Blackwell Electronics

Customer Brand preferences

Floriana Trama
Data analysis department

AGENDA

- **GENERAL OVERVIEW**

- **CUSTOMER BRAND PREFERENCES:**
  - **Customer brand preferences result**
  - **Data exploration and pre-processing**
  - **Model development and selection, parameters and performance metrics of the executed classifiers**

- **CONCLUSION & RECOMMENDATIONS**

- **APPENDIX**

GENERAL OVERVIEW

Description of the datasets
Two different datasets are used for the prediction of the favourite brand:

- **"Complete Responses"** containing **9.898 data points** representing the complete answers to a market survey of Blackwell's existing customers
- **"Survey Incomplete"** containing **5.000 data points** representing the incomplete answers to a market survey of Blackwell's existing customers.

In both datasets, each data point comprises **7 attributes**, namely: **"Salary"** tracks the money earned yearly by the customer who replied to the survey, **"Age"** records the age of the customer, **"Ed. level"** indicates the level of educations reached by the customer, **"Car"** indicates the brand of the primary car owned by the customer, **"Zip Code"** indicates the zip code of the area in which the customer lives, **"Credit"** indicates the amount of credit available to the customer, and **"Brand"** records which is the favorite brand between Acer and Sony.

The "**Complete Responses" dataset is complete**, hence all data points contain all the values of the 5 attributes. On the contrary, the **"Survey Incomplete" dataset is missing the answers** related to the brand preferences. A first overview of the main data attributes of the combined datasets can be found below:

SALARY

- Min = $ 20.000
- Max = $ 150.000
- Average = $ 85.102

AGE

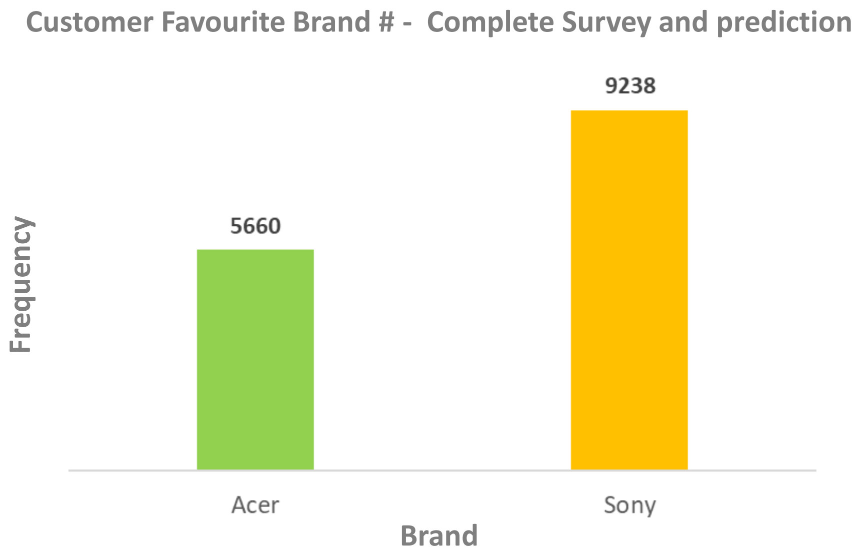
- Min = 20
- Max = 80
- Average = 49,8

CREDIT

- Min = $ 0
- Max = $ 500.000
- Average = $ 249.288

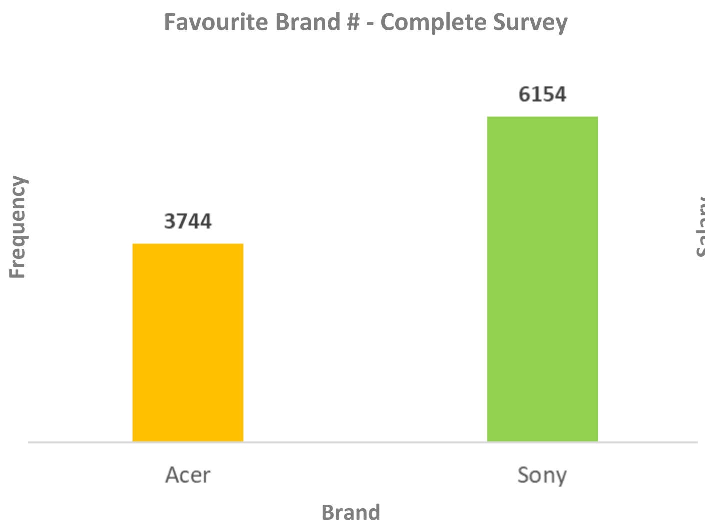
Objective of the analysis: predict the customer computer favorite brand between Acer and Sony and decide with which manufacturer pursue a deeper strategic relationship

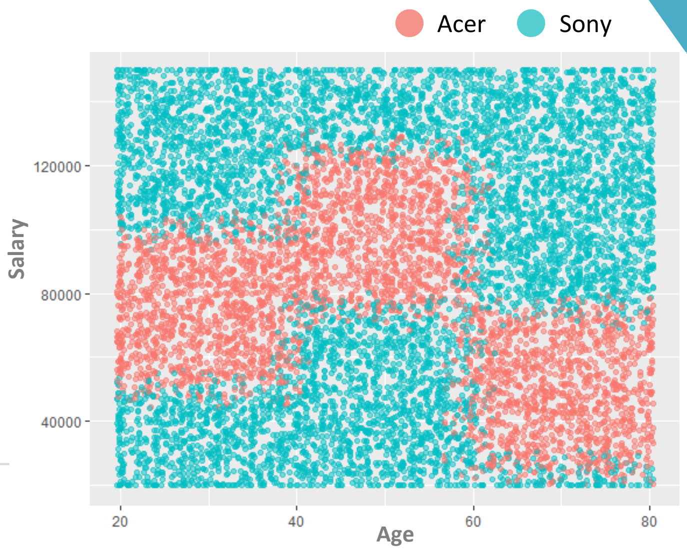
CUSTOMER BRAND PREFERENCES: Customer brand preferences result



In order to predict the customers favourite brand, a classification analysis using R software has been performed by building a model based on the "Complete Responses" data. The model is then applied to the "Incomplete Survey" dataset to make predictions on the missing values. As it can be seen from the bar chart above, the result of the analysis shows a **clear preference for Sony** as the best computer brand for Blackwell customers. For this reason it is the **recommended brand with which pursue a deeper, stronger relationship**.

CUSTOMER BRAND PREFERENCES: Data exploration and pre-processing





As per good practice, the analysis starts with the **exploration** and **pre-processing** of the data: starting with the "Complete Dataset", **several charts** are created in order to understand the distribution of the different variables and get a first insight about the correlation among them. Of particular interest, the above charts:

- The **bar chart** shows a preliminary result on the favourite brand only based on the answers from the complete survey: with **6.154 preferences**, **Sony** results the **preferred computer brand**;
- The **scatter plot** displays a clear pattern among the dependent variable, **"Brand"**, and other two independent variables, **"Age"** and **"Salary"**, making evident their importance in the brand's prediction. Despite Sony is generally the most favourite brand, Acer seems to be the preferred one for young customers aged 20-40 earning 60-100K $ per year, customers aged 40-60 earning 80-120k $ yearly and finally among older customers aged 60-80 getting a salary in the range of 20-70K $ each year.

"Brand", "Ed. Level", "Car" and "Zip Code" attributes are transformed to make them ready for the classification analysis.

