Faculté des **sciences économiques** et de **gestion**
Université de Strasbourg

UE Big Data & Management :

Amadou Khadidja

M2 APE/DS2E

Table of content

INTRODUCTION

As part of our analysis, we have data that comes from the different districts of the city of Turin in Italy. Depending on the different variables that make up the database, our objective is to predict, make correlations between the variables and estimate, thanks to the different Machine Learning techniques, the best performance models for classifying the members of a museum by based on several criteria.

Thus, this leads us to answer the seven questions that constitute the take home.

DATA PREPARATION

1. *Describe the most interesting variables by plotting distributions, correlations, co-occurrence.*

Before answering this question, we will first start by doing a series of actions to make our database “cleaner”.

a) Merge the data

Our study focuses on three databases, each with different variables. We decided initially to merge the two databases namely 'data1' and 'int13' by the client code of each user. Which gives us the total 'merge' database.

Then we created a variable 'visit' which counts the number of visits of each member and then merged this into the total 'merge' database.

To make the database more dynamic, we transformed the years of birth of each member in dd/mm/yy format which we then subtracted from our base year 2022 to obtain the age of each user of the map.

b) Rename the columns

For the sake of ease of reading, we have renamed the columns that make up the database:

```
#rename columns
colnames(Dataf)
names(Dataf)[names(Dataf) == "si2014"] = "Fidels"
names(Dataf)[names(Dataf) == "importo"] = "prix"
names(Dataf)[names(Dataf) == "sconto"] = "Remise"
names(Dataf)[names(Dataf) == "riduzione"] = "Réduction"
names(Dataf)[names(Dataf) == "agenzia_tipo"] = "Agence_tp"
names(Dataf)[names(Dataf) == "sesso"] = "Femme"
names(Dataf)[names(Dataf) == "data_nascita"] = "Age"
names(Dataf)[names(Dataf) == "comune"] = "Commune"
names(Dataf)[names(Dataf) == "nuovo_abb"] = "Nouvelle_abonne"
names(Dataf)[names(Dataf) == "tipo_pag"] = "ModePaiement"
names(Dataf)[names(Dataf) == "cap"] = "CP"
names(Dataf)[names(Dataf) == "ultimo_ing.x"] = "Date_last_v"
names(Dataf)[names(Dataf) == "abb13"] = "start_D_13"
names(Dataf)[names(Dataf) == "abb14"] = "Renewal_D_14"
```

c) Data transformation

We have also transformed some variables into categorical variables: *'Remise'*, '*nouvelle_abonne'*, *'Modedepaiement','Churner'*.

Also, we transformed the names of the cities TORINO, DATO MANCANTE, CAVALLERMAGGIORE and ROMA by their postal code to generate geographical maps to study the distribution of some variables on the different cities according to gender and age.

```
FinalData$Commune <- as.numeric(factor(FinalData$Commune))
FinalData$CP=ifelse(FinalData$Commune == "TORINO" & FinalData$CP == "XXXXX" , "10100", FinalData$CP)
FinalData$CP=ifelse(FinalData$Commune == "DATO MANCANTE" & FinalData$CP == "XXXXX" , "10098", FinalData$CP)
FinalData$CP=ifelse(FinalData$Commune == "CAVALLERMAGGIORE" & FinalData$CP == "XXXXX" , "12030", FinalData$CP)
FinalData$CP=ifelse(FinalData$Commune == "ROMA" & FinalData$CP == "XXXXX" , "10090", FinalData$CP)
FinalData<-FinalData[!(FinalData$CP == "XXXXX" | FinalData$CP<10000),]
FinalData$CP <- as.numeric(as.character(FinalData$CP))
```

Given the large number of variables that make up the database, we chose to study the correlation between a handful of variables. The result allowed us to have this correlation matrix below:

