# BRAND PREFERENCE PREDICTION

Ubiqum – DA121- Technical department

Alessandro Arnone and David Lopez Salcines

25/11/19

# BRAND PREFERENCE PREDICTION REPORT

## TABLE OF CONTENTS

## OVERVIEW

The objective of this report is to investigate whether customer responses to survey questions allow us to predict the computer brand preference and if so, to provide the sales team with a complete view of what brand our customers prefer based on the predictions made.

We will run and optimize different decision tree classification methods in order to find patterns in our dataset.

The final model choosen to predict the uncompleted surveys is svm R which produced the following results:

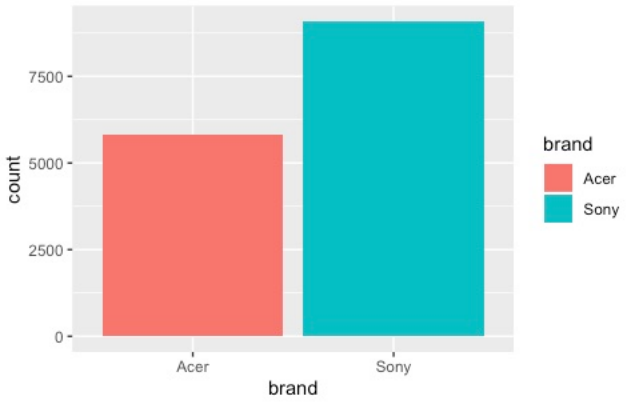



**Figure 1- brand preference**

- Around 65% of the people prefer Sony
- There is a relation between Brand, Salary and Age

## THE DATA

The sales team engaged a market research firm to conduct a survey of our existing customers, providing us with all the information gathered in a CSV file.
Specifically, as first insights we can conclude that:

- 9898 surveys have been submitted
- 7 different variables have been recorded through the survey ( Salary, Age, Eleven, Car, Zipcode, Credit, Brand)
- Around 65% of the people involved in the study prefer a Sony computer. This class difference could potentially create some inbalance in our prediction (since a model created on inbalanced classes tends to predict for the most numerous class) hence this needs to be treated. As solution, the dataset will be down-sampled ( in this case, we prefered to lose information rather then having an additional 40% of information that can be Bias ) to a total of 7488 observations ( 50% of customers preferring Acer, 50% of customers preferring Sony – see Figure 2 for the distribution before balancing the class)
- All the variables are uniformly distributed (see tab.1 for histograms and plots)
- Data looks either manipulated and altered ( credit and salary have decimals, which is uncommon) or automatically generated by a machine
- No outliers identified
- No rows duplicates found

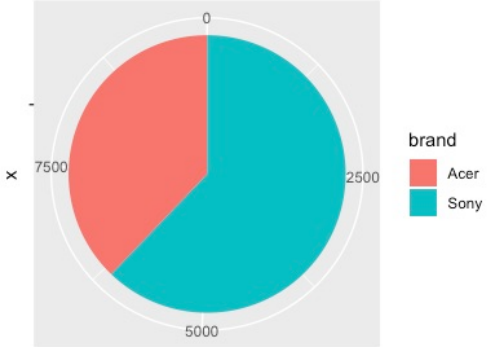
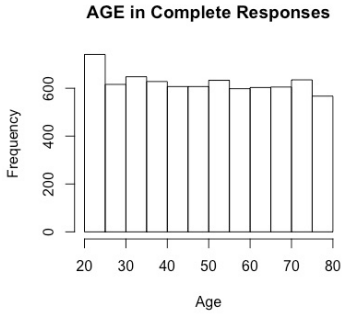
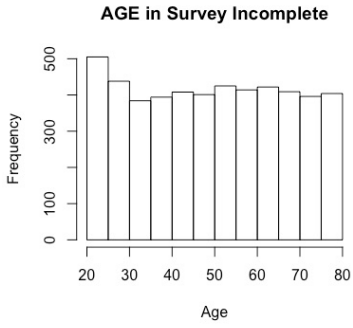
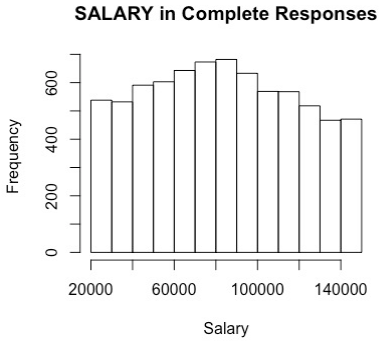


