

ETUDE DE CAS

ENTRAINEMENT

réalisé par Nicolas BEHBAHANI

Le 22 mai 2017

RÉSUMÉ. — L'objectif de cette étude de cas consiste à créer un modèle de prédiction de la variable cible à partir des variables à disposition (X1 à X15).

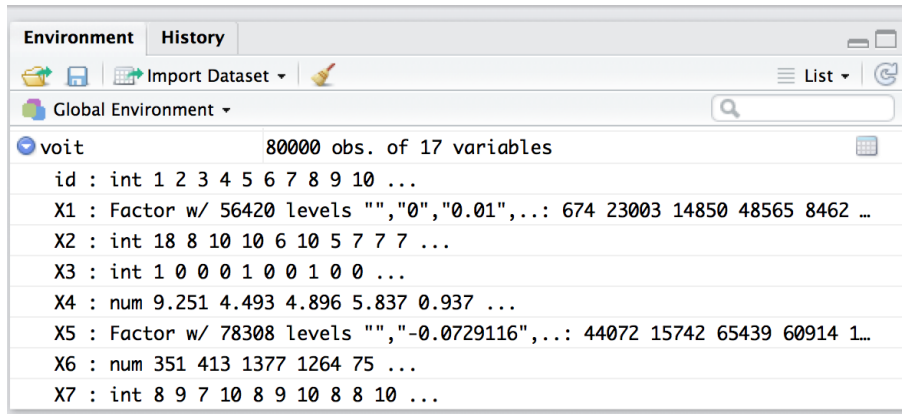
Remarque. — Les données du fichier (Data.csv) sont fictives. La variable cible représente la survenue d'un comportement (cible=1) ou pas (cible=0). La première variable est un identifiant et n'est pas à utiliser dans cette étude.

SOMMAIRE

Introduction.	3
§ 1. Description du jeu de données.	3
§ 2. Objectif de l'étude	3
§ 3. Chargement des données.	4
§ 4. Analyse exploratoire des données	4
§ 5. Recherche des variables explicatives pertinentes	6
5.1 Méthode par analyse graphique	6
§ 6. Construction des modèles	9
§ 7. Modélisation	11

INTRODUCTION

On s'intéresse pour notre étude de cas à une base de données contenant 8000 observations.



The screenshot shows the RStudio 'Environment' pane. At the top, there are tabs for 'Environment' and 'History'. Below the tabs, there are icons for file operations and a search bar. The main area displays the 'Global Environment' with a search bar. A dataset named 'voit' is listed, showing '80000 obs. of 17 variables'. Below this, the variables are listed with their data types and some sample values:

Variable	Type	Sample Values
id	int	1 2 3 4 5 6 7 8 9 10 ...
X1	Factor w/ 56420 levels	"", "0", "0.01", ...: 674 23003 14850 48565 8462 ...
X2	int	18 8 10 10 6 10 5 7 7 7 ...
X3	int	1 0 0 0 1 0 0 1 0 0 ...
X4	num	9.251 4.493 4.896 5.837 0.937 ...
X5	Factor w/ 78308 levels	"", "-0.0729116", ...: 44072 15742 65439 60914 1...
X6	num	351 413 1377 1264 75 ...
X7	int	8 9 7 10 8 9 10 8 8 10 ...

Nous constatons dès le début qu'il faut remplacer les points (.) par les virgules (,) dans le fichier de notre base de données afin de ne travailler qu'avec des variables numériques.

Le problématique et les objectifs de l'étude

Contexte de l'étude

§ 1. DESCRIPTION DU JEU DE DONNÉES

Le jeu de données contient :

- dans la première colonne les identifiants (id)
- des variables numériques allant de X1 à X15
- une variable *cible* codée 1 pour la survenue d'un comportement ou 0 dans le cas contraire.

§ 2. OBJECTIF DE L'ÉTUDE

L'objectif de cette étude de cas consiste à créer un modèle de prédiction de la variable cible à partir des variables explicatives à disposition (X1 à X15).