	Age	Money_notP	N_visits	CP	Remise	Femme	Churner	Mon_paied	ModePaiement
Age		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0008	0.0000
Money_notP	0.0000		0.0000	0.6199	0.0000	0.0000	0.0000	0.0000	0.0000
N_visits	0.0000	0.0000		0.3063	0.0000	0.0000	0.0000	0.0000	0.0000
CP	0.0000	0.6199	0.3063		0.0000	0.0545	0.0000	0.0000	0.0000
Remise	0.0000	0.0000	0.0000	0.0000		0.1963	0.0000	0.0000	0.0000
Femme	0.0000	0.0000	0.0000	0.0545	0.1963		0.0100	0.0000	0.0000
Churner	0.0000	0.0000	0.0000	0.0000	0.0000	0.0100		0.0000	0.0000
Mon_paied	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000
ModePaiement	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	

There is a strong correlation between the variables 'CP' and 'Money_notP.' In other words, the price of the entrance ticket has an influence on the municipality of origin. The higher the price and the richer the town, the more prestigious is the museum.

Conversely, the correlation between age and the 'My_paid' variable is very weak. It is the oldest people who are likely to pay a high price ticket.

We also notice that the variable 'woman' and 'discount' are correlated. In other words, the genre positively influences, but on a small scale, the fact of having a monthly subscription to a museum.

We can also use standard regression techniques, namely logit regression, to confirm the correlation between the variables. That's what we'll do later.

2. *Do you spot some problem with the variables? Are there any specific problems you should take care about?*

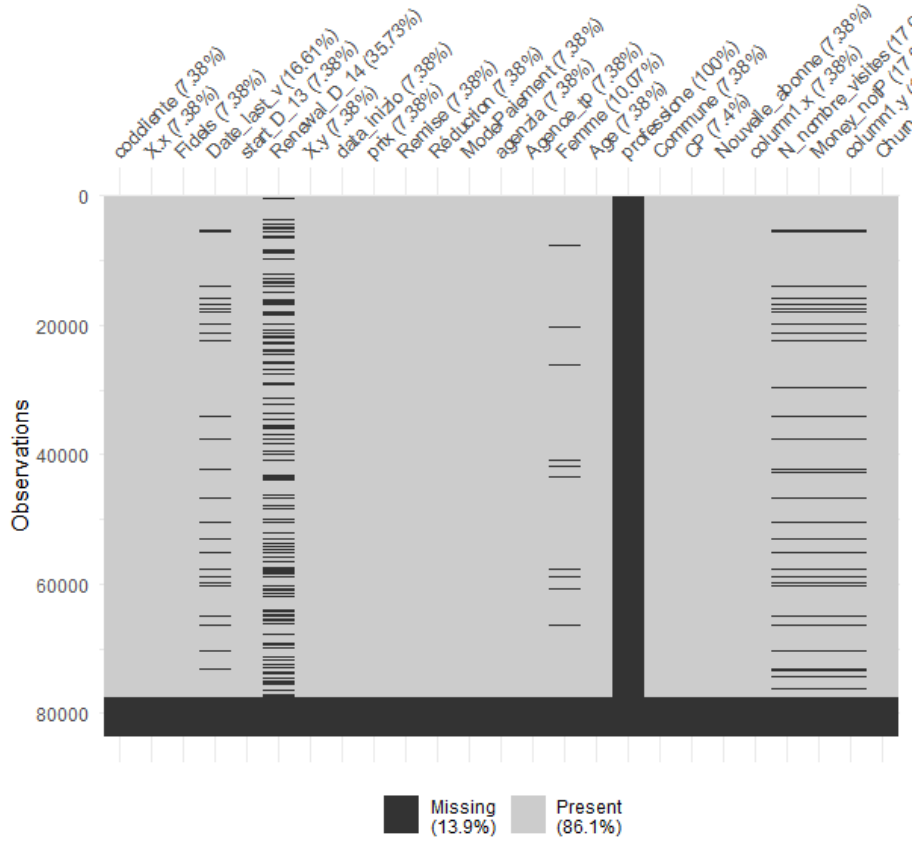
To begin with, there are 48,662 people who have not subscribed to a subscription and 18,674 subscribers.

Depending on the distribution of the database, some variables are distributed disproportionately. For example, concerning one person, the number of his visits is 243 per year. and it is a woman. You can tell that he is a museum employee and that these visits are supposed not "in line" with tourist visits.

