Predicting Customer Preferences: Acer or Sony?

Data analytics department, Blackwell electronics

Else Frijling

2019

Table of Contents

[Summary 3](#_Toc536643789)

[Introduction 4](#_Toc536643790)

[Data exploration 5](#_Toc536643791)

[Descriptive statistics 5](#_Toc536643792)

[Correlation Matrix 6](#_Toc536643793)

[Boxplots, histograms and scatterplots 6](#_Toc536643794)

[Classifier selection to make predictions 7](#_Toc536643795)

[Comparison of classifiers and final selection 8](#_Toc536643796)

[Conclusion 10](#_Toc536643797)

[Recommendations 11](#_Toc536643798)

[Appendix 1: Boxplots, Histograms and scatterplots made of CompleteResponses.csv 13](#_Toc536643799)

[Appendix 2: Modelexports from R 16](#_Toc536643802)

[C5.0 16](#_Toc536643803)

[Performance 16](#_Toc536643804)

[Calculated Performance Across Resamples 16](#_Toc536643805)

[Confusion Matrix and Statistics 16](#_Toc536643806)

[Variable importance 17](#_Toc536643807)

[Random Forest 18](#_Toc536643808)

[Performance 18](#_Toc536643809)

[Calculated Performance Across Resamples 18](#_Toc536643810)

[Confusion Matrix and Statistics 18](#_Toc536643811)

[Variable importance 19](#_Toc536643812)

[Appendix 3: Scatterplots made of Merged\_CompleteResponses\_and\_Predictions.csv 20](#_Toc536643813)

# Summary

The sales team of Blackwell Electronics engaged a market research firm to conduct a survey among existing customers. One of the objectives was to find out which computer brand the customer preferred: Acer or Sony.

In this report the CompleteRespondent dataset is analyzed to see if it is possible to predict the brand preference of the respondent in the IncompleteSurvey file. With the use of Random Forest, a machine learning algorithm, the brand preference of 5000 missing answers is predicted. The prediction is primarily based on the available data of age and salary. In total, 62% of the respondents prefer Sony and 38% prefer Acer, however, there are interesting insights connected to specific age and salary combinations and brand preference.

# Introduction

The sales team of Blackwell Electronics engaged a market research firm to conduct a survey of among existing customers. One of the objectives of the survey was to find out which of two brands of computers our customers prefer to decide with which brand the company should pursue a deeper strategic relationship.

Not all answers to the brand preference question were properly captured, therefore the question answered in this report is: Is it possible to predict the brand preference of the respondents in the incomplete survey based on the complete respondent’s data?

In order to predict the brand preference of the respondents from the incomplete survey data all available data is explored with the ggplot2 and caret package in R. In R, two models are built, trained on 75% of the complete survey set, tested on 25% of the complete survey set and ultimately used to make predictions about the preferred brand to include in the incomplete survey data.

For the analysis two classifiers are selected: C5.0 and Random Forest. Each classifier has different parameters that can be adjusted to find the most optimal settings for the algorithm. It is our goal to tune the algorithms in the best way in order for them to make predictions with the highest accuracy and the lowest error rate. The classifier, or machine learning algorithm, with the best matrix and best use of the attributes will be selected to make the final predictions with.

# Data exploration

The collected data in the CompleteRespondents.csv file and IncompleteSurvey.csv files include information on the following topics:

* + Salary
  + Age
  + Level of education
  + Brand of primary car
  + Zip code
  + Available credit amount
  + Brand of preference

The CompleteRespondents file includes the data of 9898 respondents. With this data the 5000 missing values will be predicted and included in the IncompleteSurvey file.

According to the Survey Key, the respondents are only able to choose between Acer and Sony, therefor it is not measured if they maybe prefer buying a different brand altogether, such as Apple. It also not specifies on what specifications an Acer or Sony is preferred.

The questionnaire assumes all respondents drive a car, because there is no option of “I have no car”, only “none of the above”. This means it is not clear whether the respondent has no car or owns a different brand of car. The yearly salary is filled in, but it is not clear if people entered their gross or net salary. Last but not least, how are the respondents selected?

