Predicting Brand Preference for Missing Values on Survey

Jeroen Meij

30-01-2019

Blackwell Data analytics department

1) Summary

For this report, 3 different machine learning algorithms have been evaluated to predict respondents' preferred computer brand in an incomplete survey. The method used are:

- C5.0 Decision tree
- Random Forest
- K-Nearest Neighbors

With the help of the ggplot2 package in, R the data was visualized to see how brand preference was distributed among the respondents. The caret package in R provided the tools to tune and run each machine learning algorithm efficiently. Variables that provided the most predictive power for consumers preferred brand were their yearly salary and their age.

With the predicted incomplete surveys and the complete surveys combined we know how many of the respondents prefer Sony over Acer and vice versa. As it turns out, 9,257 of the respondents prefers Sony over Acer and 5,641 of the respondents prefers Acer over Sony. Hence 62% of the respondents prefers Sony over Acer, and 38% prefers Acer over Sony.

With these numbers, it is clear that most respondents prefer Sony products. However, since this preference is both salary and age dependent, the management must first make a choice which type of customer they want to try to attract more before engaging with either Sony or Acer.

2) Full report:

2.1) Objective:

The sales team engaged a market research firm to conduct a survey of our existing customers. One of the objectives of the survey was to find out which of two brands of computers our customers prefer. Unfortunately, the data related to the brand preference was not properly captured for all of the respondents.

Therefore, I investigated whether customer responses to some survey questions (e.g. income, age, etc.) enabled us to predict the answer to the brand preference question.

2.2) Data:

Danielle Sherman provided me with 2 datasets:

- The completed surveys
- The surveys where brand prediction is missing

Table 1 contains the variables obtained from the survey:

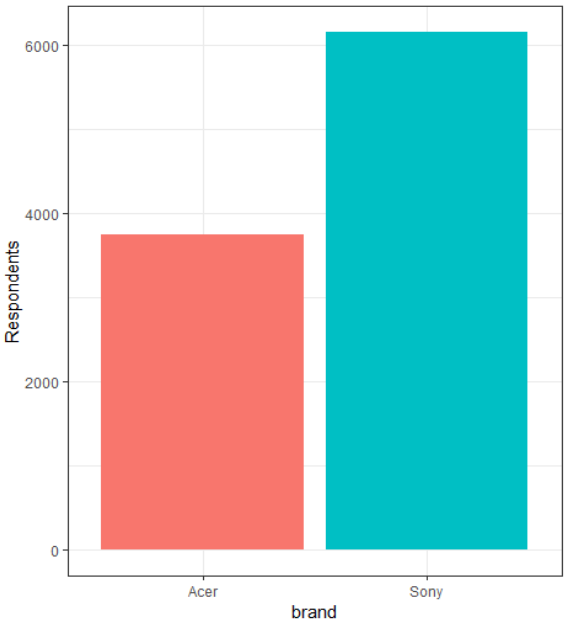
Table 1: Variables obtained from survey

Variables	Description
Yearly salary	In USD
Age	In Years
Highest level of education obtained	5 options
Primary car brand	20 options
Zip code	9 options
Amount of credit available	In USD
Preferred brand	**Variable of interest. 2 options**

Multiple plots have been evaluated to analyze the data. For reference, these can be found in the appendix of this report.

When looking at the respondents´ preferred brand in the completed surveys data (**figure 1**), we see that more people have picked Sony instead of Acer. Overall, the percentage that picked Sony was: 62% whereas the percentage that picked Acer was 38%.

Figure 1: Bar chart for brand preference



The thing that struck out the most was that there is a clear pattern observable when looking at the respondents' yearly salary and their preferred brand. **Figure 2.1** provides a histogram, and **figure 2.2** provides a density plot on this relationship.

As we can see in the figures, respondents with a relatively lower yearly salary and respondents with a relatively higher yearly salary generally prefer Sony over Acer. People that have a relatively medium yearly salary however prefer Acer over Sony.

