Introduction au machine learning

Maxime Prévôt & François Martin

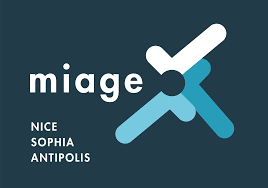


Table des matières

[Exploration et visualisation des données 3](#_Toc71303801)

[Pré-traitement des données éventuels 5](#_Toc71303802)

[Classification supervisée 5](#_Toc71303803)

[RPART 5](#_Toc71303804)

[Répartition 6](#_Toc71303805)

[Matrice de confusion 6](#_Toc71303806)

[C5.0 6](#_Toc71303807)

[Répartition 6](#_Toc71303808)

[Matrice de confusion 6](#_Toc71303809)

[TREE 7](#_Toc71303810)

[Répartition 7](#_Toc71303811)

[Matrice de confusion 7](#_Toc71303812)

[CLASSIFIERS 7](#_Toc71303813)

[RPART 7](#_Toc71303814)

[RANDOM FOREST 7](#_Toc71303815)

[KNN 8](#_Toc71303816)

[SUPPORT VECTOR MACHINES 8](#_Toc71303817)

[NAIVE BAYES 8](#_Toc71303818)

# Exploration et visualisation des données

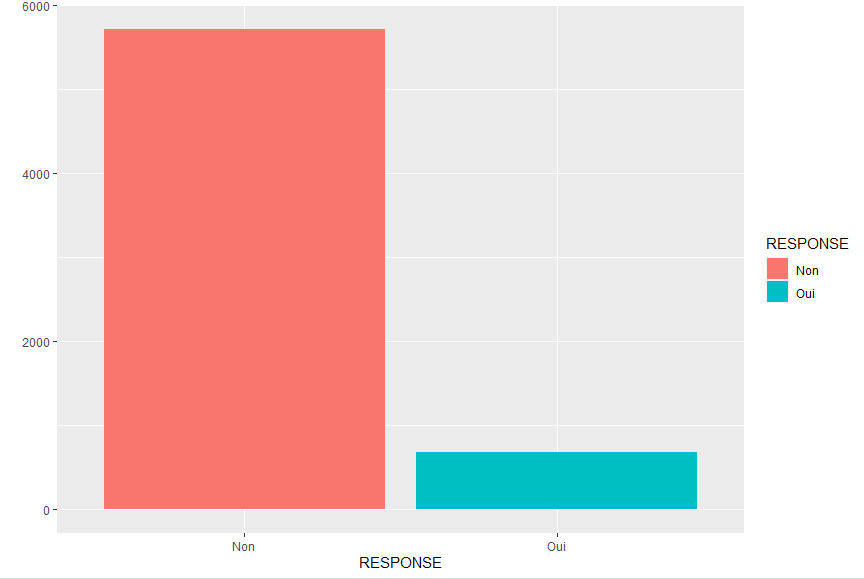
Le but de ce projet est de prédire grâce à différentes informations la réponse qu’elle soit positive ou négative à une offre promotionnelle.

Notre ensemble de données se compose de 6400 instances composées de 28 variables hétérogènes ainsi qu’une variable de classe, la réponse à l’offre.