Age is negatively correlated with being a museum subscriber, which surprised us. So, we hypothesized that being a woman reduces the number of visits to a museum. It is considered that women have many more daily responsibilities than men (children, work, home, etc.) and that culturally they are not attracted to museums.

3. *Analyze graphically the pattern of missing values. Is there any variable you should drop from the analysis?*

Our database contains many missing variables. We have about 14% of missing values against 84% of present values. The 'professional' variable contains 100% missing data, so we decided to delete it.

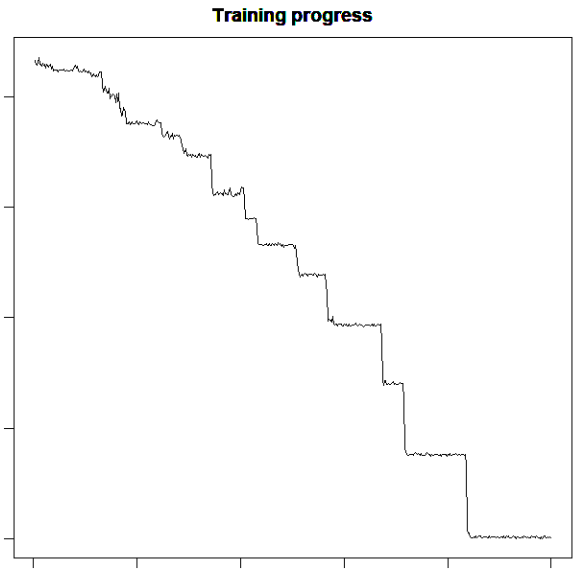


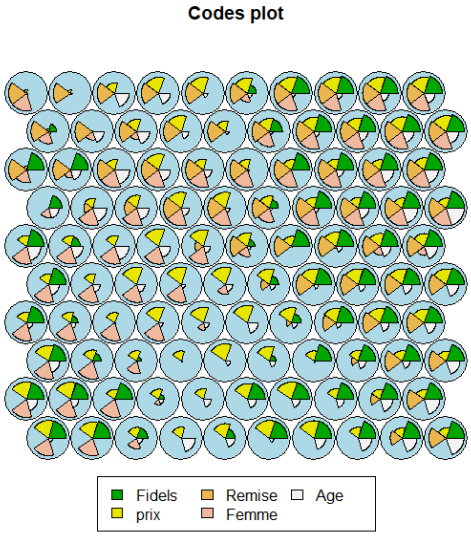
We must therefore remove the missing values because they constitute a problem for the different machine learning models that we will use later and can interfere with the learning process and the results. The cleaning allowed us to obtain our final database.

4. *Can you cluster the observations? Is there a cluster with most churners?*

The first plot shows the progress over time, the curve is continually decreasing. Hence, after 450 iterations, no more iterations are required.

The second graph on the right (code plots) allows us to see the distribution of the across map variables. We can see that the most dominant colors are orange and green, which correspond to the variables 'Fidels' and 'Woman'. There are five clusters defined. These different variables are correlated with each other since they are dependent on each other.





we can see the different classes of nodes through the graphs below. The distance between the variables is negligible hence the first graph.

