MINI PROJET - CHURN & DATA BALANCING

DATA -2- AMSD

Réalisé par :

Mohammed Erifai MAAMIR - MLSD
Ahmed Seyfeddine GOUMEIDA - MLSD

2020/2021

Tables des matières:

I.Présentation des données :

1) Informations sur l'ensemble des données :

Les données sont liées aux campagnes de marketing direct d'une institution bancaire portugaise. Les campagnes de marketing étaient basées sur des appels téléphoniques. Souvent, plus d'un contact avec le même client était nécessaire pour savoir si le produit (dépôt bancaire à terme) serait ('oui') ou non ('non') souscrit.
Il y a quatre ensembles de données :

- 1) bank-additional-full.csv avec tous les exemples (41188) et 20 entrées, classés par date (de mai 2008 à novembre 2010), très proches des données analysées dans [Moro et al., 2014].
- 2) bank-additional.csv avec 10% des exemples (4119), sélectionnés aléatoirement à partir de 1), et 20 entrées.
- 3) bank-full.csv avec tous les exemples et 17 entrées, classés par date (ancienne version de ce jeu de données avec moins d'entrées).
- 4) bank.csv avec 10% des exemples et 17 entrées, sélectionnées au hasard à partir de 3 (ancienne version de cet ensemble de données avec moins d'entrées).Les plus petits ensembles de données sont fournis pour tester des algorithmes d'apprentissage automatique plus exigeants en termes de calcul (par exemple, SVM).

Dans notre étude on a choisi "bank-additional-full.csv".
L'objectif de classification est de prédire si le client souscrira (oui/non) un dépôt à terme (variable y).

2) Informations sur les attributs :

2.1)Variables d'entrée :

Données sur les clients de la banque :
1 - âge (numérique)
2 - job : type d'emploi (catégories : 'admin.', 'col bleu', 'entrepreneur', 'femme de ménage', 'direction', 'retraité', 'indépendant', 'services', 'étudiant', 'technicien', 'sans emploi', 'inconnu')
3 - marital : état civil (catégorique : 'divorcé', 'marié', 'célibataire', 'inconnu' ; remarque : 'divorcé' signifie divorcé ou veuf)
4 - education (catégories : 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5 - défaut : le crédit est-il en défaut ? (catégorique : 'non', 'oui', 'inconnu')
6 - housing : a un prêt au logement ? (catégories : 'non', 'oui', 'inconnu')
7 - loan : a un prêt personnel ? (catégories : 'non', 'oui', 'inconnu')

2.2)Lié au dernier contact de la campagne actuelle :

8 - contact : type de communication du contact (catégorique : 'cellulaire', 'téléphone')
9 - month : mois de l'année du dernier contact (catégorique : 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week : jour de la semaine du dernier contact (catégorique : 'lun', 'tue', 'wed', 'thu', 'fri')
11 - duration : durée du dernier contact, en secondes (numérique). Note importante : cet attribut affecte fortement la cible de sortie (par exemple, si duration=0 alors y='no'). Pourtant, la durée n'est pas connue avant qu'un appel ne soit effectué. De même, après la fin de l'appel, y est évidemment connu. Ainsi, cette entrée ne devrait être incluse qu'à des fins de référence et devrait être écartée si l'intention est d'avoir un modèle prédictif réaliste.
Autres attributs :

12 - campaign : nombre de contacts effectués pendant cette campagne et pour ce client (numérique, inclut le dernier contact).

13 - pdays : nombre de jours qui se sont écoulés depuis le dernier contact du client lors d'une campagne précédente (numérique ; 999 signifie que le client n'a pas été contacté auparavant).

14 - previous : nombre de contacts effectués avant cette campagne et pour ce client (numérique)

15 - poutcome : résultat de la campagne marketing précédente (catégorique : 'échec', 'inexistant', 'succès')

2.3)Attributs du contexte social et économique

16 - emp.var.rate : taux de variation de l'emploi - indicateur trimestriel (numérique)

17 - cons.price.idx : indice des prix à la consommation - indicateur mensuel (numérique)

18 - cons.conf.idx : indice de confiance des consommateurs - indicateur mensuel (numérique)

19 - euribor3m : taux euribor à 3 mois - indicateur quotidien (numérique)

20 - nr.employed : nombre de salariés - indicateur trimestriel (numérique)

2.4)Variable de sortie (objectif souhaité) :

21 - y - le client a-t-il souscrit un dépôt à terme ? (binaire : 'oui', 'non')

