Task 5

Report: Predicting Brand

Barcelona, October 2018

Student:
Ing. Javier E. Villasmil
Villasmil@gmail.com
+58.412.2379138

ÍNDEX

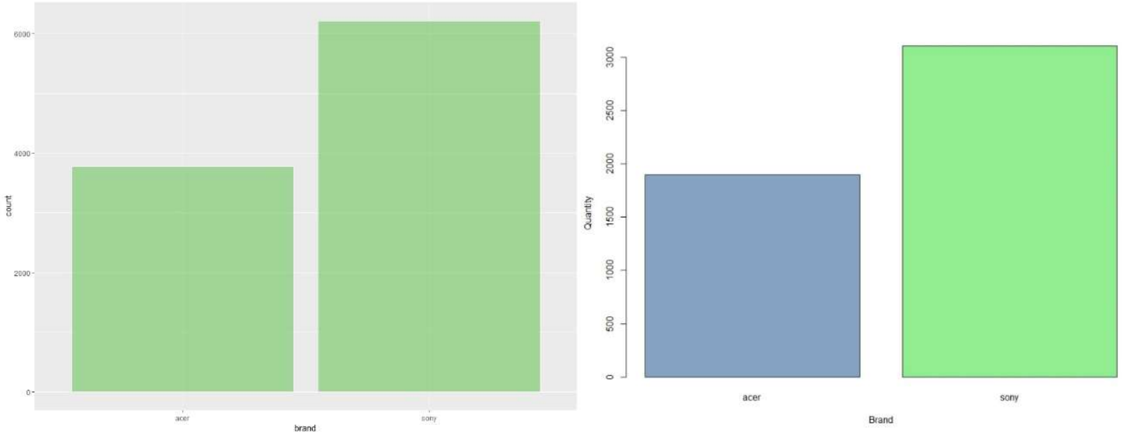
LIST OF FIGURES

LIST OF TABLES

1. EXECUTIVE SUMMARY

After analyzing the data set provided by Blackwell's eCommerce, there were found several key insights useful for the management.

- The data set was pre-processed by checking for attributes with missing values and fixing the .XLS file.
- Variables's data types had to be changed to factor or numerical according the nature of each.
- Features were removed by checking relationships and dependencies between variables.
- The key features that defines ***Brand*** (dependent variable) are ***Salary*** and ***Age.***
- There were not any Outliers or strange values.
- Algorithms and models were tuned switching parameters automatically using the *"control"* function provided by The Caret Package in R.
- Decision trees and scatter plots were used to determine the weighted importance of the predictors against the label ("Brand").
- The distribution of the variables in the dataset is uniform, meaning that our data and model might not be representative for the "real life" case, but it will work for completing the survey. This could be caused by using a wrong survey method.
- Six (6) models were evaluated, *three (3) kNN*, *three (3) RPART* – the best one was picked for the prediction.
- After evaluating the performance of both algorithms - the model ***kNN*** with the features ***salary + age*** is the one with the best metrics.
- The best performance values are Accuracy = 0.92 and Kappa 0.83 for K = 17
-



Distribution of Brand (left) and Prediction for the incomplete survey (right)

The predictions were made using the kNN against the Incomplete Survey. The results were:

For a total of *5.000 rows*

- ***3.105 (62%)*** records predicted as ***SONY.***
- ***1.895*** **(38%)** records predicted as ***ACER.***

2. SCOPE

Analyze using data mining and modeling methods, the dataset supplied by the CTO and head of Blackwell's eCommerce Team – the goal is to predict the ***preferred computer brand (Sony/Asus)*** based on a survey realized to users..

This dataset contains general information about possible costumers such as salary, age, education level, type of car, location, credit and preferred computer brand.

3. OBJECTIVES

- Clean, transform and preprocess the dataset.
- Determine relationships between variables.
- Verify and select important features used in the predictive models.
- Optimize and tune the selected algorithms.
- Predict ***the computer brand*** based on the survey and the data set supplied.

4. METHODOLOGY

The approach used to assess the dataset was to apply basics methods of data mining, descriptive statistics and simple charting to observe the distribution and relationships between variables (features) in our dataset.