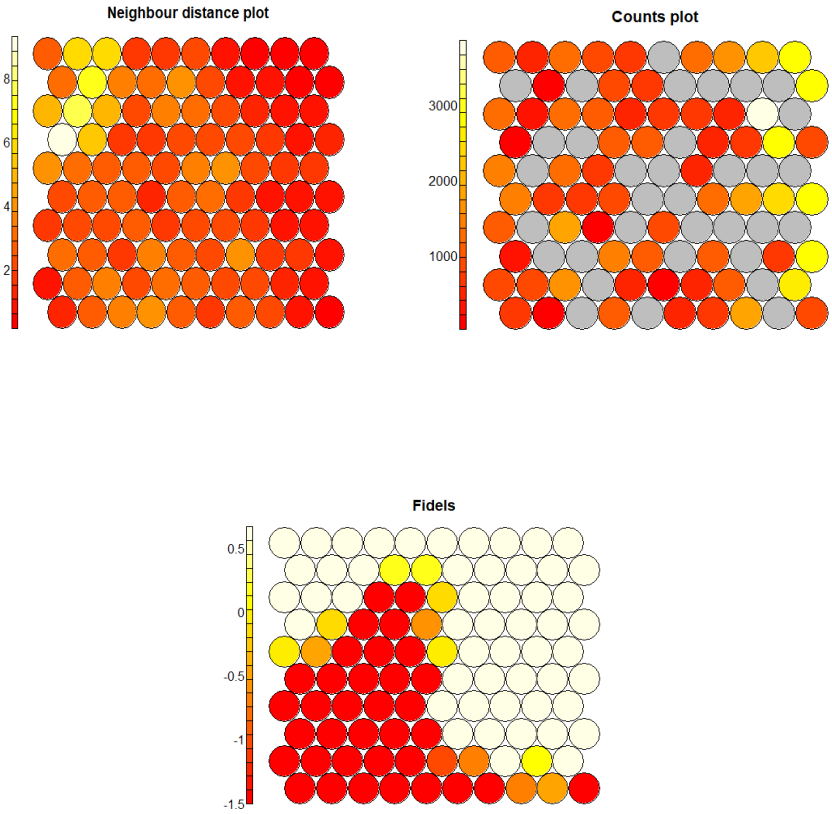


Table showing the frequency of museum patron distributions in Italy:

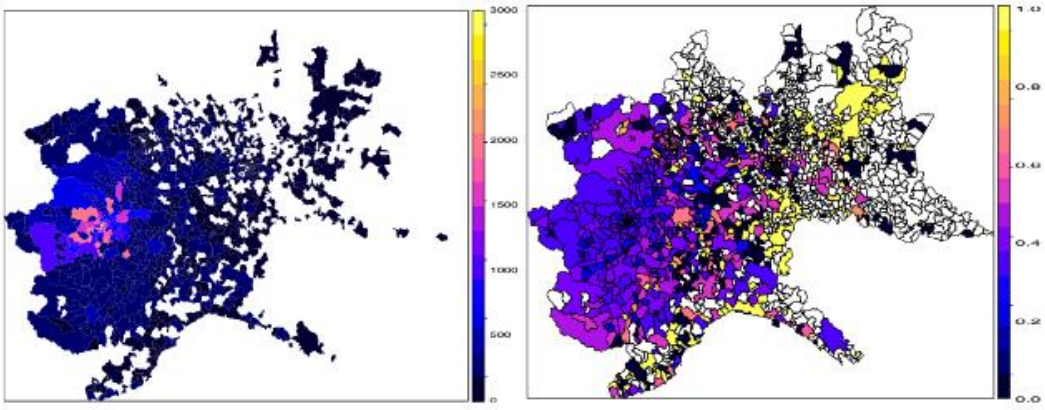


Figure 5: Frequency distribution of customers (left) and Frequency distribution of churners in percentage (right) in north west of Italy

TASK CHURN and MARKETING COMPAIGN

5. *Is there a causal impact of gender on the probability of churning? Identify a suitable model to create a counterfactual group with observational data. Present graphically the result in a meaningful way*

Since the dependent variable 'churner' was categorized above, we will now test the impact of age and gender. This study will be done using the logit function, using the non-subscriber as reference variables (churner = 0).