II. Analyse univariée :

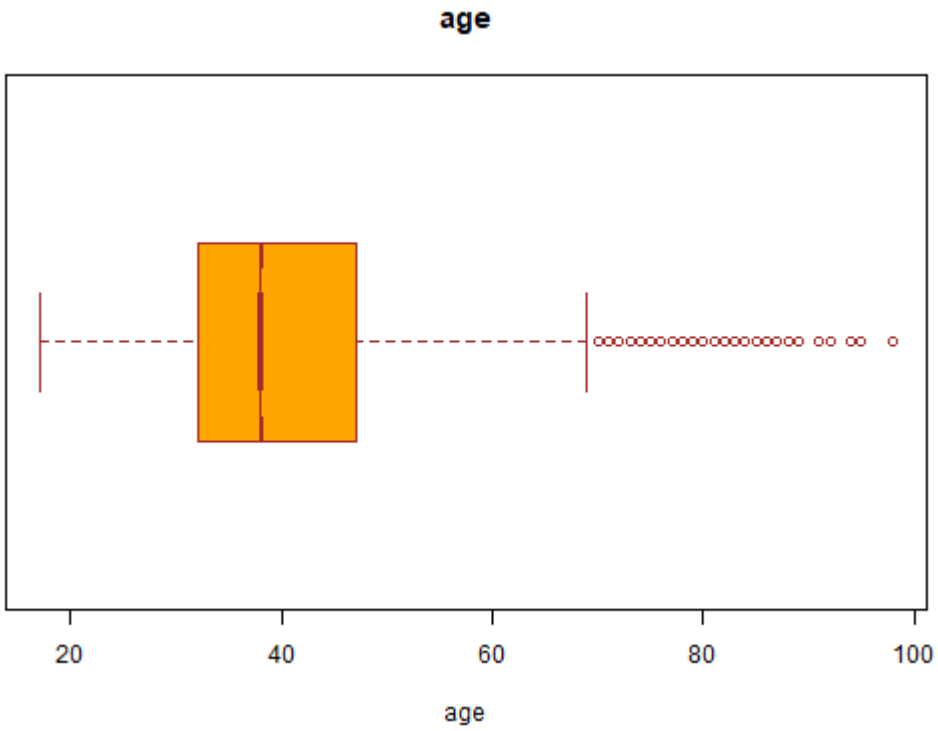
L'analyse univariée permet d'explorer une seule feature à la fois. Cette analyse se base sur les statistiques descriptives. Ces dernières permettent de tirer des indications concises sur une *feature* donnée. Parmi ces indicateurs, on retrouve la moyenne, la médiane ainsi que les mesures de dispersion de données.

1) **Summary :**

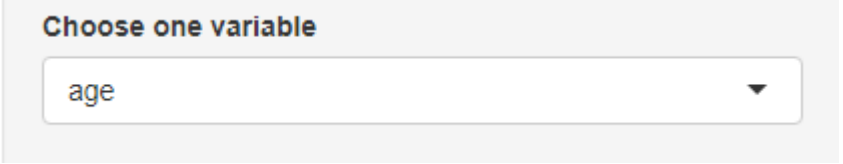
```
    campaign          pdays          previous            poutcome      emp.
 Min.   : 1.000   Min.   :  0.0   Min.   :0.000   failure    : 4252   Min.
 1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000   nonexistent:35563   1st Q
 Median : 2.000   Median :999.0   Median :0.000   success    : 1373   Media
 Mean   : 2.568   Mean   :962.5   Mean   :0.173                       Mean
 3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000                       3rd Q
 Max.   :56.000   Max.   :999.0   Max.   :7.000                       Max.

 cons.price.idx  cons.conf.idx     euribor3m      nr.employed     y
 Min.   :92.20   Min.   :-50.8   Min.   :0.634   Min.   :4964   no :36548
 1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.344   1st Qu.:5099   yes: 4640
 Median :93.75   Median :-41.8   Median :4.857   Median :5191
 Mean   :93.58   Mean   :-40.5   Mean   :3.621   Mean   :5167
 3rd Qu.:93.99   3rd Qu.:-36.4   3rd Qu.:4.961   3rd Qu.:5228
 Max.   :94.77   Max.   :-26.9   Max.   :5.045   Max.   :5228
```

