# CUSTOMER BRAND PREFERENCES REPORT

**Oloruntobi Adelokun**

14 November 2019

The CTO, Danielle Sherman requested for a report that will predict Blackwell Electronics customers' brand preferences that are missing from an incomplete survey by conducting two classification methods in R.

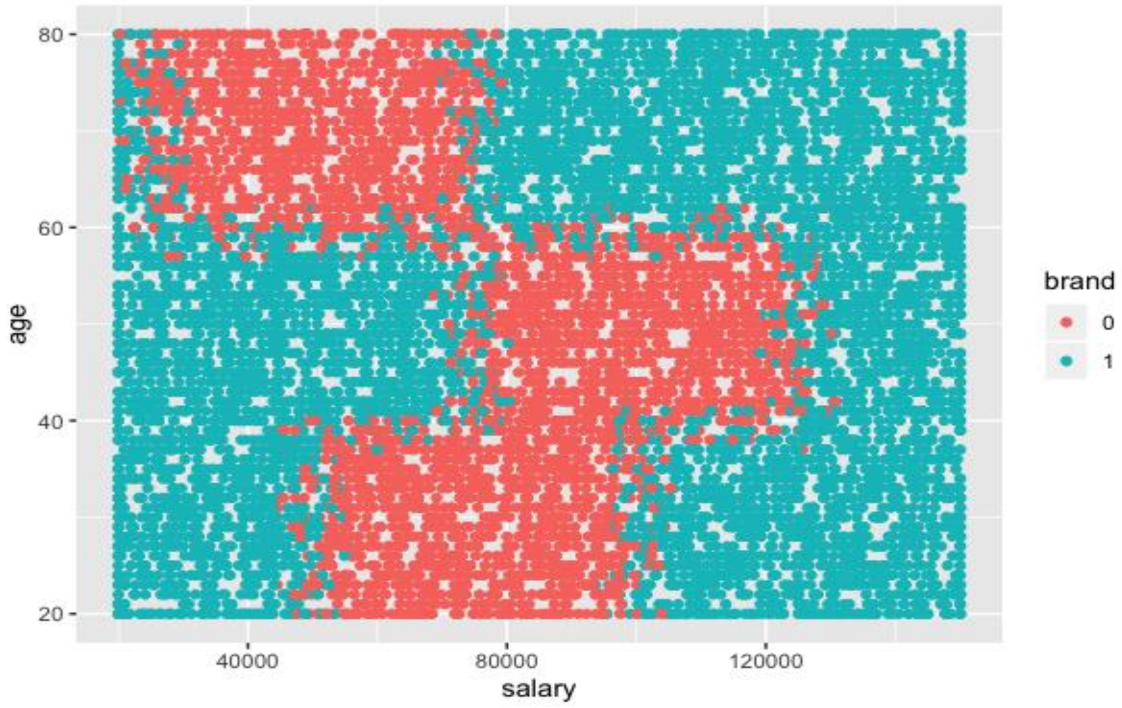
## DATA PRE-PROCESSING

Two sets of data ("Complete Responses and Incomplete Survey datasets) were provided consists of 7 variables namely "salary", "age", "elevel", "car", "zipcode", "credit" and "brand". There were no missing values in the Complete Reponses datasets.

The variables "elevel", "car", "zipcode" and "brand" were converted to factor variables.

## FEATURE ENGINEERING

A relationship was found between "age", "salary" and brand using scatterplot diagram as shown below:



Per the scatterplot above, we can visualize that age can be stratified into 3 groups so on, while salary into 5 groups. Thus, the variables "age" and "salary" is binned into 3 and 5 respectively. These are converted into factors again. I now choose the variable of interest her ("age", "salary" and "brand" using the "features" algorithm. This is now set as our new data to train. Below is an overview of the new data.

| | brand | age_bin | salary_bin |
|---|---|---|---|
| 1 | Acer | (40,60) | (9.8e+04,1.24e+05] |
| 2 | Sony | (60,80.1) | (9.8e+04,1.24e+05] |
| 3 | Acer | (19.9,40) | (7.2e+04,9.8e+04] |
| 4 | Sony | (40,60) | (4.6e+04,7.2e+04] |
| 5 | Acer | (19.9,40) | (4.6e+04,7.2e+04] |
| 6 | Sony | (40,60) | 1.24e+05,1.5e+05] |

## MODEL TRAINING

The "createDataPartition" function in caret package was used to create training and testing sets. The training set being 75% and the remaining 25% representing the testing set.

