Classification: Predict which Brand of Products Customers Prefer

Introduction

The sales team at Blackwell electronics had engaged a marketing research firm to conduct survey of the existing customers. The main purpose of this survey was to find out which brand of computers customers prefer, which will help Blackwell electronics in deciding which manufacturer they should pursue a deeper strategic relationship. But when the survey took place brand preference question was not properly captured for all the respondents.

Danielle Sherman – CTO of Blackwell electronics has asked to utilize R and RStudio to investigate if customer responses to some survey questions like income, ages, salary, etc. that can help in predicting answer to customer brand preference question which are missing in incomplete surveys. To perform this investigation he would like us to run and optimize two different decision tree classification methods (1) R – C5.0 and (2) Random Forest and compare which one works better and use this trained model to make preferred brand predictions for incomplete surveys.

Method

In this study analysis was performed on complete survey data using common data mining approach containing data exploration, pre-processing/feature engineering, training, testing, and optimizing models, prediction, and evaluation.
For the purpose of this analysis the 'caret' package for statistical programming language R is used applying C5.0 and Random Forest. These models were trained, and performance evaluation was done in terms of accuracy, kappa, and confusion matrix. After performing model comparison and selecting best performing model; selected model was utilized to make preferred brand predictions for the incomplete survey data.

Initial Data Exploration:

Complete Responses Data Structure:

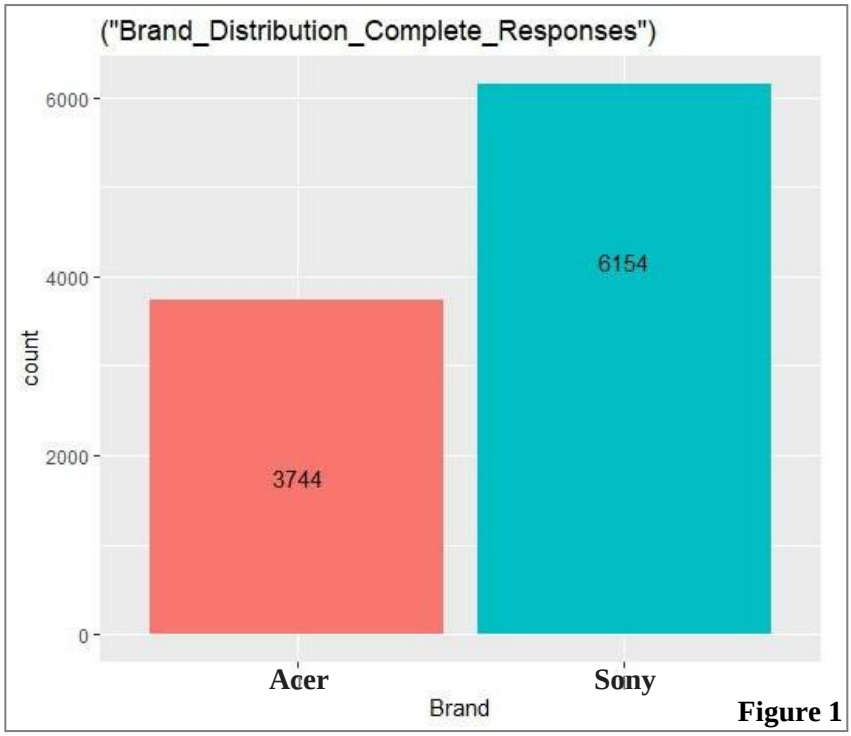

Figure 1

The complete survey data shows that about 3744 customers prefer Acer while 6154 customers prefer Sony brand (Figure 1), and this preference is influenced by selective variables in data set including Yearly Salary, Age, Education level, Zip code (area), Available Credit and Car make.