In addition, R (Rstudio) helped with the evaluation of histograms, scatter plots, decision trees, and with the cleaning / transformation of the data.

On the other hand, for the classification and regression modeling, the algorithms used were ***k-Nearest Neighbors*** *(kNN)* and a ***Tree Based Model*** *(CART)* found in the Caret R package, these algorithms were tuned according to each own parameters.

5. ANALYSIS

5.1 PRE-PROCESSING

The first step to start analyzing the data was to transform and clean the dataset. After revising the provided *excel* file, it was necessary to change the values of columns, assign new data types, and check missing values.

- The feature **Brand** was changed to a *factor* data type, and their values replaced with
  - 0 = “sony”
  - 1 = “acer”
- The feature **Zipcode** was changed to a *factor* data type, and their values replaced with:
  - "0" = "New England"
  - "1" = "Mid-Atlantic"
  - "2" = "East North Central"
  - "3" = "West North Central"
  - "4" = "South Atlantic"
  - "5" = "East South Central"
  - "6" = "West South Central"
  - "7" = "Mountain"
  - "8" = "Pacific"
- The feature **Elevel** was changed to a *factor* data type, and their values replaced with:
  - "0" = "Less than High School"
  - "1" = "High School Degree"
  - "2" = "Some College"

- "3" = "4-year college degree"
- "4" = "Masters or Doctoral"

- The feature **Cars** was changed to a *factor* data type, and their values left as provided.
- The features **Salary** and **Credit** were changed to a *numeric* type and their valued left as provided.

The data set was complete and ***there were not missing values***.

5.2 FEATURE SELECTION

Selecting which columns are representative for our model to predict the brand is a task where the analyst needs to compare different relationships between your label and predictors.

To do this, two tools were used: Random Forest, Scatter charts.

5.2.1 RANDOM FOREST

The first approach was to check the dependency of all the variables against the label and their respective weights. The parameters for the random forest operator were *number of tree: 500*, *maximal depth: 10* and *gain ratio as a criterion* (because our label is a numeric variable). Pruning was applied.

Attributes	Weight
Salary	0.499
Age	0.202
Credit	0.190
E. level	0.060
Zip code	0.033
Car	0.017

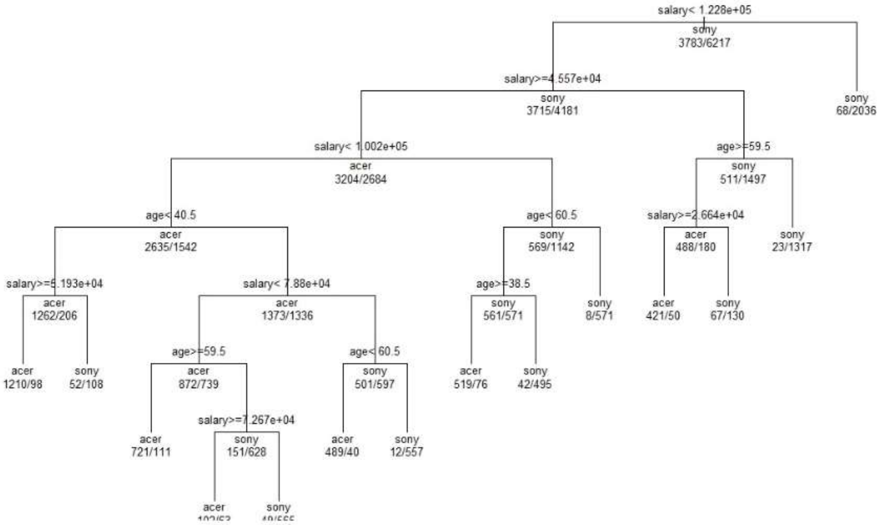


Fig 1. Classification tree for *Brand* feature and feature importance by weight.

As seen in Fig 1, the best variables to use in our modelling according to classification and regression trees are ***Salary*** and ***Age***.

The feature ***credit*** seems important according to the weighted values, but it does not appear in the classification tree as a main variable. Further investigation should be doing with plots and charts.