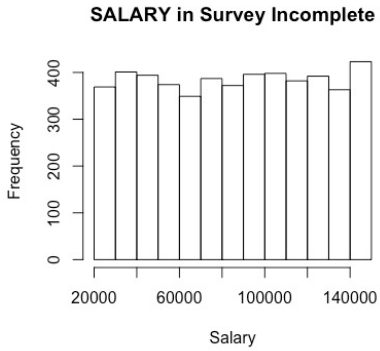

**Figure 2 - Inbalanced classes**

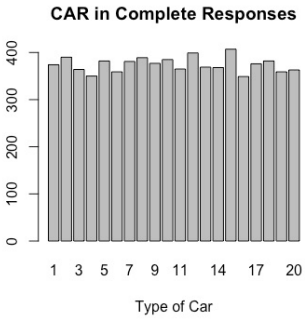
**Table 1- Variables Distribution**

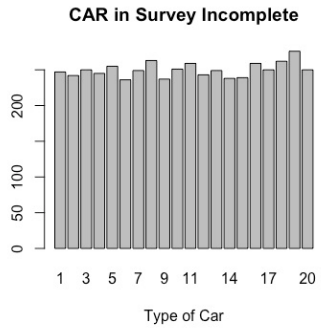


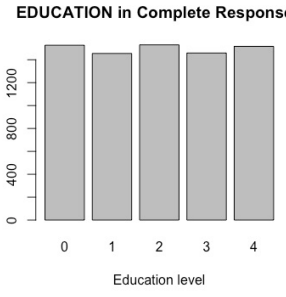



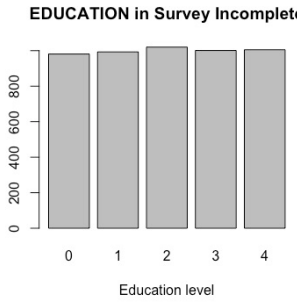



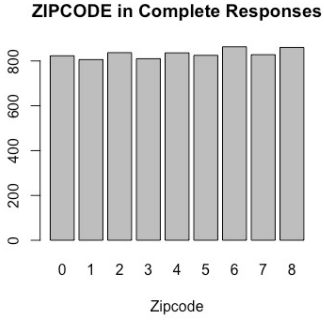



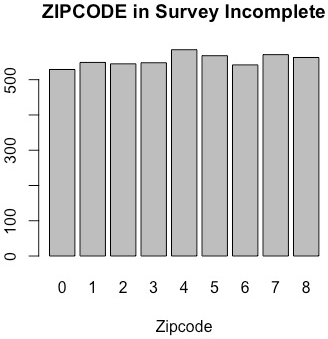



## ANALYSIS

### SELECT THE RIGHT VARIABLES

In order to predict the Brand Preference, the variables that most influence this factor need to be identified. For the selection of predictors we will use both common sense and statistic tools such as regression analysis, the function varimp(),statistical hypothesis tests (such as Chi-squared test ) ,logistic regression and anova.

The variable to exclude are all the variables that:

1. logically are not linked with the Brand Preference such as **Car**
2. have p-value > 0.99 during statistical hypothesis tests (Chi-squared test) and low Cramer's V such as **Education level** (see Tab. 2)
3. have no statistical relevance when modelling the variables with the logistic regression such as **Credit** (see Tab.3)
4. do not contribute heavly to the model during the estimation of the contribution of each variable with the var.imp() function such as **ZIPCODE** ( see Figure 3)

**Table 2 - Chi-squared test and Cramer's V**

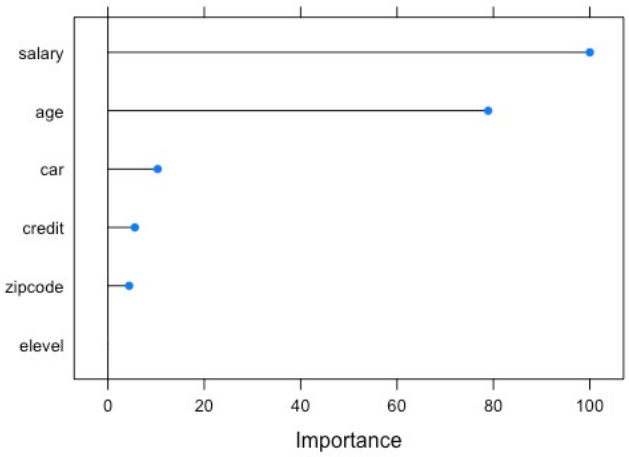
| | p-value | Cramer's V | Comments | |
|---|---|---|---|---|
| Level of education | 0.99 | 0.006 | p-value almost 1. Since our condition of dependency is p<0.05, such a big p with such a low Cramer can definitely suggest us that the 2 variables are independent. | Independent |
| Car | 0.52 | 0.049 | p-value not small enough to reject the hypothesis of independency and Cramer's V really low. Brand and car are indipendend | Independent |
| Zip | 0.19 | 0.03 | p-value not low enough to reject the hypothesis of independency but not big enough to define hypotesis of dependencies | Need additional information |