§ 3. CHARGEMENT DES DONNÉES

Nous chargeons d'abord l'ensemble des données afin de sélectionner les colonnes qui nous intéressent dans notre étude de cas :

```
> names(voit)
[1] "id"      "X1"      "X2"      "X3"      "X4"      "X5"      "X6"      "X7"      "X8"      "X9"
[11] "X10"     "X11"     "X12"     "X13"     "X14"     "X15"     "cible"
```

Nous prenons uniquement les colonnes qui nous intéressent en supprimant la colonne identifiant (id) qui n'est pas utile dans notre analyse :

Observation des 6 premières lignes de notre nouvelle base de données :

```
Console ~/Desktop/Test_job/
> head(base)
  cible    X1 X2 X3    X4    X5    X6 X7    X8    X9    X10 X11    X12
1     0 1015.25 18 1 9.251 1.2609082 351.38 8 290461 0 -0.46 1.11 289953.38
2     1  248.86 8 0 4.493 1.1061443 413.00 9 34228 0 0.64 1.19 36046.58
3     0  179.88 10 0 4.896 1.4805655 1376.72 7 244860 0 0.52 1.17 245050.07
4     0  719.93 10 0 5.837 1.4019217 1263.60 10 5401 0 -0.46 1.16 5426.54
5     0   14.3 6 1 0.937 1.1067053 75.00 8 781155 0 0.61 1.17 781147.86
6     0 1102.86 10 0 6.682 1.2016039 0.00 9 3321072 0.01 0.57 1.15 3320520.58
      X13 X14 X15
1 290460.94 1 10
2 37724.79 4 16
3 245363.97 39 16
4  6088.62 2 7
5  781155 47 3
6 3321069.34 33 7
```

Observation des 6 dernières lignes de notre nouvelle base de données :

```
> tail(base)
  cible    X1 X2 X3    X4    X5    X6 X7    X8    X9    X10 X11    X12
79995  0 283.27 10 0 6.092 1.1038129 0.00 10 69401 0 -0.78 1.17 69259.37
79996  0 586.53 9 0 3.339 1.0532532 433.50 8 456930 0.09 -0.34 1.13 457021.74
79997  0  24.18 11 0 5.097 1.2581011 132.28 10 275969 0 0.51 1.13 275956.92
79998  0 1299.16 12 1 4.087 1.1065218 597.75 7 10471761 0 -0.31 1.12 10471111.43
79999  0 951.14 17 0 8.842 1.3328117 19.46 10 300451 0.09 -0.3 1.15 299975.44
80000  0  779.82 10 0 5.509 1.1928193 881.85 10 1615035 0.1 0.91 1.36 1614645.1
      X13 X14 X15
79995 69400.49 102 16
79996 457620.78 24 4
79997 275969 15 16
79998 10471757.92 4 7
79999 300441.56 2 4
80000 1615033.09 15 16
```

§ 4. ANALYSE EXPLORATOIRE DES DONNÉES

Affichage de la structure de la base de données :

```

Console ~/Desktop/Test_job/
> str(voit)
'data.frame': 80000 obs. of 17 variables:
 $ id : int 1 2 3 4 5 6 7 8 9 10 ...
 $ X1 : num 1015.2 248.9 179.9 719.9 14.3 ...
 $ X2 : int 18 8 10 10 6 10 5 7 7 7 ...
 $ X3 : int 1 0 0 0 1 0 0 1 0 0 ...
 $ X4 : num 9.251 4.493 4.896 5.837 0.937 ...
 $ X5 : num 1.26 1.11 1.48 1.4 1.11 ...
 $ X6 : num 351 413 1377 1264 75 ...
 $ X7 : int 8 9 7 10 8 9 10 8 8 10 ...
 $ X8 : int 290461 34228 244860 5401 781155 3321072 61460 335579 29933297 3042109 ...
 $ X9 : num 0 0 0 0 0 0.01 0 0 0 0.01 ...
 $ X10 : num -0.46 0.64 0.52 -0.46 0.61 0.57 -0.38 -0.75 0.71 0.7 ...
 $ X11 : num 1.11 1.19 1.17 1.16 1.17 1.15 1.1 1.15 1.19 1.2 ...
 $ X12 : num 289953 36047 245050 5427 781148 ...
 $ X13 : num 290461 37725 245364 6089 781155 ...
 $ X14 : int 1 4 39 2 47 33 0 1 38 3 ...
 $ X15 : int 10 16 16 7 3 7 4 4 16 16 ...
 $ cible: int 0 1 0 0 0 0 0 0 0 0 ...