Attributes such as education level, zip code and type of car could be removed without affecting our predictive model.

It appears that costumers prefer a computer brand based on their salary and age group.

5.2.2 CHARTING

A scatter plot is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis, it is possible to add another variable as a color to each dot.

By doing this, it allowed us to see which trio of attributes have a visual pattern. See Fig 2 and Fig. 3.

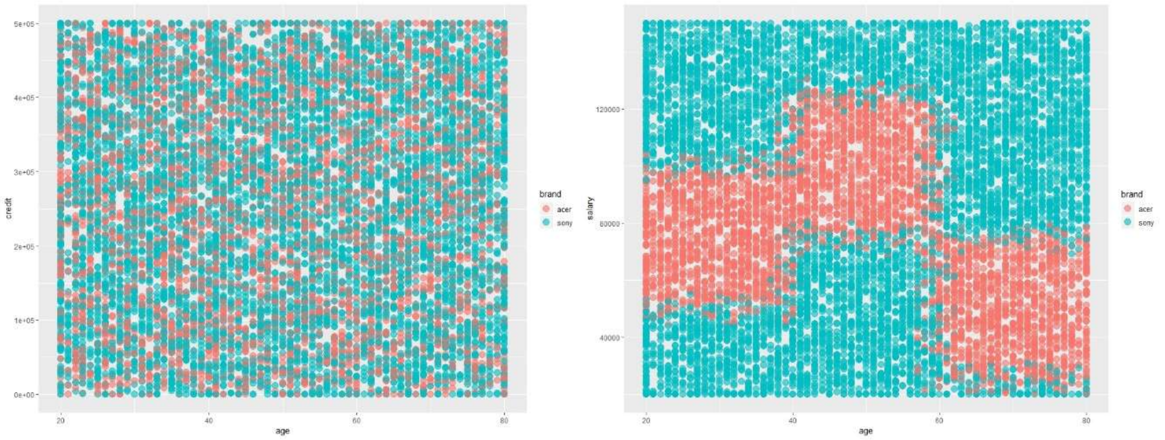


Fig 2. Scatter plot age against credit (left) and age against salary (right) – using brand as legend.

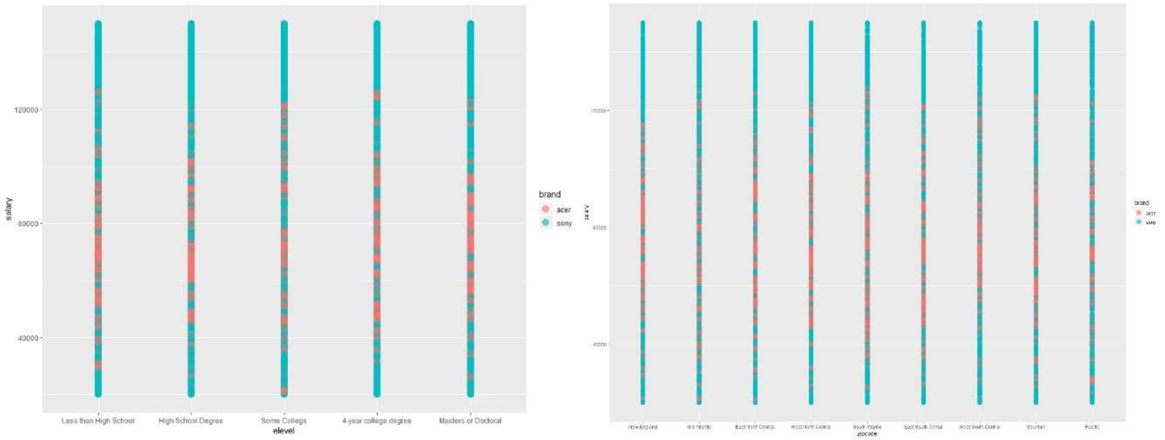


Fig 3. Scatter plot E.level against salary (left) and Zip code against salary (right) – using brand as legend.