### C5.0 Model

Using decision tree, C5.0 on the training set with 10 fold cross validation and an Automatic Tuning Grid with a tuneLength of 2. The output of this model is shown below:

```
7424 samples
  2 predictor
  2 classes: 'Acer', 'Sony'

Pre-processing: centered (6), scaled (6)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 6683, 6682, 6682, 6682, 6681, 6681, ...
Resampling results across tuning parameters:

  model  winnow  trials  Accuracy   Kappa
  rules  FALSE    1      0.9066560  0.8036694
  rules  FALSE   10      0.9043219  0.7980327
  rules  FALSE   20      0.9030201  0.7950848
  rules   TRUE    1      0.9066560  0.8036694
  rules   TRUE   10      0.9043219  0.7980327
  rules   TRUE   20      0.9030201  0.7950848
  tree   FALSE    1      0.9066560  0.8036694
  tree   FALSE   10      0.9057575  0.8014350
  tree   FALSE   20      0.9066560  0.8036694
  tree    TRUE    1      0.9066560  0.8036694
  tree    TRUE   10      0.9057575  0.8014350
  tree    TRUE   20      0.9066560  0.8036694

Accuracy was used to select the optimal model using the
 largest value.
The final values used for the model were trials = 1, model =
 rules and winnow = TRUE.
```

**C5.0 variable importance**

```
                             Overall
salary_bin (7.2e+04,9.8e+04]   100.00
salary_bin (4.6e+04,7.2e+04]    50.51
age_bin( 60,80.1]               50.26
age_bin( 40,60]                 26.96
salary_bin (1.24e+05,1.5e+05]   25.43
salary_bin (9.8e+04,1.24e+05]    0.00

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were trials = 1, model = rules and winnow =
TRUE.
```

**Random Forest**

Using Random Forest with 10-fold cross validation and manually tuning 5 different mtry values (1-5), the model returns this output

```
7424 samples
   2 predictor
   2 classes: 'Acer', 'Sony'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 6681, 6681, 6681, 6682, 6681, 6683, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  1     0.7272782  0.3270384
  2     0.8491396  0.6648566
  3     0.9046778  0.7988311
  4     0.9066545  0.8035993
  5     0.9066545  0.8035993

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 4.
```

Comparing the accuracy and Kappa for the models, it can be deduced that the best model to pick for the prediction is C5.0 because of the high accuracy and Kappa values.

## PREDICTION

The C5.0 model was used to predict the brands preference of the incomplete data using the “predict” function.

First of all, the data was pre-processed by binning the variables “age” and “salary” into 3 and 4 groups respectively as done with the complete data. These variables are then selected for further analytics. The new data looks like this