Box Plots – Checking for any Outliers

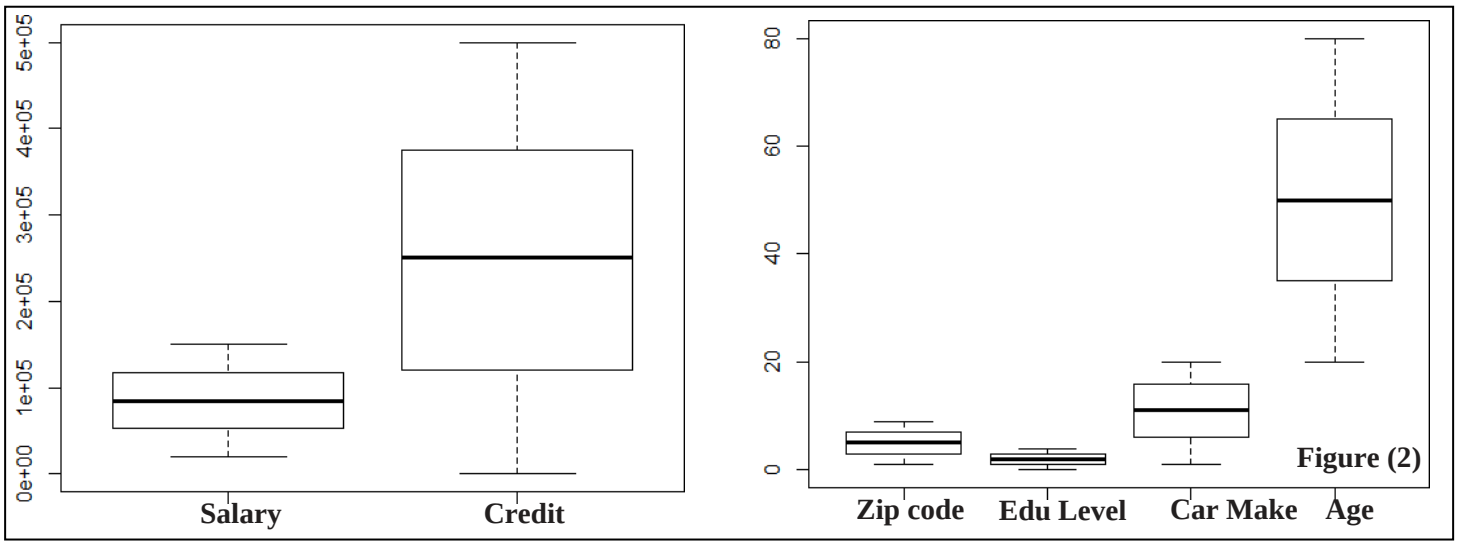

Figure (2)

Figure (2) shows boxplot with no outliers for all features/variables of Complete Responses survey data.

Date type and data transformation:

```
'data.frame':   9898 obs. of  7 variables:
 $ Yearly_Salary   : num  119807 106880 78021 63690 50874 ...
 $ Age             : int  45 63 23 51 20 56 24 62 29 41 ...
 $ Edu_Level       : int  0 1 0 3 3 3 4 3 4 1 ...
 $ Car_Make        : int  14 11 15 6 14 14 8 3 17 5 ...
 $ Zipcode         : int  4 6 2 5 4 3 5 0 0 4 ...
 $ Available_Credit: num  442038 45007 48795 40889 352951 ...
 $ Brand           : int  0 1 0 1 0 1 1 1 0 1 ...
```

Figure (3)

Of the 7 attributes, salary, age and available credit should be numeric while brand, zip code, car make, and education level were transformed as factors. (Figure 3 & 4).

```
$ Yearly_Salary    : num  119807 106880 78021 63690 50874 ...
$ Age              : num  45 63 23 51 20 56 24 62 29 41 ...
$ Edu_Level        : Factor w/ 5 levels "0","1","2","3",..: 1 2 1 4 4 4 5 4 5 2 ...
$ Car_Make         : Factor w/ 20 levels "1","2","3","4",..: 14 11 15 6 14 14 8 3 17 5 ...
$ Zipcode          : Factor w/ 9 levels "0","1","2","3",..: 5 7 3 6 5 4 6 1 1 5 ...
$ Available_Credit : num  442038 45007 48795 40889 352951 ...
$ Brand            : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 2 2 1 2 ...
```

Figure (4)