## Descriptive statistics

First, descriptive statistics of both datasets are explored. Both datasets have similar values, there are no significant differences between the datasets.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Salary** | **Age** | **Level of education** | **Car** | **Zip Code** | **Credit** | **Brand** |
| Min | $20.000 | 20 | 0 | 1 | 0 | 0 | 0 |
| 1st Qu.: | $52.082 | 35 | 1 | 6 | 2 | $120.807 | 0 |
| Median : | $84.950 | 50 | 2 | 11 | 4 | $250.607 | 1 |
| Mean | $84.871 | 50 | 2 | 11 | 4 | $249.176 | 1 |
| 3rd Qu.: | $117.162 | 65 | 3 | 16 | 6 | $374.640 | 1 |
| Max | $150.000 | 80 | 4 | 20 | 8 | $500.000 | 1 |

Figure Descriptive statistics Complete Survey

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Salary** | **Age** | **Level of education** | **Car** | **Zip Code** | **Credit** | **Brand** |
| Min | $20.000 | 20 | 0 | 1 | 0 | 0 | Na |
| 1st Qu.: | $52.242 | 35 | 1 | 6 | 2 | $121.878 | Na |
| Median : | $85.969 | 50 | 2 | 11 | 4 | $250.870 | Na |
| Mean | $85.560 | 50 | 2 | 11 | 4 | $249.510 | Na |
| 3rd Qu.: | $118.380 | 65 | 3 | 16 | 6 | $375425 | Na |
| Max | $150.000 | 80 | 4 | 20 | 8 | $500.000 | Na |

Figure 2 Descriptive statistics Incomplete Survey

## Correlation Matrix

A correlation matrix is made to see if there are correlations between the attributes. Salary and brand have a correlation of 0.21, all other attributes have very low correlations.

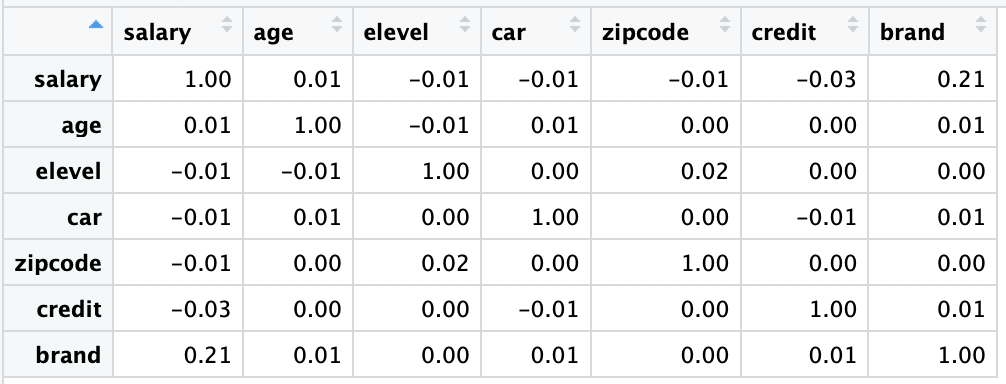


Figure Correlation Matrix of CompleteResponses

## Boxplots, histograms and scatterplots

Boxplots are made of all the attributes to see if there are important outliers that must be dealt with. There are no outliers and the boxplots are quite evenly distributed. Several scatterplots are made to see the relation between the attributes and brand, but no clear correlation is discernable, except a small one between salary and brand. Respondents with high salaries tend to prefer Sony a bit more often. When the attribute Age is added an interesting pattern appears.