In this case, the feature *"Credit"* was discarded because its distribution against the age/brand seems random and it's not relate to our label - leaving this column would cause data noise in the model. Additionally, it is important to check the distribution of each variable. Having features equally distributed in all the ranges means that the survey could have had problems when joining the data. See Fig 4 and Fig. 5.

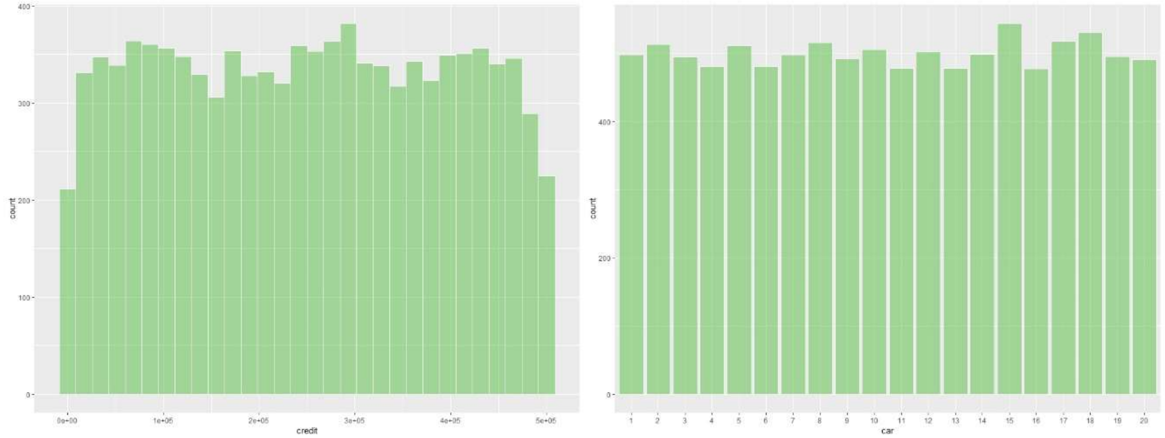


Fig 4. Distribution of credit (left) and car type (right).

