

# PROJET MACHINE LEARNING

Marlène CHEVALIER et Olga SILVA

Juin 2020

## Sujet d'étude : succès d'une campagne marketing

L'analyse porte sur une campagne marketing conduite auprès d'un ensemble de clients résidant en France. Sur chaque client, le résultat de la campagne est mesurée par sa "réussite (success)" ou son "échec (failure)". L'objectif du projet consiste à prédire le succès de la campagne marketing en fonction des caractéristiques des clients.

### Jeu de données

Pour cela nous disposons de 2 fichiers, qui pour chaque ligne d'un fichier décrit un client.

- le fichier d'apprentissage/d'entraînement (10 000 clients et 12 variables) : *projet-app-13-learn.csv*
- le fichier test (10 000 clients et 11 variables) : *projet-app-13-test.csv*

Les variables caractérisant chaque client sont les suivantes :

- **age** : âge du client;
- **sex** : genre du client;
- **f\_name** : prénom du client ;
- **last name** : nom du client;
- **commune** : nom de la commune de résidence du client ; (Lyon, Paris et Marseille sont découpées en arrondissements et que le nom de commune est alors celui de la ville suivi de l'indication d'arrondissement)
- **insee code** : code insee de la commune de résidence ;
- **city type** : type de la commune de résidence ;
- **department** : numéro du département de résidence ;
- **reg** : code de la région de résidence ;
- **catégorie** : code de catégorie socio-professionnelle du client
- **revenu** : salaire mensuel en équivalent temps plein
- **cible** : résultat de la campagne codé par "success" pour un résultat considéré comme positif et "failure" dans le cas contraire. Cette variable est absente du fichier test.

## **Données annexes**

Des informations annexes peuvent être utilisées pour cette analyse :

- la *table des catégories socio-professionnelles* associe le code et le nom de chaque catégorie ;
- le fichier *cities-gps.csv* contient les coordonnées géographiques des villes de métropole, identifiées par leur code insee ;
- le fichier *cities-population.csv* contient la population de chaque ville de métropole ;
- le fichier *departments.csv* contient l'association entre le numéro de département et son nom, ainsi que la région dans laquelle chaque département est situé ;
- le fichier *regions.csv* contient l'association entre le numéro de région et son nom ;

## **Contenu du Rapport (A garder pour mémoire - et supprimer avant transmission du projet )**

1. Analyse exploratoire minimale des données (statistiques univariées, dépendances, etc.) ;
2. Justification du modèles prédictif choisi ;
3. Description précise de la chaîne de traitement : prétraitements éventuels, ajustement des modèles, choix du modèle, évaluation de ses performances attendues (le rapport doit impérativement contenir un tableau indiquant la qualité numérique attendue pour les prévisions sur le fichier test) ;
4. Analyse de l'importance des variables : cela peut être fait avant l'ajustement des modèles, pendant celui-ci ou après le choix du modèle final. Dans tous les cas, le rapport doit discuter de l'opportunité de construire des modèles sur une partie seulement des variables. Si c'est le cas, les prévisions finales et les performances attendues doivent concerner les modèles n'utilisant que les variables pertinentes ;
5. Interprétation du modèle retenu : si cela est possible, une interprétation de la façon dont les décisions du modèle retenu sont prises fournira un complément très important au reste de l'analyse.

## **Préparation des données**

### **Chargement des données et préparation des variables**

Retraitements nécessaires :

- Nous allons renommer la variable catégorie, pour enlever l'accent : *catégorie* devient *category*
- Certaines variables numériques sont à interpréter comme des codes, nous les convertissons en facteur : *reg* et *category*.
- Nous convertissons la variable cible en variable logique : “failure” devient FALSE et “success” devient TRUE

### **Compléter avec les données géographiques**

Ajoutons les coordonnées géographiques et la population aux jeux de données d'entraînement et de test.

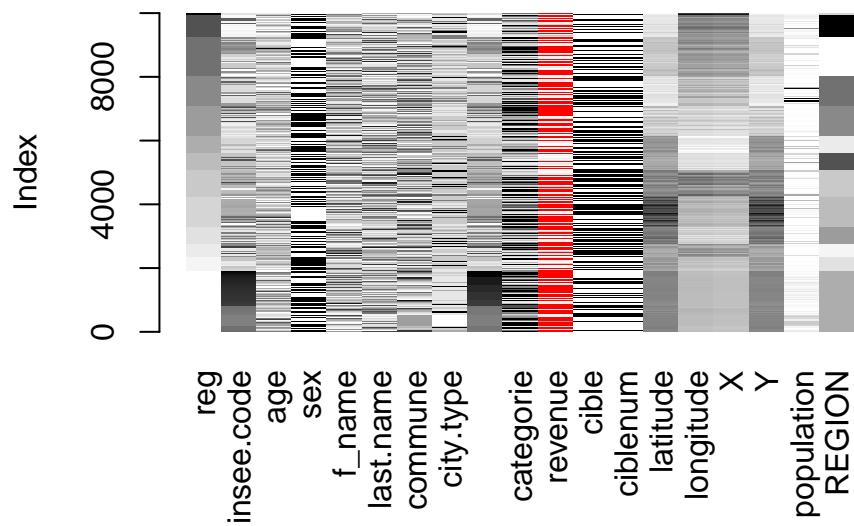
## Traitemet des données manquantes

### Recensement des données manquantes

### Jeu d'apprentisage

Dans le jeu d'apprentissage, seule la variable "revenue" est manquante pour 5320 valeurs, soit 53% de valeurs manquantes.

Le graphique de suivant confirme cette observation :



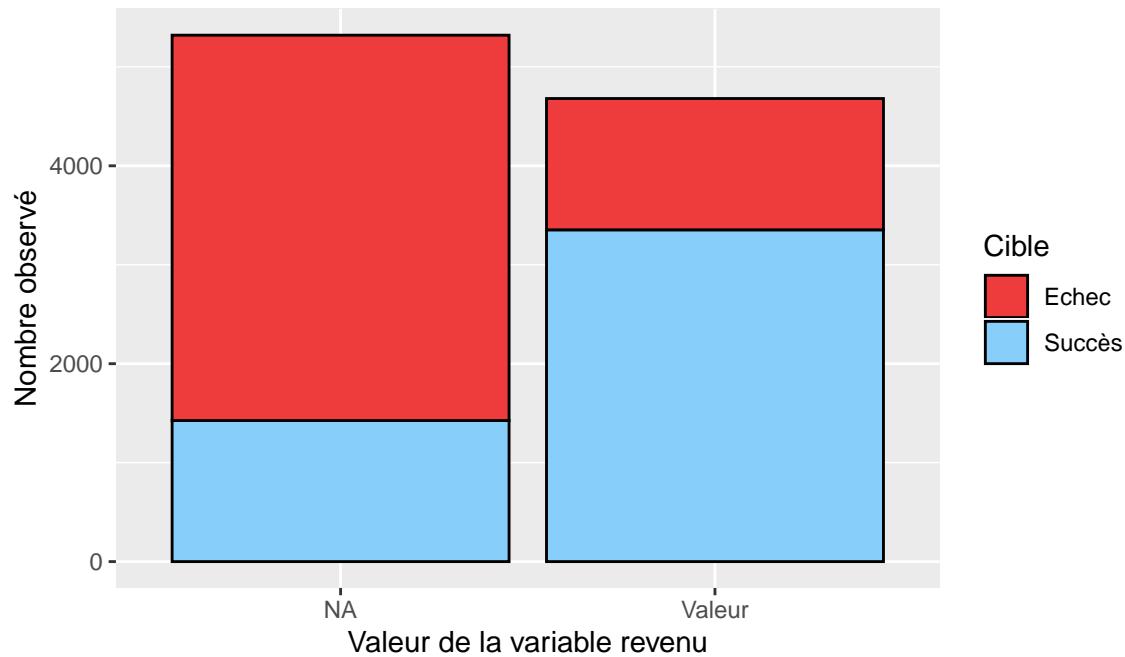
### Jeu de test

Dans le jeu test, seule la variable "revenue" est manquante pour 5175 valeurs, soit 52% de valeurs manquantes.

### Résultat de campagne et les revenus manquants

Regardons s'il y a un lien entre le résultat de la campagne et les revenus manquants.

## Répartition du succès/échec de la cible en fonction de la disponibilité de la donnée revenu



La proportion d'échec de la campagne est plus forte lorsque la variable "revenu" n'est pas renseignée que lorsque les personnes ont indiqué la valeur de leur revenu.

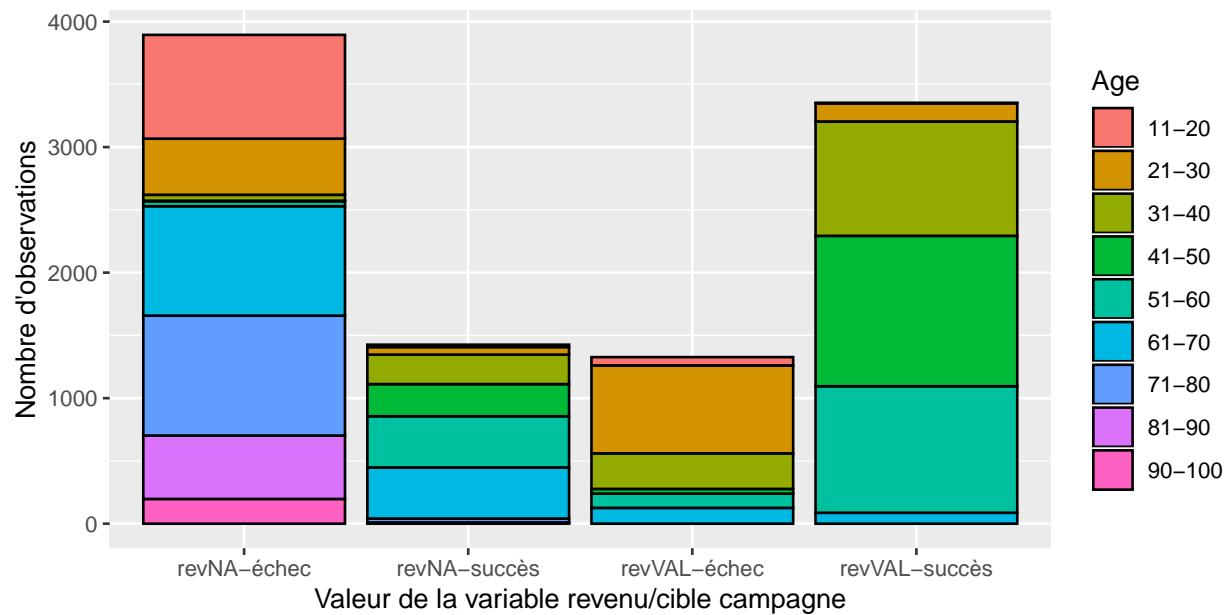
### Identifier la cause des revenus manquants et l'influence sur la campagne

Composons une variable qui dépend de la disponibilité (prend une valeur) ou non (NA) de la variable revenu, et du succès ou de l'échec de la campagne (variable cible). Cette nouvelle variable *revmv.cible* prend donc valeurs :

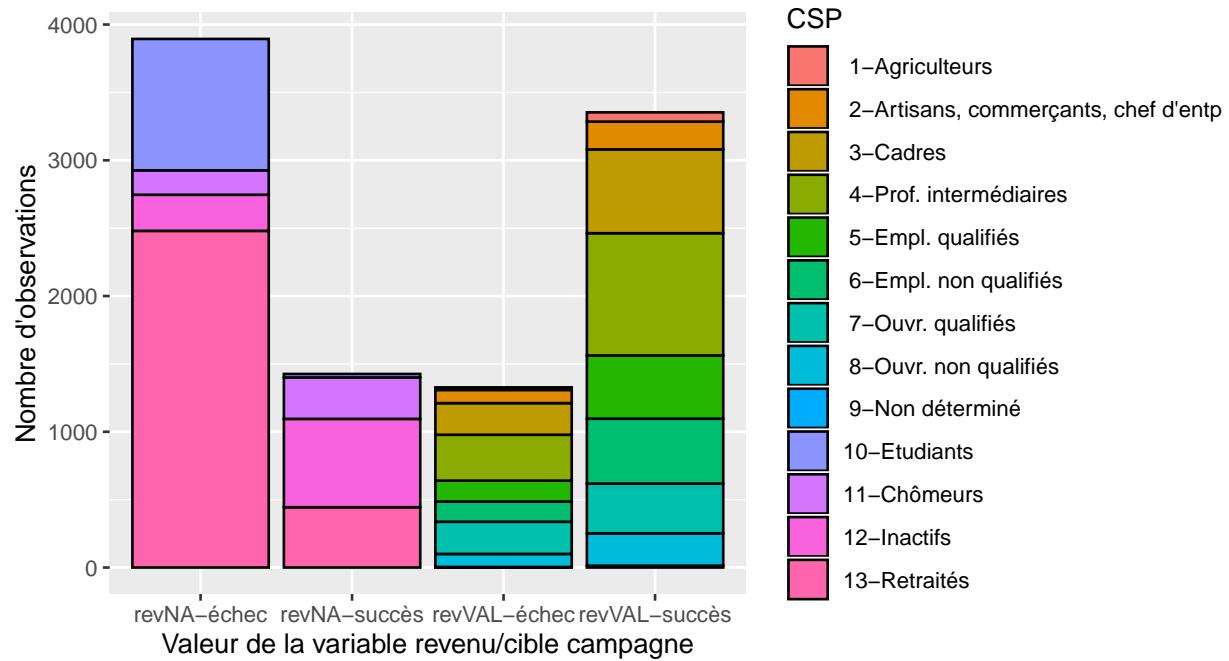
Variable Revenu	Cible = Succès	Cible = Echec
Valeur numérique	revVAL-succès	revVAL-échec
NA	revNA-succès	revNA-échec