```

Affichage d'un résumé de la base de données :

```

Console ~/Desktop/Test_job/
> summary(voit)
      id      X1      X2      X3      X4
Min.   : 1    Min.   : 0.0    Min.   : 0.00    Min.   :0.0000    Min.   : 0.074
1st Qu.:20001 1st Qu.: 85.0    1st Qu.: 8.00    1st Qu.:0.0000    1st Qu.: 3.984
Median :40000 Median : 328.0    Median :11.00    Median :0.0000    Median : 5.491
Mean   :40000 Mean   : 1168.8    Mean   :10.82    Mean   :0.3226    Mean   : 5.624
3rd Qu.:60000 3rd Qu.: 974.4    3rd Qu.:13.00    3rd Qu.:1.0000    3rd Qu.: 7.113
Max.   :80000 Max.   :755300.4    Max.   :30.00    Max.   :1.0000    Max.   :19.509
NA's   :128

      X5      X6      X7      X8
Min.   :-37.217 Min.   : 0.0    Min.   : 1.000    Min.   : 0
1st Qu.: 1.118 1st Qu.: 60.3    1st Qu.: 8.000    1st Qu.: 357752
Median : 1.231 Median : 303.6    Median : 9.000    Median : 1248377
Mean   : 1.225 Mean   : 1220.5    Mean   : 8.242    Mean   : 3596878
3rd Qu.: 1.374 3rd Qu.: 971.7    3rd Qu.: 9.000    3rd Qu.: 3851951
Max.   : 44.599 Max.   :1078829.6    Max.   :11.000    Max.   :67636004
NA's   :629      NA's   :75

      X9      X10      X11      X12
Min.   :0.000000 Min.   : -13.1300 Min.   : 1.000    Min.   : 0

```

La variable cible *cible* doit être traitée en type *facteur* et non en variable numérique, c'est pourquoi nous le transformons ainsi :

```

Console ~/Desktop/Test_job/
> base$cible <- factor(base$cible)
> summary(base)
      cible      X1      X2      X3      X4
0:70506 Min.   : 0.0    Min.   : 1.00    Min.   :0.0000    Min.   : 0.074
1: 7323 1st Qu.: 86.5    1st Qu.: 8.00    1st Qu.:0.0000    1st Qu.: 3.984
      Median : 331.1    Median :11.00    Median :0.0000    Median : 5.491
      Mean   : 1168.0    Mean   :10.83    Mean   :0.3174    Mean   : 5.624
      3rd Qu.: 978.3    3rd Qu.:13.00    3rd Qu.:1.0000    3rd Qu.: 7.110
      Max.   :755300.4    Max.   :30.00    Max.   :1.0000    Max.   :19.509

      X5      X6      X7      X8
Min.   :-37.217 Min.   : 0.0    Min.   : 1.000    Min.   : 0
1st Qu.: 1.118 1st Qu.: 63.2    1st Qu.: 8.000    1st Qu.: 367441
Median : 1.231 Median : 309.1    Median : 9.000    Median : 1283582
Mean   : 1.224 Mean   : 1231.3    Mean   : 8.236    Mean   : 3635837
3rd Qu.: 1.374 3rd Qu.: 982.2    3rd Qu.: 9.000    3rd Qu.: 3990656
Max.   : 44.599 Max.   :1078829.6    Max.   :11.000    Max.   :67636004

      X9      X10      X11      X12
Min.   :0.000000 Min.   : -13.1300 Min.   : 1.000    Min.   : 0
1st Qu.:0.000000 1st Qu.: -0.4700 1st Qu.: 1.140    1st Qu.: 367419

```

Nous supprimons ensuite les valeurs manquantes de notre base de données avec la commande suivante :

```

Console ~/Desktop/Test_job/
> base <- na.omit(base)
> attach(base)
The following objects are masked from base (pos = 3):

  cible, X1, X10, X11, X12, X13, X14, X15, X2, X3, X4, X5, X6, X7, X8, X9

The following objects are masked from voit (pos = 4):

  cible, X1, X10, X11, X12, X13, X14, X15, X2, X3, X4, X5, X6, X7, X8, X9

The following objects are masked from base (pos = 6):

  cible, X1, X10, X11, X12, X13, X14, X15, X2, X3, X4, X5, X6, X7, X8, X9

The following objects are masked from voit (pos = 7):

  cible, X1, X10, X11, X12, X13, X14, X15, X2, X3, X4, X5, X6, X7, X8, X9