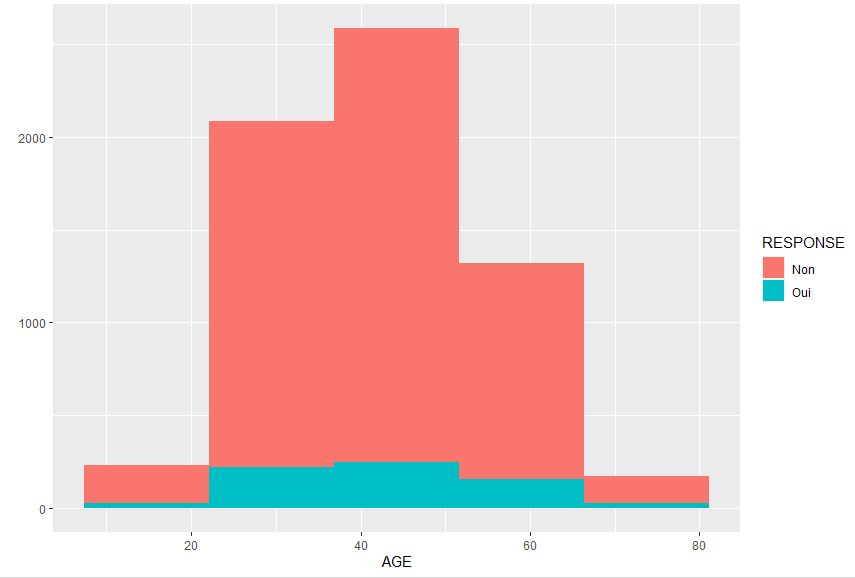
Les variables sont définies comme suit :

|  |  |  |  |
| --- | --- | --- | --- |
| VARIABLE | TYPE | DESCRIPTION | VALEURS |
| AGE | Numérique | Age en années | [18, 77] |
| MARITAL | Booléen | Statut marital | 0 = non marié, 1 = marié |
| ADDRESS | Numérique | Nombre d'années à l'adresse actuelle | [0, 56] |
| INCOME | Numérique | Revenu du foyer en milliers ($) | [9, 1116] |
| INCCAT | Ordinal | Catégories de revenu en milliers ($) | 1 = [0, 25[, 2 = [25, 50[, 3 = [50, 74[, 4 = [75, +?[ |
| CAR | Numérique | Prix du premier véhicule en milliers ($) |  |
| CARCAT | Ordinal | Catégorie de prix du véhicule premier | 1 = économique, 2 = standard, 3 = luxueux |
| ED | Ordinal | Nombre d'années d'éducation | 1 = inf. au bac, 2 = niveau bac, 3 = bac + 2, 4 = bac + 3/4, 5 = bac +5 et + |
| EMPLOY | Numérique | Nombre d'années chez l'employeur actuel | [0, 57] |
| RETIRE | Booléen | Retraité(e) | 0 = non, 1 = oui |
| EMPCAT | Ordinal | Nombre d'années chez l'employeur actuel (catégorie) | 0 = [0, 5[, 1 = [5, 15[, 2 = [15, +?[ |
| JOBSAT | Ordinal | Satisfaction Emploi | 1 = très insatisfait, 2 = insatisfait, 3 = neutre, 4 = satisfait, 5 = très satisfait |
| GENDER | Nominal | Sexe | F = femelle, M = mâle |
| RESIDE | Ordinal | Nombre de personnes dans le foyer | [1, 9] |
| WIRELESS | Booléen | Service sans fil | 0 = non, 1 = oui |
| MULTLINE | Booléen | Lignes multiples | 0 = non, 1 = oui |
| VOICE | Booléen | Boîte vocale | 0 = non, 1 = oui |
| PAGER | Booléen | Service Paging | 0 = non, 1 = oui |
| INTERNET | Catégoriel | Internet | 0 = non, 1 = oui, 8 = ne sais pas, 9 = pas de réponse |
| CALLID | Booléen | Affichage numéro d'appel | 0 = non, 1 = oui |
| CALLWAIT | Booléen | Double appel | 0 = non, 1 = oui |
| OWNTV | Booléen | Possesseur d'un téléviseur | 0 = non, 1 = oui |
| OWNVCR | Booléen | Possesseur d'un magnétoscope | 0 = non, 1 = oui |
| OWNCD | Booléen | Possesseur de matériel HiFi | 0 = non, 1 = oui |
| OWNPDA | Booléen | Possesseur d'un agenda électronique | 0 = non, 1 = oui |
| OWNPC | Booléen | Possesseur d'un ordinateur | 0 = non, 1 = oui |
| OWNFAX | Booléen | Possesseur d'un fax | 0 = non, 1 = oui |
| NEWS | Booléen | Abonnement à un magasine | 0 = non, 1 = oui |
| RESPONSE | Booléen | Réponse positive à l'offre | 0 = non, 1 = oui |