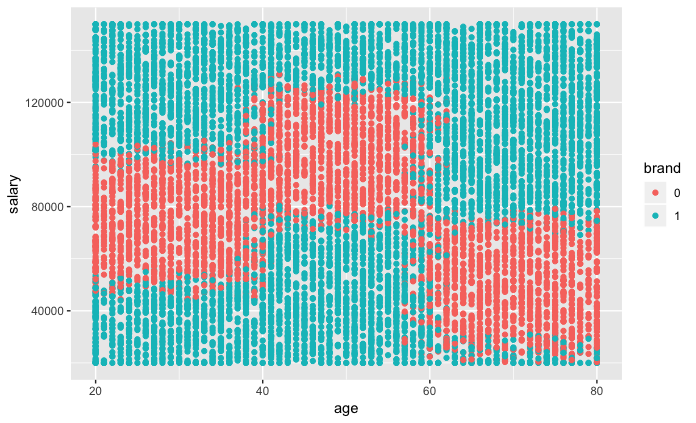


Figure CompleteResponses investigation, 0=Acer 1= Sony

The scatterplot shows that it must be possible to train an algorithm to recognize this pattern. As the data of the IncompleteSurvey is similar to the CompleteResponse data it must be possible to make predictions of the brand preference.

# Classifier selection to make predictions

Two algorithms will be trained to recognize the pattern of brand preference. In order to increase the performance of the classifiers C5.0 and Random Forest, the most optimal parameters must be selected. The two algorithms produced the performance metrics listed, in the table below, with their most optimal parameters on the resampling of the test-data. In this analysis, finding the best parameters for C5.0 was done using the Automated Grid search and with Random Forest the Manual Grid was used.

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Parameter(s) tuned** | **Accuracy** | **Error** |
| C5.0 | trials = 95  model = tree winnow = FALSE | 91,63% | 0.8229 |
| Random Forest | mtry = 4 | 91,84% | 0.8270 |

In general, the metrics of the model on the test-data is slightly lower compared to the metrics generated by the training-data, because the model is based on the training data and did not see the test-data before. Therefore, post resampling, often has a lower accuracy and higher error rate. A detailed export on all the metrics and performance of the models is available for review in the appendix.

The table below show the results of the variable importance on a scale of 0 to 100. These are the attributes that are used by the algorithms and how important they are in making predictions about the brand preference.

|  |  |  |
| --- | --- | --- |
| **Attribute** | **C5.0** | **Random Forest** |
| Salary | 100 | 100 |
| Age | 100 | 62 |
| Elevel.Q | 71 | na |
| Elever^4 | 71 | na |
| Credit | 70 | 12 |
| Car | 70 | 6 |
| Zipcode | 49 | 4 |
| Elevel.C | 20 | na |

It is interesting that the attribute “level of education” (elevel) is not on the Random Forest list. The C5.0 algorithm, however, seems to have divided the attribute into sub-attributes “elevel.Q”, “elevel^4”, and “elevel.C”. For future endeavours it is advised to research the data on education level and brand preference in more depth and how C5.0 is using and adjusting it. For now, we can conclude that the algorithm C5.0 includes more attributes to make its predictions on brand preference. The Random Forest algorithm bases its predictions almost solely on salary and age.

## Comparison of classifiers and final selection

The results for the predictions are summarized in the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Prediction**  **C5.0** | **P%** | **Prediction**  **Random Forest** | **RF%** | **Complete**  **Responses** | **C%** |
| **Acer** | 1941 | 39% | 1904 | 38% | 3744% | 38% |
| **Sony** | 3059 | 61% | 3096 | 62% | 6154 | 62% |
| Total | 5000 | 100% | 5000 | 100% | 9898 | 100% |

The two models predicted the preferences of 4840 respondents equally and predicted 160 preferences differently. In the scatterplot below dark blue and the light blue colours are showing the different predictions. The medium blue colour (0.0) shows the amount of predictions where Random Forest and C5.0 both predicted the same brand. Light blue (1.0) and dark blue (-1.0) show the instances where the algorithms made a different prediction.