```

Ainsi nous obtenons au final plus que 77829 observations :

```

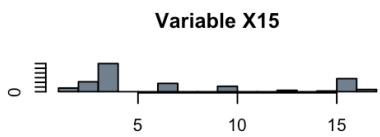
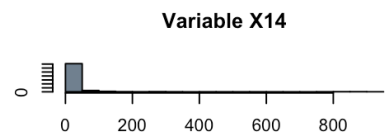
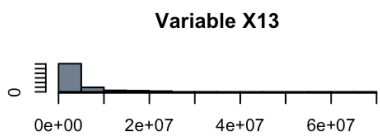
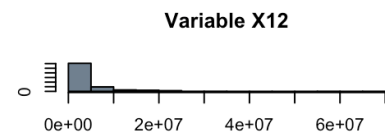
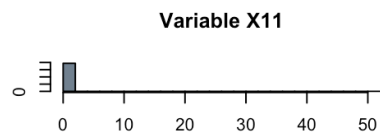
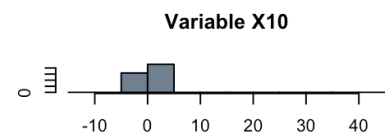
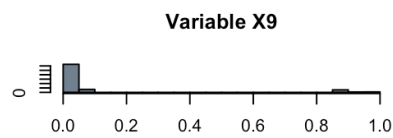
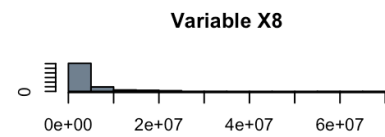
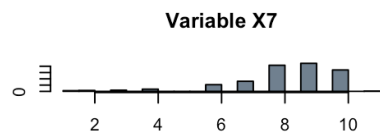
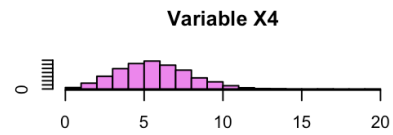
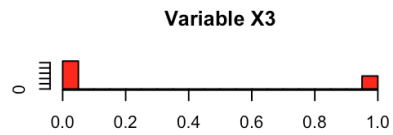
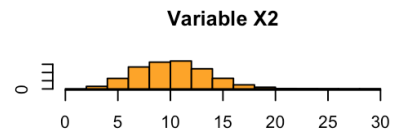
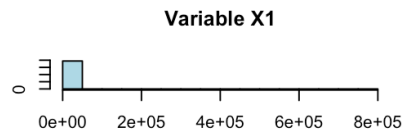
Console ~/Desktop/Test_job/
> str(base)
'data.frame': 77829 obs. of 16 variables:
 $ cible: Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 ...
 $ X1 : num 1015.2 248.9 179.9 719.9 14.3 ...
 $ X2 : int 18 8 10 10 6 10 5 7 7 7 ...
 $ X3 : int 1 0 0 0 1 0 0 1 0 0 ...
 $ X4 : num 9.251 4.493 4.896 5.837 0.937 ...
 $ X5 : num 1.26 1.11 1.48 1.4 1.11 ...
 $ X6 : num 351 413 1377 1264 75 ...
 $ X7 : int 8 9 7 10 8 9 10 8 8 10 ...
 $ X8 : int 290461 34228 244860 5401 781155 3321072 61460 335579 29933297 3042109 ...
 $ X9 : num 0 0 0 0 0 0.01 0 0 0 0.01 ...
 $ X10 : num -0.46 0.64 0.52 -0.46 0.61 0.57 -0.38 -0.75 0.71 0.7 ...
 $ X11 : num 1.11 1.19 1.17 1.16 1.17 1.15 1.1 1.15 1.19 1.2 ...
 $ X12 : num 289953 36047 245050 5427 781148 ...
 $ X13 : num 290461 37725 245364 6089 781155 ...
 $ X14 : int 1 4 39 2 47 33 0 1 38 3 ...
 $ X15 : int 10 16 16 7 3 7 4 4 16 16 ...
 - attr(*, "na.action")=Class 'omit' Named int [1:2171] 24 25 91 114 160 194 196 223 230 304