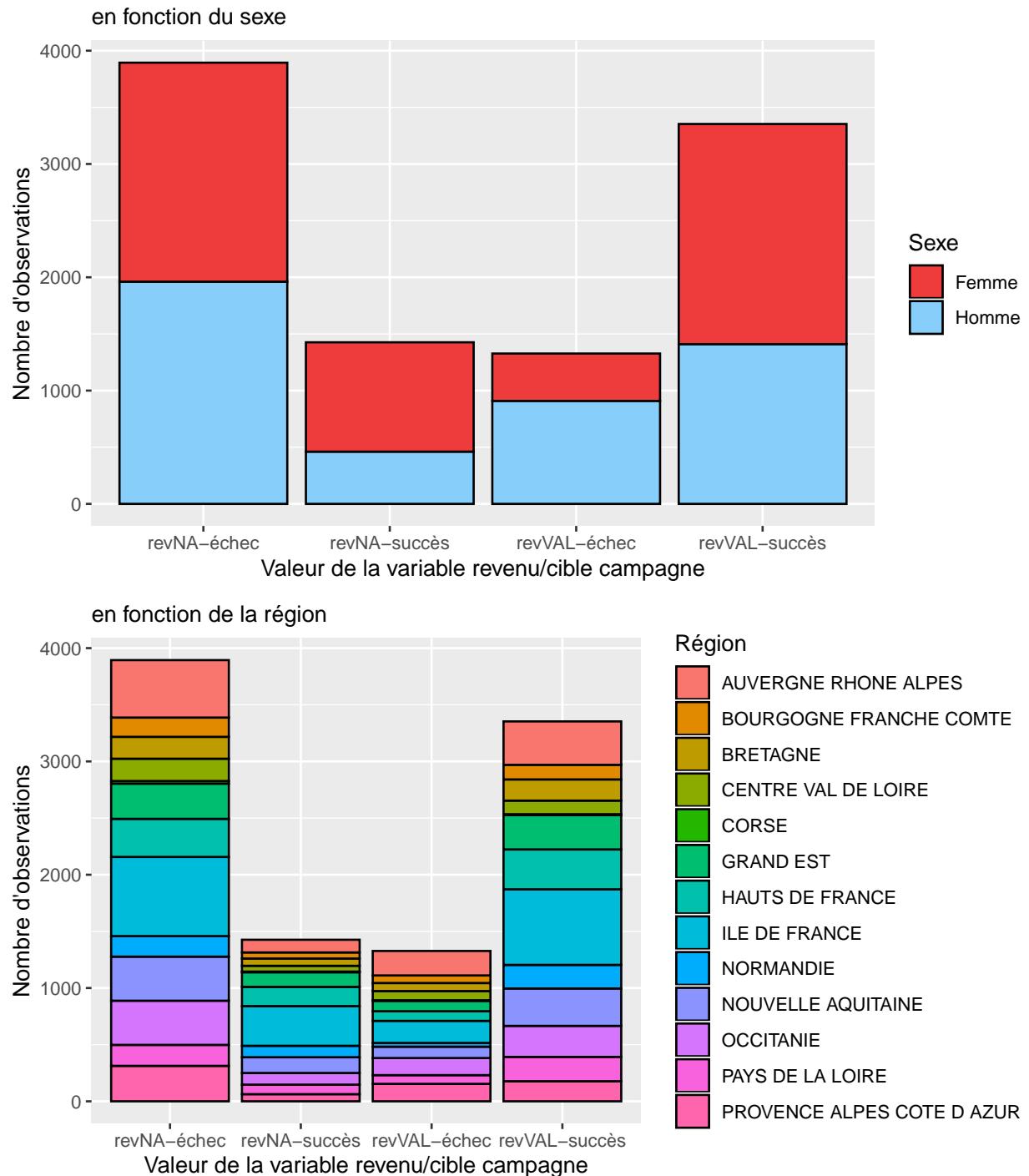
Cherchons s'il se dégage une influence marquée, entre cette variable composée et une des covariables de notre jeu de donnée : l'age, la catégorie socio-professionnelle, le genre, la région ou le type de ville.

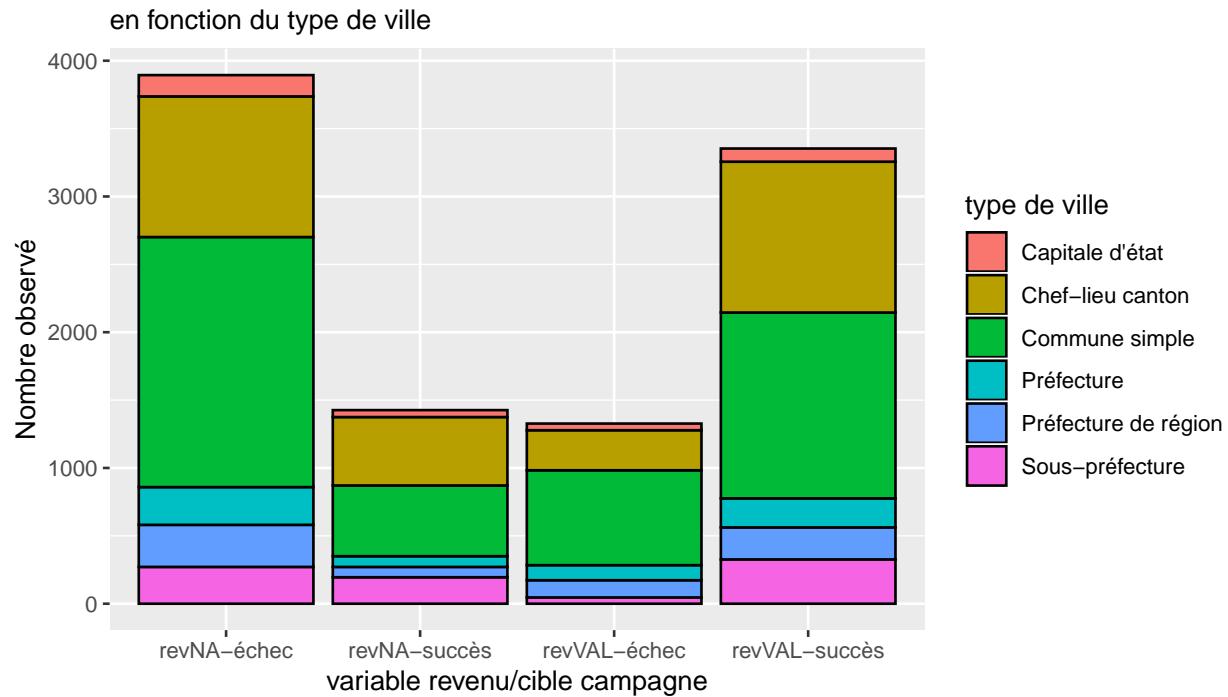
Répartition du succès/échec et de la disponibilité de revenu  
en fonction de l'age



en fonction de la CSP







Il apparaît que les valeurs du genre, de la région ou du type de ville sont représentées indifféremment que le revenu soit manquant ou non.

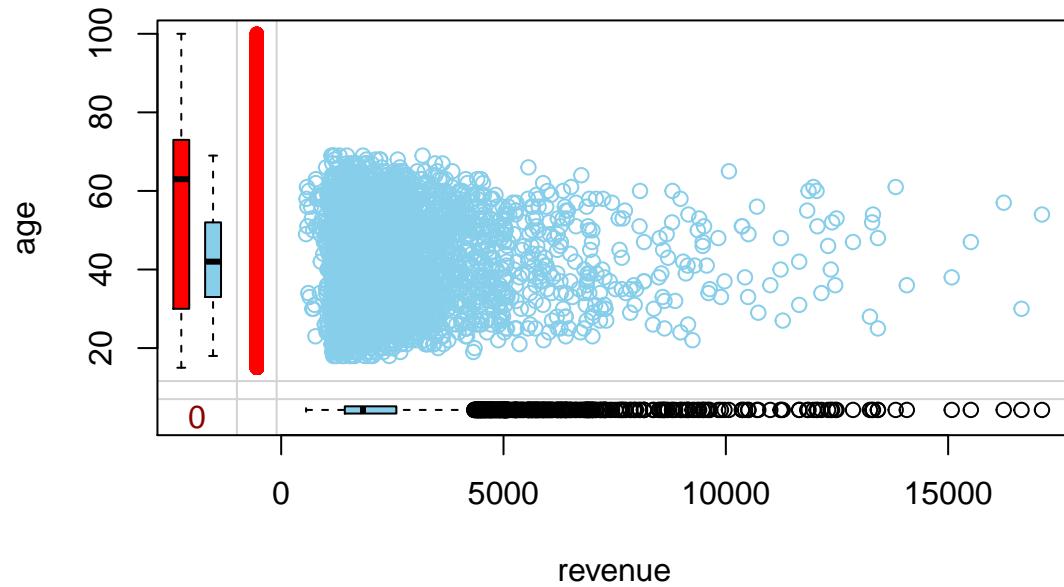
L'âge présente une particularité : en effet les personnes de plus de 70 ans et à forte majorité celles de moins de 20 ans n'ont pas indiqué leur revenu et ne sont pas la cible de la campagne.

C'est la catégorie socio-professionnelle qui est corrélée à l'absence d'indication de revenu : en effet, il apparaît que les étudiants, les chômeurs, les inactifs et les retraités soient les seuls à n'avoir pas renseigné leur revenu. **Les valeurs manquantes du revenu viennent donc de 4 CSP : 10 Etudiants - 11 Chômeurs - 12 Inactifs et 13 Retraités.**

— Annexe —

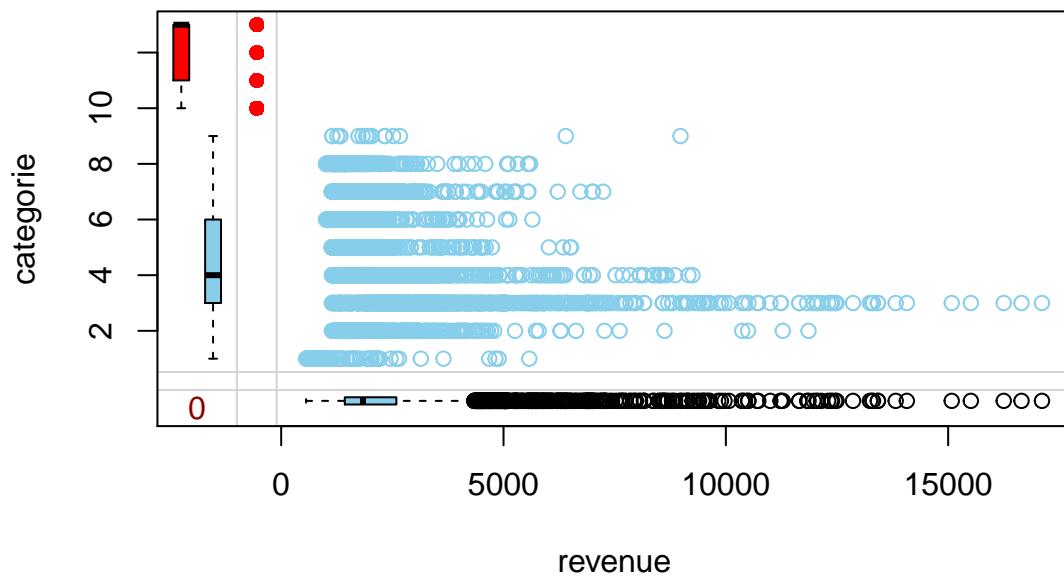
Afin d'avoir plus d'information sur les données manquantes, nous allons utiliser le graphique “Marginplot” et observé chaque variable avec la variable *revenue*. Dans ce graphique, les points sans valeurs manquantes apparaissent dans le scatterplot. On observe qu'il manque des valeurs pour tous les âges (ligne rouge verticale, de 15 à 100 ans). Nous avons également les boxplots des distributions des valeurs : en rouge quand la valeur est manquante et en bleu quand la valeur est observée.

#### Le revenu et l'âge



Les valeurs des revenus sont observées uniquement sur la tranche d'âge des 18 ans-69 ans, qui correspond à la période d'activité professionnelle.

#### Le revenu et la catégorie socio-professionnelle

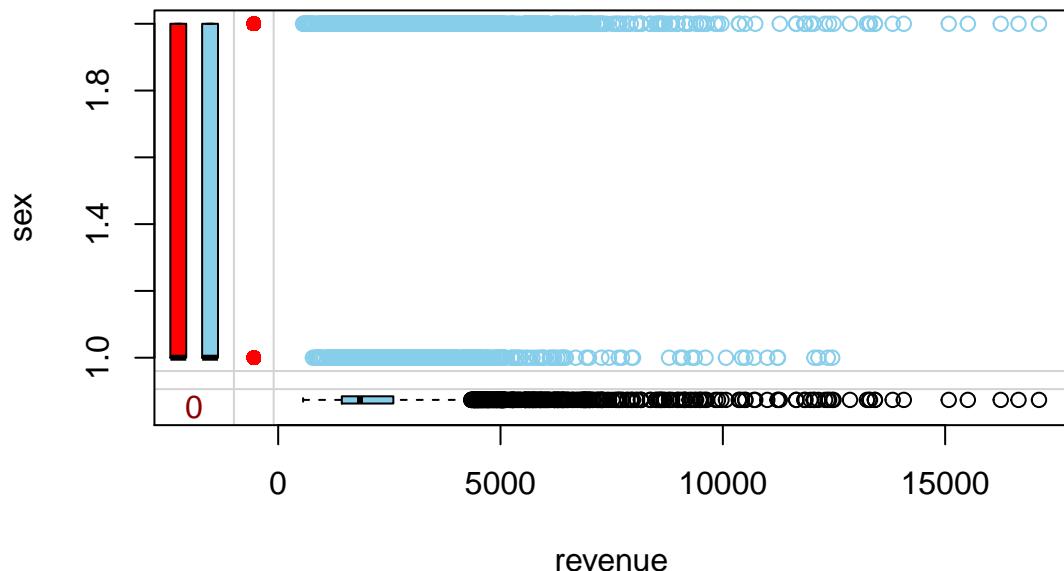


Selon le graphique, il y a que trois catégories sans revenu (10, 11 et 12) : chômeurs, inactifs et retraités. Sur

le reste des catégories, nous ne trouverons pas des données absentes.