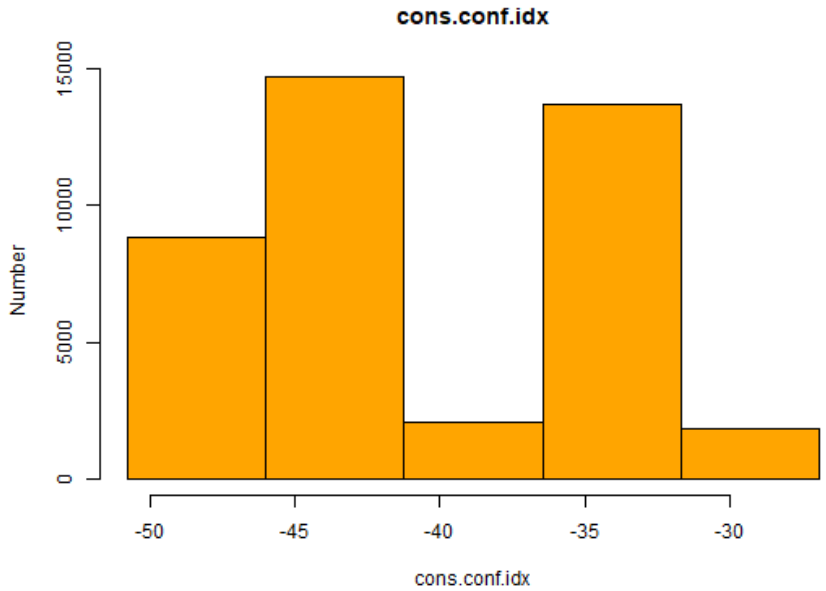
2) **Boite à moustache(boxplot) :**

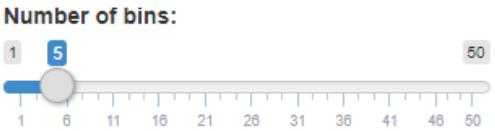


choix dynamique de la variable via l'entrée suivante :



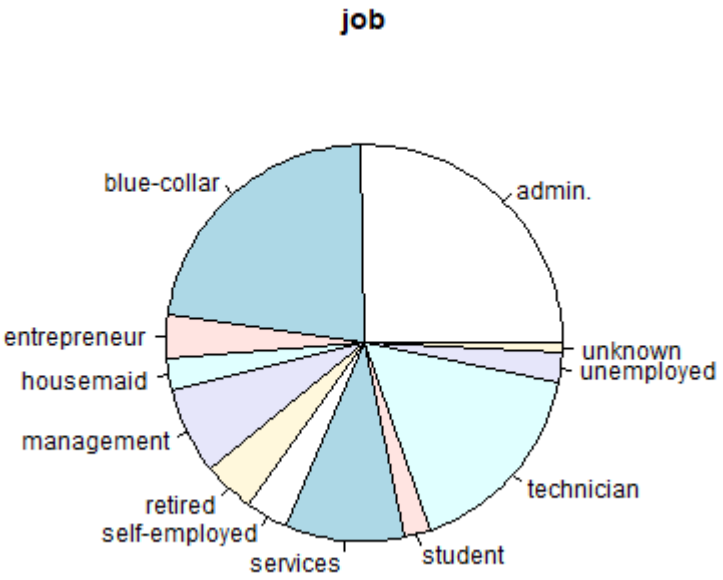
3) **Histogramme :**



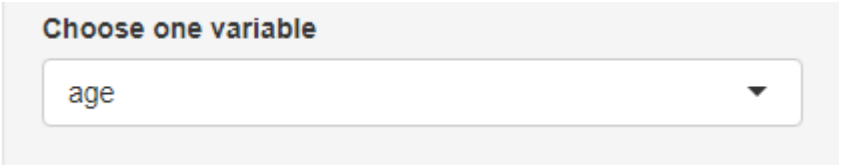


choix dynamique de la variable via l'entrée suivante :

4) **Pie** :



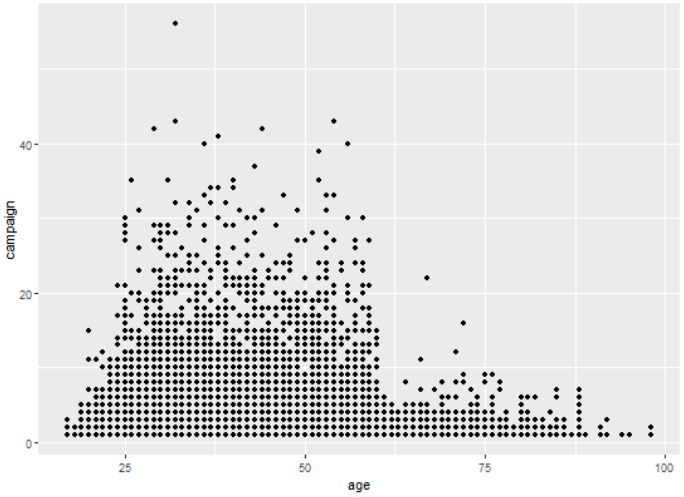
choix dynamique de la variable via l'entrée suivante :