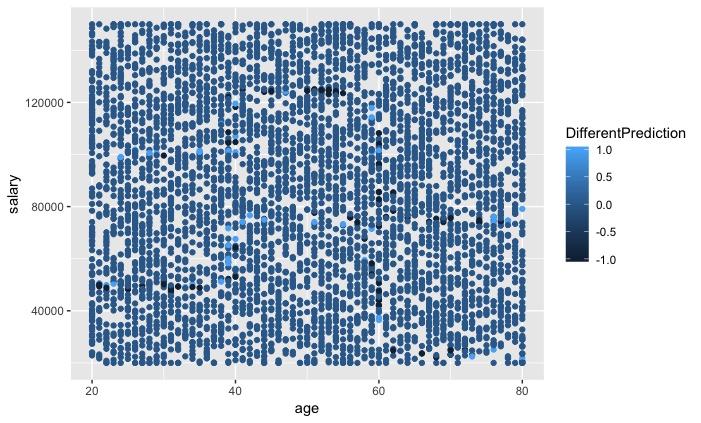


Figure 1 = different prediction, 0= same prediction, 1 = different prediction

The model that is used to produce the final list of predictions is Random Forest, because it has a higher accuracy and a lower error rate, even though it is only a small difference with C5.0. Also, C5.0 includes more attributes compared to Random Forest, however, these are not necessarily adding value to the prediction because they showed no correlation.

In the scatterplot below, it is visible among what lines the brand preference shifts from the Complete Responses dataset:

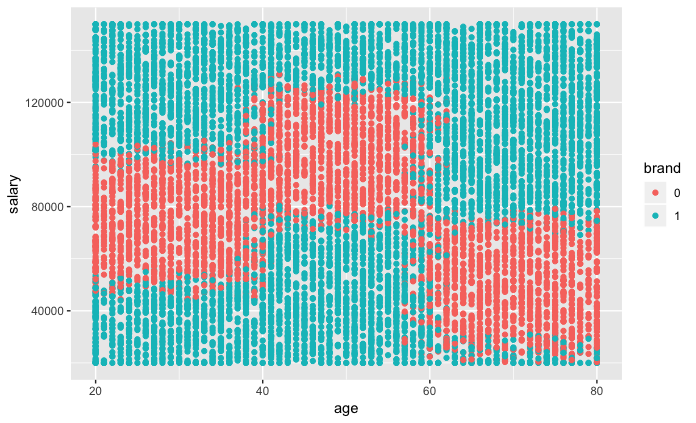


Figure CompleteResponses 0 = Acer, 1 = Sony

In the scatterplot below, it is visible that the Random Forest algorithm learned to see the pattern from the Complete Responses dataset and along what lines to predict the brand preferences in the Incomplete dataset:

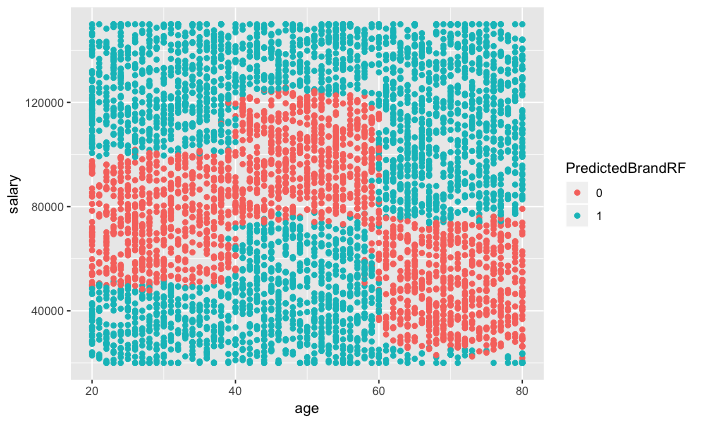
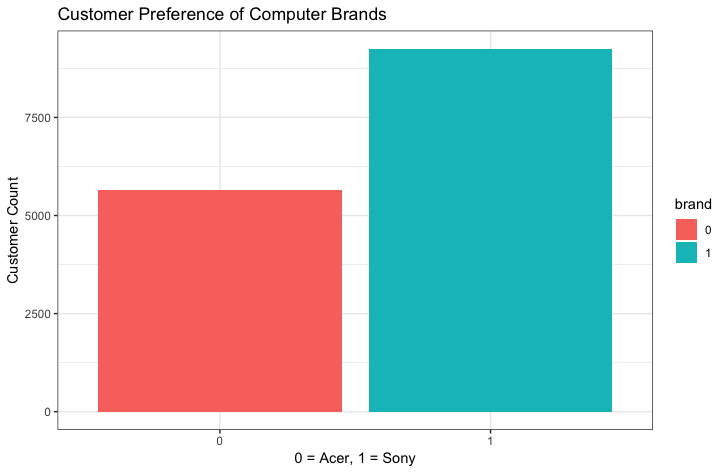