Figure 2.1: Histogram on sales and brand preference

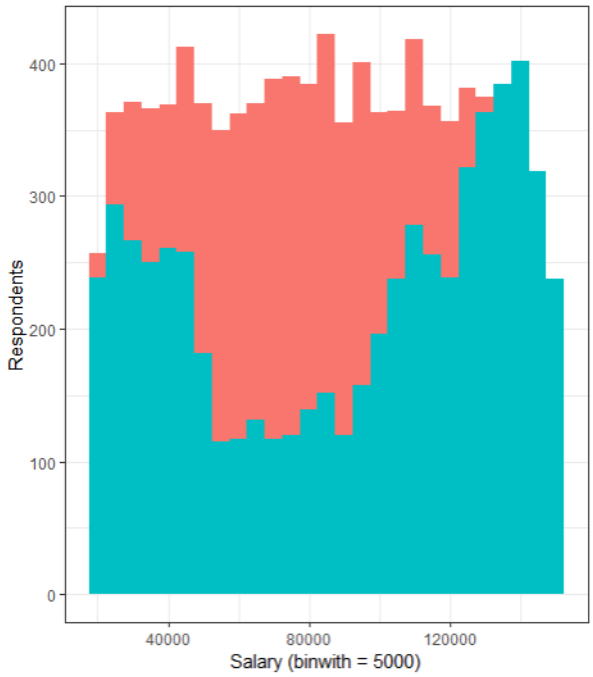


Figure 2.2: Density plot on sales and brand preference

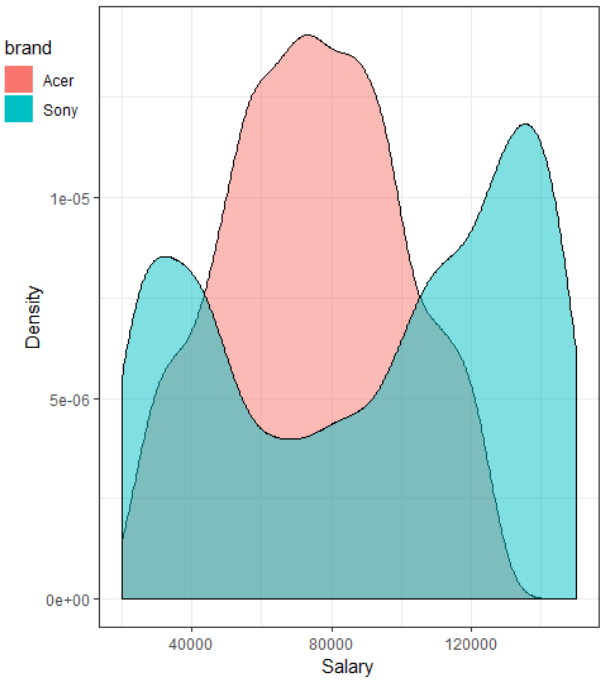


Figure 3 is a scatterplot with the respondents' age and salary on respectively the x- and y-axis. The different colors provide their brand preference. Firstly, it provides us with the same insight as the previous figures have given: on average, respondents with a lower yearly salary and respondents a higher yearly salary generally prefer Sony over Acer and people that have a medium yearly salary prefer Acer over Sony.

Figure 3: Scatterplot with age, salary and brand preference

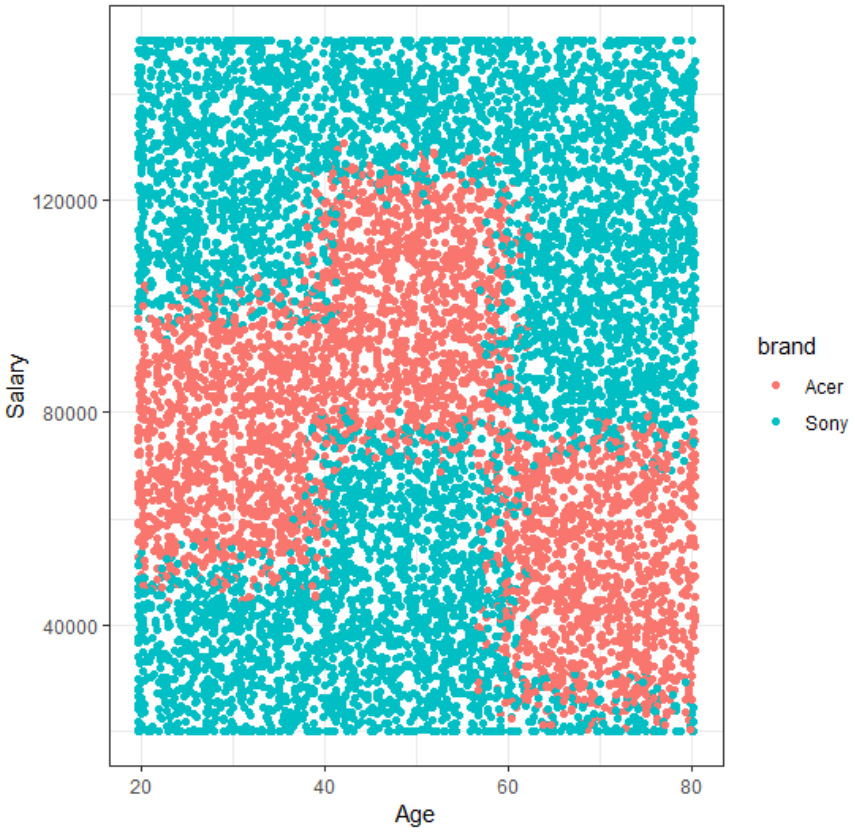


Figure 3 however provides a more accurate overview regarding the distribution of these values in comparison to the previous figures. As it turns out, people aged 20 to 40 with lower-medium incomes prefer Acer over Sony. People aged 40 to 60 with higher-medium incomes prefer Acer over Sony. People aged 60 to 80 with low/lower-medium incomes prefer Acer over Sony. Although it was not apparent from the figures with age and brand preference as portrayed in the Appendix, age does seem to have an impact on the respondents' preferred brand.