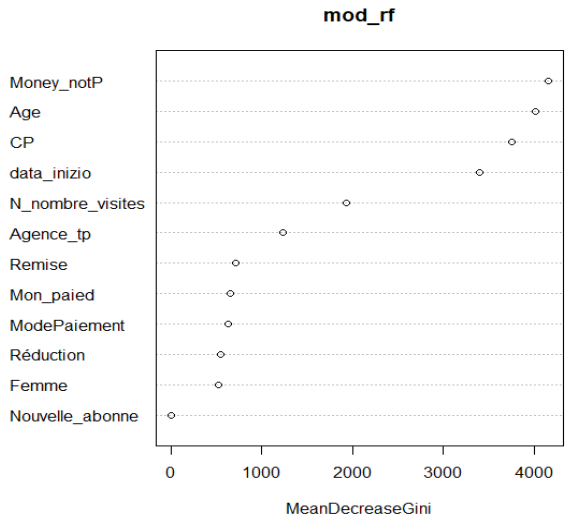
The figure below shows us that the feminine gender is negatively correlated with being a museum subscriber. The 'Age' variable is significant according to the Student law at 99% while the 'Woman' variable is not significant. We can conclude that being a woman reduces the probability of being a member of a museum. This may be because women in a certain way have more responsibility than men and prefer to allocate their pleasure to other activities than coming to a museum.

```
Call:
glm(formula = Churner ~ Age + Femme, family = "binomial", data = test_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4052  -0.8110  -0.6578   1.2136   2.3241

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 61.7688254  1.8748322  32.946  < 2e-16 ***
Age         -0.0318335  0.0009529 -33.406  < 2e-16 ***
Femme       -0.1283479  0.0326394  -3.932 8.41e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23939  on 20208  degrees of freedom
Residual deviance: 22760  on 20206  degrees of freedom
AIC: 22766

Number of Fisher Scoring iterations: 4
```

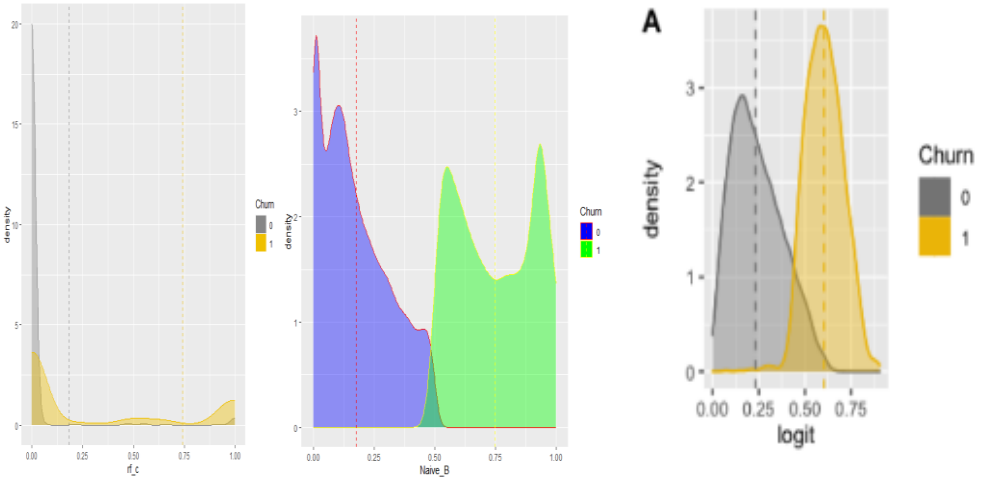
6. *Which models could you use to predict churners?*

Predictions allow us to see the percentage of the model that was correctly predicted.

First, we notice that thanks to the Random Forest model, the classification and the order of the variables are obtained thanks to the meandecreasegini below. The most important variable is 'Money_notP' followed by the age and lastly there is the museum subscription.



We chose to use three prediction models the Naive, Random Forest and logit. By making the data prediction graph, we notice that the density of having subscribed to a subscription is higher in the Naïve distribution. This gives us a better performance of the model compared to the Random Forrest model.

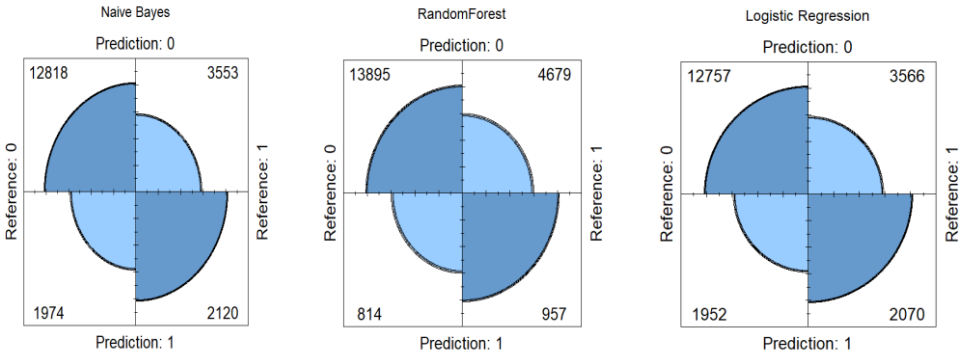


Next, we round our results by creating a confusion matrix to compare the number of true/false positives and negatives. We will form a confusion matrix with the training data.