Feature Selection

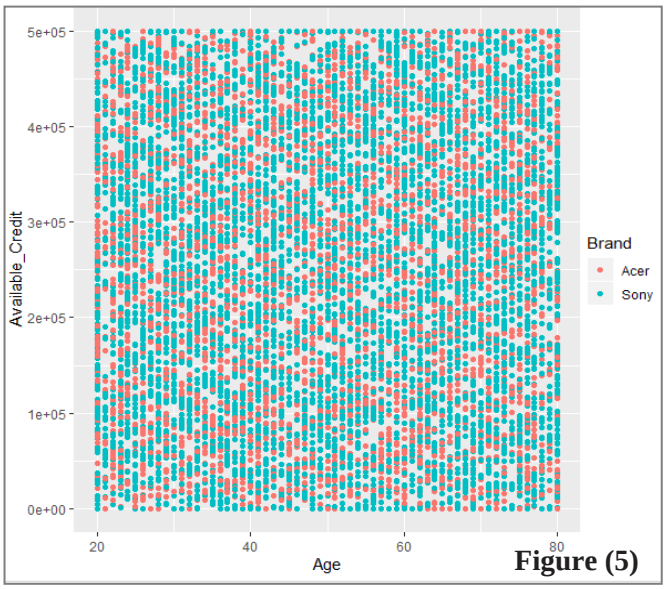

Figure (5)

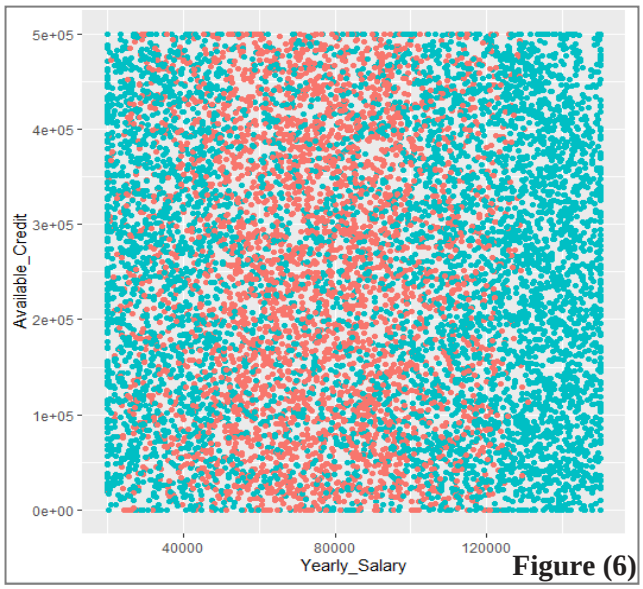

Figure (6)

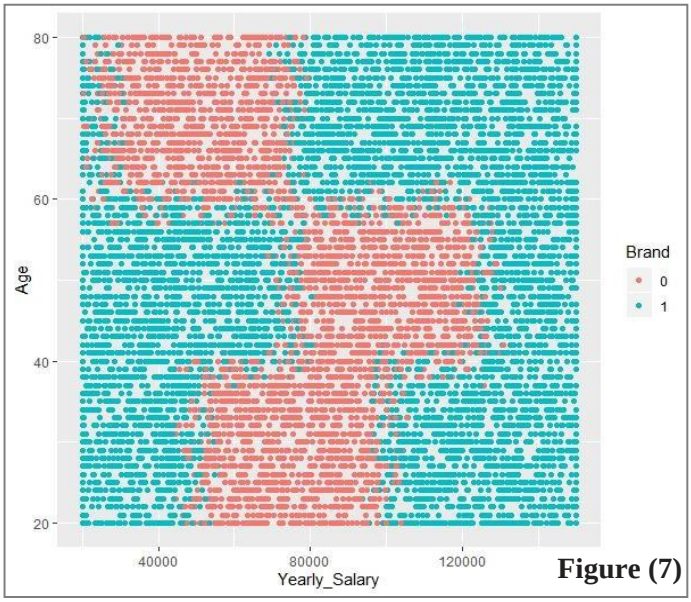

Figure (7)