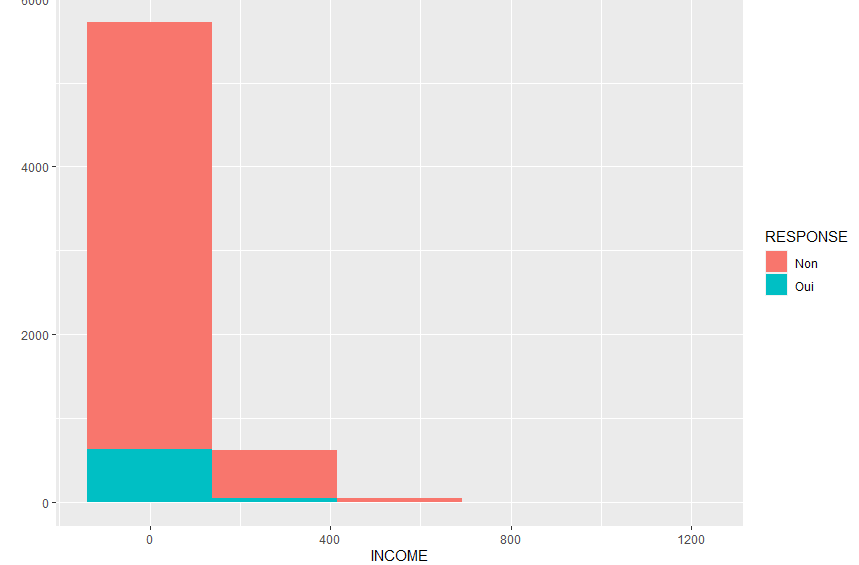
Notre ensemble de données est composé de 5721 Non et 679 Oui en réponse à l’offre promotionnelle.