```

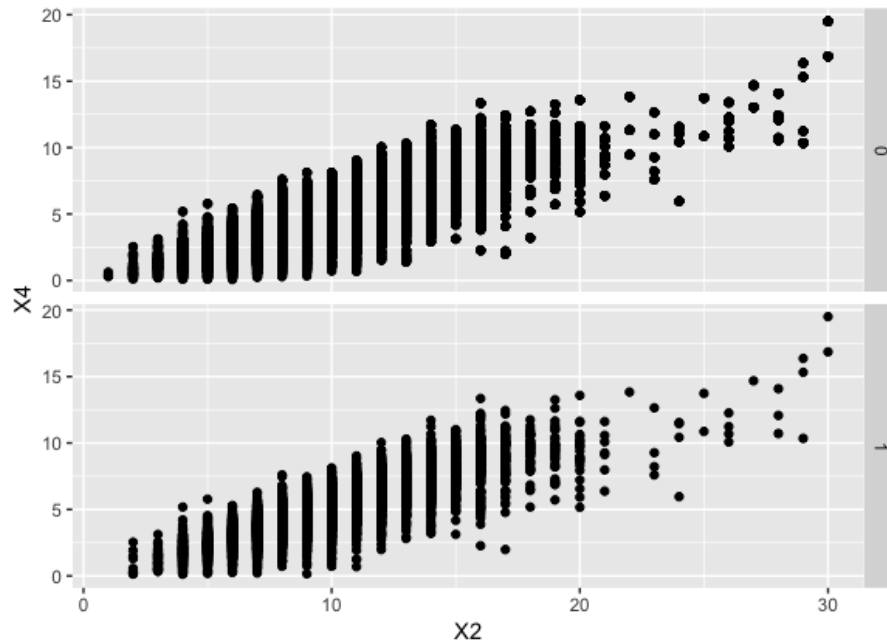
§ 5. RECHERCHE DES VARIABLES EXPLICATIVES PERTINENTES

5.1 Méthode par analyse graphique

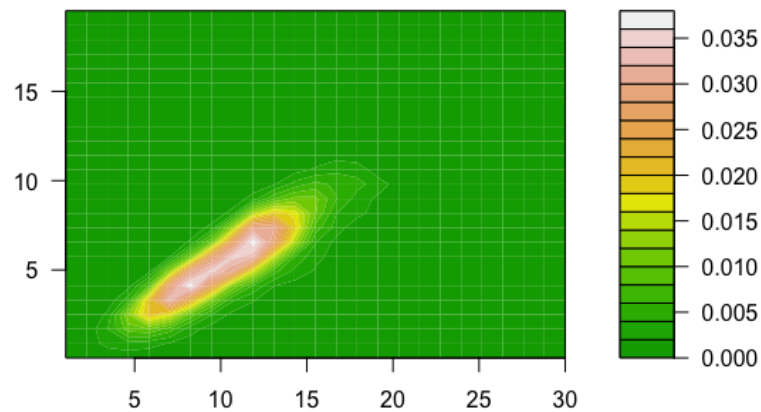
On représente graphiquement l'ensemble des variables (de X1 à X15) afin d'identifier des similitudes dans la distribution :



On constate alors que les deux variables X2 et X4 semblent être distribuées de la même façon.
D'autre part, on peut également utiliser un autre outil d'analyse qui est le nuage de points pour toutes les variables. On l'applique pour les deux variables X2 et X4 qui ont la même distribution.



D'autre part, la représentation de la densité de ces deux variables donne le graphique suivant :



Cette régression linéaire par lecture graphique, nous conduit donc à calculer le coefficient de corrélation entre ces deux variables X2 et X4 :


```
> cor(base$X2, base$X4)
[1] 0.8474524
```

Nous obtenons alors que le résultat est proche de 1 donc nous pouvons effectuer une régression linéaire complète.