2.3) Building an accurate predictive model:

With help from the *caret* package in R, the data was split, the models were cross validated, parameters for each algorithm were finetuned, predictions were made and the outcomes were evaluated.

The complete survey with 9898 observations was split in 2 separate sets: 75% of the total observations was used to generate a set to train the models on, and the remaining 25% was used to generate a set to test the models on. The Caret package stratifies each set, so the training and testing set have similar distributions across the variables. Three algorithms have been used to generate a predictive model, all with very accurate results. The algorithms are:

- C5.0 Decision tree
- Random Forest
- K-Nearest Neighbors

Each algorithm had its own parameters to finetune, which was done as follows:

- For the C5.0 Decision Tree, I let the caret package automatically choose for the best parameters from a total of 40 tries.
- For the Random Forest and K-Nearest Neighbors I ran the parameters through a grid to determine the parameter which provided the most accurate model. The Random Forest parameter *Mtry*, (*the number of variables randomly sampled as candidates at each split)* was tested for values ranging from 1 to 20.
- The k-NN parameter k (*the number of nearest datapoints an unknown point will base its value on*) was tested for values ranging from 1 to 150.

In the k-NN prediction predictive variables were reduced to two: age and yearly salary. Both variables were normalized. The number of folds for cross validation was set to 10. The process was not repeated multiple times. The models with the best performing parameters were tested on the test-set and afterwards accuracy and Cohen´s kappa were calculated. The results are displayed in the next chapter.

2.4) Model estimates:

Complete tables of outcomes can be found in the Appendix. This chapter will give the parameters which performed the best, and the outcomes related to these parameters.

Table 2.1, 2.2 and 2.3 contain the values optimizing the predictive model for the C5.0 Decision Tree, the Random Forest, and k-Nearest Neighbors, together with their cross validated accuracy and Kappa:

Table 2.1: Parameter values for C5.0 DT

Model	**Model**	**Winnow**	**Trials**	*Accuracy*	*Kappa*
C5.0 DT	Rules	FALSE	40	*0.922954*	*0.836243*

Table 2.2: Parameter values for Random Forest

Model	**Ntrees**	**Mtry**	*Accuracy*	*Kappa*
Random Forest	500	14	*0.924031*	*0.838935*

Table 2.3:: Parameter values for k-NN

Model	**K**	*Accuracy*	*Kappa*
k-NN	69	*0.92713*	*0.84535*

Using these parameters on each model, the test-set's preferred brand was predicted and evaluated against its real value. **Table 3.1, 3.2** and **3.3** contain the accuracy and kappa of each model, together with their confusion matrix.

Figures 4.1, 4.2 and **4.3** show the wrongly predicted values per preferred brand, plotted on an age/salary scatter like the one in figure 3 on page 4. It is clearly observable that each model is accurate and fails around the same area: the borders of the age/salary groups where the demographics are split between a Sony preference and an Acer preference.

There are no areas distinguishable where one algorithm outshines the others. C5.0 Decision Tree puts most importance in variables that do not seem to have an extreme importance. Furthermore, k-NN uses only salary to predict preferred brand, whereas age also seems to matter. The Random Forest model will be used to evaluate the incomplete survey as it has the highest accuracy and kappa (although with values this close to each other, that doesn't say anything).

Table 3.1: Confusion matrix for the C5.0 predicted model and overall important variables.

	Reference		
Prediction	**Acer**	**Sony**	*Prediction accuracy*
Acer	841	105	0.889
Sony	95	1433	0.938
Reference accuracy	*0.899*	*0.932*	
Test model accuracy	0.919		
Kappa	0.826		

Top 3 variables of importance	
age	100
salary	100
Chrysler	32.2

Table 3.2: Confusion matrix for the Random Forest predicted model and overall important variables

	Reference		
Prediction	**Acer**	**Sony**	*Prediction accuracy*
Acer	844	105	*0.889*
Sony	92	1433	*0.940*
Reference accuracy	*0.902*	*0.932*	
Test model accuracy	0.920		
Kappa	0.831		

Top 3 variables of importance	
salary	100
age	60.7
credit	12.1

Table 3.3: Confusion matrix for the k-NN predicted model and overall important variables

	Reference		
Prediction	**Acer**	**Sony**	*Prediction accuracy*
Acer	846	111	*0.884*
Sony	90	1427	*0.941*
Reference accuracy	*0.904*	*0.928*	
Test model accuracy	0.919		
Kappa	0.828		