CUSTOMER BRAND PREFERENCES: Model development and selection, parameters and performance metrics of the executed classifiers

After exploring the data and implementing indispensable pre-processing activities, a **model including all the explanatory variables** is built **by using a Random Forest algorithm with 10-fold cross validation**. The 75% of the dataset is used for the training and the remaining 25% for the testing.
Despite the good metrics (Accuracy and Kappa) obtained with this model (Fig. 1 Appendix), by using the **"VarImp"** function to ascertain how the model prioritizes each feature, it is possible to see that **"Salary"** and **"Age"** are the **most important features for** making the **predictions**, confirming the insights gained through the previous scatter plot.

```
rf variable importance

  only 20 most important variables shown (out of 34)

           Overall
salary    100.0000
age        47.9959
credit     18.2261
elevel1     0.9917
elevel2     0.9389
elevel3     0.9230
elevel4     0.8480
zipcode4    0.5962
zipcode6    0.5495
zipcode5    0.5449
zipcode1    0.5326
zipcode3    0.5299
zipcode2    0.4672
zipcode7    0.4467
zipcode8    0.3154
car10       0.2642
car15       0.2540
car7        0.1781
car17       0.1779
car8        0.1671
```

Given these results, **features selection is performed on** the dataset, leaving only two attributes, "Salary" and "Age" (apart from "Brand" as the dependent variable) for building the classification models.
Two new classification models are built by **training and testing 2 different algorithms**, namely **Random Forest** and **C5.0** in order to **identify the best performing model** in terms of Accuracy, Kappa and Confusion Matrix and make realistic brand predictions.

Random Forest

```
Random Forest

7424 samples
   2 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     0.9203934  0.8314483
  2     0.9121783  0.8136557
  3     0.9123140  0.8140002

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 1.
> varImp(rfFitm1)
rf variable importance

       Overall
salary     100
age          0
```

C5.0

```
C5.0

9898 samples
   2 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 9898, 9898, 9898, 9898, 9898, 9898, ...
Resampling results across tuning parameters:

  model  winnow  trials  Accuracy   Kappa
  rules  FALSE    1      0.9053851  0.8019988
  rules  FALSE   10      0.9165829  0.8223215
  rules   TRUE    1      0.9053851  0.8019988
  rules   TRUE   10      0.9165829  0.8223215
  tree   FALSE    1      0.9025832  0.7958103
  tree   FALSE   10      0.9158840  0.8213953
  tree    TRUE    1      0.9025832  0.7958103
  tree    TRUE   10      0.9158840  0.8213953

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were trials = 10, model = rules and winnow = TRUE.
> varImp(C5model_Brand)
C5.0 variable importance

       Overall
salary     100
age          0
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  845  126
         1   91 1412

               Accuracy : 0.9123
                 95% CI : (0.9004, 0.9231)
    No Information Rate : 0.6217
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8149
 Mcnemar's Test P-Value : 0.021

            Sensitivity : 0.9028
            Specificity : 0.9181
         Pos Pred Value : 0.8702
         Neg Pred Value : 0.9395
             Prevalence : 0.3783
         Detection Rate : 0.3416
   Detection Prevalence : 0.3925
      Balanced Accuracy : 0.9104

       'Positive' Class : 0
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  814   75
         1  122 1463

               Accuracy : 0.9204
                 95% CI : (0.909, 0.9307)
    No Information Rate : 0.6217
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.829
 Mcnemar's Test P-Value : 0.001048

            Sensitivity : 0.8697
            Specificity : 0.9512
         Pos Pred Value : 0.9156
         Neg Pred Value : 0.9230
             Prevalence : 0.3783
         Detection Rate : 0.3290
   Detection Prevalence : 0.3593
      Balanced Accuracy : 0.9104

       'Positive' Class : 0
```

Comparing the **two models**, there is **very little differences** in the main metrics: both the models are good for making predictions as their Accuracy, Kappa and Confusion matrix show really high values and low errors. It is decided to **use the Random forest** model with 1 mtry as it presents the **highest Accuracy (0,92) and Kappa (0,83)**.

Finally, brand predictions are calculated by **applying the selected model to the "Incomplete survey" data**.

CONCLUSIONS & RECOMMENDATIONS

This classification analysis was conducted with the objective to inform Blackwell' decisions about customers' favourite computer brand and in order to strengthen the relationship with that company. The results show that the consumers' preferred brand is Sony, hence this is the manufacturer with which is recommended to pursue a stronger relationship.