Example of the logit model:

The logistic model generates 12,757 true negatives (0), 2,070 true positives (1), while there are 1,952 false negatives and 3,566 false positives.

Below are the graphs of the confusion matrices of the three models:



Now, calculate the misclassification error (for the training data) which {1 – misclassification}

```
1 - sum(diag(table)) / sum(table)
```

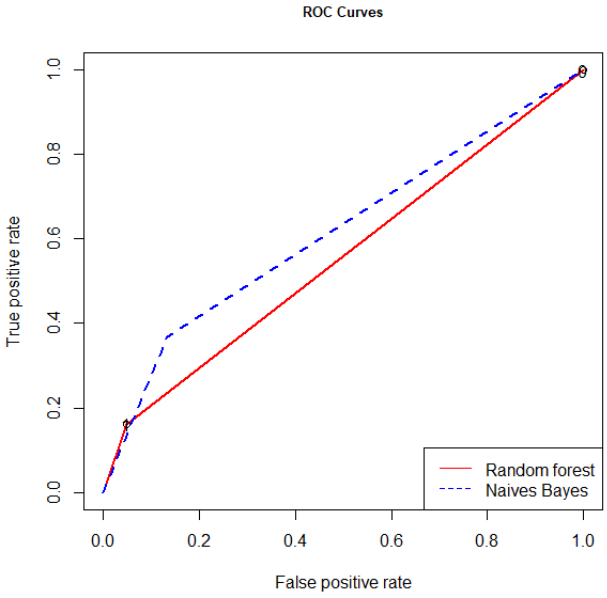
The misclassification error is 0.5691833, which means that the logit model predicts about 53% true values. In this, we can use regression techniques with categorical variables on various other data.

Table of predictions :

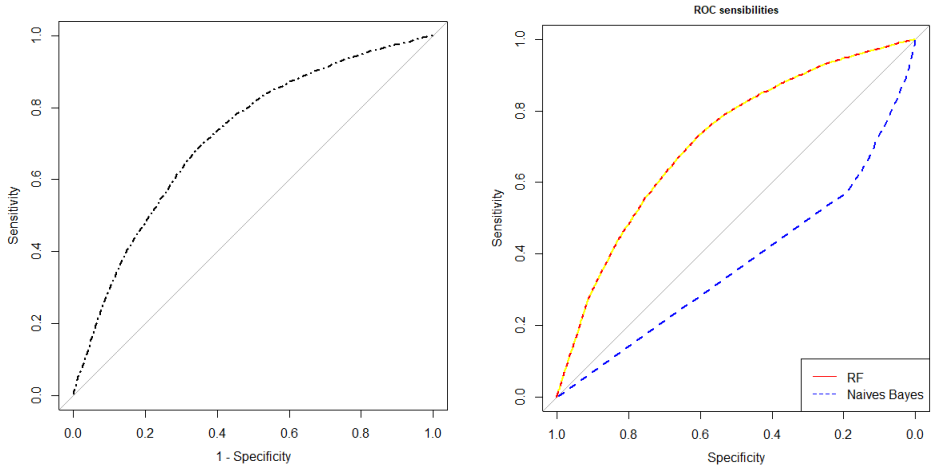
Modèles de machine Learning	Prédictions
Logit	0.56918
Naïves Bayes	0.72556
Random Forest	0.73299

The Random Forest model produces good results in terms of model performance.

The ROC curve allowed us to arrive at this graph: the true predicted values are higher in the Random Forrest model but as they increase, these values tend towards those of the prediction of the naive Bayes model.



The closer the curve is to the upper left corner, the more accurate and sensitive the model. It can be concluded that both models have corrected sensitivity.



7. *Cumulative profit*

We can see from the two graphs below that the profit is much better if we increase the ticket price to 5 euros. The Random Forrest and naive bayes models predict lower profits than the initial data.