Customer Intelligence and Big Data

Alan Rijnders and Lorenzo Severi

11/4/2021

We start reading in the data to perform the analysis

```
#read in data
data <- read.csv("ch.csv")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

We now start exploring the dataset

```
#summary of the dataset
summary(data)
```

```
##        X           CLIENTNUM         Attrition_Flag      Customer_Age
##  Min.   :    1   Min.   :708082083   Length:10127       Min.   :26.00
##  1st Qu.: 2532   1st Qu.:713036770   Class :character   1st Qu.:41.00
##  Median : 5064   Median :717926358   Mode  :character   Median :46.00
##  Mean   : 5064   Mean   :739177606                      Mean   :46.33
##  3rd Qu.: 7596   3rd Qu.:773143533                      3rd Qu.:52.00
##  Max.   :10127   Max.   :828343083                      Max.   :73.00
##     Gender          Dependent_count Education_Level    Marital_Status
##  Length:10127       Min.   :0.000   Length:10127       Length:10127
##  Class :character   1st Qu.:1.000   Class :character   Class :character
##  Mode  :character   Median :2.000   Mode  :character   Mode  :character
##                     Mean   :2.346
##                     3rd Qu.:3.000
##                     Max.   :5.000
##  Income_Category    Card_Category      Months_on_book  Total_Relationship_Count
##  Length:10127       Length:10127       Min.   :13.00   Min.   :1.000
##  Class :character   Class :character   1st Qu.:31.00   1st Qu.:3.000
##  Mode  :character   Mode  :character   Median :36.00   Median :4.000
##                                        Mean   :35.93   Mean   :3.813
##                                        3rd Qu.:40.00   3rd Qu.:5.000
##                                        Max.   :56.00   Max.   :6.000
##  Months_Inactive_12_mon Contacts_Count_12_mon  Credit_Limit   Total_Trans_Amt
```

```
##  Min.   :0.000          Min.   :0.000         Min.   : 1438   Min.   :  510
##  1st Qu.:2.000          1st Qu.:2.000         1st Qu.: 2555   1st Qu.: 2156
##  Median :2.000          Median :2.000         Median : 4549   Median : 3899
##  Mean   :2.341          Mean   :2.455         Mean   : 8632   Mean   : 4404
##  3rd Qu.:3.000          3rd Qu.:3.000         3rd Qu.:11068   3rd Qu.: 4741
##  Max.   :6.000          Max.   :6.000         Max.   :34516   Max.   :18484
##  Total_Trans_Ct   Avg_Utilization_Ratio
##  Min.   : 10.00   Min.   :0.0000
##  1st Qu.: 45.00   1st Qu.:0.0230
##  Median : 67.00   Median :0.1760
##  Mean   : 64.86   Mean   :0.2749
##  3rd Qu.: 81.00   3rd Qu.:0.5030
##  Max.   :139.00   Max.   :0.9990
```

```
head(data, 5)
```

```
##   X CLIENTNUM    Attrition_Flag Customer_Age Gender Dependent_count
## 1 1 768805383 Existing Customer           45      M               3
## 2 2 818770008 Existing Customer           49      F               5
## 3 3 713982108 Existing Customer           51      M               3
## 4 4 769911858 Existing Customer           40      F               4
## 5 5 709106358 Existing Customer           40      M               3
##   Education_Level Marital_Status Income_Category Card_Category Months_on_book
## 1     High School        Married     $60K - $80K          Blue              39
## 2        Graduate         Single  Less than $40K          Blue              44
## 3        Graduate        Married    $80K - $120K          Blue              36
## 4     High School        Unknown  Less than $40K          Blue              34
## 5      Uneducated        Married     $60K - $80K          Blue              21
##   Total_Relationship_Count Months_Inactive_12_mon Contacts_Count_12_mon
## 1                        5                      1                     3
## 2                        6                      1                     2
## 3                        4                      1                     0
## 4                        3                      4                     1
## 5                        5                      1                     0
##   Credit_Limit Total_Trans_Amt Total_Trans_Ct Avg_Utilization_Ratio
## 1        12691            1144             42                 0.061
## 2         8256            1291             33                 0.105
## 3         3418            1887             20                 0.000
## 4         3313            1171             20                 0.760
## 5         4716             816             28                 0.000
```

We are interested to see how many customers have remained at the company and how many have left.