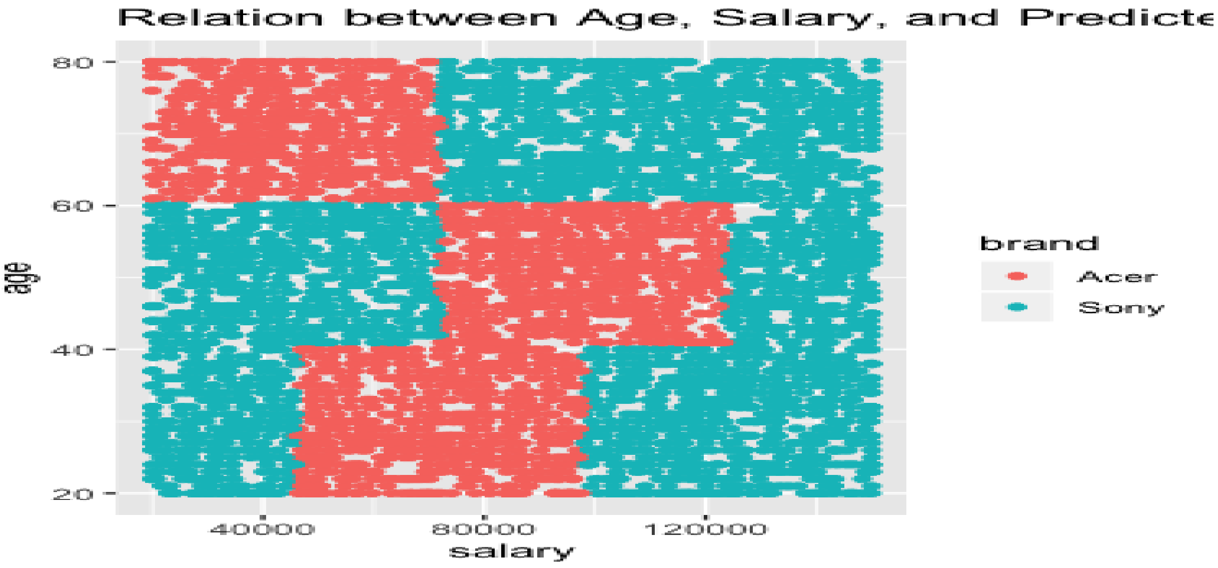
```
      brand  age_bin      salary_bin

1       0     (40,60]     (9.8e+04,1.24e+05]
2       0     (40,60]     (1.24e+05,1.5e+05]
3       0     (40,60]     (9.8e+04,1.24e+05]
4       0     (40,60]      (1.99e+04,4.6e+04]
5       0     (40,60]       (7.2e+04,9.8e+04]
6       0     (60,80.1]     (1.99e+04,4.6e+04]
```

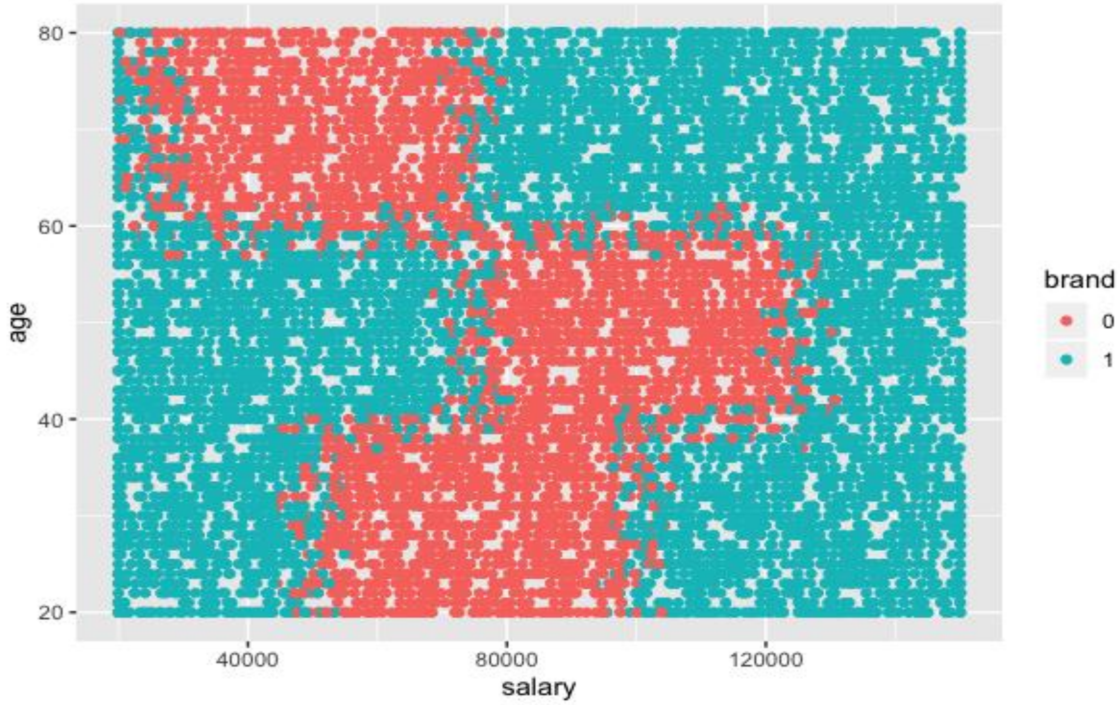
A summary of the prediction after applying the model is shown below:

```
Acer Sony
1967 3033
```