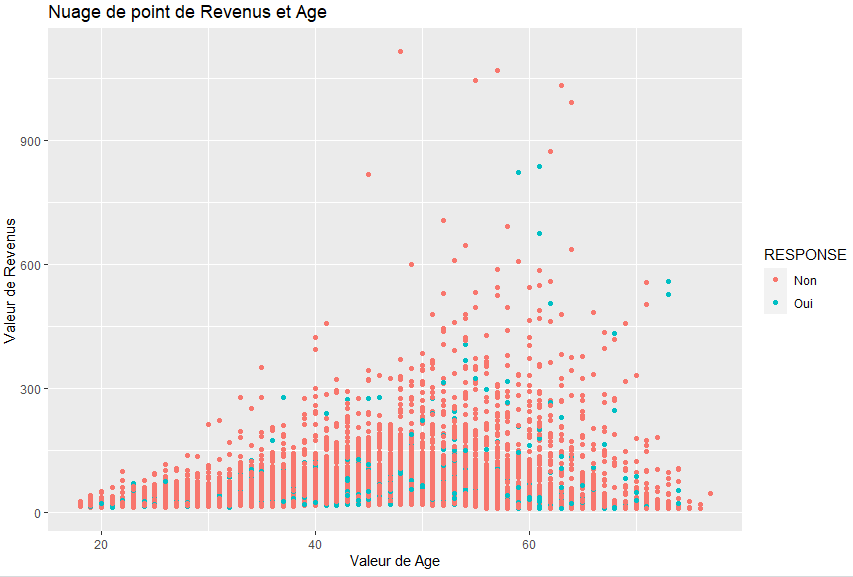
### Age en fonction de la réponse



### Revenus en fonction de la réponse



Nuage de points revenus x âge

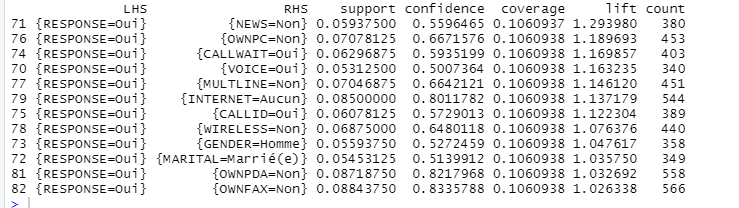


# 

# Pré-traitement des données éventuels

Afin de prédire les résultats de la réponse à l’appel d’offre nous avons dû nous concentrer sur les variables ayant le plus d’impact dans la réponse. Globalement de trop nombreuses variables sont ressorti avec une réponse négative en antécédant comme en conséquence pour permettre une évaluation. Nous nous sommes donc appliqués à la recherche de variables impactante dans une réponse positive.

Nous avons trouvé ces variables avec une réponse positive en antécédant :



# Classification supervisée

Afin de définir les classifiers les plus efficaces nous avons testé plusieurs possibilités.

Pour commencer nous avons créer trois arbres de décision différents :

* RPART
* C5.0
* TREE

Dont voici les prédictions :

## RPART

### Répartition

|  |  |
| --- | --- |
| Non | Oui |
| 2119 | 281 |

Avec un taux de succès de 0.81875

### Matrice de confusion

|  |  |  |
| --- | --- | --- |
|  | Non | Oui |
| Non | 1914 | 230 |
| Oui | 205 | 51 |

Rappel

[1] 0.1992188

Spécificité

[1] 0.8927239

Précision

[1] 0.1814947

Taux de vrai négatif

[1] 0.9032563

## 

## C5.0

### Répartition

|  |  |
| --- | --- |
| Non | Oui |
| 2400 | 0 |

Avec un taux de succès de 0.8933333

### Matrice de confusion

|  |  |  |
| --- | --- | --- |
|  | Non | Oui |
| Non | 2144 | 0 |
| Oui | 256 | 0 |

## TREE

### Répartition

|  |  |
| --- | --- |
| Non | Oui |
| 2400 | 0 |

Avec un taux de succès de 0.8933333

### Matrice de confusion

|  |  |  |
| --- | --- | --- |
|  | Non | Oui |
| Non | 2144 | 0 |
| Oui | 256 | 0 |

## 

## CLASSIFIERS

Afin de définir les classifiers les plus efficaces nous les avons évalués à l’aide de l’indice AUC et de courbe ROC.

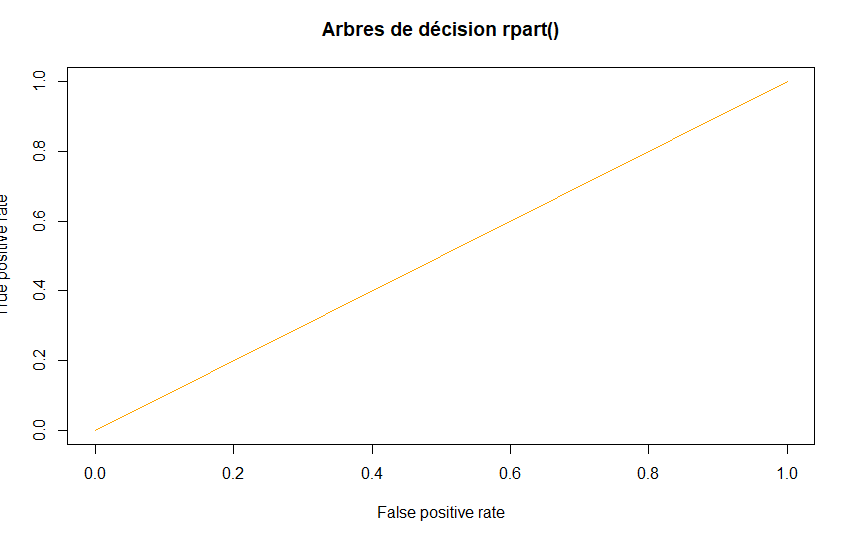
### RPART

Pour les prédictions :

|  |  |  |
| --- | --- | --- |
|  | Non | Oui |
| Non | 2144 | 0 |
| Oui | 256 | 0 |

