty much as many customers aged 30 as there are aged 40,50,60, 70 or 80. Similarly, the number of customers driving one make of car does not differ greatly from those that drive any other make of car. And so on across all the variables.

While it appears that the model works well at predicting brand preferences, the similarity between the data in the completed and incompletely surveys renders the model redundant.

***Area for further investigations***

Even though the data isn’t suitable for a predictive model, it does suggest a customer’s brand preference is dependent on their age and salary, as shown in Figure 1. The plot indicates a very strong preference for Acer or Sony depending on salary and age. While the plot is a little ‘too perfect’ with very distinct banding of preference, it does suggest this relationship should be investigated further.

Figure 1: Brand preference according to salary and age



Key: 0=Acer and 1=Sony

**The predictions**

The predictions for the incomplete surveys are shown in Figure 2 and a combination of brand preferences for the complete and incomplete surveys are in Figure 3. However, these results should not contribute to business decisions for the reasons outlined in the Recommendation.

Figure 2: predictions for brand preference in the incomplete survey

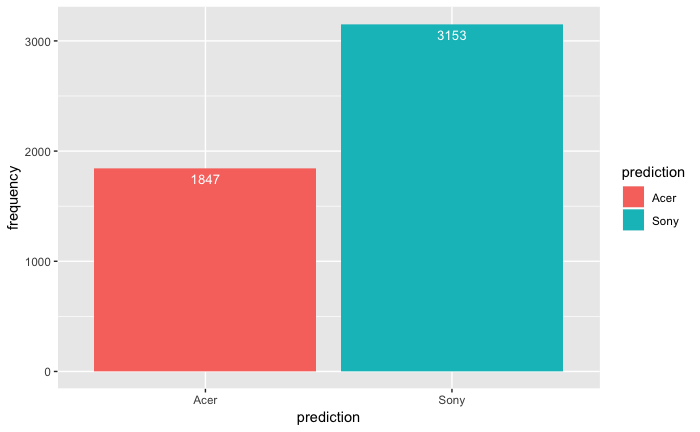
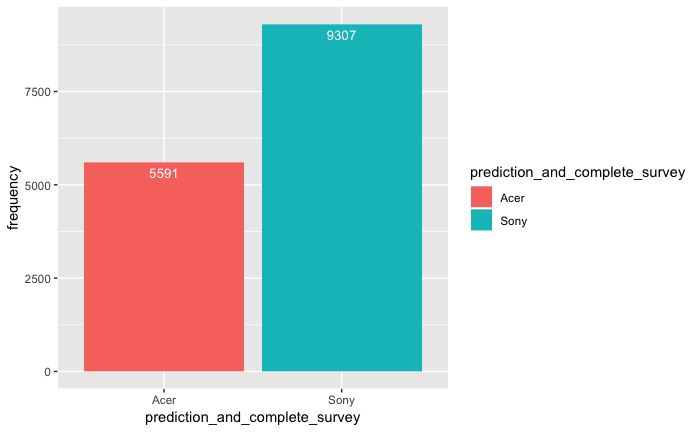


Figure 3: combination of brand preference in the complete and incomplete surveys



**Appendix: The process**

The stages in the analysis were:

1. Data preparation and pre-processing
2. Set up base models using C5.0, KNN and random forest classification methods in R. Train and tune the models.
3. Assess which model has the best accuracy
4. Assess the importance of the independent variables
5. Refine the selected model based on the important variables
6. Make the predictions
7. Asses the accuracy of the predictions

***Data preparation and pre-processing***

* Changed elevel,car, zipcode and brand to factors in both the Complete and Incomplete Surveys
* Checked for missing values
* Plotted the independent variables with the dependent variable
* Checked the numerical variables (salary, age and credit) have a normal distribution

***Set up base models***Created a training set with 75% of the Complete Survey data and a testing set with the remaining 25%.

Three models were used:

* C5.0 decision tree with 10-fold cross validation and an Automatic Tuning Grid with a tuneLength of 2. The results for the model are in Appendix 1.
* Random forest with 10-fold cross validation and manually tune 5 different mtry values. The results for the model are in Appendix 2.
* KNN with the variables normalized.

***Assessing the models***

Each model was trained using all independent variables and then the accuracy of making predictions was measured by making the predictions using the known brand preferences in the test set. The C5.0 model had the best accuracy rate and kappa score.

## Accuracy Kappa   
## 0.9086500 0.8053946

***Relationships between the variables***

With the base model decided, investigated the Complete Survey data and the relationships between the variables to see which are the best independent variables to be included in the model.

Histograms and barplots of the variables show all variables have a uniform distribution in the Complete and Incomplete Survey data with similar proportions for each variable.

A recursive feature elimination identifies the top 5 variables as salary, age, credit, elevel and zipcode.

control <- rfeControl(functions=rfFuncs,method="cv",number=10)  
results <- rfe(CompleteResponses[,1:6],CompleteResponses[,7],sizes=c(1:6),rfeControl = control)  
print(results)

##   
## Recursive feature selection  
##   
## Outer resampling method: Cross-Validated (10 fold)   
##   
## Resampling performance over subset size:  
##   
## Variables Accuracy Kappa AccuracySD KappaSD Selected  
## 1 0.6407 0.2369 0.010252 0.02118   
## 2 0.9167 0.8235 0.012669 0.02686   
## 3 0.9208 0.8319 0.009616 0.02032   
## 4 0.9219 0.8343 0.011199 0.02340   
## 5 0.9224 0.8353 0.009792 0.02048 \*  
## 6 0.9210 0.8318 0.009251 0.01957   
##   
## The top 5 variables (out of 5):  
## salary, age, credit, elevel, zipcode

predictors(results)

## [1] "salary" "age" "credit" "elevel" "zipcode"

An anova test between brand and salary shows the importance of salary.

an\_test<-aov(salary~brand,data=CompleteResponses)  
summary(an\_test)

## Df Sum Sq Mean Sq F value Pr(>F)   
## brand 1 6.002e+11 6.002e+11 440.7 <2e-16 \*\*\*  
## Residuals 9896 1.348e+13 1.362e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

A chi test between brand and car shows that car is not important.

chi\_test.data<-table(CompleteResponses$brand,CompleteResponses$car)  
print(chisq.test(chi\_test.data))

##   
## Pearson's Chi-squared test  
##   
## data: chi\_test.data  
## X-squared = 12.574, df = 19, p-value = 0.8596

An anova test between brand and age suggests age is less important.

an\_test<-aov(age~brand,data=CompleteResponses)  
summary(an\_test)

## Df Sum Sq Mean Sq F value Pr(>F)  
## brand 1 576 576.3 1.861 0.173  
## Residuals 9896 3063964 309.6

***Assess the importance of the independent variables***

The C5.0 models was first assessed with just salary and then the model repeated a number of times, each time adding another variable. Found that the model with salary, age, credit, elevel and zipcode had the best accuracy and kappa scores:

## Accuracy Kappa   
## 0.9122878 0.8125552

This was the model selected for making the predictions.