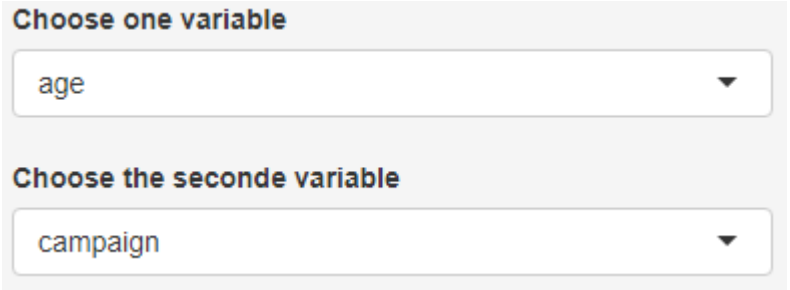
III. Analyse bivariée :

L'analyse bivariée est l'une des formes les plus simples d'analyse quantitative.
Elle implique l'analyse de deux variables dans le but de déterminer la relation empirique entre elles.
Faire une analyse bivariée, c'est étudier la relation entre deux variables : sont-elles liées ? les valeurs de l'une influencent-elles les valeurs de l'autre ? ou sont-elles au contraire indépendantes ?
Et donc pour répondre à ces questions on utilise :

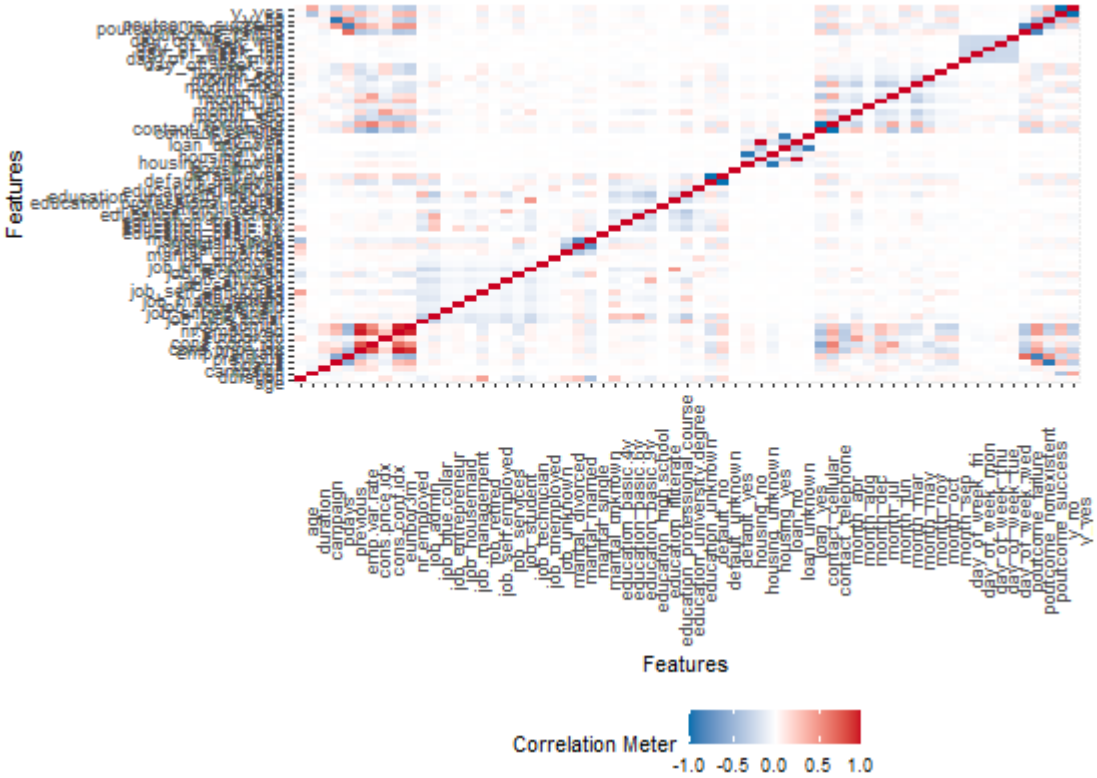
1) **Nuage de points:**



pour choisir les variables qui vont être utilisées à représenter le nuage de points on utilises les deux entrées dans le navbar à gauche :



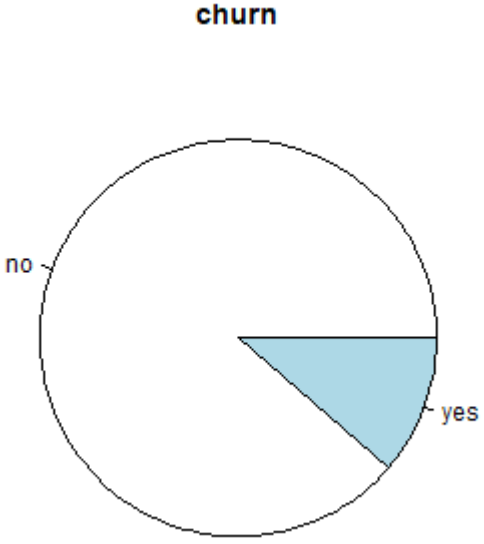
2) **matrice de corrélation :**



On remarque que nos variables sont indépendantes.