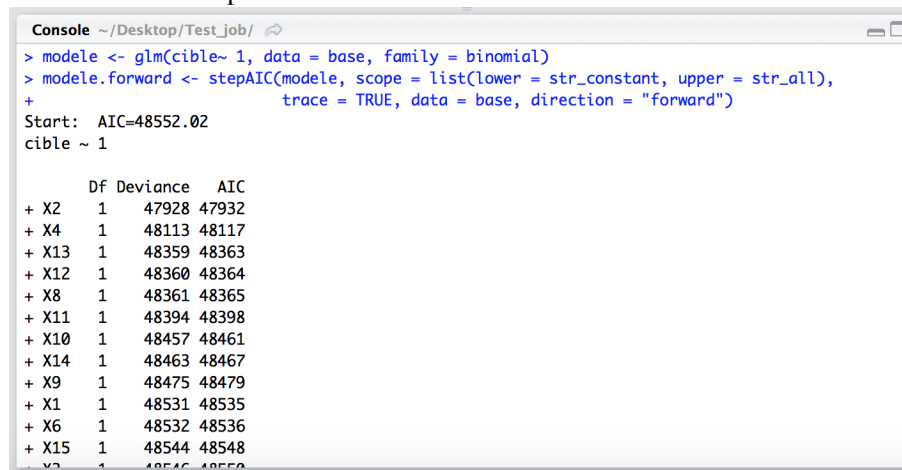
Dans cette étude, nous n'effectuons pas de discrétisation car nous n'avons aucune connaissance des variables qui restent inconnues.

§ 6. CONSTRUCTION DES MODÈLES

On souhaite modéliser la variable *cible*. Un modèle avec peu de variables sera plus facilement généralisable en terme de robustesse.

Pour sélectionner le « meilleur modèle », on va s'appuyer sur une mesure qui permet de comparer les modèles entre eux notamment par le critère d'information d'Akaike (AIC) :

- forward : on part du modèle avec uniquement une constante, et on ajoute les variables unes à unes jusqu'à ce que l'ajout d'une variable supplémentaire se solde par un modèle jugé moins bon en fonction du critère de comparaison sélectionné.



```
Console ~/Desktop/Test_job/
> modele <- glm(cible~ 1, data = base, family = binomial)
> modele.forward <- stepAIC(modele, scope = list(lower = str_constant, upper = str_all),
+                             trace = TRUE, data = base, direction = "forward")
Start: AIC=48552.02
cible ~ 1
```

	Df	Deviance	AIC
+ X2	1	47928	47932
+ X4	1	48113	48117
+ X13	1	48359	48363
+ X12	1	48360	48364
+ X8	1	48361	48365
+ X11	1	48394	48398
+ X10	1	48457	48461
+ X14	1	48463	48467
+ X9	1	48475	48479
+ X1	1	48531	48535
+ X6	1	48532	48536
+ X15	1	48544	48548

Et l'affichage du modèle final s'affiche ainsi :

```

Console ~/Desktop/Test_job/
> summary(modele.forward)

Call:
glm(formula = cible ~ X2 + X13 + X8 + X7 + X6 + X12 + X1 + X3 +
      X9 + X11 + X15 + X10 + X14 + X5 + X4, family = binomial,
      data = base)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.3826  -0.2627  -0.1289   8.4904

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.104e+01  2.201e-01 -50.176 < 2e-16 ***
X2           -1.065e-01  9.248e-03 -11.516 < 2e-16 ***
X13          1.917e-03  2.824e-05  67.886 < 2e-16 ***
X8           1.608e-05  1.293e-05   1.243  0.21387
X7           9.674e-01  2.034e-02  47.568 < 2e-16 ***
X6          -5.002e-04  1.377e-05 -36.330 < 2e-16 ***
X12          1.032e-03  2.712e-05  37.870 < 2e-16 ***
X1           7.105e-01  2.061e-02  34.454 < 2e-16 ***
X3           1.032e-03  2.712e-05  37.870 < 2e-16 ***
X9           1.032e-03  2.712e-05  37.870 < 2e-16 ***
X11          1.032e-03  2.712e-05  37.870 < 2e-16 ***
X15          1.032e-03  2.712e-05  37.870 < 2e-16 ***
X10          1.032e-03  2.712e-05  37.870 < 2e-16 ***
X14          1.032e-03  2.712e-05  37.870 < 2e-16 ***
X5           1.032e-03  2.712e-05  37.870 < 2e-16 ***
X4           1.032e-03  2.712e-05  37.870 < 2e-16 ***