APPENDIX

Fig. 1

```
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, 6682, 6681, 6682, 6681, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  1     0.6217673  0.0000000000
  2     0.6219019  0.0004421814
  3     0.7269719  0.3429706969
  4     0.8479316  0.6688493287
  5     0.8915693  0.7689609978
  6     0.9101575  0.8093233713
  7     0.9159505  0.8217661138

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 7.
```

Random Forest codes

```
# Customer brand preferences --------------------------------------------
# Floriana Trama --------------------------------------------------------
# Data analysis department ----------------------------------------------
# Y = Brand -------------------------------------------------------------
# Random forest - Manual grid -------------------------------------------

# Libraries -------------------------------------------------------------
library(readr)
library(caret)

# Data exploration ------------------------------------------------------
CompleteDataset <- read.csv("C:/Users/T450S/Desktop/Floriana/Ubiqum/Data Analytics II/Task 2/Database/CompleteResponses.csv")
summary(CompleteDataset)
str(CompleteDataset)
hist(CompleteDataset$salary)
hist(CompleteDataset$age)
hist(CompleteDataset$credit)

# Pre-processing data ---------------------------------------------------

CompleteDataset$brand<-as.factor(CompleteDataset$brand)
plot(CompleteDataset$brand)
CompleteDataset$elevel<-as.factor(CompleteDataset$elevel)
CompleteDataset$car<-as.factor(CompleteDataset$car)
CompleteDataset$zipcode<-as.factor(CompleteDataset$zipcode)
plot(CompleteDataset$brand,CompleteDataset$salary)
plot(CompleteDataset$brand,CompleteDataset$age)
ggplot(CompleteDataset, aes(age, salary, color = as.factor(brand)))+ geom_jitter(alpha = 0.5)

# Features selection ----------------------------------------------------
CompleteDataSubset <- CompleteDataset[c(1,2,7)]

# Set seed --------------------------------------------------------------
set.seed(123)

# Create 75%/25% training and test sets ---------------------------------
inTraining <- createDataPartition(CompleteDataSubset$brand, p = .75, list = FALSE)
training <- CompleteDataSubset[inTraining,]
testing <- CompleteDataSubset[-inTraining,]

# 10 fold cross validation ----------------------------------------------
fitControl <- trainControl(method = "repeatedcv", number = 10, repeats = 1)

# Dataframe for manual tuning of mtry -----------------------------------
rfGrid <- expand.grid(mtry=c(1,2,3))

# Train Random Forest Regression model ----------------------------------
system.time(rfFitm1 <- train(brand~., data = training, method = "rf“, trControl=fitControl, tuneGrid=rfGrid))
```

```
# Ttraining results -------------------------------------------------------
rfFitm1
varImp(rfFitm1)

# Predictions --------------------------------------------------------------
prediction <- predict(rfFitm1, testing)
confusionMatrix(prediction, testing$brand)
postResample(prediction, testing$brand)
Specialtable <- cbind(testing, prediction )

# Incomplete survey dataset -----------------------------------------------
IncompleteDataset <- read_csv("Floriana/Ubiqum/Data Analytics II/Task 2/Database/SurveyIncomplete.csv")
IncompleteDataSubset <- IncompleteDataset[c(1,2,7)]

# Prediction on Incomplete dataset ----------------------------------------
prediction <- predict(rfFitm1, IncompleteDataSubset)
prediction
Specialtable <- cbind(IncompleteDataSubset, prediction )
summary(prediction)
```

C5.0 codes

```
# 10 fold cross validation ------------------------------------------------
fitControl <- trainControl(method = "repeatedcv", number = 10, repeats = 1)
C5model_Brand <- train(brand~.,
                data = CompleteDataSubset,
                method = "C5.0",
                trainControl = fitControl,
                metric = "Accuracy",
                tuneLength = 2)

# Training results -------------------------------------------------------
C5model_Brand
varImp(C5model_Brand)
prediction <- predict(C5model_Brand, testing)
confusionMatrix(prediction, testing$brand)
```