Avec un indice AUC de 0.5

Voici un graphique des courbes ROC

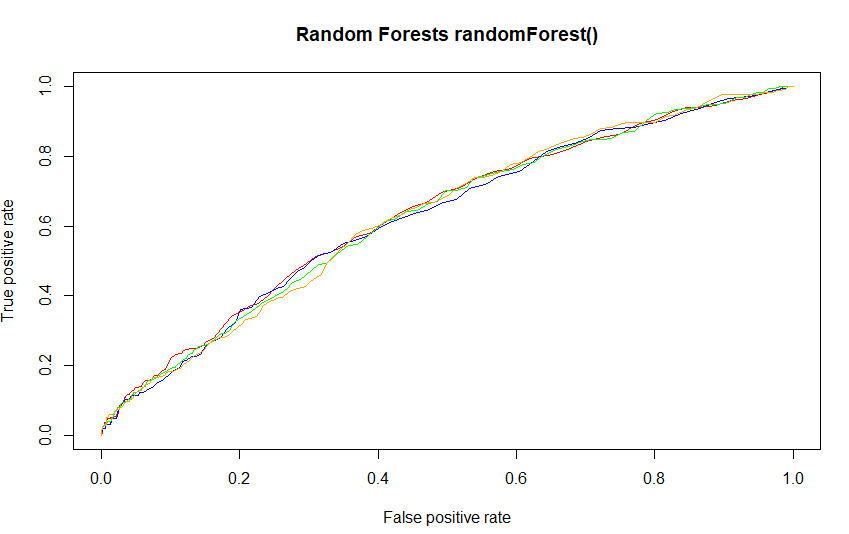


### RANDOM FOREST

Pour les prédictions :

|  |  |  |
| --- | --- | --- |
|  | Non | Oui |
| Non | 2144 | 0 |
| Oui | 256 | 0 |

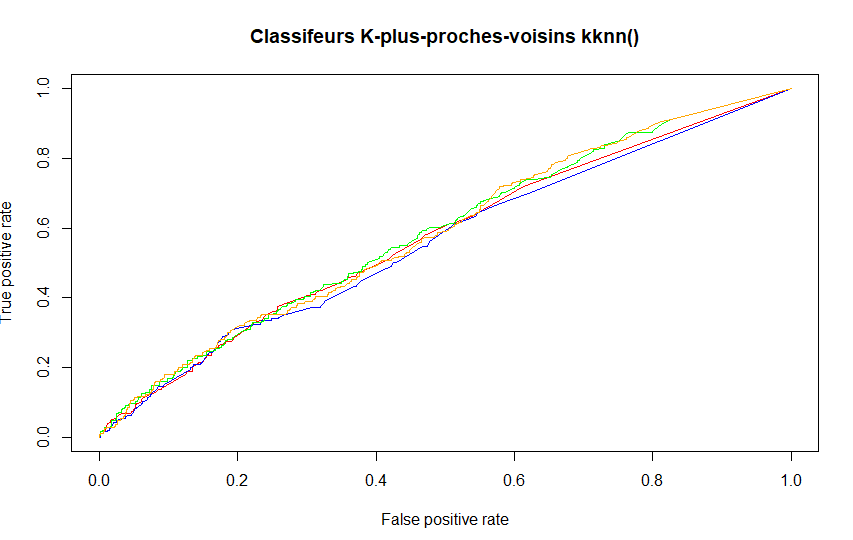
Avec un indice AUC de 0.634613492595616



### KNN

Pour les prédictions :

|  |  |  |
| --- | --- | --- |
|  | Non | Oui |
| Non | 2139 | 5 |
| Oui | 252 | 4 |