Top 2 variables of importance	
salary	100
age	0

Figure 4.1: Errors from C5.0 Decision Tree model per brand

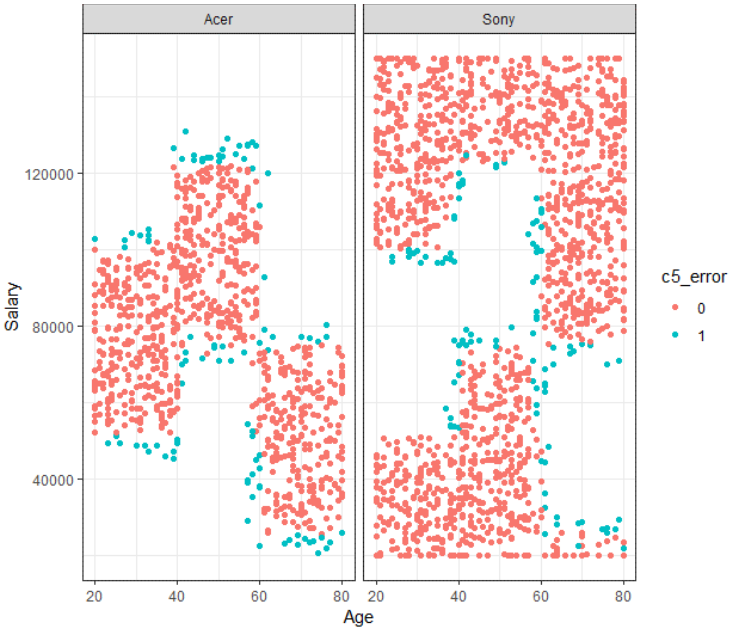


Figure 4.2 Errors from Random Forest model per brand

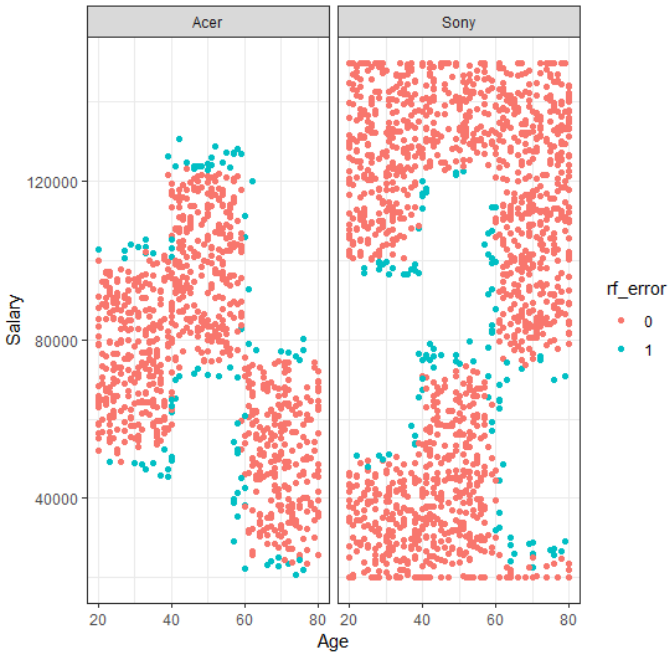
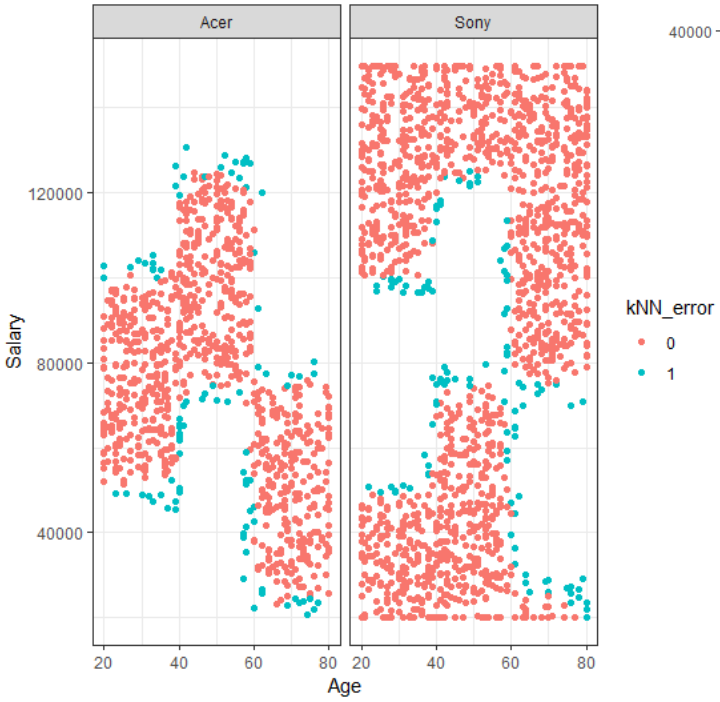


Figure 4.3: Errors from kNN model per brand



*the pink dots in the figure above display the correct predictions of a respondent's preferred brand, the blue dots show the mis predicted brand preferences