Figure Figure 1 IncompleteSurvey 0 = Acer, 1 = Sony

# Conclusion

It is possible to predict the brand preference of customers based on salary and age with an accuracy of 92% and an error rate of 0.8270. In general, Sony is more often preferred instead of Acer, therefore it is advised to pursue a deeper strategic relationship with Sony.



Based on the analysis 62% of the total respondents prefer Sony and 38% prefer Acer. The scatterplot below visualizes the conclusion in more detail:

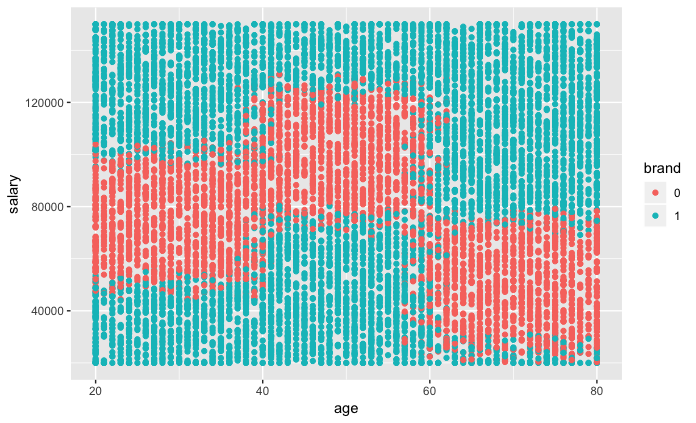


Figure CompleteResponses, 0 = Acer, 1 = Sony

Most customers prefer Sony, but we can also group them into six segments:

* Sony customer group 1a: age 30, with a salary of $30.000
* Sony customer group 1b: age 30, with a salary above$90.000
* Sony customer group 2a: age 50, with a salary of $45.000
* Sony customer group 2b: age 50, with a salary above $120.000
* Sony customer group 3a: age 70, with a salary of $20.000
* Sony customer group 3b: age 70, with a salary above $80.000

However, customers who prefer Acer can be grouped into three clear segments which could be useful information for the marketing department in order to target the right product to the right customer:

* Acer customer group 1: age 30, with a salary of $75.000
* Acer customer group 2: age 50, with a salary of $100.000
* Acer customer group 3: age 70, with a salary of $50.000

The errors made by the predictive model are around the edges of the three Acer-preferring customer groups. Especially difficult to predict the brand preference is when the respondent falls into one of the following categories:

* when a customer between 20-40 years old earns around $50.000 **or** $100.000
* when a customer between 40-60 years old earns around $80.000 **or** $120.000
* when a customer between 60-80 years old earns around $20.000 **or $**80.000

The two models predicted 160 preferences differently and these were all scattered around the lines of the Acer-preferring customer groups. Bearing in mind that the yearly salary is filled in, but it is unclear whether people entered their gross or net salary, the conclusion and borders between the customer groups could be different.

## Recommendations

In order to make the most use of this analysis it is best to pursue a more strategic relationship with Sony as most customers prefer that brand, but it is also useful to focus the marketing efforts of Acer computers on three clear types of customers:

* Acer customer group 1: age 30, with a salary of $75.000
* Acer customer group 2: age 50, with a salary of $100.000
* Acer customer group 3: age 70, with a salary of $50.000

Future marketing research could be conducted in order to investigate what kind of customers the Sony preferers are, if they actually buy Sony computers, what their hobbies are, what sports they like, what they find important in life, whether or not they are married or have children, thereby creating customer profiles to use in future marketing campaigns.