```

- backward : L'opération consiste donc à partir du modèle complet, de retirer une variable et de voir quel retrait entraîne la plus forte baisse de l'AIC. Si le retrait d'une variable ne se solde pas par une diminution de l'AIC, on s'arrête. Sinon, on recommence le processus de retrait.
- stepwise : un mélange des méthodes forward et backward, basée sur le F-partiel. On vérifie que l'ajout d'une variable ne provoque pas la suppression d'une variable déjà introduite

```

Console ~/Desktop/Test_job/
> summary(modele.stepwise)

Call:
glm(formula = cible ~ X2 + X13 + X7 + X6 + X12 + X1 + X3 + X9 +
      X11 + X15 + X10 + X14 + X5 + X4, family = binomial, data = base)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.3827  -0.2625  -0.1289   8.4904

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.108e+01  2.183e-01 -50.743 < 2e-16 ***
X2           -1.062e-01  9.244e-03 -11.493 < 2e-16 ***
X13          1.897e-03  2.327e-05  81.546 < 2e-16 ***
X7           9.702e-01  2.016e-02  48.119 < 2e-16 ***
X6          -4.942e-04  1.243e-05 -39.742 < 2e-16 ***
X12          -1.897e-03  2.327e-05 -81.545 < 2e-16 ***
X1          -9.618e-04  1.435e-05 -67.017 < 2e-16 ***
X3           7.105e-01  2.061e-02  34.454 < 2e-16 ***
X9           1.032e-03  2.712e-05  37.870 < 2e-16 ***
X11          1.032e-03  2.712e-05  37.870 < 2e-16 ***
X15          1.032e-03  2.712e-05  37.870 < 2e-16 ***
X10          1.032e-03  2.712e-05  37.870 < 2e-16 ***
X14          1.032e-03  2.712e-05  37.870 < 2e-16 ***
X5           1.032e-03  2.712e-05  37.870 < 2e-16 ***
X4           1.032e-03  2.712e-05  37.870 < 2e-16 ***

```

- L'algorithme sélectionne au final le meilleur modèle suivant :
- ```
glm(formula = cible ~ X2 + X13 + X7 + X6 + X12 + X1 + X3 + X9 + X11 + X15 + X10 + X14 +
X5 + X4, family = binomial, data = base)
```
- L'algorithme retient tous les variables à l'exception de X8.

## § 7. MODÉLISATION

Notre fonction pour réaliser une régression logistique sous R :

```

Console ~/Desktop/Test_job/ ↗
> summary(m.logit)

Call:
glm(formula = formula, family = binomial(link = lien), data = data)

Deviance Residuals:
 Min 1Q Median 3Q Max
-8.4904 -0.3826 -0.2627 -0.1289 8.4904

Coefficients:
 Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.104e+01 2.201e-01 -50.176 < 2e-16 ***
X1 -9.804e-04 2.073e-05 -47.300 < 2e-16 ***
X2 -1.065e-01 9.248e-03 -11.516 < 2e-16 ***
X3 7.161e-01 3.870e-02 18.504 < 2e-16 ***
X4 -1.914e-02 1.378e-02 -1.389 0.16471
X5 2.250e-02 9.310e-03 2.417 0.01564 *
X6 -5.002e-04 1.377e-05 -36.330 < 2e-16 ***
X7 9.674e-01 2.034e-02 47.568 < 2e-16 ***
X8 1.608e-05 1.293e-05 1.243 0.21387
X9 4.517e-01 5.488e-02 8.232 < 2e-16 ***
X10 1.085e-01 2.512e-02 4.317 1.58e-05 ***
X11 4.184e-01 8.526e-02 4.907 9.24e-07 ***
X12 -1.933e-03 3.712e-05 -52.079 < 2e-16 ***
X13 1.917e-03 2.824e-05 67.886 < 2e-16 ***
X14 6.638e-04 1.498e-04 4.430 9.40e-06 ***
X15 -8.591e-03 2.883e-03 -2.980 0.00288 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

 Null deviance: 48550 on 77828 degrees of freedom
Residual deviance: 30088 on 77813 degrees of freedom
AIC: 30120

Number of Fisher Scoring iterations: 16

```