## Checking the Prediction



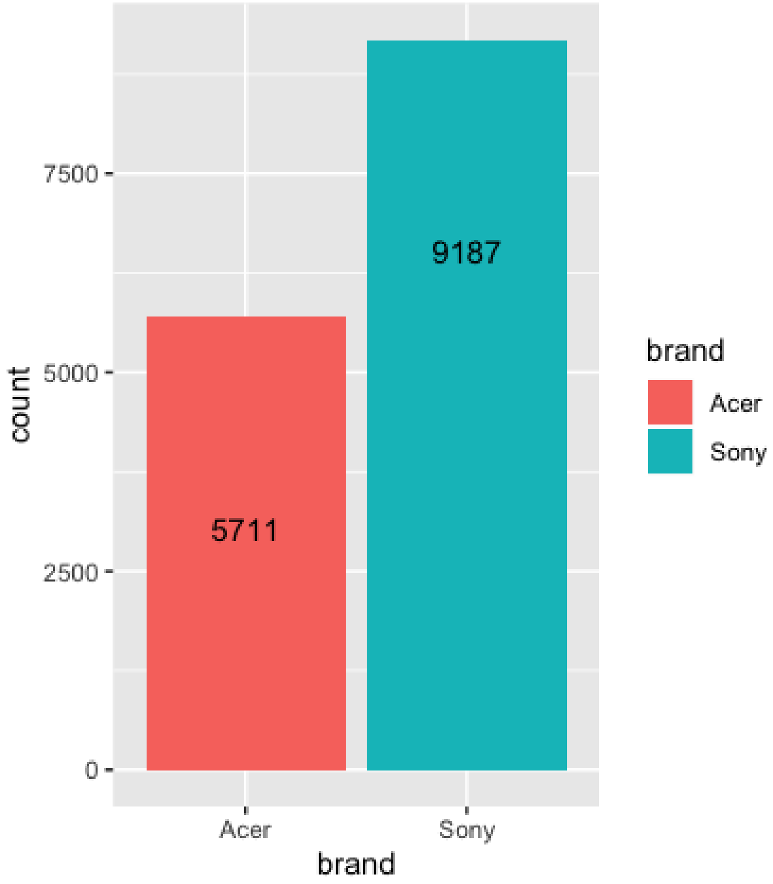
Plot: Relation between Age, Salary, and Brand (Predictions)



Plot: Relation between Age, Salary, and Brand (CompleteResponses)

Per both plots, we can observe that the prediction and the real survey responses look almost the same thus giving us comfort over our model.

## CUSTOMERS’ PREFERENCE (REAL AND PREDICTED)



## RECCOMMENDATIONS

- Blackwell’s customers prefer to Sony to Acer; thus, we should prioritize stocking our inventory with Sony products.
- Also, per the scatterplot diagram, it’s obtainable that the customers that prefer Sony products are those
    - aged 20 to 40 and with salary of 60,000 to 100,000
    - aged 40 to 60 with salary range of 80,000 to 130,000
    - aged 60 to 80 with salary of 10,000 to 70,000.

We should tailor advertising to focus on this age groups for Sony products.