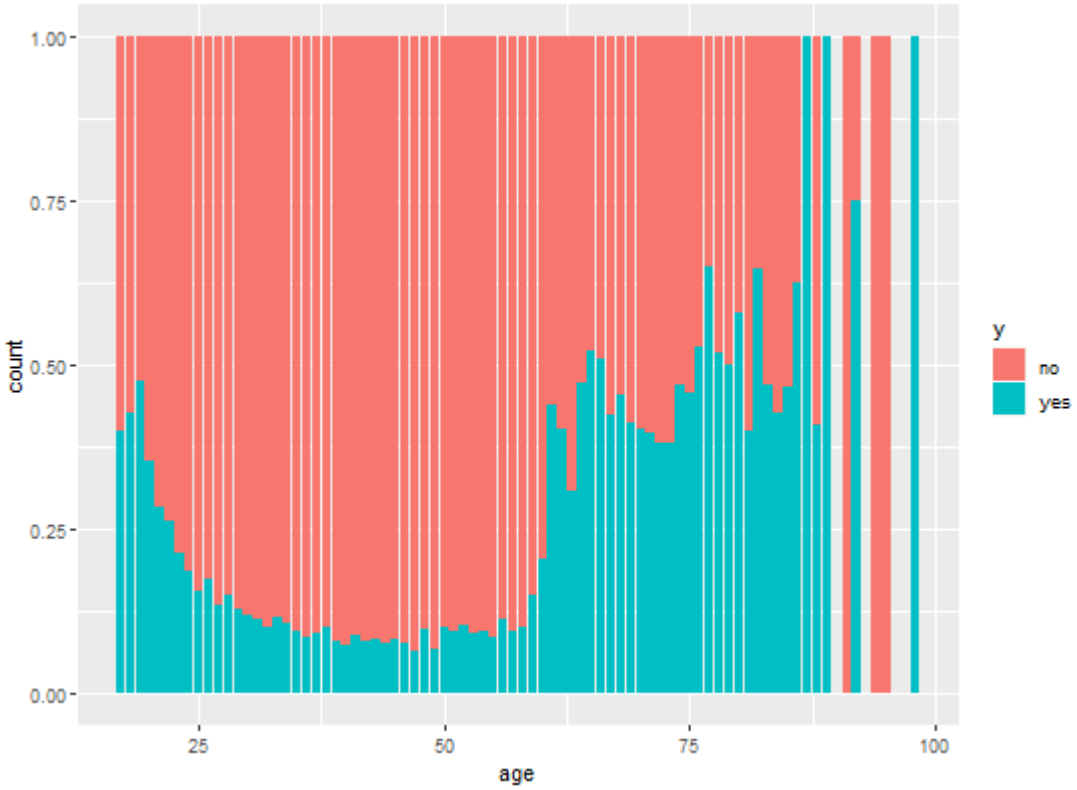
IV. L'exploration des variables vis-`à-vis du churn :

1) **Affichage de Customer attrition in data :**

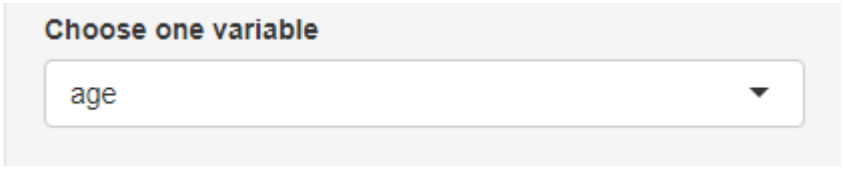


on remarque un grand déséquilibre dans notre dataset (forte domination de la classe ''no'')