In previous datasets the regional data was arranged according to four regions: “East”, “West”, “Central” and “South”, in this dataset the zip codes are grouped into nine different regions. It could be fruitful to look if it is better to divide the customers into nine regions or four regions to see what segmentation creates the most useful insights for the business. The previous analysis showed that Blackwell’s customers tend to have different spending behaviour per age and region. Therefore, the next step will be a comparison between the two analyses to see whether or not we should deepen our strategic relationship with a different brand per region.

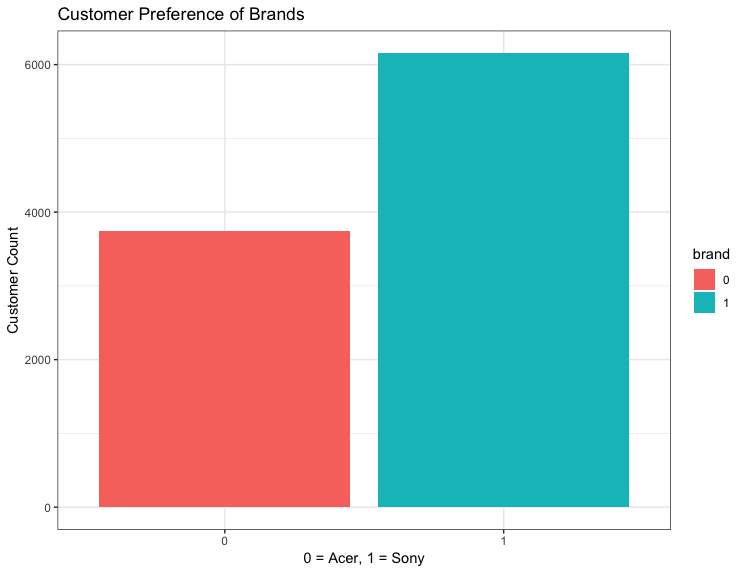
# Appendix 1: Boxplots and histograms of CompleteResponses.csv

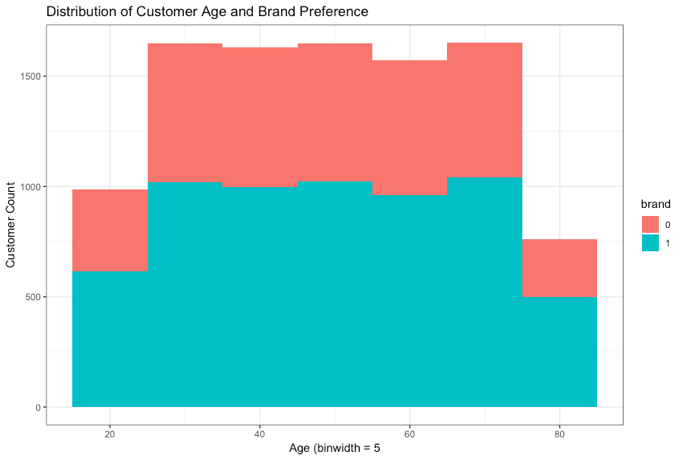
# 

Figure Salary Figure 9 Age Figure 10 Credit

# 

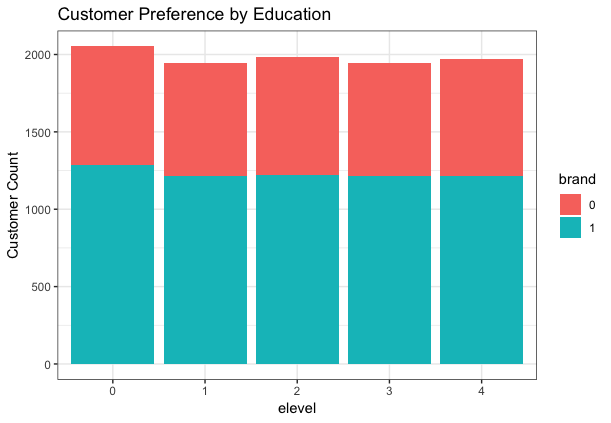
Figure 11 Zipcode Figure 12 Car Figure 13 Education level