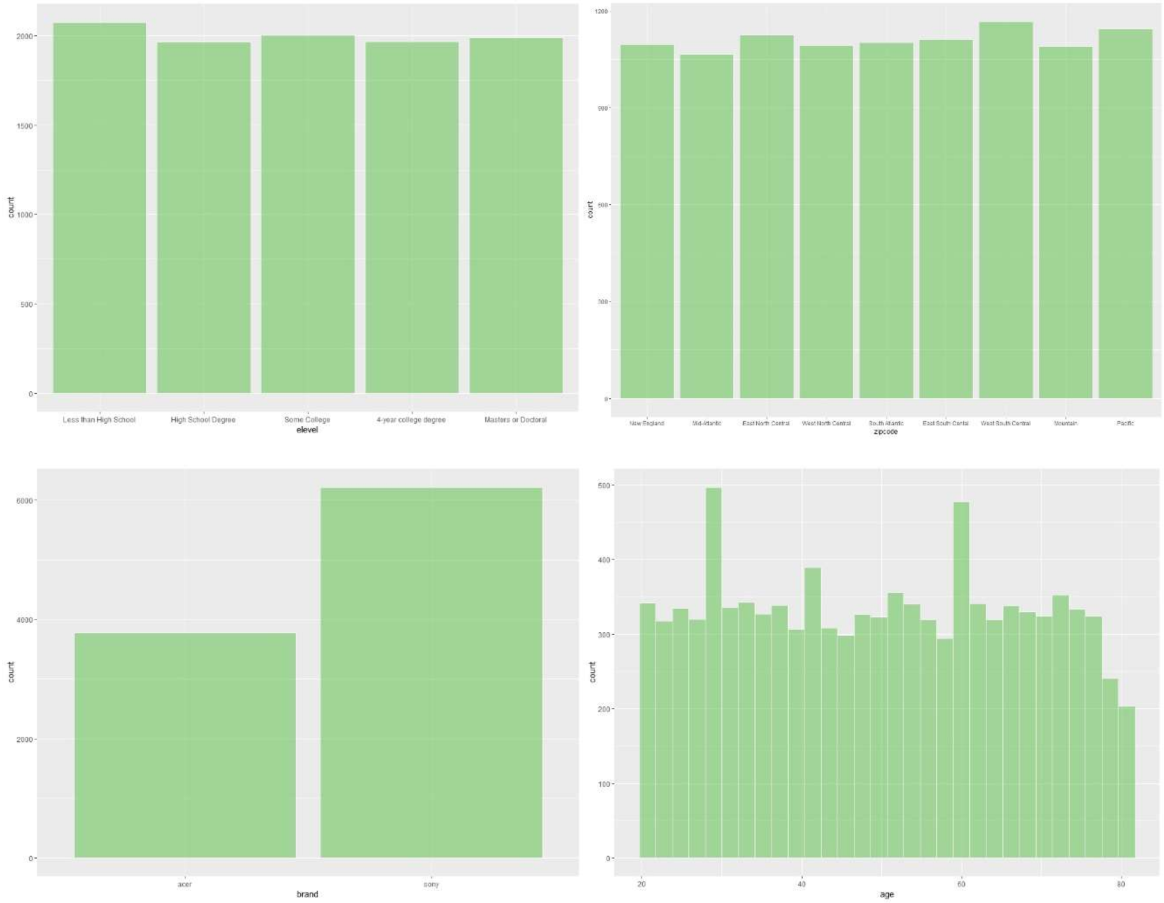


Fig 5. Distribution of E level (top left), Zip code (top right), Brand (bottom left) and Age (bottom right).

Furthermore, while checking the values for Brand it was noted that our dataset is divided 62/38, meaning that 62% of the values (6217 records) are from people who picked sony and 38% (3783) are from people who picked acer. See fig 5.

This is important because when evaluating our classification algorithms, our minimum “acceptable” accuracy should be set at 62%, because if you force the most repeated value as a prediction the model will guess correctly at least 62% of the time and the kappa performance will be near zero or even negative (unreliable model).

Finally, after evaluating the charts, the initial hypothesis was supported and our main variables are **Salary** and **Age**. The features others can be removed for our modeling process.

Attributes	Selection
Salary	✓
Age	✓
Credit	
E. level	
Zip code	

Table 1. Selected attributes for data modeling

5.2.3 OUTLIERS

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In this case, there were not outliers or missing values detected. The data set was clean.

5.3 PREDICTIVE MODELS

The algorithms used to make models that could predict the brand are ***k-Nearest Neighbors*** *(kNN)*, and a ***Tree Based Model*** *(rPart)*.

These algorithms were tuned according to each own parameters, using R's package caret and its functions.

Caret let the user evaluate a range of values for each parameter picked and select the best one for the given combination.

Three (3) models were made for each algorithm. The first considers all the features, the second one considers only salary and the third one considers the main attributes salary and age.

The models and algorithms were compared against each other (using performance metrics and confusion matrixes); the best one was selected to make the predictions.

5.3.1 K-NEAREST NEIGHBORS (kNN)

All Features			Salary			Salary + Age		
K	Accuracy	Kappa	K	Accuracy	Kappa	K	Accuracy	Kappa
17	0.60142	0.06715	17	0.71617	0.39241	17	0.92013	0.83051

CM – All Features			CM – Salary			CM – Salary + Age		
	Acer	Sony		Acer	Sony		Acer	Sony
Acer	237	294	Acer	518	319	Acer	832	92
Sony	708	1260	Sony	364	1235	Sony	107	1492

Table 2. Performance and confusion matrix for the kNN models.

- The best model using the kNN algorithm is the one who take into account two (2) features. ***Salary and Age.***
- The kNN model that uses *Salary + Age* predicts correctly Acer 90% of the time and Sony 93% of the time.
- The best performance values are Accuracy = 0.92 and Kappa 0.83 for K = 17
- The kNN model that uses *all the features* predicts correctly Acer 44% of the time and Sony 64% of the time.
- The kNN model that uses *only salary* predicts correctly Acer 64% of the time and Sony 77% of the time.