**Table 3 - Logistic regression**

| | Residual Deviance | p-value | Comments | |
|---|---|---|---|---|
| Null ( only intercept) | 10380 | | | |
| Salary | 9965 | < 2e-16 | p-value really small. It suggests that there is a statistical significance and we can reject the hypothesis of independence. More over we see a drop in the residual deviance compared to the Null model which is a good parameter to understand if the variable is contributing to the model | Dependent |
| Age | 9960 | 0.02454 | p-value really small. It suggests that there is a statistical significance and we can reject the hypothesis of independence. It contributes less to the model | Dependent |
| Credit | 9960 | 0.30852 | p-value big. There is no statistical significance and we cannot reject the hypothesis of independencies. Also it does not contribute to the model since the Residual Deviance keep on being the same | Independent |

In fact, an additional check can be done through the varimp() function of caret which evaluate the importance of a variable using a model-based approach. In our case we used this function when trying to apply a random forest model to our result. The data can be sum-up in the graph below where Salary and Age are the one with the biggest impact on the prediction compared to Car, Credit, Zipcode and elevel.

Figure 3 - Plot varimp()




To sum up the variables picked as predictors are Salary and Age. A first prediction of the model can be seen plotting the decision tree where we can see that the only two variable shown are the one mentioned earlier.

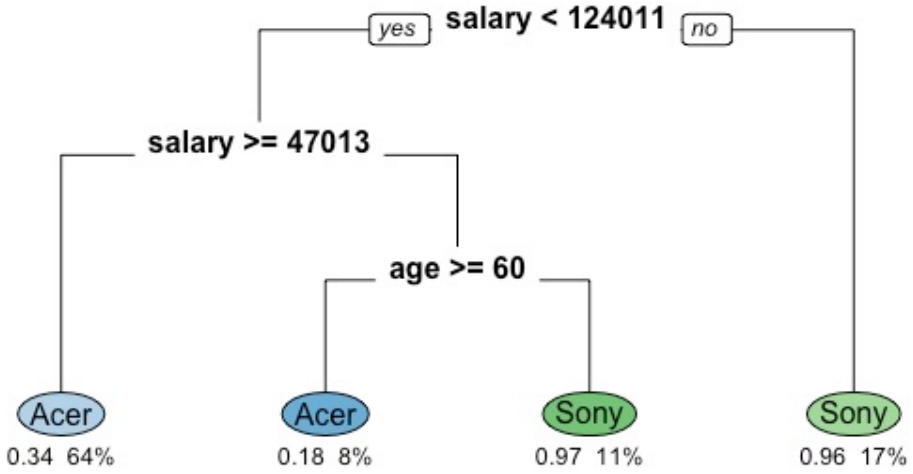



Figure 4 - Decision Tree

## SELECT THE RIGHT MODEL

Once selected the variables to include in our predictive model, multiple tests will be run in order to find the model that maximize the different accuracy metrics used to assess the model such as: **Accuracy** and **Kappa**

Specifically, the predictive Model used are: C5.0, Random Forest, K-nn, SVM Radial and SVM linear.

The following steps have been run in order to perform this task:

1. **Model tuning**: Tune each of the model with different hyperparameters and find the ones that maximize the accuracies metrics

2. **Model selection:** Choose the best model

## MODEL TUNING

A cross-validation 10-folder has been applied to all the models. The best parameters for each model are underlined in yellow.