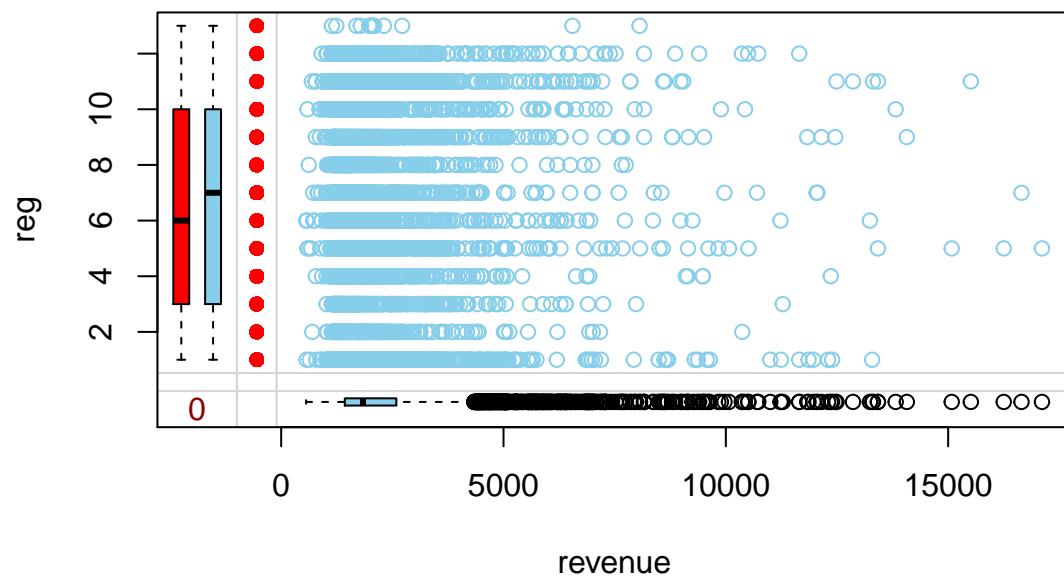
Dans la distribution on note que des personnes aux revenus se trouvent dans la catégorie cadres.

#### Le revenu et le genre



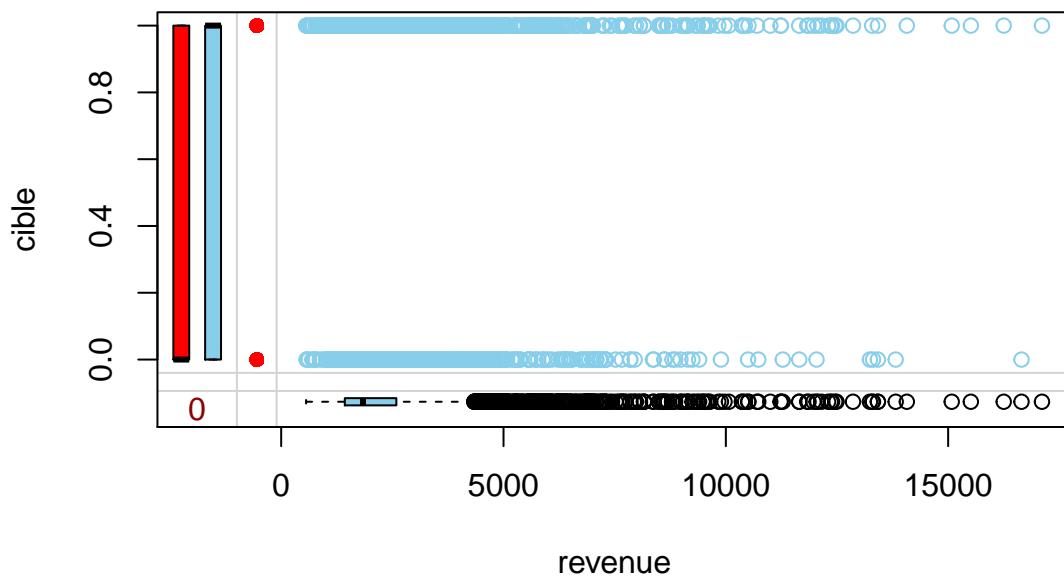
En ce qui concerne le genre, nous avons des données absentes pour les deux catégories, cependant les hommes ont les salaires le plus élevés de l'échantillon.

#### Le revenu et la région



Pour la région aussi, il semblerait aussi que nous avons des données manquantes pour toutes les régions.

#### Le revenu et le resultat de la campagne



```
## [1] 0
```

```
## [1] 0
```

‘ Pour la cible, on observe qu'il y a des données manquantes pour les deux catégories : Failure et success. Mais leur absence n'est pas équilibrée, on retrouve plus des données absentes pour l'échec que pour le succès.

— fin annexe—

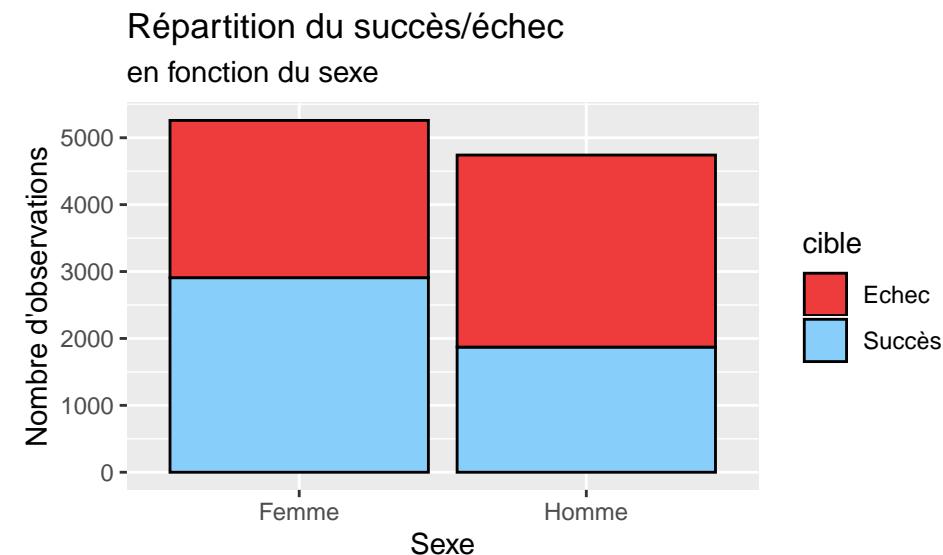
## Analyse exploratoire des données

Regardons maintenant graphiquement les **principales particularités du jeu de données d'apprentissage**.

Sur l'ensemble de la population d'apprentissage :

- la majorité des individus, soit 5221, sont une **cible “échec” de la campagne, soit 52% de cette population.\***
- 4779 individus sont associés à une **campagne réussie, soit 48% de cette population.**

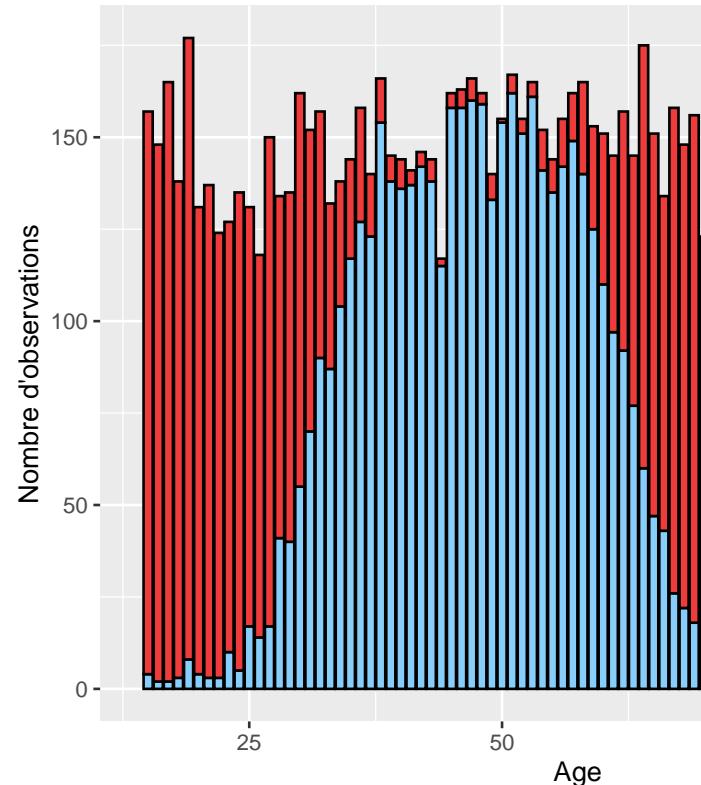
##Résultat de la campagne selon le sexe des personnes interrogées



L'ensemble d'apprentissage comporte 5259 femmes (53%) et 4741 hommes (47%), soit 11% de femmes de plus que les hommes; ce qui est en excès par rapport à la population globale française qui comprend 6% de plus de femmes seulement.

Le taux de réussite de la campagne chez les femmes de 55 % est plus important que chez les hommes 39 %. *Les femmes sont donc majoritairement en réussite sur la campagne alors que les hommes en échec sur la cible de campagne, soit 61%.*

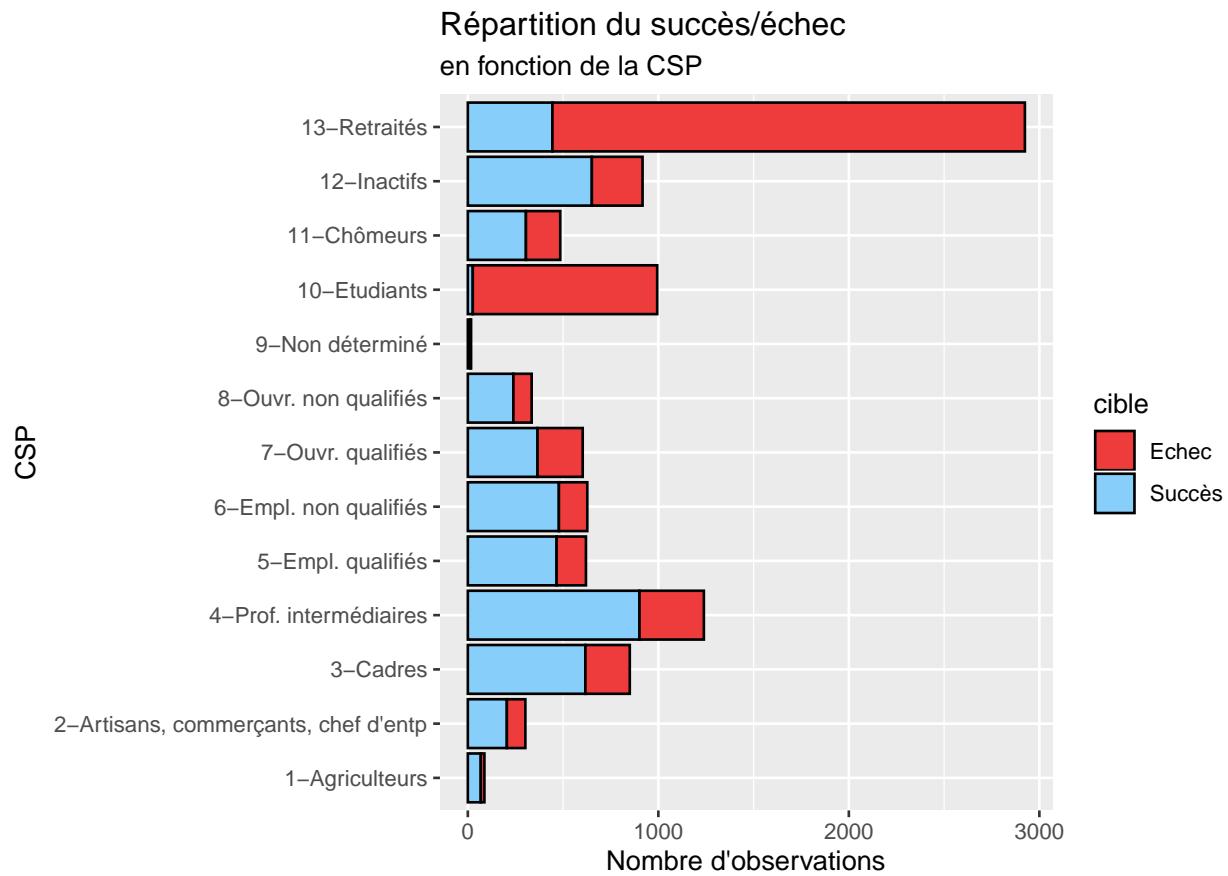
## Répartition du succès/échec en fonction de l'âge



##Résultat de la campagne selon l'âge des personnes interrogées

Il apparaît que les populations les plus jeunes (avant 25 ans) et les plus âgées (après 74 ans) sont quasi-totalement en échec sur la campagne. La courbe du nombre d'observations de réussite de la campagne forme une cloche entre 25 ans et 74 ans : pour les 35-55 ans, la réussite de la campagne observée est quasi-totale.