```
table(data$Attrition_Flag)
```

```
##
## Attrited Customer Existing Customer
##              1627              8500
```

In other words, out of total 10127 customers in the database we have 8500 customers that have remained at the company ,whereas 1627 customers have left.

```
table(data$Gender)
```

```
##
##    F    M
## 5358 4769
```

```
table(data$Dependent_count)
```

```
##
##    0    1    2    3    4    5
##  904 1838 2655 2732 1574  424
```

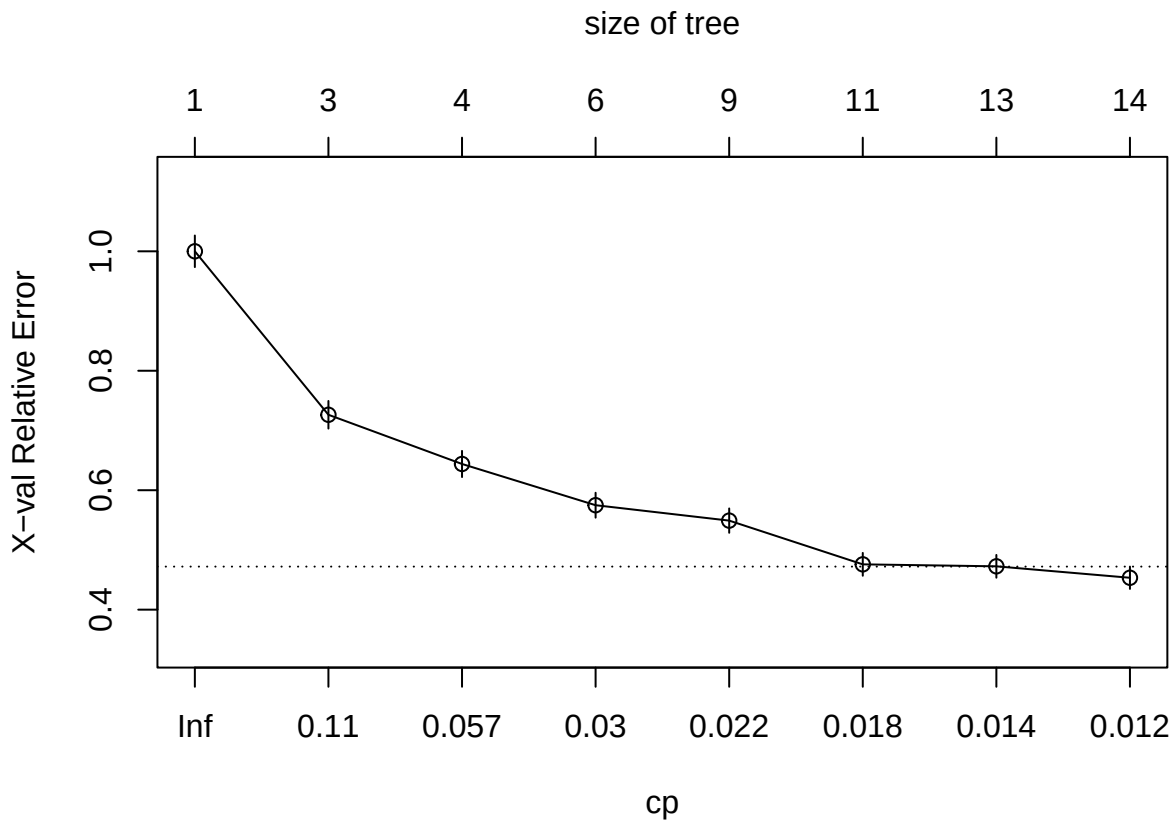
Since Attrition_Flag is a character variable with two possible values: either “Existing Customer” or “Attrited Customer” for modeling purposes we recode this variable into a factor with levels 0 and 1, where 1 represents a customer that has left the company when the person is still a customer at the company.