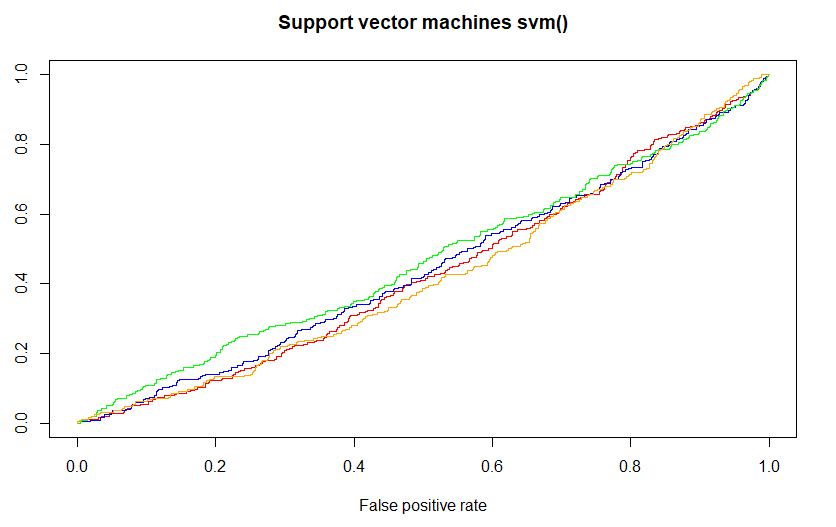
Avec un indice AUC de 0.586350170534048

### SUPPORT VECTOR MACHINES

Pour les prédictions :

|  |  |  |
| --- | --- | --- |
|  | Non | Oui |
| Non | 2144 | 0 |
| Oui | 256 | 0 |

Avec un indice AUC de 0.569228442747202

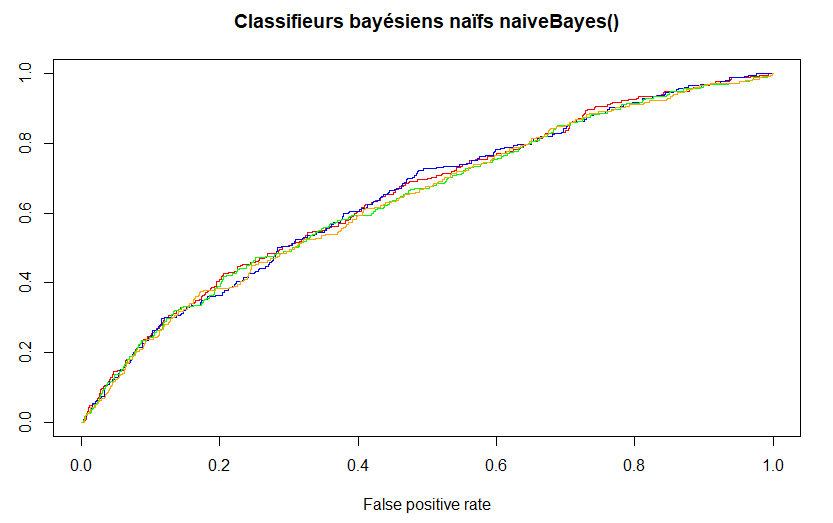


### NAIVE BAYES

Pour les prédictions :

|  |  |  |
| --- | --- | --- |
|  | Non | Oui |
| Non | 2076 | 68 |
| Oui | 230 | 26 |

Avec un indice AUC de 0.649616298390859



Nous avons donc un classement des classifiers qui définit comme suit :

|  |  |
| --- | --- |
| Classifier | AUC |
| NAIVE BAYES | 0.649616298390859 |
| RANDOM FOREST | 0.634613492595616 |
| KNN | 0.586350170534048 |
| SUPPORT VECTOR MACHINE | 0. 569228442747202 |
| RPART | 0.5 |

Nous allons donc nous tourner sur le classifier Naive Bayes étant le plus performant. Les résultats de la prédiction via le classifier Naive Bayes sont à retrouver dans le fichier « PREVOT\_MAXIME\_MARTIN\_François.csv » aux dernières colonnes « PREDICTION » , « PROBABILITE\_OUI» et « PROBABILITE\_NON».