##Résultat de la campagne selon la catégorie socio-professionnelle des personnes interrogées

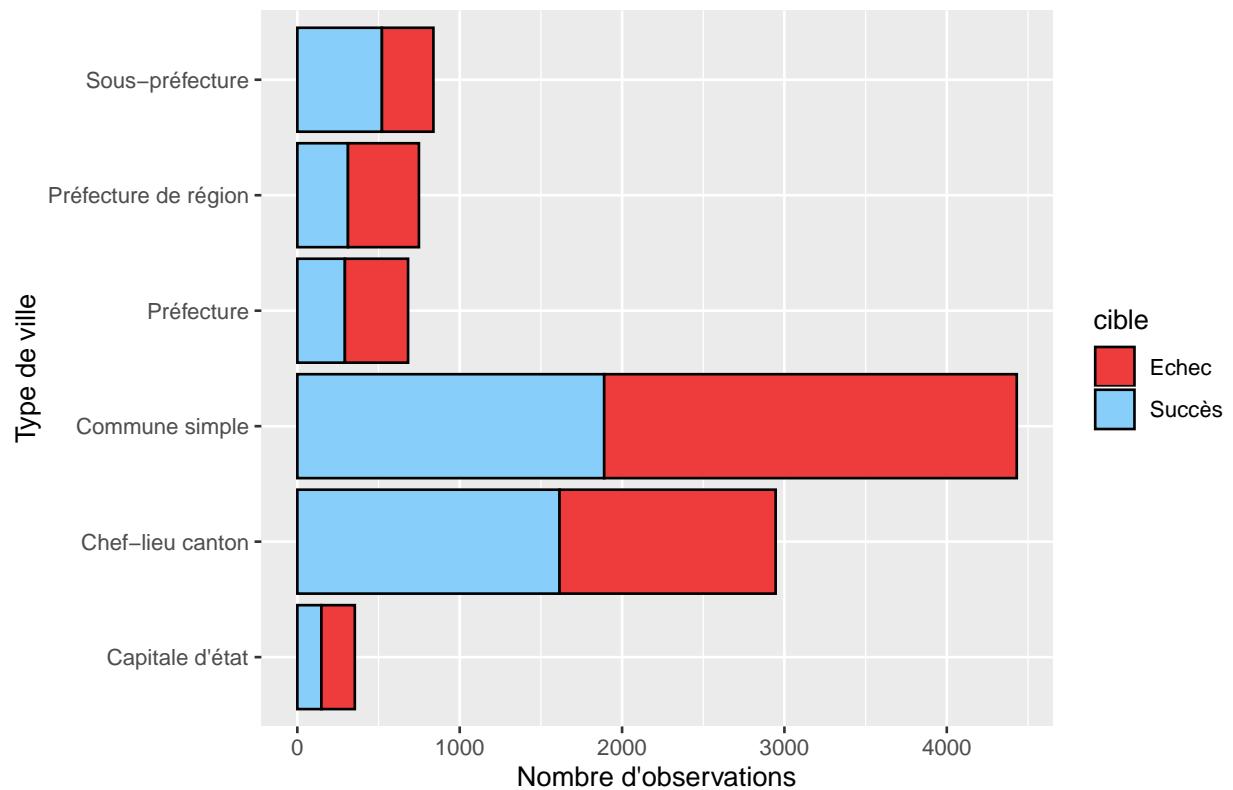


Les catégories socio-professionnelles les plus représentées dans cette étude sont les retraités (à presque 30%), puis les professions intermédiaires (12%), les étudiants(10%) et inactifs (9%) ; les plus faiblement représentés sont les agriculteurs (1,5%).

Les catégories socio-professionnelles présentent toutes une majorité de succès, sauf les étudiants et les retraités, ce qui se rapproche des résultats de la campagne par age (ci-dessous). Il apparaît que les retraités et les étudiants, les 2 CSP les plus fortement représentées, fassent basculer le résultat global de la campagne marketing vers l'échec.

##Résultat de la campagne selon le type de ville

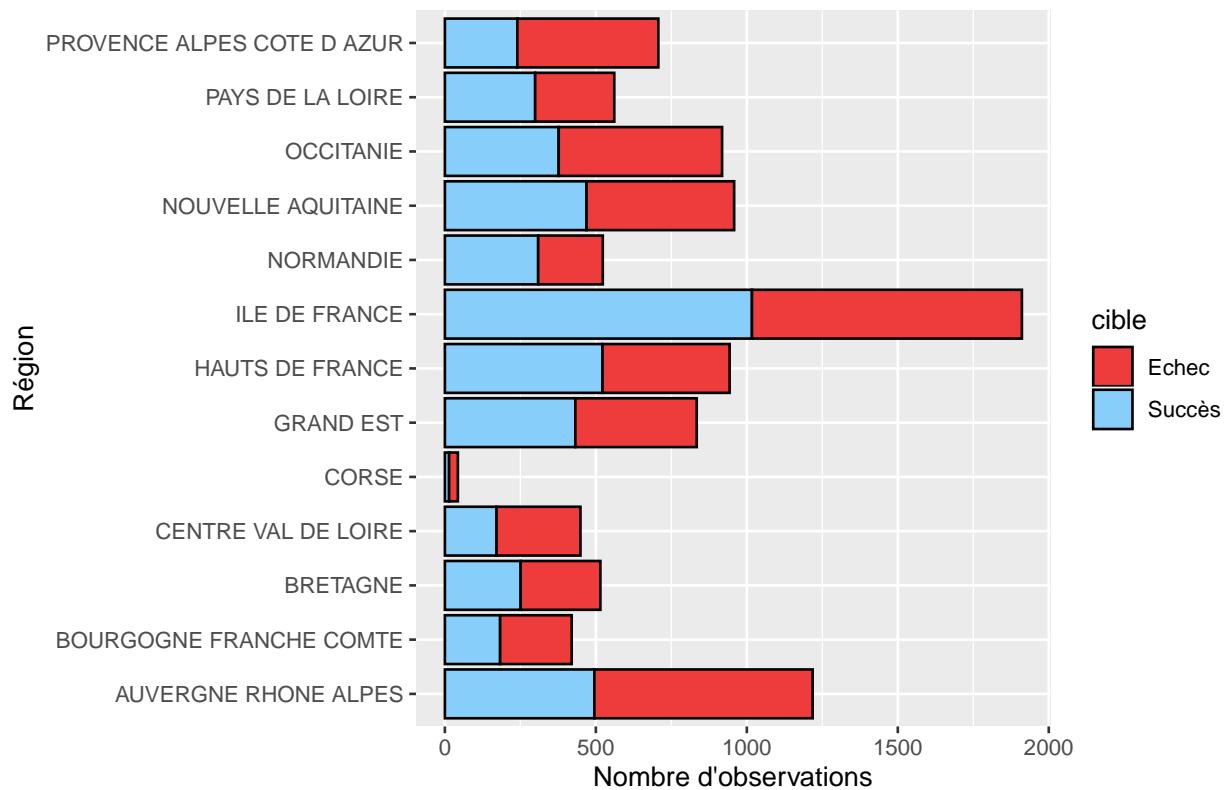
### Répartition du succès/échec en fonction du type de ville



Le type de villes principalement interrogées sont les communes simples et les chefs lieu de canton (73% des réponses). Les échecs de la campagne sont majoritaires dans tous les types de ville sauf 2: les sous-préfectures et les chefs-lieux de canton.

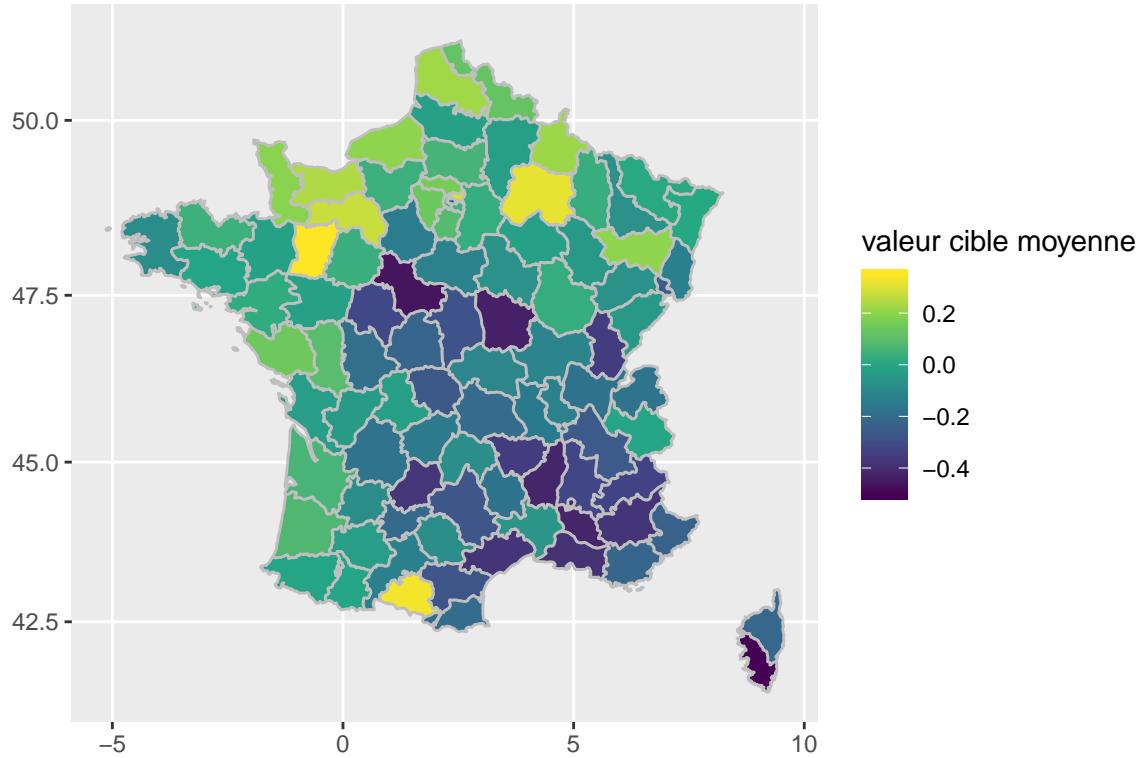
##Résultat de la campagne selon les régions

### Répartition du succès/échec en fonction des régions



Les régions dans lesquelles le plus de personnes ont été interrogées sont l'Ile de France (où le succès l'emporte) et l'Auvergne Rhône Alpes (où l'échec l'emporte). Les régions individuellement sont proches du résultat national (52% échec contre 48% de succès), sauf en Normandie, Ile de France, Hauts de France, régions pour lesquelles le succès est majoritaire.

## Succes/Echecs de la campagne par département



Sur cette carte, on a attribué la valeur “1” aux individus pour lesquels la campagne est un succès et la valeur “-1” en cas d’échec. La valeur cible moyenne est la moyenne de ces valeurs pour chaque département.

On retrouve sur la carte les zones de succès (valeur moyenne cible >0) au Nord et à l’Ouest ; et du centre au quart sud Est, les régions enregistrent majoritairement des échecs.

(3 départements à fort succès : Mayenne, Ariège et Marne / 3 départements à fort échec : Corse du Sud, Loire et Cher, Nièvre).

##Recapitulatif de l’analyse exploratoire : caractéristiques des résultats de la campagne

Succès 48% : sont plutôt	Echec 52% : sont plutôt
des femmes	des hommes
age 35-55 ans	- de 25 ans ou + de 70 ans
professions intermédiaires, inactifs, cadres	étudiants, retraités
résidents du quart Nord-Ouest	résidents du Centre au quart Sud-Est

### analyse Olga

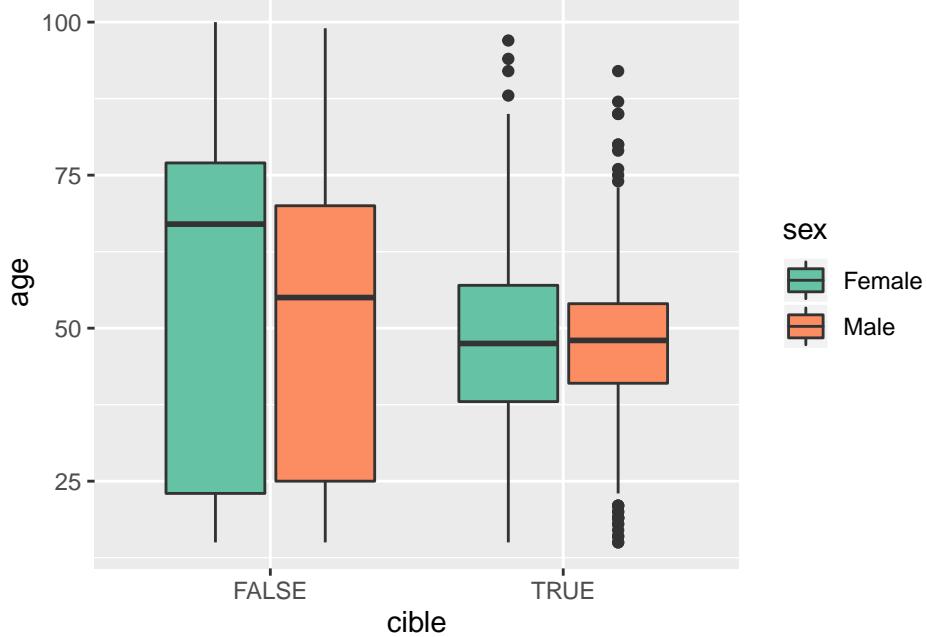
Les résultats semblent bien équilibrés entre les deux catégories à prédire : Echec et Succès

Nous observons une distribution très différente selon l’âge : l’échec est bien répartie entre 25 et 75 ans, par contre le succès est concentré entre 30 et 55 ans, avec quelques valeurs extrêmes supérieures à 75 ans. L’âge est sûrement une variable qui sera pertinente pour la prédiction.

Pour la population et les revenus, il ne semble pas y avoir des différences.