## C5.O AUTOMATIC GRID

- TuneLength = 3
- PreProcess= Center, Scale

Table 4 - C5.0 accuracy table

| model | winnow | trials | Accuracy | Kappa |
|---|---|---|---|---|
| rules | FALSE | 1 | 0.839 | 0.678 |
| rules | FALSE | 10 | 0.918 | 0.836 |
| rules | FALSE | 20 | 0.920 | 0.840 |
| rules | TRUE | 1 | 0.839 | 0.678 |
| rules | TRUE | 10 | 0.918 | 0.836 |
| rules | TRUE | 20 | 0.920 | 0.840 |
| tree | FALSE | 1 | 0.839 | 0.678 |
| tree | FALSE | 10 | 0.919 | 0.838 |
| tree | FALSE | 20 | 0.920 | 0.839 |
| tree | TRUE | 1 | 0.839 | 0.678 |
| tree | TRUE | 10 | 0.919 | 0.838 |
| tree | TRUE | 20 | 0.920 | 0.839 |

Additional Consideration:

- No rules have been defined
- Winnow results having the same value since the algorithm uses only 2 variables ( the ones not important for the model have already been excluded)
- There in an increase of accuracy increasing the number of trials ( as we can see from the graph below)


Figure 5 - accuracy plot

## C5.0 MANUAL GRID

- Winnow= False
- Model= Tree
- Trials= 25, 35, 45

**Table 5 - C5.0 Accuracy Table**

| Trials | Accuracy | Kappa |
|---|---|---|
| 25 | 0.922 | 0.844 |
| 35 | 0.922 | 0.843 |
| 45 | 0.922 | 0.843 |

Additional Consideration:

- The increase of boosting iteration does not bring any increase in the accuracy

## RANDOM FOREST - AUTOMATIC GRID

- Tune Length=3
- Mtry= 2,4,6
- For academic purpose the variable excluded has been reintroduced to study the effect of different variables on the models

**Table 6 - Accuracy Table**

| mtry | Accuracy | Kappa |
|---|---|---|
| 2 | 0.914 | 0.828 |
| 4 | 0.922 | 0.845 |
| 6 | 0.918 | 0.835 |

Additional Consideration:

- there is an accuracy improvement if we include other 2 variables to the model. This could potentially create problem since there might be a risk of overfitting the model and an increase of computational time with only a 0.08% gain

## K-NN – AUTOMATIC GRID:

- TuneLength= 5
- PreProcess= Center, Scale
- Value k = 5,7,9,11,13
- Computational time is really low

**Table 7- K-NN Accuracy Table**

| k | Accuracy | Kappa |
|---|---|---|
| 5 | 0.919 | 0.837 |
| 7 | 0.920 | 0.840 |
| 9 | 0.922 | 0.844 |
| 11 | 0.923 | 0.847 |
| 13 | 0.923 | 0.845 |

## SVMRADIAL – AUTOMATIC GRID

- TuneLength= 5
- PreProcess= Center, Scale
- Value C = {0.25 , 0.50 , 1.00 , 2.00 , 4.00 }
- Gamma = 1.25

**Table 8 - Accuracy Table SVM Radial**

| C | Accuracy | Kappa |
|---|---|---|
| 0.25 | 0.920 | 0.839 |
| 0.50 | 0.923 | 0.847 |
| 1.00 | 0.924 | 0.848 |
| 2.00 | 0.925 | 0.849 |
| 4.00 | 0.926 | 0.853 |

## SVMLINEAR

- TuneLength= 5
- PreProcess= Center, Scale
- Tuning parameter 'C' was held constant at a value of 1

**Table 9 - Accuracy Table SVM Linear**

| C | Accuracy | Kappa |
|---|---|---|
| 1 | 0.682 | 0.363 |

## MODEL SELECTION

The best tuning parameters are summarized in the following tab:

| Model | Parameters | Accuracy | K |
|---|---|---|---|
| C5.0 | Trials=20 | 0.920 | 0.839 |
| C5.*0 Manual* | Trials=25 | 0.922 | 0.844 |
| Random Forest | mtry= 4 | 0.922 | 0.845 |
| Random Forest | mtry=2 | 0.921 | 0.842 |
| K-nn | 11 | 0.923 | 0.847 |
| SVM Linear | C=1 | 0.682 | 0.363 |
| SVM radial | C=4 Sigma = 1.23 | 0.926 | 0.853 |

The model to exclude are:

- SVM linear since It presents the lower accuracy
- Random Forest with mtry= 4 since the computational time and the risk of overfitting do not justify the choice

Moreover, if we check the accuracy of our prediction on our Test Set:

| Model | Parameters | Accuracy | K |
|---|---|---|---|
| C5.0 | Trials=20 | 0.923 | 0.845 |
| Random Forest | mtry=2 | 0.922 | 0.844 |
| kNN | N=11 | 0.928 | 0.857 |
| svmRadial | C=4 sigma = 1.23 | 0.929 | 0.858 |