3) Conclusion

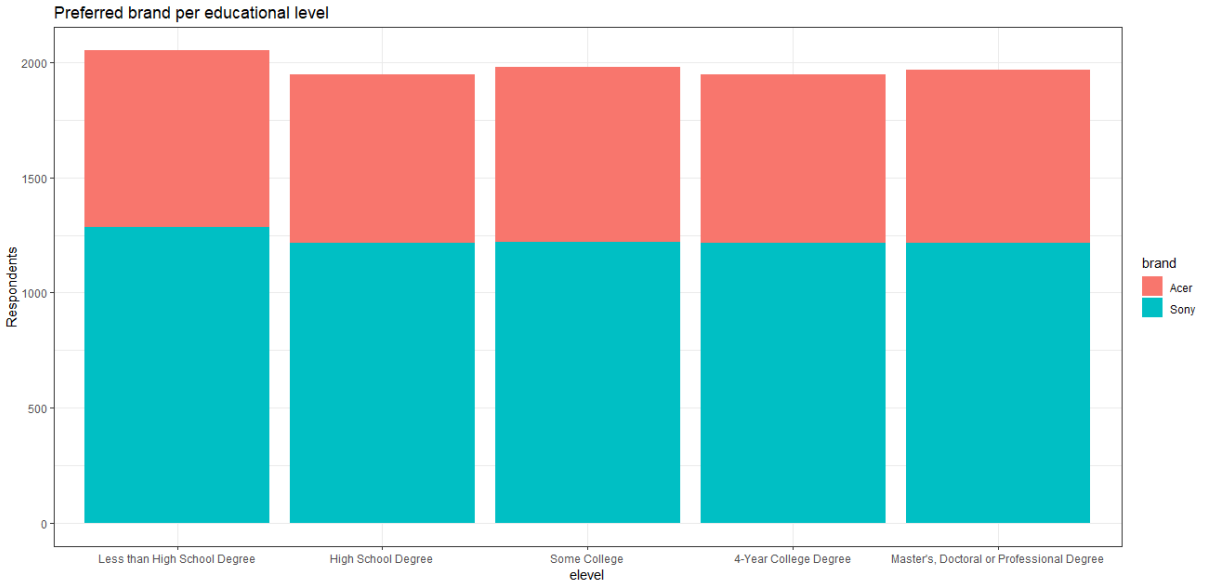
The Random Forest model predicted the missing brand preferences for the incomplete survey data. Therefore, we now have a complete survey. However, it needs to be double checked to see whether the incomplete survey data does not consist of too many respondents in the age/salary ranges where all models had difficulties in predicting the correct preference. For now, the assumption is made that this is not the case and that we can rely on the outcome.

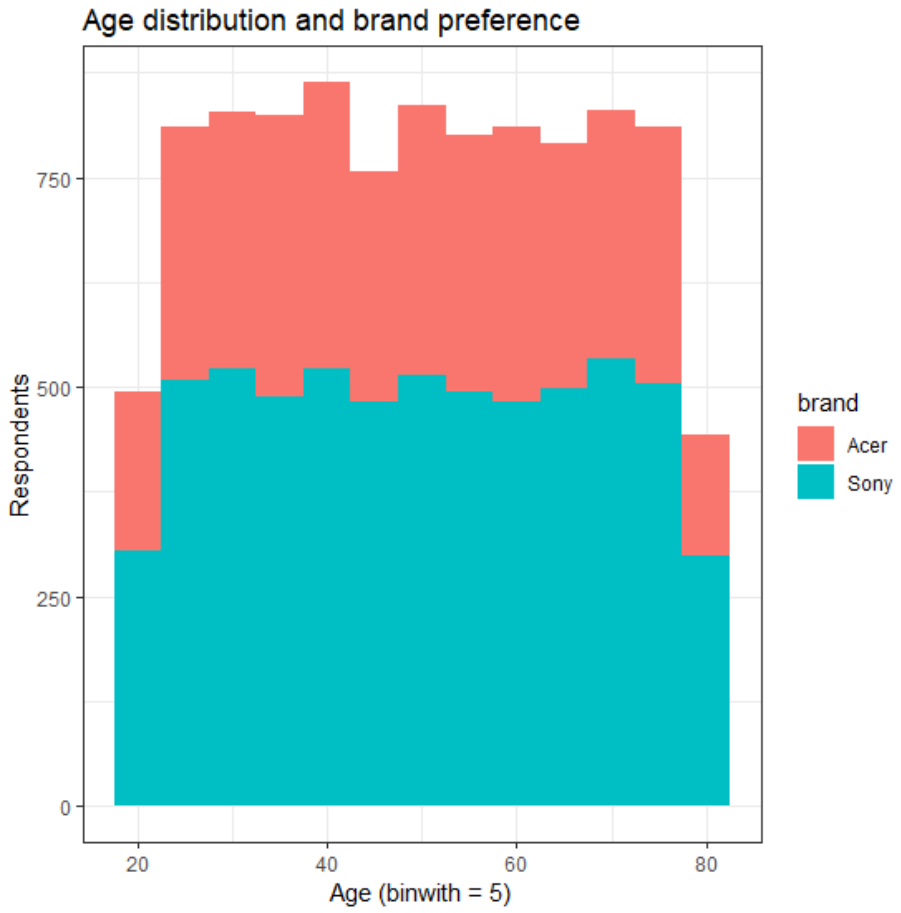
In the incomplete survey, 1897 respondents were predicted to prefer Acer over Sony. 3103 respondents were predicted to prefer Sony over Acer. This corresponds 38% of the respondents preferring Acer and 62% of the respondents preferring Sony. When combining these results with the complete survey, a total of 9,257 of the respondents prefers Sony and 5,641 of the respondents prefers Acer. Again, that is 62% of the respondents that prefers Sony over Acer, and 38% that prefers Acer over Sony.