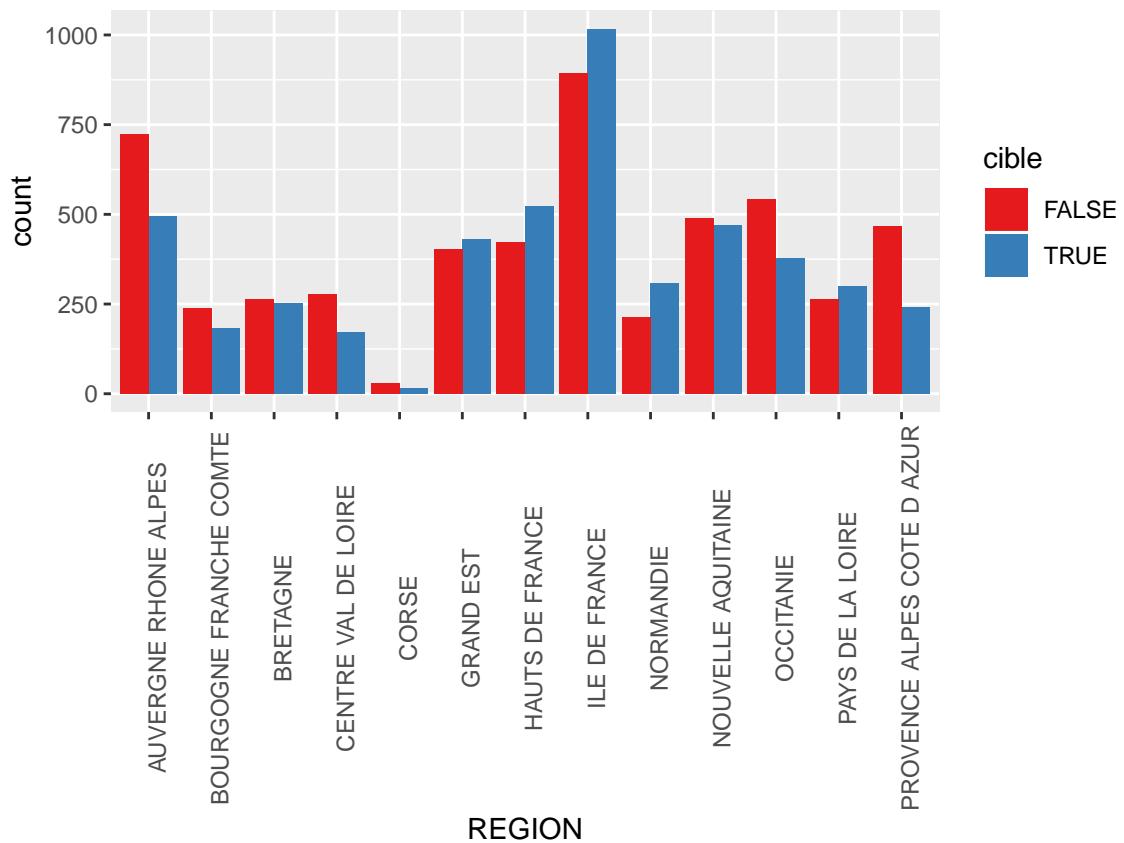
Les hommes ont un taux d'échec (60%) plus élevé que les femmes (45%). Pour les âges, la mediane de l'échec des femmes est autour de 70 ans et pour les hommes 50 ans.

L'échantillon est composé d'un plus grand nombre de femmes que d'hommes (5259 vs 4741)



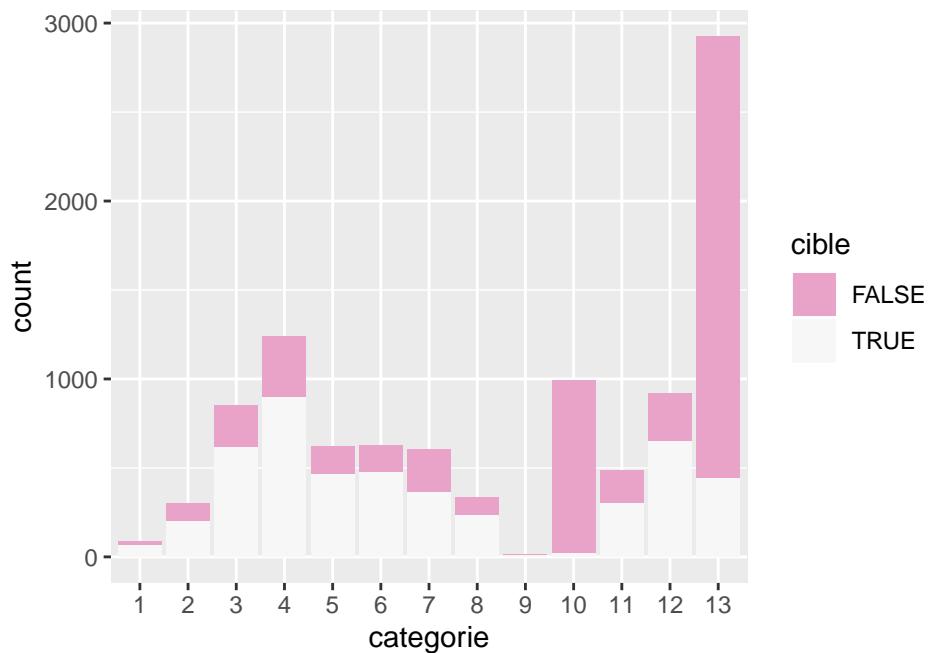
```
##  
##          FALSE  TRUE  
##  Female  2351 2908  
##  Male    2870 1871
```

La région Ile de France cumule une bonne partie des résultats (concentration de succès), mais en dehors ils sont bien équilibrés. La région Corse a une participation très basse.



```
##  
##  
##          FALSE   TRUE  
##  AUVERGNE RHONE ALPES      723   495  
##  BOURGOGNE FRANCHE COMTE  237   183  
##  BRETAGNE                  264   251  
##  CENTRE VAL DE LOIRE     278   171  
##  CORSE                     29    14  
##  GRAND EST                 402   432  
##  HAUTS DE FRANCE           421   522  
##  ILE DE FRANCE              894  1017  
##  NORMANDIE                 214   309  
##  NOUVELLE AQUITAINE        489   469  
##  OCCITANIE                  541   377  
##  PAYS DE LA LOIRE            262   299  
##  PROVENCE ALPES COTE D AZUR 467   240
```

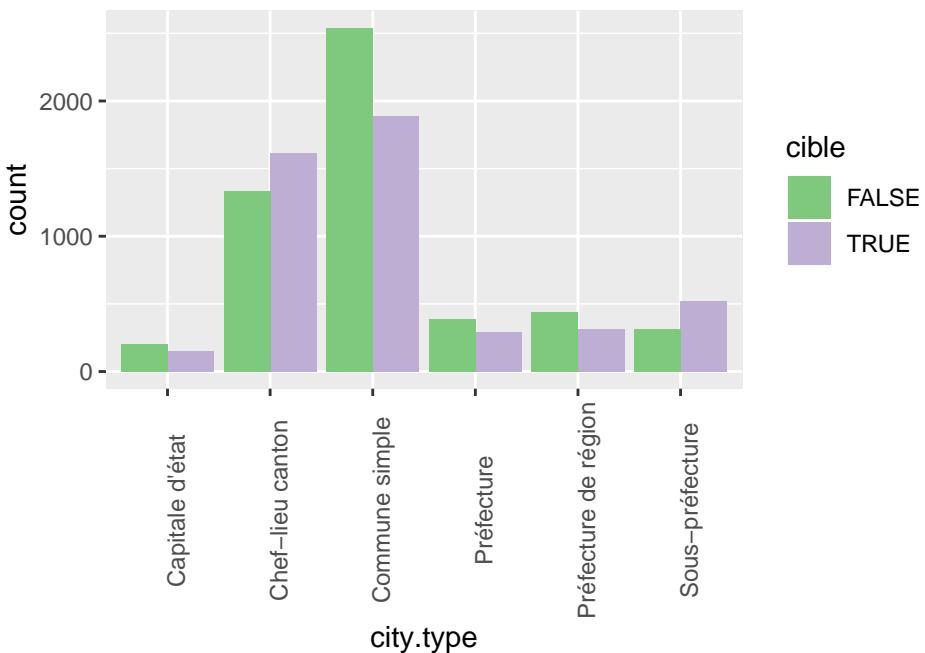
Les retraités et les étudiants (catégorie 13 et 10), enregistrent la plupart des échecs, tandis que les catégories 2 à 8, 11 et 12 (chômeurs et inactifs) ont la plupart des succès, sûrement s'agit-il d'une des variables à garder pour la prédiction.



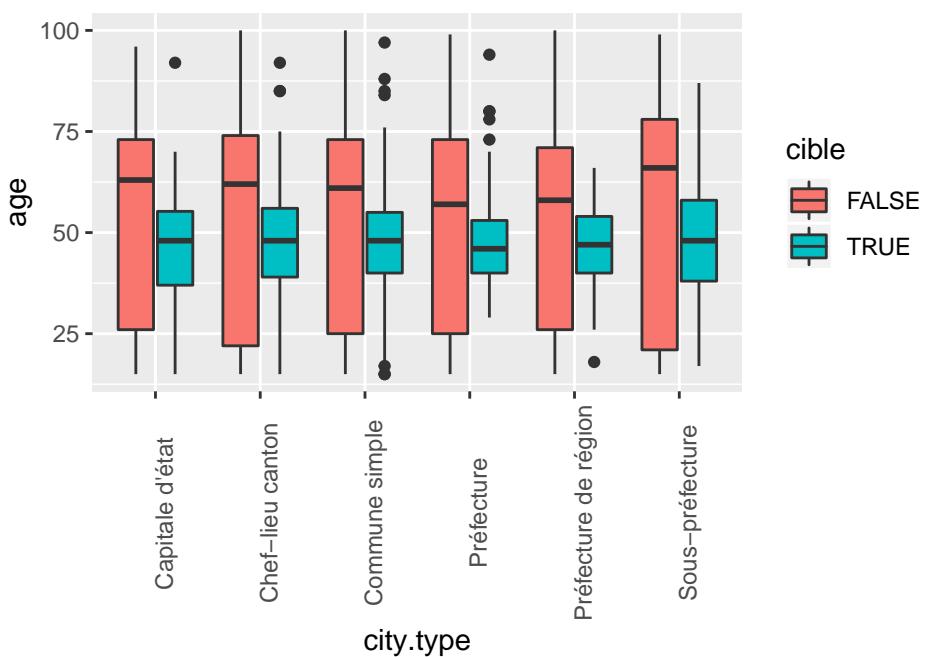
```
##          FALSE TRUE
## 1      19   68
## 2      97  205
## 3     233  617
## 4     338  901
## 5     154  466
## 6     149  478
## 7     237  366
## 8      96  239
## 9      4   13
## 10    968  26
## 11    180  305
## 12    266  651
## 13   2480 444
```

Paris, en tant que capitale, enregistre très peu de résultats, en sachant que la région Ile de France est la plus représentée dans l'échantillon.

Les âges sont bien repartis sur tous les types de communes.

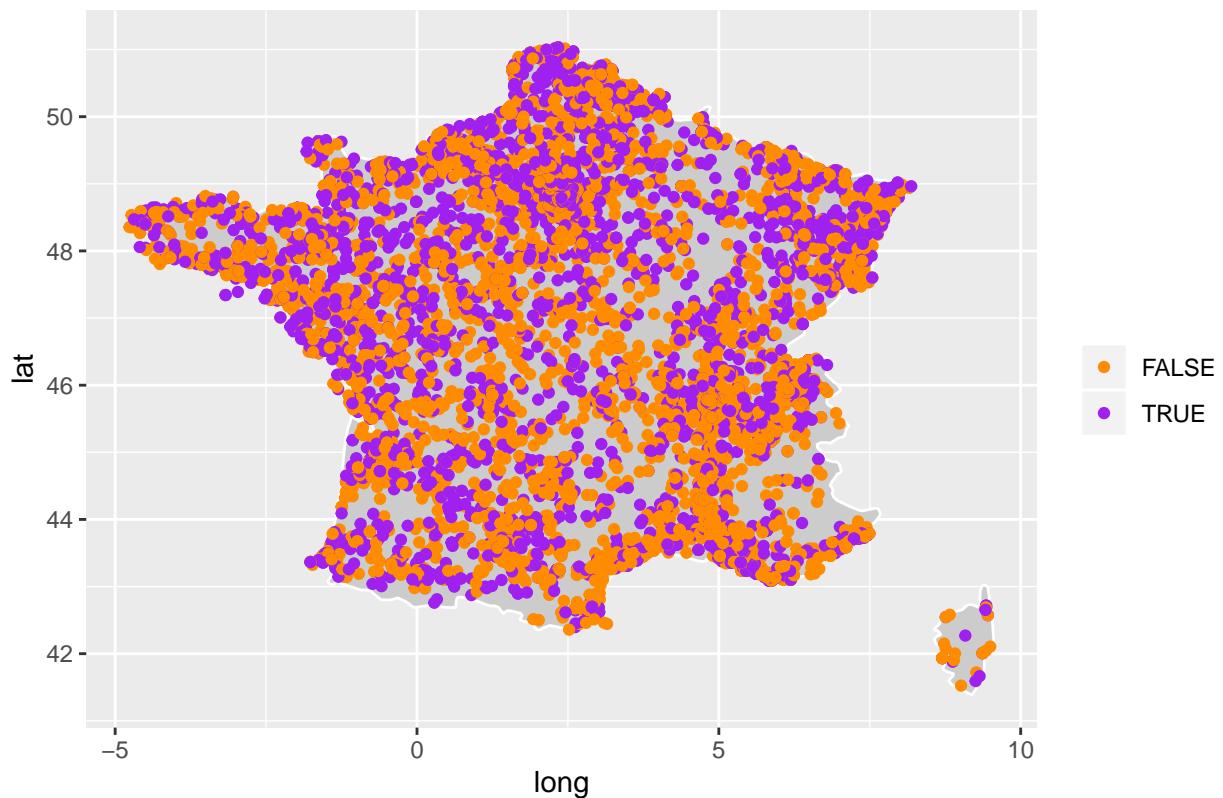


```
##  
##  
##                                     FALSE  TRUE  
##   Capitale d'état           206   148  
##   Chef-lieu canton        1331  1615  
##   Commune simple         2541  1890  
##   Préfecture              389   293  
##   Préfecture de région    437   312  
##   Sous-préfecture         317   521
```