- svmRadial present the higher level of accuracy and k
- the 4 models only differentiates by less then 7 centesimal. Hence the decision of a model simply based on a variable like Accuracy is risky. That said, we will continue our analyis with svmRadial

The confusion matrix of the model choosed applied to the test Dataset is :

Table 10 - Confusion matrix

| Observed \ Prediction | Acer | Sony | |
|---|---|---|---|
| Acer | 893 | 47 | 91% |
| Sony | 85 | 847 | 94% |
| | 91% | 94% | |

A high % of each class justify the high K accuracy we found in the model. Moreover, since the classes have been balanced, the number of Acer and the number of Sony is pretty similar.

## CONCLUSION

Applying the model choosen (svm Radial) to the uncompleted surveys we have the following distribution:

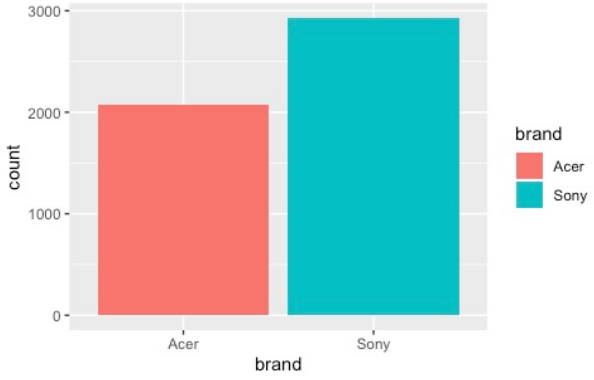



Figure 7- Brand prediction

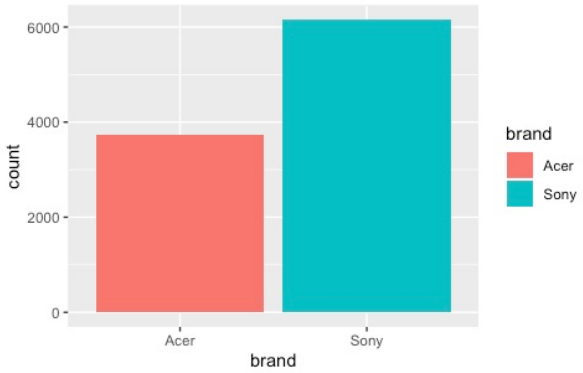
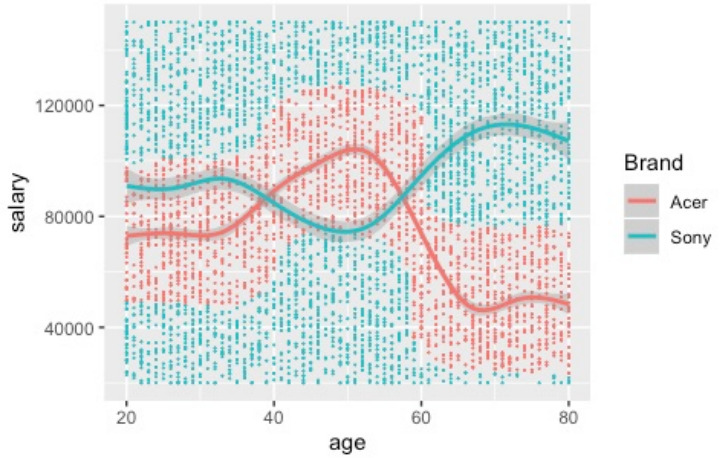



Figure 6 - Unbalanced Class in the original Dataset

We have 2072 people prefering Acer and 2928 people prefering Sony. Exactly the same proportion of the dataset before the class balance. Moreover, as we can see from the graphes below:




not only we have the same behaviour in both dataset analyzed (Complete Response before balanced, Incomplete Survey) but also a similiar proportion of 65/35 of people prefering either Sony or Acer respectively. That means:

- Dataset has been altered to keep the proportion 65%-35%
- It seems that the Survey Incomplete dataset it's partially a subset of the Complete Survey Dataset
    - some salaries are the same ( included the decimals)
    - if we round up salaries to the closest integer, we find around 1800 rows identical in terms of salaries and age
- Whilst the prediction shoulds reliable based on the model performances, we need to take into account the fact that the reliability depends on the quality of the dataset