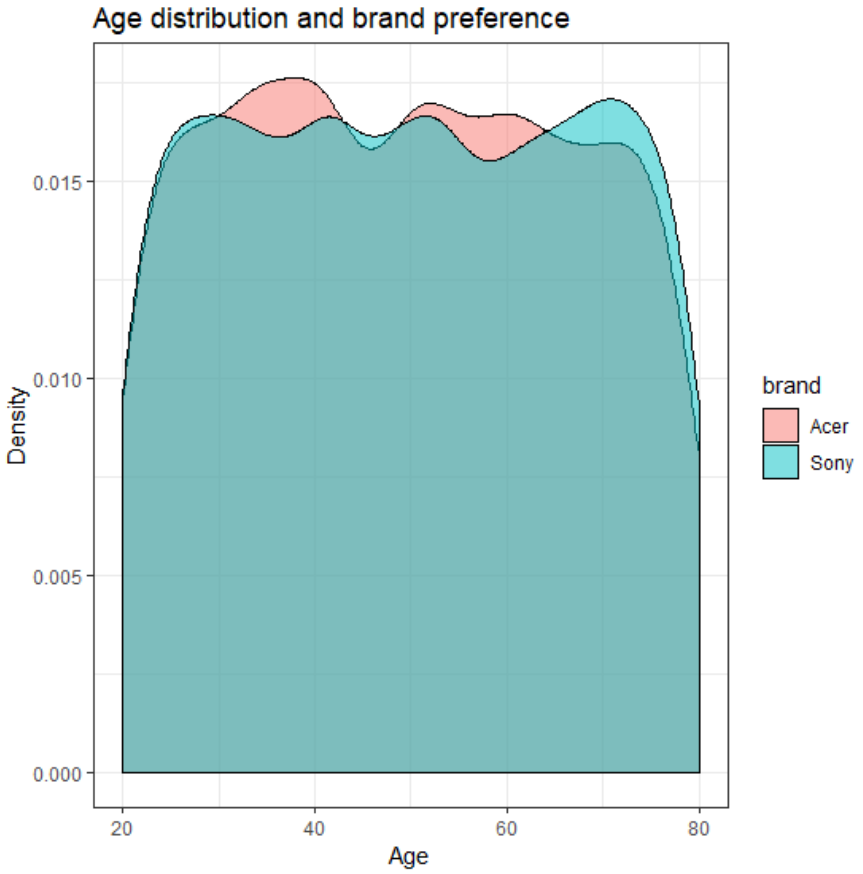
This indicates that Sony is clearly in general the most preferred brand. However as has been shown in chapter 2 of this report, there are clear distinct groups within the respondents' demography which prefer on over the other. Therefore, the choice of which brand to we should contact for more cooperation depends on which types of consumers Blackwell wants to attract the most. My suggestion for the management team are therefore as follows:

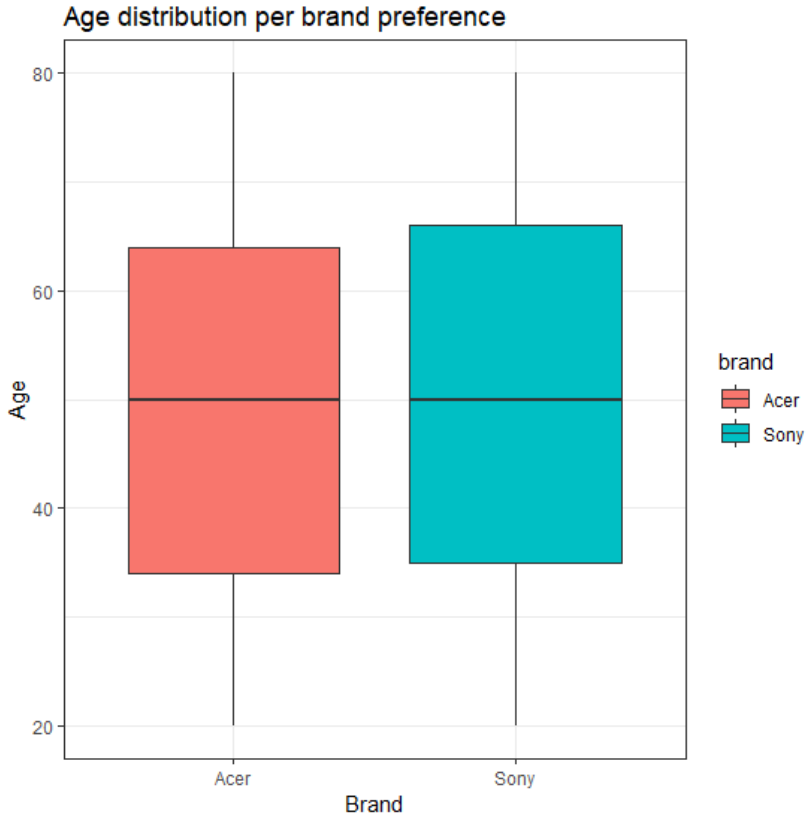
1) Make the data department and the marketing department do a thorough investigation on the differences between the specific consumer groups so we can answer the following questions:
    - Which groups are the most profitable?
    - Which groups does Blackwell want to affiliate with the most?
    - Which groups are the easiest to influence using marketing techniques?
    - Which groups tend to buy at Blackwell more often?
    - Etc.
2) With these questions answered, go back to the results in this data
3) Choose the brand to engage with more intensively based on the outcomes of all the research