5.3.2 TREE BASED MODEL (RPART)

All Features			Salary			Salary + Age		
Cp	Accuracy	Kappa	Cp	Accuracy	Kappa	Cp	Accuracy	Kappa
0.00211	0.91680	0.82247	0.00070	0.72777	0.41636	0.00211	0.91645	0.82156

CM – All Features			CM – Salary			CM – Salary + Age		
	Acer	Sony		Acer	Sony		Acer	Sony
Acer	815	78	Acer	592	312	Acer	815	78
Sony	130	1476	Sony	353	1242	Sony	130	1476

Table 3. Performance and confusion matrix for the Rpart models.

- Two models performed the same: ***All features*** and ***Salary + Age***. One hypothesis for this matter is that the algorithm ignored the unnecessary predictors and left only the two main attributes (salary and age).
- The best model using the RPART model is the one who take into account two (2) features. ***Salary and Age.***
- The best performance values are Accuracy = 0.92 and Kappa 0.82 for Cp = 0.00211
- The RPART model that uses *Salary + Age* predicts correctly Acer 91% of the time and Sony 92% of the time.

- The RPART model that uses only *Salary* predicts correctly Acer 65% of the time and Sony 77% of the time.

To evaluate the models that performed the same both classification trees were plotted. IT happens that it was the same tree, meaning that our initial hypothesis was correct and the algorithm ignored the unnecessary predictors leaving only two features: age and salary.

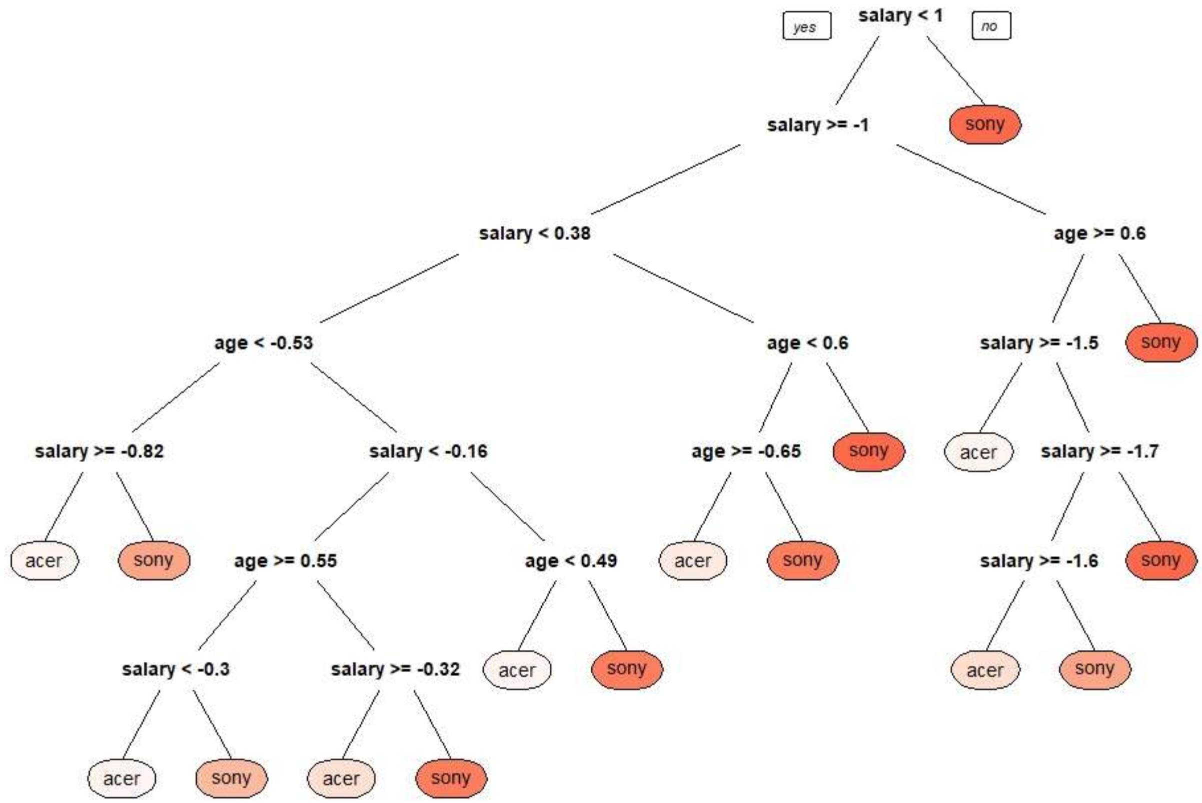


Fig 6. Classification tree for two RPART models – *all variables* and *salary + age*.