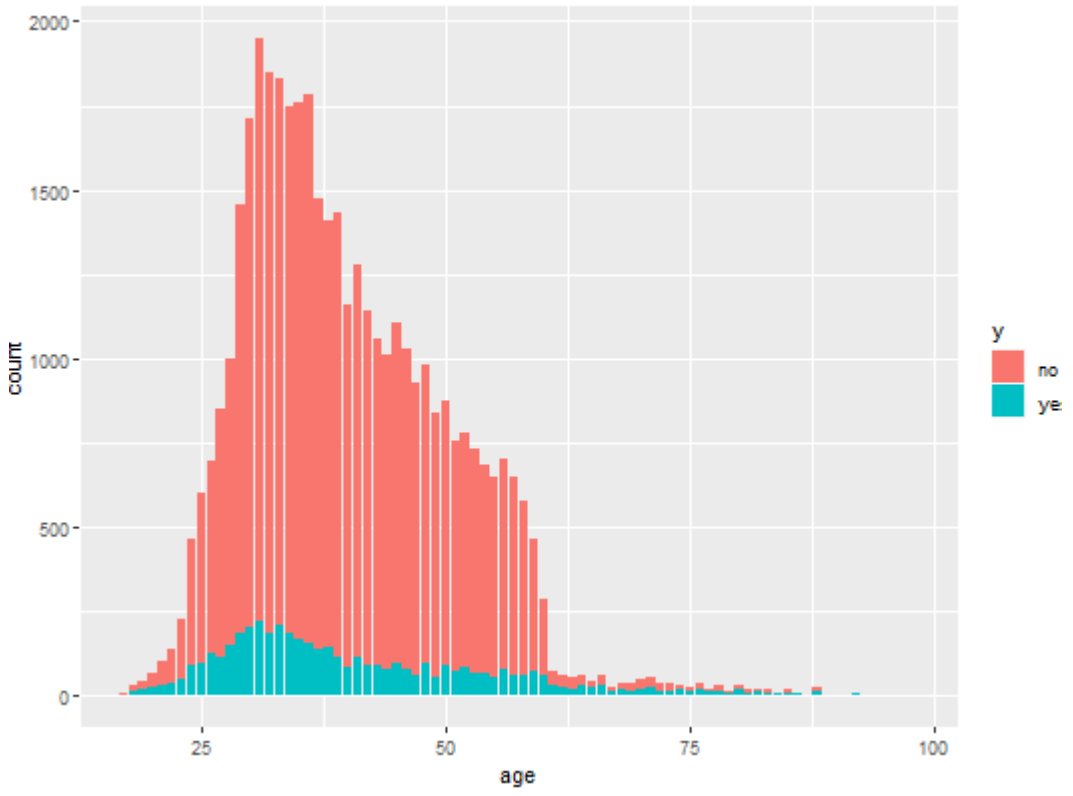
2) **Variables distribution in customer attrition:**



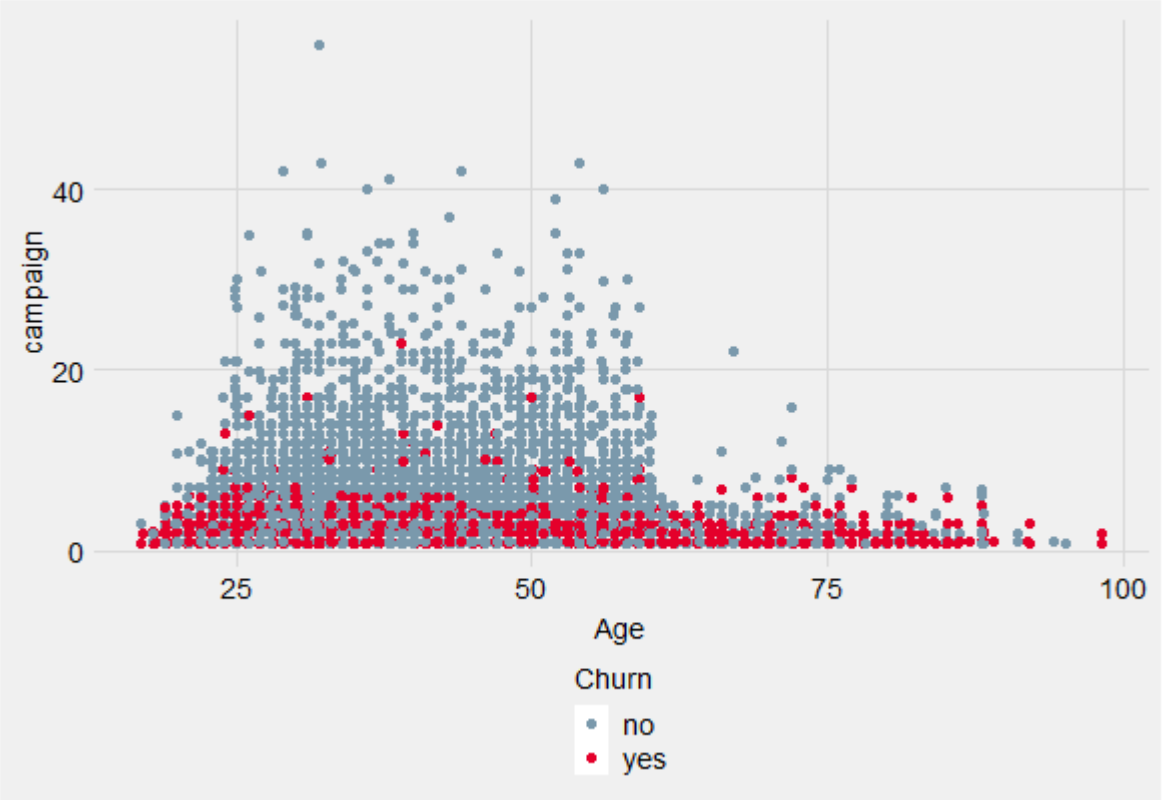
pour sélectionner la variable on utilise l'entrée suivante:



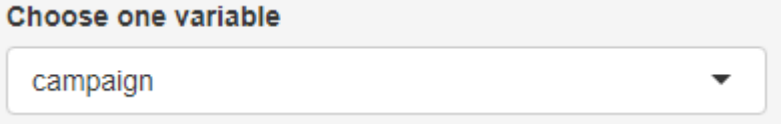
3) **Affichage ciblé pour la variable âge :**



- Affichage du nuage de points par rapport à la variable âge + une deuxième variable (choix dynamique)

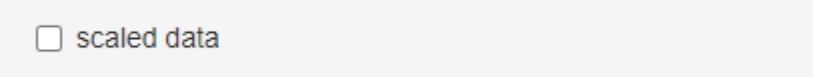


pour le choix dynamique :



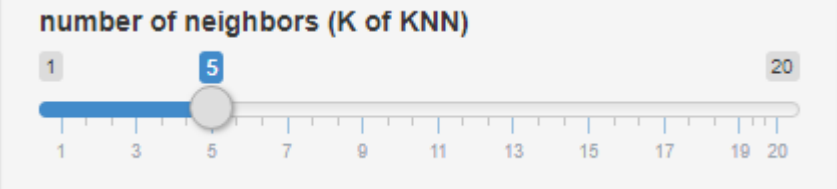
V. Entraînement de modèles :

- Paramétrage :
  - check box pour data scaling :



1) KNN :

- choix dynamique de k (nombre de voisin) :



- matrice de confusion :

KNN With original data :

Outcome	Ref-FALSE	Ref-TRUE
Pre-FALSE	6997	497
Pre-TRUE	291	452

- matrice de confusion (scaled data) :

KNN With original data :

Outcome	Ref-FALSE	Ref-TRUE
Pre-FALSE	7017	593
Pre-TRUE	271	356

2) LOGISTIC REGRESSION (LR) :

```
[1] "---------------LOGISTIC REGRESSION SUMMARY----------------"

Call:
glm(formula = dataset$y ~ . - y, family = "binomial", data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.9610  -0.3180  -0.1940  -0.1404   3.1704

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.581e+00  1.937e+01  -0.185 0.853312
age             5.473e-03  1.871e-03   2.925 0.003450 **
job             8.151e-03  5.551e-03   1.469 0.141956
marital         1.006e-01  3.612e-02   2.786 0.005330 **
education       5.145e-02  9.852e-03   5.222 1.77e-07 ***
default        -3.817e-01  6.544e-02  -5.832 5.47e-09 ***
housing        -5.615e-03  2.037e-02  -0.276 0.782805
loan           -3.401e-02  2.809e-02  -1.211 0.225934
contact        -7.038e-01  6.431e-02 -10.943  < 2e-16 ***
month          -1.123e-01  9.351e-03 -12.014  < 2e-16 ***
day_of_week     5.641e-02  1.451e-02   3.887 0.000101 ***
duration        4.584e-03  7.281e-05  62.966  < 2e-16 ***
campaign       -3.312e-02  1.142e-02  -2.899 0.003741 **
pdays          -1.019e-03  1.584e-04  -6.432 1.26e-10 ***
previous       -6.680e-02  5.558e-02  -1.202 0.229449
poutcome        4.595e-01  7.623e-02   6.028 1.66e-09 ***
emp.var.rate   -9.165e-01  6.814e-02 -13.450  < 2e-16 ***
cons.price.idx  7.157e-01  1.198e-01   5.975 2.31e-09 ***
cons.conf.idx   2.030e-02  6.807e-03   2.982 0.002860 **
euribor3m       6.402e-01  1.027e-01   6.231 4.63e-10 ***
nr.employed    -1.310e-02  1.844e-03  -7.106 1.20e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28999  on 41187  degrees of freedom
Residual deviance: 17563  on 41167  degrees of freedom
AIC: 17605

Number of Fisher Scoring iterations: 6

[1] "-----------------------CONFUSION MATRIX------------------------"

glm.pred     1     2
       1 35598  2747
       2   950  1893
[1] "-----------------------SCORE------------------------"
[1] 0.9102408
```

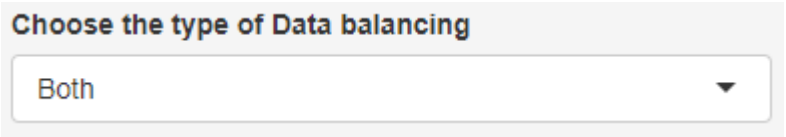
En regardant les matrices de confusion pour les deux méthodes KNN et LR on remarque que la classification est bonne pour la classe dominante mais ce n'est pas le cas pour l'autre classe (minoritaire).