```
data <- data %>% mutate(Attrition_Flag = recode(Attrition_Flag, 'Attrited Customer' = 1 , 'Existing Cus
```

We now split the data into a training and a test sample, we use the training sample to train our model and the test sample to test our predictions to assess the power of our models.

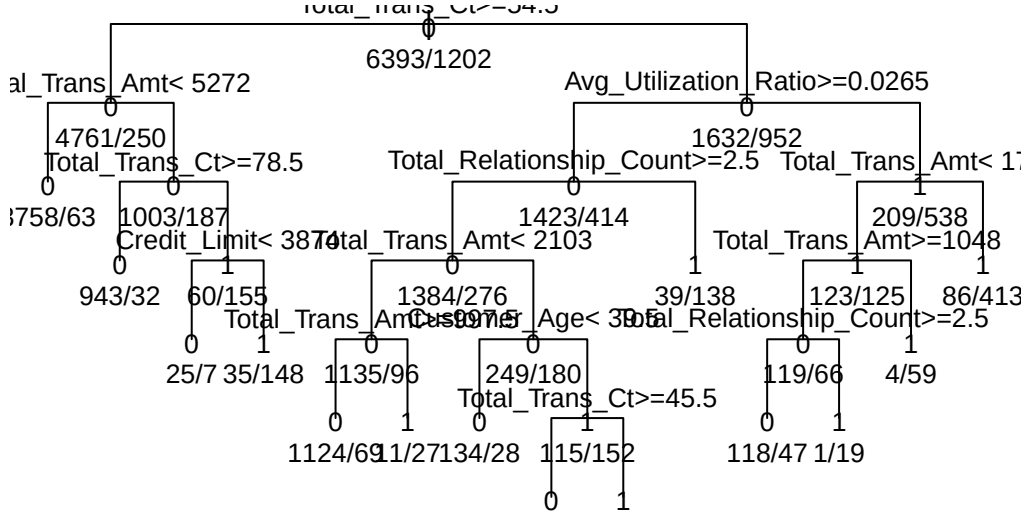
```
smp_size <- floor(0.75 * nrow(data))

## set the seed to make your partition reproducible
set.seed(12345)
train_ind <- sample(seq_len(nrow(data)), size = smp_size)

train <- data[train_ind, ]
test <- data[-train_ind, ]
```

```
library(rpart)
tree <- rpart(Attrition_Flag ~ . - X , method = "class", data = train)
printcp(tree)
plotcp(tree)
```



```
summary(tree)
```

```
# plot tree
plot(tree, uniform=TRUE,
   main="Classification Tree for Attrition")
text(tree, use.n=TRUE, all=TRUE, cex=.8)
```




```
#library caret is a comprehensive library support all sorts of model analysis
library(caret)
```

```
## Loading required package: lattice
```

```
options(digits=4)
# assess the model's accuracy with train dataset by make a prediction on the train data.
Predict_model1_train <- predict(tree, train, type = "class")
#build a confusion matrix to make comparison
conMat <- confusionMatrix(as.factor(Predict_model1_train), as.factor(train$Attrition_Flag))
#show confusion matrix
conMat$table
```

```
##           Reference
## Prediction    0    1
##          0 6190  291
##          1  203  911
```

```
#install.packages('caret', dependencies = TRUE)
#library(caret)
sensitivity(conMat$table)
```

```
## [1] 0.9682
```

```
specificity(conMat$table)
```

```
## [1] 0.7579
```

```
print(accuracy <- (6190+911)/(6190+911+291+203))
```

```
## [1] 0.935
```

The model looks to do a decent job, our sensitivity seems to be quite higher than our specificity, which implies that our model is better at correctly classifying clients that left than at finding true loyal customers. This could be because of ...

Now that we have constructed the model we proceed by predicting the values in the test set in order to assess the suitability of the model.