5.3.3 MODEL SELECTION

After evaluating the performance of both algorithms - the model ***kNN*** with the features ***salary + age*** is the one with the best metrics, therefore it was picked to make the predictions of the incomplete survey. See Table 4.

	KNN	RPART
Sensitivity	0.8868	0.8624
Specificity	0.9408	0.9498
Acer Pred Value	0.9011	0.9127
Sony Pred Value	0.9318	0.9191
Prevalence	0.3782	0.3782
Detection Rate	0.3353	0.3261
Detection Prevalence	0.3721	0.3573
Balanced Accuracy	0.9138	0.9061

Table 4. Comparison between best kNN and RPART model.

6. PREDICTIONS

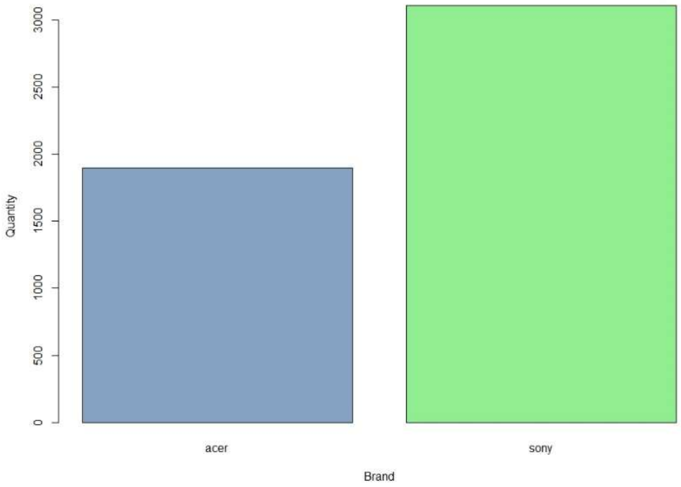


Fig 7. Prediction for the incomplete survey using the model selected.

The predictions were made using the model selected in part 5.3.3 against the Incomplete Survey. The results were:

For a total of *5.000 rows*

- ***3.105 (62%)*** records predicted as ***SONY.***
- ***1.895*** **(38%)** records predicted as ***ACER.***

It is important to note that the distribution of predictions for the attribute brand is similar to the distribution of the original dataset, therefore our model is behaving accordingly.

7. CONCLUSIONS

- The data set was pre-processed by checking for attributes with missing values and fixing the .XLS file.
- Variables's data types had to be changed to factor or numerical according the nature of each.
- Features were removed by checking relationships and dependencies between variables.
- The key features that defines ***Brand*** (dependent variable) are ***Salary*** and ***Age.***
- There were not any Outliers or strange values.
- Algorithms and models were tuned switching parameters automatically using the *"control"* function provided by The Caret Package in R.
- Decision trees and scatter plots were used to determine the weighted importance of the predictors against the label ("Brand").
- The distribution of the variables in the dataset is uniform, meaning that our data and model might not be representative for the "real life" case, but it will work for completing the survey. This could be caused by using a wrong survey method.
- Six (6) models were evaluated, *three (3) kNN*, *three (3) RPART* – the best one was picked for the prediction.
- After evaluating the performance of both algorithms - the model ***kNN*** with the features ***salary + age*** is the one with the best metrics.
- The best performance values are Accuracy = 0.92 and Kappa 0.83 for K = 17
- For a total of *5.000 rows*
  - ***3.105 (62%)*** records predicted as ***SONY.***
  - ***1.895*** **(38%)** records predicted as ***ACER.***