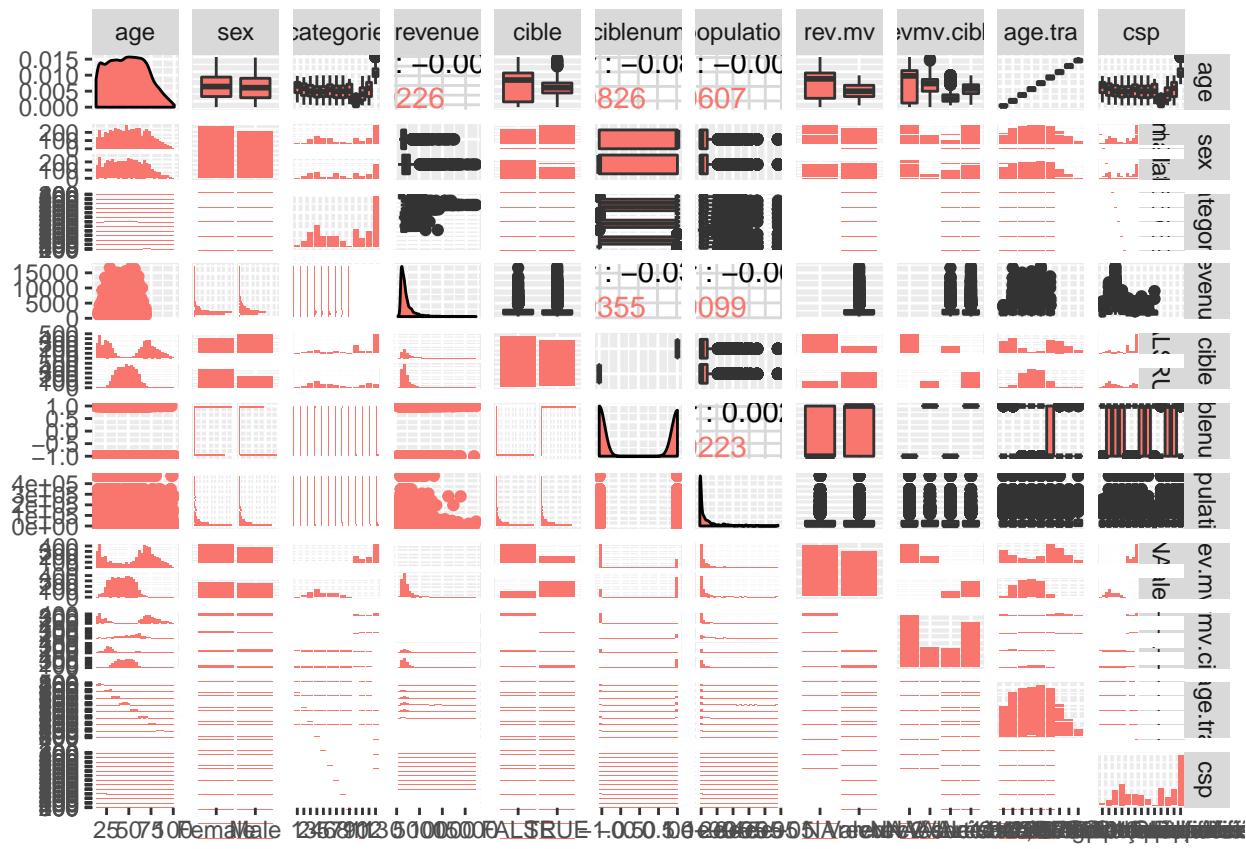


Voici la répartition des réponses sur toute la France, on observe plus de succès au nord et plus d'échecs au sud (région Auvergne-Rhône-Alpes).

Carte



Ce graphique montre que les variables qui vont beaucoup influencer notre prédiction seront sûrement l'âge et la catégorie socio-professionnelle.



## Ajustement des modèles

Il s'agit de construire un modèle prédictif, en testant plusieurs méthodes (régression logistique, random forests), de fournir des prévisions du succès de la campagne à partir du profil client et une évaluation des performances attendues (Le critère de qualité principal retenu est le taux d'erreur) D'autre part, on tentera de sélectionner les variables les plus pertinentes (importance des variables) et d'interpréter le modèle (évaluation basée sur la courbe ROC).

## Régression logistique

## Validation croisée

Nous créons d'abord la partition de données, avec 70% pour l'entraînement et 30% pour la validation. Nous sortons les variables qui ne semblent pas apporter d'information et la variable revenu qui contient des valeurs manquantes.

Ensuite nous utilisons la fonction GLM, pour faire une régression logistique

```
##  
## Call:  
## glm(formula = cible ~ categorie + age + sex + REGION + city.type,  
##       family = binomial(logit), data = train)  
##  
## Deviance Residuals:
```

```

##      Min     1Q   Median     3Q    Max
## -3.0675 -0.5942 -0.1248  0.6455  3.1890
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.566988  0.426579 -3.673 0.000239 ***
## categorie2                 -0.378315  0.379689 -0.996 0.319064
## categorie3                 -0.113965  0.360875 -0.316 0.752153
## categorie4                 -0.193547  0.356439 -0.543 0.587130
## categorie5                 -0.365385  0.368574 -0.991 0.321516
## categorie6                 -0.351235  0.370750 -0.947 0.343453
## categorie7                 -0.338330  0.362825 -0.932 0.351085
## categorie8                 -0.026037  0.381537 -0.068 0.945593
## categorie9                  0.416452  0.983075  0.424 0.671841
## categorie10                -3.956910  0.422147 -9.373 < 2e-16 ***
## categorie11                -0.358697  0.368775 -0.973 0.330716
## categorie12                -0.636809  0.359923 -1.769 0.076845 .
## categorie13                -5.659625  0.370713 -15.267 < 2e-16 ***
## age                         0.060895  0.003018 20.178 < 2e-16 ***
## sexMale                     -1.347140  0.074481 -18.087 < 2e-16 ***
## REGIONBOURGOGNE FRANCHE COMTE 0.290287  0.186283  1.558 0.119160
## REGIONBRETAGNE              0.388629  0.169423  2.294 0.021799 *
## REGIONCENTRE VAL DE LOIRE  -0.169438  0.184160 -0.920 0.357542
## REGIONCORSE                  -0.142140  0.451085 -0.315 0.752681
## REGIONGRAND EST              0.827234  0.153159  5.401 6.62e-08 ***
## REGIONHAUTS DE FRANCE        1.105253  0.146371  7.551 4.32e-14 ***
## REGIONILE DE FRANCE          0.937179  0.130027  7.208 5.70e-13 ***
## REGIONNORMANDIE              1.555923  0.181271  8.583 < 2e-16 ***
## REGIONNOUVELLE AQUITAINE    0.780405  0.142113  5.491 3.99e-08 ***
## REGIONOCCITANIE               0.137704  0.142837  0.964 0.335012
## REGIONPAYS DE LA LOIRE       0.695074  0.166862  4.166 3.11e-05 ***
## REGIONPROVENCE ALPES COTE D AZUR -0.519587  0.152830 -3.400 0.000674 ***
## city.typeChef-lieu canton    0.983998  0.189898  5.182 2.20e-07 ***
## city.typeCommune simple       0.223310  0.194873  1.146 0.251826
## city.typePréfecture          0.490373  0.226468  2.165 0.030364 *
## city.typePréfecture de région 0.315831  0.225201  1.402 0.160784
## city.typeSous-préfecture     1.747167  0.223080  7.832 4.80e-15 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 9691.8 on 7000 degrees of freedom
## Residual deviance: 5768.4 on 6969 degrees of freedom
## AIC: 5832.4
##
## Number of Fisher Scoring iterations: 6

```

Le modèle qu'il a choisi avec le meilleur AIC est : cible ~ categorie + age + sex + REGION + city.type. Il semblerait dans ce choix que la variable "population" ne soit pas pertinente.

AIC: 5832.4

## Prediction

Voici la matrice de confusion quand le seuil est fait à 50% :

```

##          FALSE TRUE
## FALSE    1243  208
## TRUE     323 1225

```

L'erreur de prédiction que nous obtenons est de :17.71 %

Nous pouvons optimiser cette prédiction, en trouvant le seuil optimal :59.03 %. La matrice de confusion devient :

```

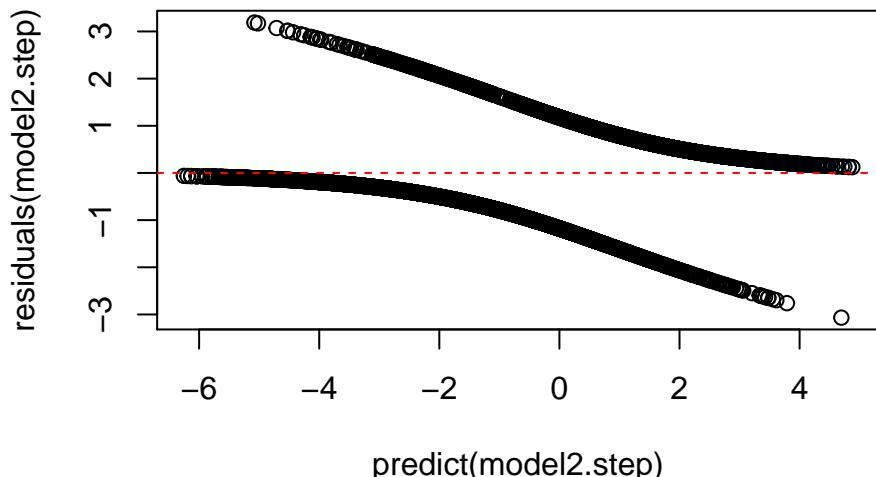
##          FALSE TRUE
## FALSE   1566 1433

```

La nouvelle erreur de prédiction est plus basse : 47.78 % .

L'accuracy de 52.22 %.

### Ajustement du modèle

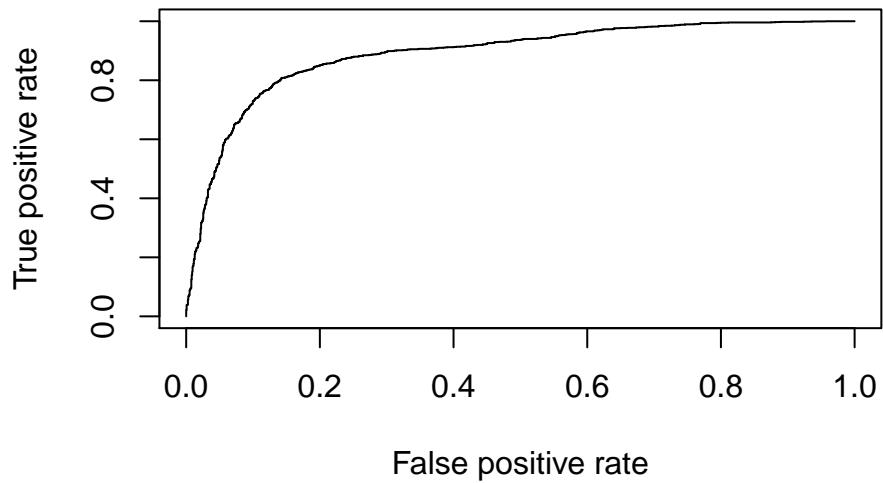


Graphique des résidus

**ROC :**

Comme dernière étape nous allons construire la courbe ROC et calculer l'AUC (area under the curve), qui sont des mesures de performance : la courbe ROC est la représentation des TPR (True positive rate, "Vrais positifs") en fonction des FPR (False positive rate, 'faux positifs), selon plusieurs seuils. La courbe de ROC doit être au dessus de la première bissectrice. Plus la courbe ROC de l'alogarithme tend vers 1 (100%) rapidement, meilleure est la qualité de la prédiction.

L'AUC est l'aire sous la courbe ROC. Elle doit être comprise entre 0.5 et 1 (seuil AUC donnant la meilleure qualité de la prédiction).



Ici on obtient un AUC de 0.8894564.

## Arbres de décision

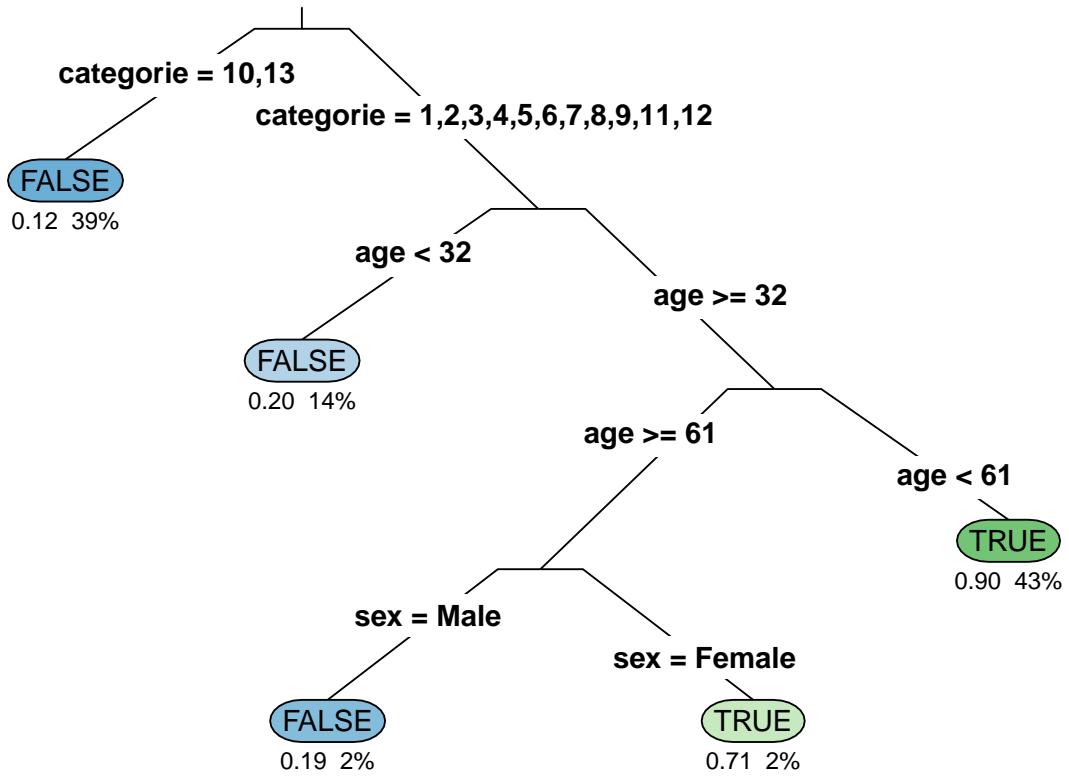
### Validation croisée

Comme pour la régression logistique, nous allons faire nos partitions avec 70% pour l'entraînement et 30% pour la validation. Nous retirons les variables de type identifiant et avec trop de modalités, pour éviter de bloquer les algorithmes :insee.code,f\_name,last.name,commune, department, latitude, longitude,X,Y,reg.