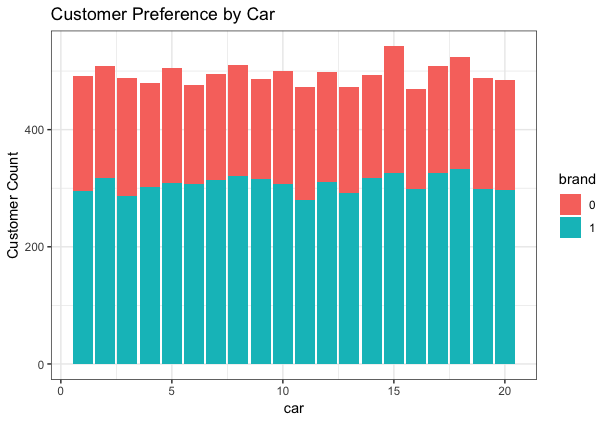


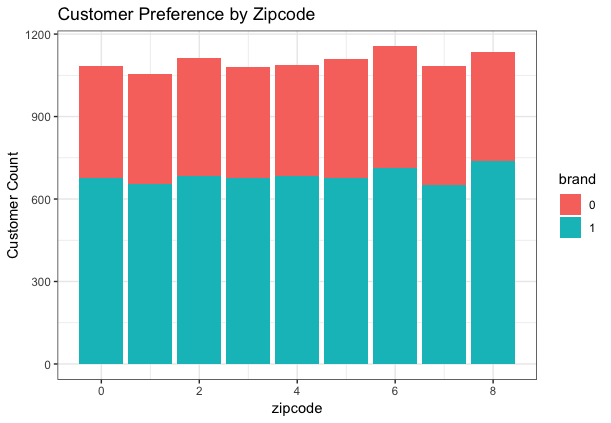


Age (bin = 5)









# Appendix 2: Modelexports from R

## C5.0

### Performance

7424 samples

6 predictor

2 classes: '0', '1'

Pre-processing: scaled (9), centered (9)

Resampling: Cross-Validated (10 fold, repeated 1 times)

Summary of sample sizes: 6682, 6681, 6682, 6681, 6682, 6681, ...

Resampling results across tuning parameters:

winnow trials Accuracy Kappa

FALSE 95 0.9224142 0.8355817

TRUE 91 0.9217431 0.8343297

Tuning parameter 'model' was held constant at a value of tree

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were trials = 95, model = tree and winnow = FALSE.

### Calculated Performance Across Resamples

Accuracy: 0.9163298

Kappa: 0.8229055

### Confusion Matrix and Statistics

|  |  |  |
| --- | --- | --- |
|  | **Reference = 0** | **Reference = 1** |
| **Prediction = 0** | 843 | 114 |
| **Prediction = 1** | 93 | 1424 |

Accuracy : 0.9163

95% CI : (0.9047, 0.9269)