- ***pour le KNN*** :
  - parmi 7288 valeurs **"no"** : 7017 prédites correctement et 291 valeurs mals prédites donc ***96%*** des valeurs sont correctement prédites.
  - parmi 949 valeurs **"yes"** : 356 prédites correctement et 593 valeurs mals prédites donc seulement ***37.5%*** des valeurs sont correctement prédites.
- ***pour le LR :***
  - parmi 36548 valeurs **"no"** : 35598 prédites correctement et 950 valeurs mals prédites donc ***97.4%*** des valeurs sont correctement prédites.
  - parmi 4640 valeurs **"yes"** : 1893 prédites correctement et 2747 valeurs mals prédites donc seulement ***40%*** des valeurs sont correctement prédites.

Ce qui justifie la nécessité d'un Data balancing.

VI. Entraînement de modèles (Balanced Data) :

Choix de type de Data balancing :(Both , Under , Over)



1) KNN : (balanced mode choisi : Both)

KNN With balanced data :

Outcome	Ref-FALSE-Balanc	Ref-TRUE-Balanc
Pre-FALSE-Balanc	6150	16
Pre-TRUE-Balanc	1260	9049

2) **LR :**

```
[1] "---------------LOGISTIC REGRESSION SUMMARY----------------"

Call:
glm(formula = dataset$y ~ . - y, family = "binomial", data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.1196  -0.4152  -0.1165   0.5084   2.6622

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    84.9629657 13.7197655   6.193 5.91e-10 ***
age             0.0085754  0.0014565   5.888 3.92e-09 ***
job             0.0171707  0.0041968   4.091 4.29e-05 ***
marital         0.1459102  0.0271833   5.368 7.98e-08 ***
education       0.0793617  0.0074768  10.614  < 2e-16 ***
default        -0.4704697  0.0475063  -9.903  < 2e-16 ***
housing         0.0055706  0.0154259   0.361   0.7180
loan           -0.0863928  0.0211726  -4.080 4.50e-05 ***
contact        -0.5405039  0.0467395 -11.564  < 2e-16 ***
month          -0.1849110  0.0073640 -25.110  < 2e-16 ***
day_of_week     0.0342237  0.0108429   3.156   0.0016 **
duration        0.0066852  0.0000768  87.050  < 2e-16 ***
campaign       -0.0175325  0.0081428  -2.153   0.0313 *
pdays          -0.0007627  0.0001557  -4.897 9.71e-07 ***
previous        0.0296269  0.0566070   0.523   0.6007
poutcome        0.5832815  0.0709184   8.225  < 2e-16 ***
emp.var.rate   -1.3817945  0.0514062 -26.880  < 2e-16 ***
cons.price.idx  0.3846145  0.0848102   4.535 5.76e-06 ***
cons.conf.idx  -0.0265201  0.0048470  -5.471 4.46e-08 ***
euribor3m       1.4659597  0.0747145  19.621  < 2e-16 ***
nr.employed    -0.0250772  0.0013543 -18.517  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 57099  on 41187  degrees of freedom
Residual deviance: 28275  on 41167  degrees of freedom
AIC: 28317

Number of Fisher Scoring iterations: 6

[1] "--------------------CONFUSION MATRIX--------------------"

glm.pred     1     2
       1 17603  2638
       2  3030 17917
[1] "----------------------SCORE-----------------------------"
[1] 0.8623871
```

En regardant les matrices de confusion pour les deux méthodes KNN et LR on remarque que la classification a été améliorée pour la classe minoritaire.

- ***pour le KNN*** :
  - parmi 7410 valeurs **“no”** : 6150 prédites correctement et 1260 valeurs mals prédites donc ***83%*** des valeurs sont correctement prédites.
  - parmi 9065 valeurs **“yes”** : 9049 prédites correctement et 16 valeurs mals prédites donc ***99.8%*** des valeurs sont correctement prédites.

- ***pour le LR :***
    - parmi 20633 valeurs **"no"** : 17603 prédites correctement et 3030 valeurs mals prédites donc ***85.3%*** des valeurs sont correctement prédites.
    - parmi 20555 valeurs **“yes”** : 17917 prédites correctement et 2638 valeurs mals prédites donc ***87.1%*** des valeurs sont correctement prédites.

Ce qui montre l’effet du Data balancing sur la performance de nos modèles.