No relationship is observed between age/available credit and yearly salary/available credit (Figures 5 & 6). A clear pattern is observed between age, salary, and brand preference. From Figure (7) Sony is preferred brand by customers of age range 20 to 40 (high income), 40 to 60 (low income) and 60 to 80 (mid to high income).While Acer is preferred brand by customers of ages 20 to 40 (mid income), 40 to 60 (mid – high income) and 60 to 80 (low income). This suggests that age and salary are two important features with meaningful values and will have most influence on predicting preferred customer brand. Looking at relation between age and salary, binning has been used to divide age into 3 bins (20 to 40, 40 to 60, and 60 to 80) and salary into 5 bins. ($20k to $46k, $46k to $72k, $72k – $98k, $98k – $124k, $124k – $150k)

Data Partition/Split

The main purpose of this study is to predict preferred customer brand using C5.0 and Random Forest classification models. First set in doing so is splitting "Complete Responses" data set into training (75%) and testing (25%) sets. Secondly, the model is trained using the training set and this model is applied to testing set. Accuracy and Kappa the performance measures used to compare and select best fit model. This best fit model is then applied to "Survey Incomplete" data set to make brand predictions.

Classification Model: Tuning, Training and Assessment

Both C5.0 and Radom Forest were tuned using 10-fold cross validation method using repeated method (repeat =1).

C5.0 – Classification Model

Figure (8)

```
C5.0

7424 samples
   6 predictor
   2 classes: '0', '1'

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

  model  winnow  trials  Accuracy   Kappa
  rules  FALSE    1      0.9089433  0.8081801
  rules  FALSE   10      0.9069245  0.8033066
  rules  FALSE   20      0.9063854  0.8021638
  rules   TRUE    1      0.9089433  0.8081801
  rules   TRUE   10      0.9059809  0.8013665
  rules   TRUE   20      0.9059793  0.8012832
  tree   FALSE    1      0.9089433  0.8081801
  tree   FALSE   10      0.8987076  0.7843315
  tree   FALSE   20      0.8993815  0.7859240
  tree    TRUE    1      0.9089433  0.8081801
  tree    TRUE   10      0.9042287  0.7970078
  tree    TRUE   20      0.9080003  0.8058669

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.
```

Figure (9)

```
C5.0 variable importance

  only 20 most important variables shown (out of 38)

                              Overall
SalaryBin(7.2e+04,9.8e+04]   100.00
AgeBin(60,80.1]               99.31
AgeBin(40,60]                 88.36
SalaryBin(4.6e+04,7.2e+04]    77.63
SalaryBin(1.24e+05,1.5e+05]   64.89
SalaryBin(9.8e+04,1.24e+05]   53.15
```

C5.0 model was applied on training set to train the model yields best Accuracy of 90% and Kappa of 81% Figure to the left.

Figure (9) shows variables of importance. It is noticeably clear from this list that Yearly Salary and Age are the two most important variables and the model trained is highly influenced by these two variables. Implying that brand predictions for incomplete survey data set will also be affected by these two variables.

Random Forest – Classification Model

```
Random Forest

7424 samples
   6 predictor
   2 classes: '0', '1'

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

  mtry  Accuracy   Kappa
  1     0.6217673  0.0000000
  2     0.7549852  0.4140221
  3     0.8627471  0.6980949
  4     0.8921062  0.7690893
  5     0.9005936  0.7892179

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 5.
```

Figure (10)

```
rf variable importance
  only 20 most important variables shown (out of 38)

                                Overall
SalaryBin(7.2e+04,9.8e+04]      100.000
AgeBin(60,80.1]                  95.818
AgeBin(40,60]                    91.318
SalaryBin(1.24e+05,1.5e+05]      88.765
SalaryBin(4.6e+04,7.2e+04]       71.747
SalaryBin(9.8e+04,1.24e+05]      63.061
```

Figure (11)

Random Forest classification model also shows similar results to C5.0 model, with best Accuracy of 90% and Kappa of 78%. (Figure 10)

Figure (11) shows list of top 20 variables of importance for Random Forest model. Like C5.0 model, Yearly Salary and Age are also two most important variables and that trained model is affected by these two variables. Again, implying that brand predictions will also influenced by these two variables.