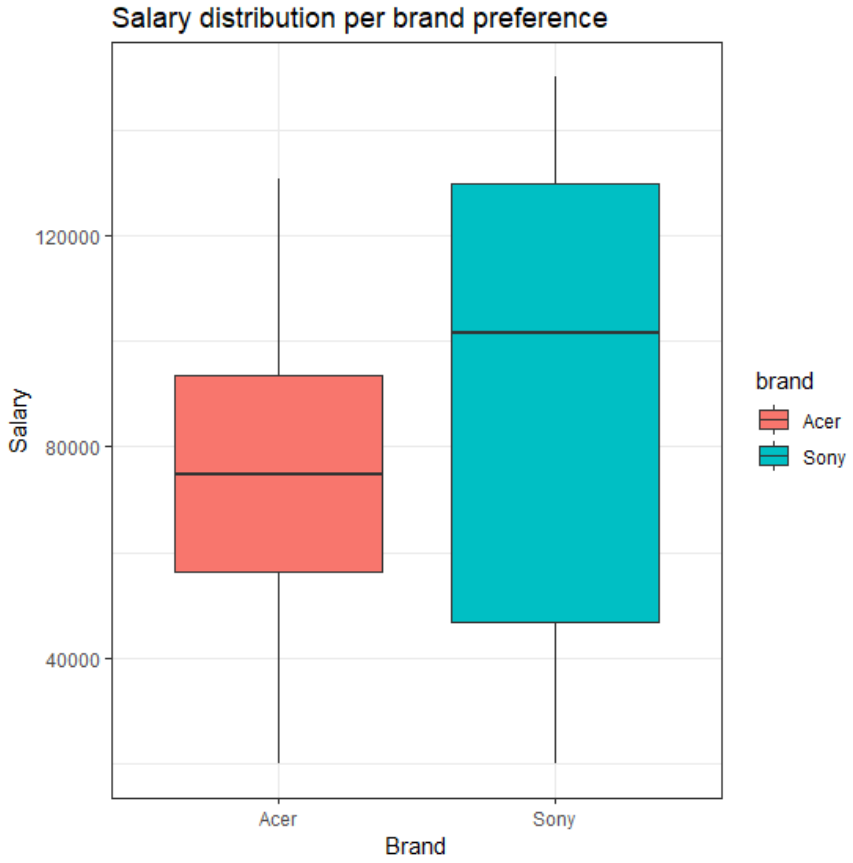
Appendix


Preferred brand per educational level
2000
1500
1000
500
0
Respondents
Less than High School Degree
High School Degree
Some College
4-Year College Degree
Master's, Doctoral or Professional Degree
elevel
brand
Acer
Sony


Age distribution and brand preference
750
500
250
0
Respondents
20
40
60
80
Age (binwith = 5)
brand
Acer
Sony


Age distribution and brand preference
Density
0.015
0.010
0.005
0.000
20
40
60
80
Age
brand
Acer
Sony


Age distribution per brand preference
Age
80
60
40
20
Acer
Sony
Brand
brand
Acer
Sony


Salary distribution per brand preference
120000
80000
40000
Salary
Acer
Sony
Brand
brand
Acer
Sony