No Information Rate : 0.6217

P-Value [Acc > NIR] : <2e-16

Kappa : 0.8229

Mcnemar's Test P-Value : 0.1645

Sensitivity : 0.9006

Specificity : 0.9259

Pos Pred Value : 0.8809

Neg Pred Value : 0.9387

Prevalence : 0.3783

Detection Rate : 0.3407

Detection Prevalence : 0.3868

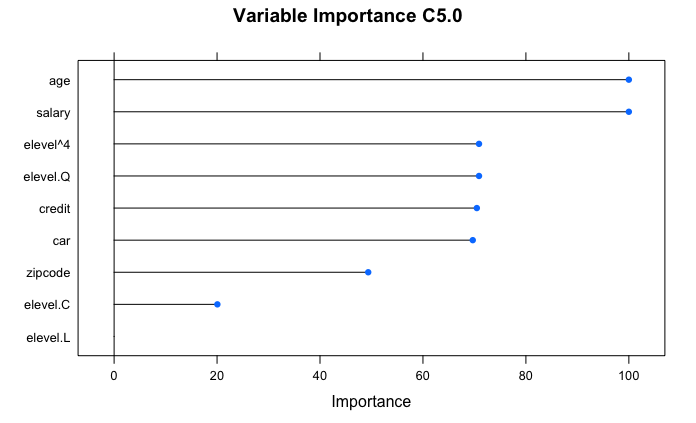
Balanced Accuracy : 0.9133

'Positive' Class : 0

### Variable importance

The top 8 variables are ranked according to their importance on a scale from 0-100. It is interesting that this variable made sub-variables of the education levels. Therefore, for future research it is advised to research the role of education level more. The attributes salary and age are the most important ones used by this model. In other words, the algorithm C5.0 bases its prediction of brand preference mostly on the salary and age of the respondent. However, Credit and Car are also quite high.

|  |  |
| --- | --- |
| **Attribute** | **Scale = 0-100** |
| Salary | 100 |
| Age | 100 |
| Elevel.Q | 71 |
| Elever^4 | 71 |
| Credit | 70 |
| Car | 70 |
| Zipcode | 49 |
| Elevel.C | 20 |



## Random Forest

### Performance

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, 6682, 6681, 6682, 6681, ...

Resampling results across tuning parameters:

mtry Accuracy Kappa

1 0.6219019 0.0004421814

2 0.8930546 0.7714677326

3 0.9229558 0.8366806264

4 0.9236293 0.8380590496

5 0.9229558 0.8366412487

6 0.9217429 0.8340187215

7 0.9208002 0.8320109496

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 4.

### Calculated Performance Across Resamples

Accuracy: 0.9183508

Kappa: 0.8270031

### Confusion Matrix and Statistics

|  |  |  |
| --- | --- | --- |
|  | **Reference = 0** | **Reference = 1** |
| **Prediction = 0** | 843 | 109 |
| **Prediction = 1** | 93 | 1429 |

Accuracy : 0.9184

95% CI : (0.9069, 0.9288)

No Information Rate : 0.6217

P-Value [Acc > NIR] : <2e-16

Kappa : 0.827

Mcnemar's Test P-Value : 0.2912

Sensitivity : 0.9006

Specificity : 0.9291

Pos Pred Value : 0.8855

Neg Pred Value : 0.9389

Prevalence : 0.3783

Detection Rate : 0.3407

Detection Prevalence : 0.3848

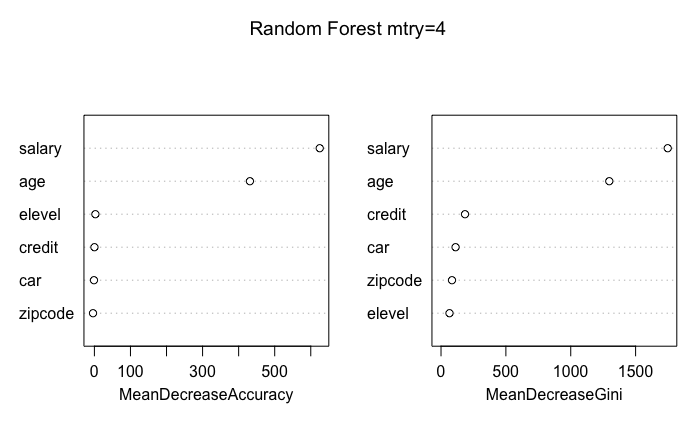
Balanced Accuracy : 0.9149

'Positive' Class : 0

### Variable importance

The top 5 variables are ranked according to their importance on a scale from 0-100. The attribute salary is the most important one in this model. In other words, the algorithm Random Forest bases it’s prediction of brand preference mostly on the salary of the respondent.

|  |  |
| --- | --- |
| **Attribute** | **Scale = 0-100** |
| Salary | 100 |
| Age | 62 |
| Credit | 12 |
| Car | 6 |
| Zipcode | 4 |



# Appendix 3: Scatterplots made of Merged\_CompleteResponses\_and\_Predictions.csv