### Création du premier arbre

Nous allons faire notre premier arbre sans ajuster les paramètres de contrôle de l'arbre, avec les paramètres par défaut de la fonction rpart (minsplit = 20, minbucket = round(minsplit/3), cp = 0.01)

Voici le graphique de l'arbre qui est produit avec rpart:



rpart utilise l’impureté Gini pour sélectionner les divisions lors de la classification. L’impureté Gini est la probabilité de classer incorrectement un élément choisi au hasard dans l’ensemble de données, s’il était étiqueté de manière aléatoire en fonction de la distribution des classes. L’impureté Gini de 0 est le plus bas possible, il est possible quand toutes les données sont dans la même classe.

Sans surprise, comme nous l’avions vu avec l’analyse initiale, la catégorie, l’âge et le sexe sont les critères utilisés pour faire les splits. Les catégories 10 et 13 font le premier split, ensuite l’âge 32 et 61, pour finir avec le sexe. Aucune autre variable n’est utilisée pour faire des splits.

Avec ce graphique, chaque noeud montre :

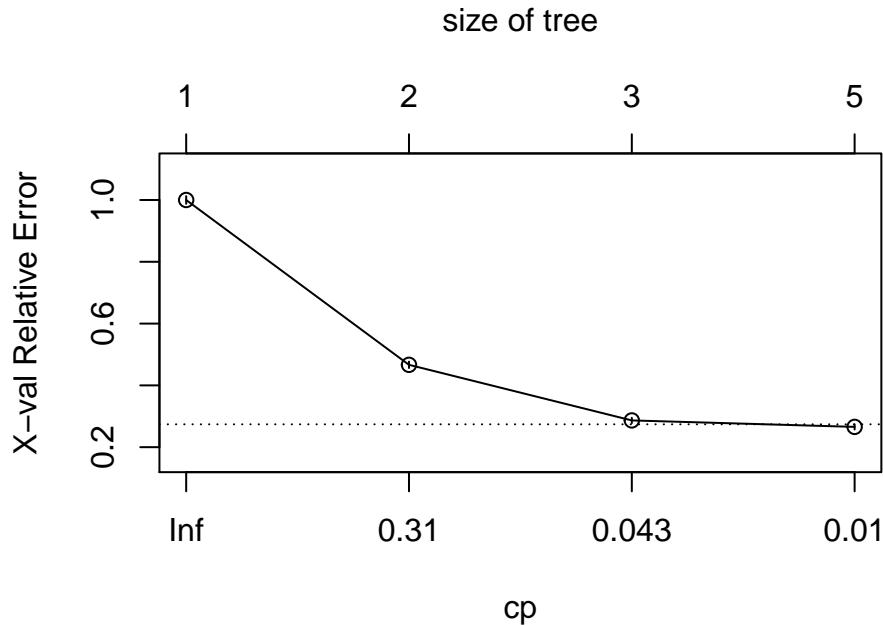
- La classe prédite (TRUE ou FALSE)
- La probabilité d’être VRAI
- Pourcentage d’observations dans chaque noeud

### Zoom sur le CP

Le CP (complexity parameter) est utilisé pour contrôler la croissance de l’arbre et sélectionner la taille optimale de l’arbre. Si le coût d’ajouter une variable est plus élevé que la valeur du CP, la croissance s’arrête.

Les fonctions **printcp** et **plotcp** fournissent l’erreur de validation croisée pour chaque nsplitt et peuvent être utilisées pour élaguer l’arbre. Celui avec le moins d’erreur de validation croisée (xerror) est la valeur optimale de CP.

Avec plotcp, il vous montre également l’endroit optimal pour élaguer l’arbre. Un bon choix de CP pour l’élagage est souvent la valeur la plus à gauche pour laquelle la moyenne se situe en dessous de la ligne horizontale, dans notre cas 0,01, la valeur par défaut.



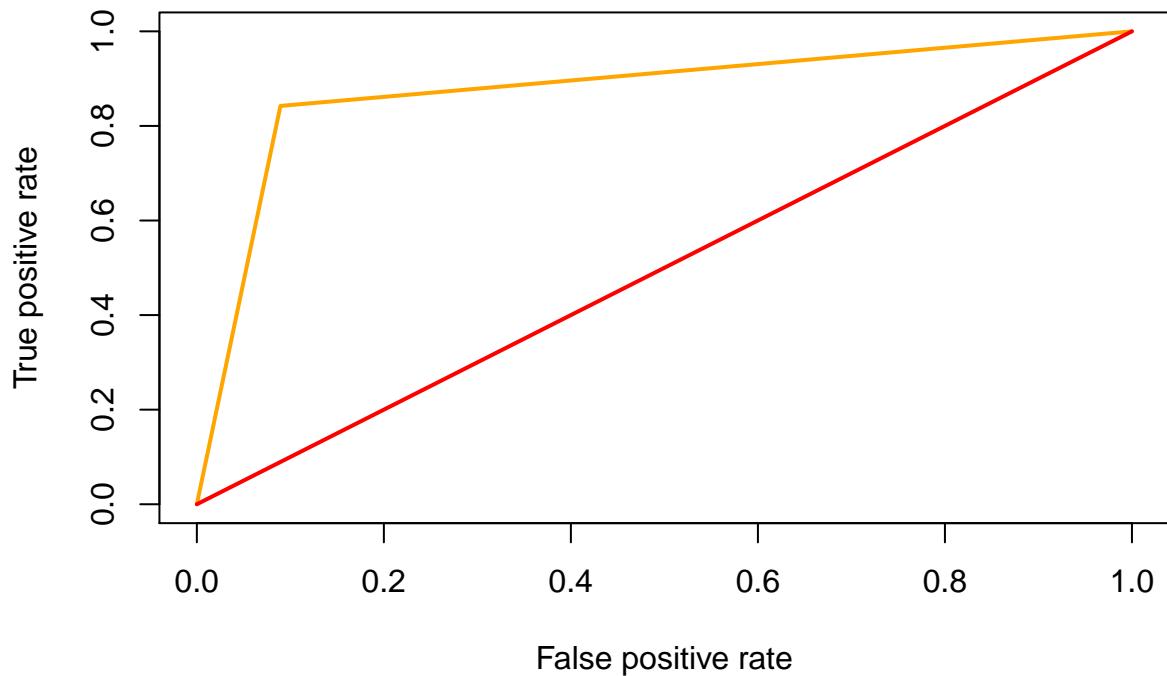
Pour la fonction **printcp**, chaque ligne représente une hauteur différente de l'arbre. En général, plus de niveaux dans l'arbre signifient qu'il a moins d'erreur de classification sur l'entraînement. Cependant, il existe le risque d'overfitting. Souvent, l'erreur de validation croisée augmentera à mesure que l'arbre atteindra plus de niveaux (au moins, après le niveau «optimal»).

```
##
## Classification tree:
## rpart(formula = cible ~ ., data = train_arbre, method = "class")
##
## Variables actually used in tree construction:
## [1] age      categorie sex
##
## Root node error: 3346/7001 = 0.47793
##
## n= 7001
##
##          CP nsplit rel error xerror      xstd
## 1 0.533473     0 1.00000 1.00000 0.0124911
## 2 0.182905     1 0.46653 0.46653 0.0104087
## 3 0.010012     2 0.28362 0.28661 0.0085979
## 4 0.010000     4 0.26360 0.26569 0.0083260
```

sans surprise la meilleure valeur est 0,01

### Matrice de confusion et ROC

```
##
## tree.pred FALSE TRUE
##   FALSE 1426 226
##   TRUE   140 1207
```

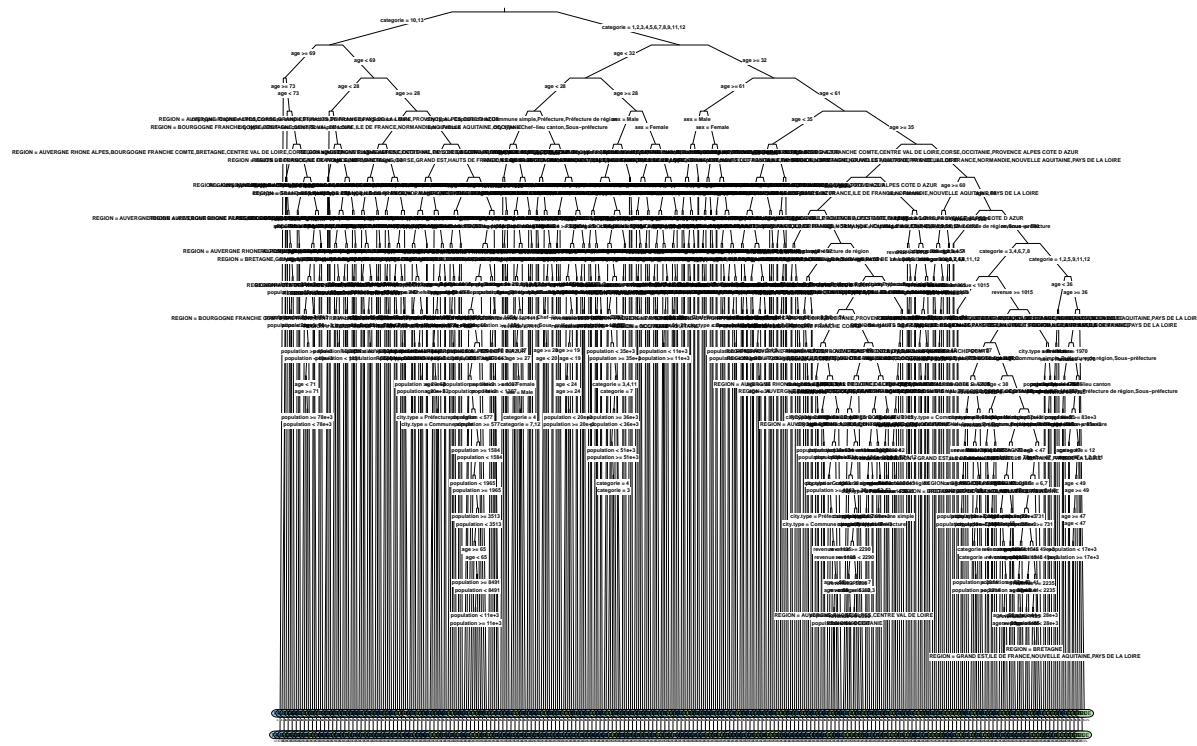


```
## [1] 0.8764446
```

### Tuner l'arbre

Nous allons aussi modifier le paramètre “minsplit”, il contrôle le nombre minimum d’observations qui doit exister dans un nœud pour qu’une tentative de division soit faite, minbucket et CP

Nous allons démarrer avec le CP = 0.0001, et minsplit =2



## Elaguer l'arbre

##Méthode ensembliste : Random Forest

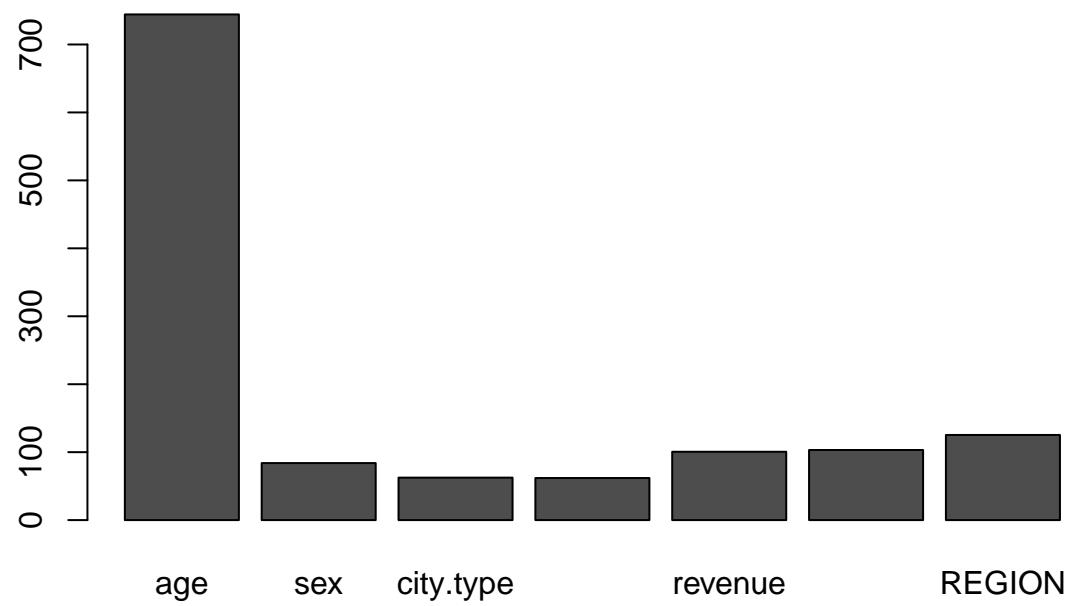
La méthode repose sur l'apprentissage du modèle prédition sur plusieurs arbres. Les différentes prédictions sont comparées selon leur qualité. Et une seule d'entre elles est retenue par vote majoritaire.