C5.0 values with Accuracy and Kappa

model	winnow	trials	Accuracy	Kappa
rules	FALSE	1	0.817216	0.635935
rules	FALSE	10	0.921338	0.832118
rules	FALSE	20	0.922279	0.834744
rules	FALSE	30	0.922685	0.835612
rules	**FALSE**	**40**	**0.922954**	**0.836243**
rules	FALSE	50	0.922954	0.836243
rules	FALSE	60	0.922954	0.836243
rules	FALSE	70	0.922954	0.836243
rules	FALSE	80	0.922954	0.836243
rules	FALSE	90	0.922954	0.836243
rules	TRUE	1	0.812364	0.627283
rules	TRUE	10	0.908132	0.80693
rules	TRUE	20	0.908805	0.809052
rules	TRUE	30	0.908671	0.808784
rules	TRUE	40	0.90894	0.809415
rules	TRUE	50	0.90894	0.809415
rules	TRUE	60	0.90894	0.809415
rules	TRUE	70	0.90894	0.809415
rules	TRUE	80	0.90894	0.809415
rules	TRUE	90	0.90894	0.809415
tree	FALSE	1	0.810747	0.617911
tree	FALSE	10	0.920396	0.830939
tree	FALSE	20	0.921339	0.833293
tree	FALSE	30	0.920799	0.832121
tree	FALSE	40	0.920799	0.832121
tree	FALSE	50	0.920799	0.832121
tree	FALSE	60	0.920799	0.832121
tree	FALSE	70	0.920799	0.832121
tree	FALSE	80	0.920799	0.832121
tree	FALSE	90	0.920799	0.832121
tree	TRUE	1	0.811556	0.619308
tree	TRUE	10	0.922688	0.83539
tree	TRUE	20	0.920263	0.830647
tree	TRUE	30	0.92161	0.833474
tree	TRUE	40	0.92161	0.833474
tree	TRUE	50	0.92161	0.833474
tree	TRUE	60	0.92161	0.833474
tree	TRUE	70	0.92161	0.833474
tree	TRUE	80	0.92161	0.833474
tree	TRUE	90	0.92161	0.833474

Random Forest, grid for mtry = 1 to mtry = 20 with Accuracy and Kappa

mtry	Accuracy	Kappa
1	0.621767	0
2	0.621902	0.000442
3	0.734645	0.367761
4	0.84469	0.66204
5	0.886314	0.758304
6	0.908539	0.806214
7	0.916082	0.822374
8	0.918239	0.826769
9	0.920933	0.832371
10	0.921876	0.834461
11	0.921606	0.833823
12	0.923762	0.838289
13	0.924031	0.838796
14	**0.924031**	**0.838935**
15	0.924031	0.838901
16	0.922954	0.836488
17	0.923357	0.837416
18	0.922012	0.834615
19	0.923627	0.838077
20	0.923357	0.837444

k-NN: grid from k = 2 to k = 100 with Accuracy and Kappa

k	Accuracy	Kappa
2	0.889147	0.764641
3	0.907333	0.803063
4	0.904371	0.796732
5	0.907468	0.803506
6	0.907467	0.803186
7	0.909621	0.807874
8	0.910294	0.809454
9	0.913122	0.815333
10	0.911776	0.812587
11	0.91339	0.816166
12	0.913124	0.815518
13	0.914604	0.8188
14	0.916356	0.822443
15	0.915951	0.821485
16	0.916759	0.823274
17	0.917162	0.824412
18	0.914739	0.819107
19	0.916489	0.822949
20	0.914604	0.819013
21	0.916622	0.823366
22	0.915681	0.82142
23	0.915815	0.821814
24	0.915545	0.821242
25	0.917026	0.824269
26	0.916892	0.823922
27	0.915546	0.821309
28	0.916489	0.823377
29	0.916623	0.823485
30	0.915679	0.821368
31	0.915949	0.822027
32	0.915142	0.820359
33	0.915006	0.820061
34	0.915949	0.822012
35	0.915949	0.822108
36	0.914737	0.819598
37	0.915949	0.822074
38	0.915275	0.820651
39	0.916353	0.823014
40	0.915948	0.82211
41	0.915814	0.821871
42	0.916353	0.822979
43	0.915545	0.82122
44	0.914602	0.819136
45	0.914736	0.81957
46	0.914737	0.819438
43	0.915545	0.82122
44	0.914602	0.819136
45	0.914736	0.81957
46	0.914737	0.819438
47	0.915005	0.820024
48	0.914332	0.81854
49	0.914467	0.818863
50	0.914198	0.818317
51	0.915007	0.820062
52	0.914332	0.818635
53	0.913659	0.817245
54	0.913524	0.816975
55	0.913524	0.816857
56	0.91312	0.815987
57	0.912312	0.814275
58	0.912985	0.815831
59	0.912448	0.814669
60	0.925514	0.841998
61	0.92686	0.84475
62	0.926187	0.843333
63	0.926053	0.843111
64	0.926322	0.843604
65	0.926726	0.844493
66	0.92686	0.844764
67	0.92713	0.84541
68	0.926726	0.844497
69	**0.92713**	**0.84535**
70	0.92686	0.844847
71	0.926457	0.844009
72	0.926053	0.843067
73	0.926053	0.843084
74	0.925109	0.841064
75	0.924301	0.839356
76	0.925513	0.841979
77	0.924571	0.839993
78	0.92538	0.841755
79	0.92538	0.841715
80	0.92484	0.840563
81	0.925514	0.841956
82	0.925379	0.841691
83	0.924572	0.840033
84	0.924437	0.839736
85	0.924572	0.840099
86	0.924842	0.840616
87	0.925111	0.841254
88	0.925111	0.841307
89	0.925516	0.842028
90	0.924841	0.840634
91	0.92538	0.841754
92	0.924975	0.840984
93	0.924841	0.840647
94	0.924303	0.839462
95	0.924841	0.840697
96	0.924168	0.839205
97	0.924841	0.840692
98	0.924168	0.839389
99	0.923629	0.838235
100	0.923226	0.837425