We built and trained two models using the survey with complete responses. The predictive model that performed best was our decision tree classifier, C5.0. The next step is to apply our model to the test set, the survey with incomplete responses.

Incomplete Survey Data Structure

The figure (xx) shows brand distribution in incomplete survey data set. This data set looks very imbalanced since the data in incomplete and by default customer preference for brand is set to Acre as evident from distribution plot (figure xx). In this data set 4937 customers prefer Acer vs. Sony.

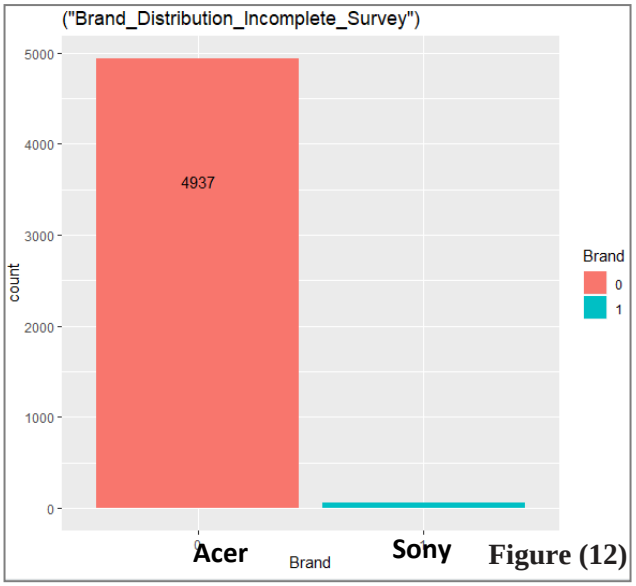

Figure (12)

Box Plots – Checking for any Outliers

Box plots for incomplete survey data set also show that there are no outliers in dataset for any variables.

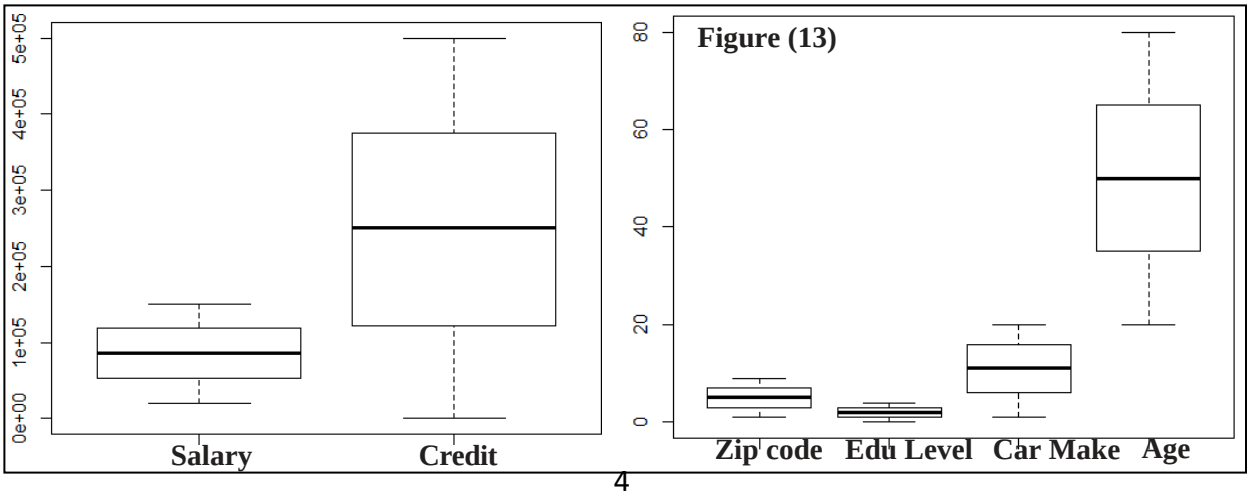

Figure (13)

Data type and data transformation:

Like complete responses data set, Zip code, Car make, education level and Brand were set as factor. Salary, Age and Available credit are numeric (Figure 14 and 15).

```
Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':        5000 obs. of  7 variables:
 $ Yearly_Salary   : num  150000 82524 115647 141443 149211 ...
 $ Age             : num  76 51 34 22 56 26 64 50 26 46 ...
 $ Edu_Level       : num  1 1 0 3 0 4 3 3 2 3 ...
 $ Car_Make        : num  3 8 10 18 5 12 1 9 3 18 ...
 $ Zipcode         : num  3 3 2 2 3 1 2 0 4 6 ...
 $ Available_Credit: num  377980 141658 360980 282736 215667 ...
 $ Brand           : num  1 0 1 1 1 1 1 1 1 0 ...
```

Figure (14)

```
$ Yearly_Salary    : num  150000 82524 115647 141443 149211 ...
$ Age              : num  76 51 34 22 56 26 64 50 26 46 ...
$ Edu_Level        : Factor w/ 5 levels "0","1","2","3",..: 2 2 1 4 1 5 4 4 3 4 ...
$ Car_Make         : Factor w/ 20 levels "1","2","3","4",..: 3 8 10 18 5 12 1 9 3 18 ...
$ Zipcode          : Factor w/ 9 levels "0","1","2","3",..: 4 4 3 3 4 2 3 1 5 7 ...
$ Available_Credit : num  377980 141658 360980 282736 215667 ...
$ Brand            : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 1 ...
```

Figure (15)

Like Complete Responses data set, binning was also applied for Age and Yearly Salary variables. Age variable was divided into 3 bins (20 to 40, 40 to 60, and 60 to 80) and salary into 5 bins. ($20k to $46k, $46k to $72k, $72k – $98k, $98k – $124k, $124k – $150k).

Conclusions and Recommendations

It can be concluded from this study that Sony is preferred brand by customers. Therefore, Blackwell Electronics should concentrate and build deeper strategic relationship and think about acquiring more Sony products. However, it should not be ignored that Acer is a preferred brand among (Figure 20):

- Age: 20-40, mid-level income
- Age: 40 -60, low-mid level income
- Age: 60-80, low income

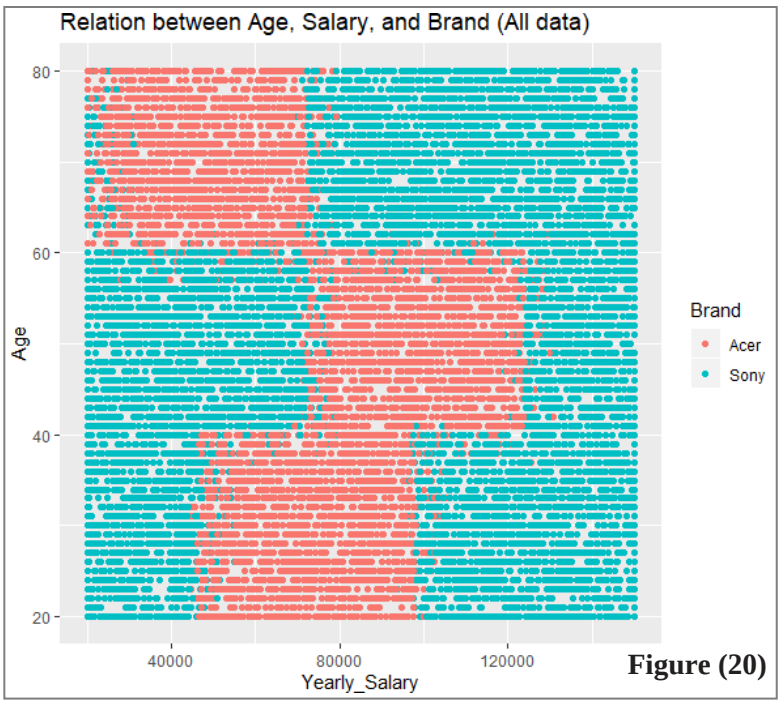

Figure (20)

Blackwell Electronics marketing team should consider promoting Acer brand among these age/income level groups

Also, the marketing team can make some efforts to promote Acre brand among customer who prefer Sony by special promotional offers and by showing any significate upside or benefits of considering Acer