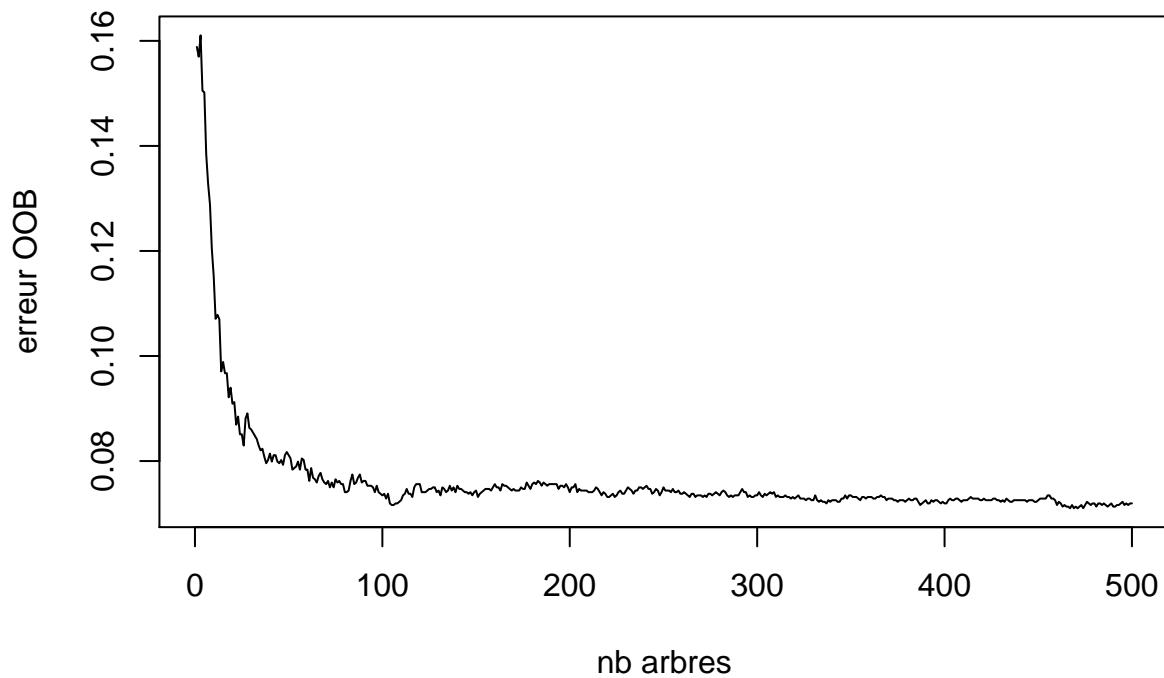
Nous reprenons les ensembles d'entraînement et de validation déjà utilisés pour les arbres de décision.

*Random forest à 500 arbres*

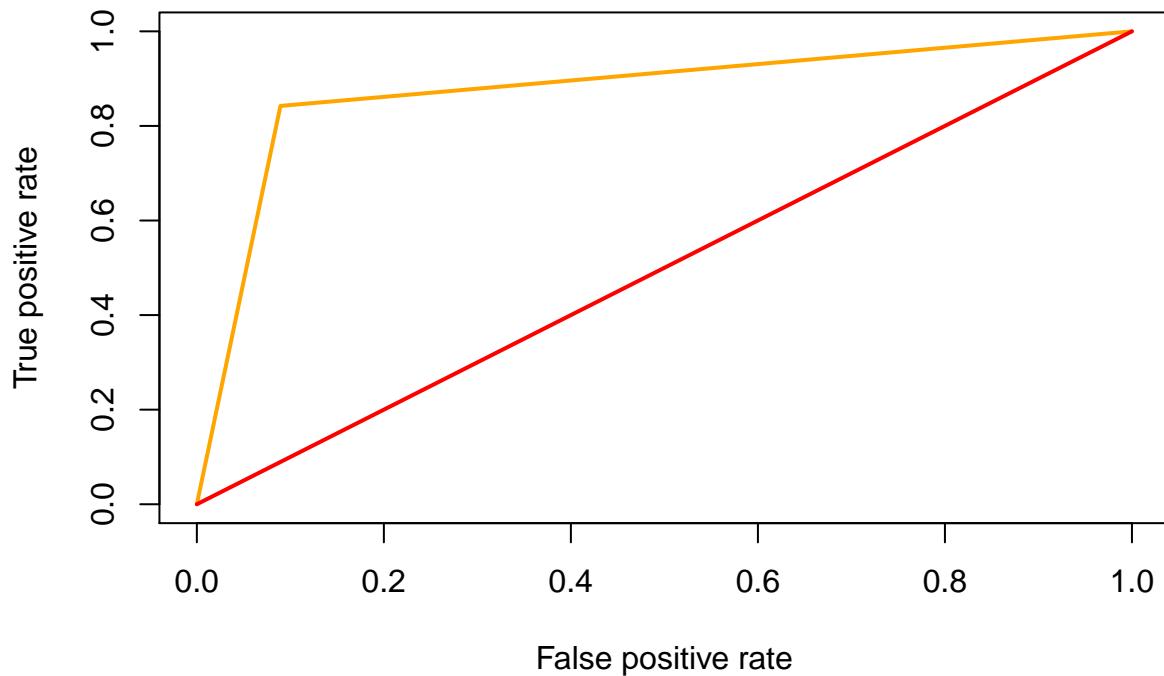
```
##
## Call:
##   randomForest(formula = cible ~ ., data = train_arbre, na.action = na.omit)
##     Type of random forest: classification
##       Number of trees: 500
## No. of variables tried at each split: 2
##
##       OOB estimate of error rate: 7.19%
## Confusion matrix:
##   -1    1 class.error
## -1 744 182  0.1965443
##  1  53 2288  0.0226399
```

L'importance des variables explicatives selon le critère de Gini : plus le critère est bas, plus la valeur de la variable a une influence significative sur la variable réponse.





```
##  
## rforest1.pred -1 1  
##           -1 316 21  
##           1   85 991
```

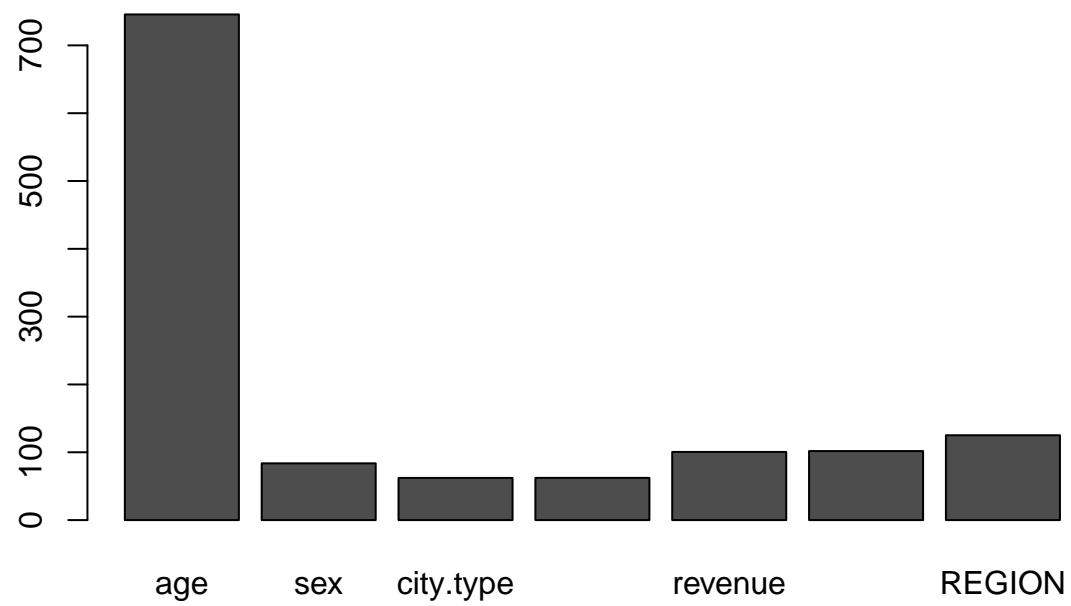


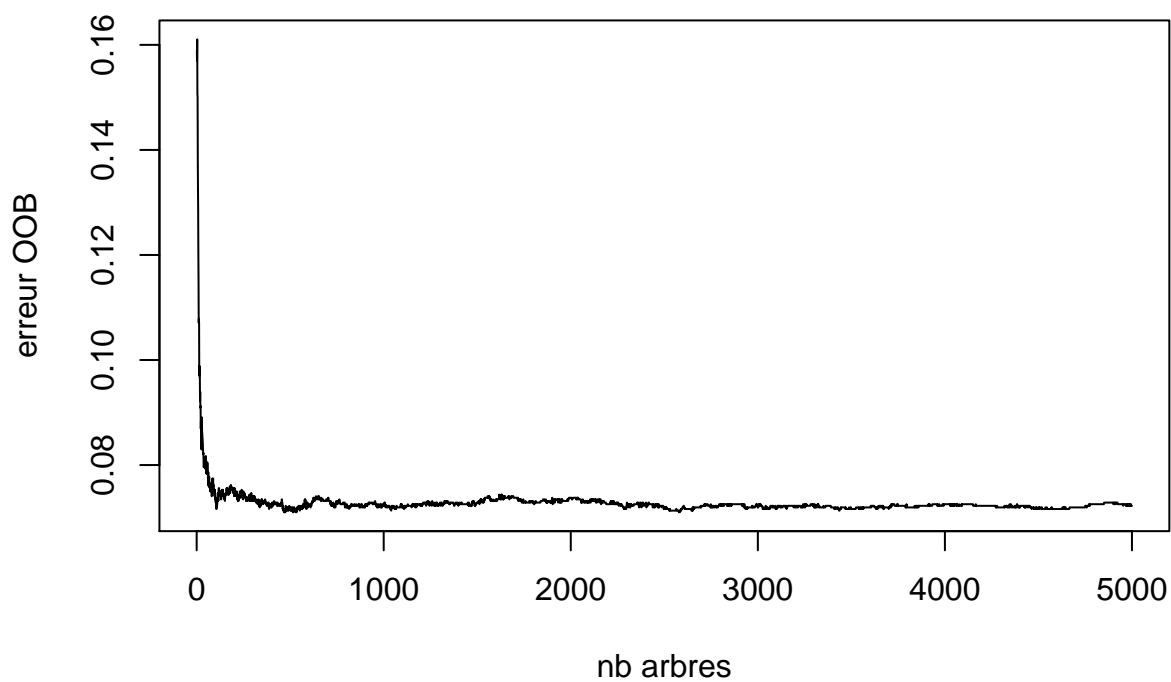
```
## [1] 0.8836395
```

*Random forest à 500 arbres*

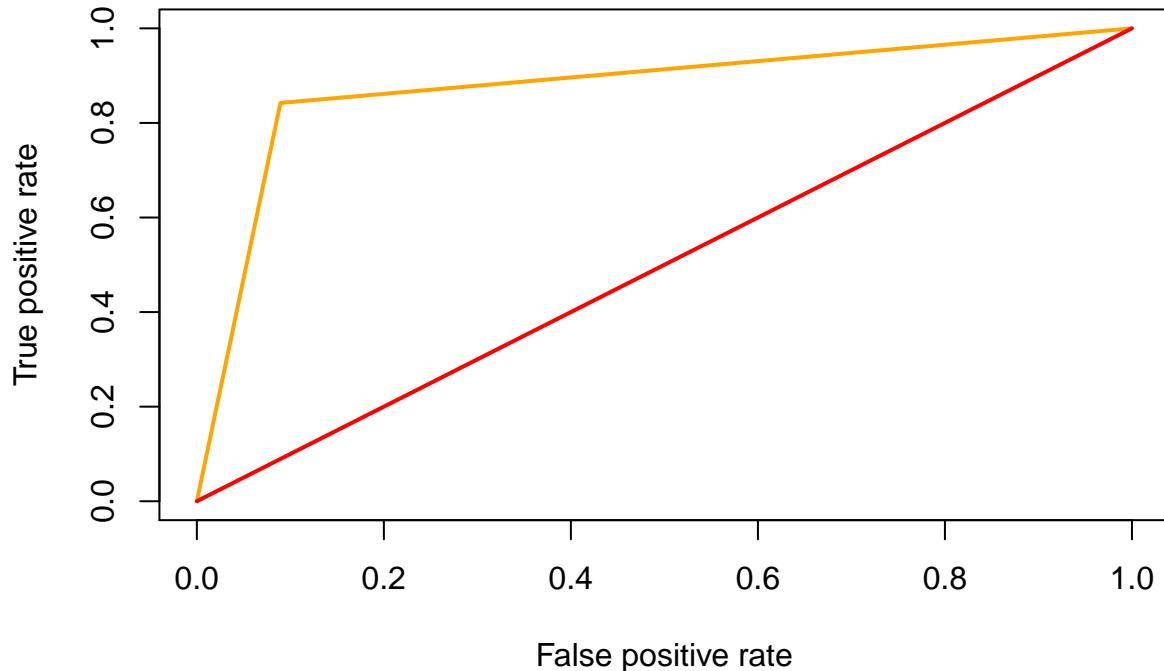
```
##
## Call:
##   randomForest(formula = cible ~ ., data = train_arbre, ntree = 5000,      mtry = 2, na.action = na.o
##                 Type of random forest: classification
##                           Number of trees: 5000
##   No. of variables tried at each split: 2
##
##             OOB estimate of  error rate: 7.22%
##   Confusion matrix:
##     -1    1 class.error
## -1 745 181  0.19546436
## 1   55 2286  0.02349423
```

L'importance des variables explicatives selon le critère de Gini : plus le critère est haut, plus la variable aura de l'influence pour distinguer le succès de l'échec de la campagne : ici l'âge la variable la plus discriminante.





```
##  
## rforest2.pred -1 1  
##                 -1 315 19  
##                  1   86 993
```



```
## [1] 0.8833807
```

##Méthode ensembliste : le boosting - XGBOOST

**Leave one out**

#### Solution - Imputation des données

Pour que l'imputation fonctionne, il faut enlever des colonnes avec trop de catégories, et celles qui ne devraient pas apporter plus d'information comme le prénom et le nom.

L'imputation avec mice, sera faite avec rf : “Random forest imputation”, en utilisant les données d'age, sex, region, categorie et population pour trouver le revenue correspondant.

Nous vérifions que l'imputation respecte bien la structure des données orginal, et c'est bien le cas :

## Modèles prédictifs (à suivre)

La tâche de prédiction consiste à déterminer le succès de la campagne marketing en fonction des caractéristiques des clients. Le critère de qualité principal retenu est le taux d'erreur, mais il est intéressant dans ce type d'application de trier les clients en fonction de leur appétence supposée à la campagne. Une évaluation basée sur la courbe ROC pourra donc être envisagée. Le projet devra mettre en œuvre au moins deux méthodes prédictives différentes comme par exemple la régression logistique et les random forests.

## Annexes