Étude de l'impact du lancement d'un nouveau produit sur un public cible

Avant-propos

Cette étude est réalisé à partir du jeu de données fictif disponible en téléchargement via le lien suivant: https://www.mediafire.com/file/tg4n4vvb72rbrnf/For%25C3%25AAtR.csv/file (https://www.mediafire.com/file/tg4n4vvb72rbrnf/For%25C3%25AAtR.csv/file) Ce fichier contient une variable Age représentant l'âge de la personne, une variable Dépense représentant le montant moyen dépenser par la personne lors de son passage en magasin, une variable Revenus représentant les revenus sur un mois de la personne et une variable YesNo prenant la valeur "yes" si la personne à acheter le produit et "no" sinon. On a connaissance de ces variables pour 4000 personnes. L'entreprise et les données sont fictives. Merci de m'ajouter sur Linkedin pour toutes questions ou remarques: https://fr.linkedin.com/in/romain-vimont__a296131b8 (https://fr.linkedin.com/in/romain-vimont%02a296131b8) (https://fr.linkedin.com/in/romain-vimont-a296131b8 (https://fr.linkedin.com/in/romain-vimont-a296131b8))

Mise en situation

Une enseigne spécialisé dans la vente de téléviseur souhaite attirer une nouvelle clientèle correspondant à de jeunes cadres. Pour cela, elle a lancé il y a 1 mois dans un de ses magasins un nouveau téléviseur avec un casque de réalité virtuelle intégré au prix de 149.99€. Durant le mois suivant le lancement, elle a administré un sondage à 4000 de ses clients grâce aux adresses e-mail renseigner lors de la création d'un compte fidélité. A partir de ces données, le service marketing souhaite savoir si une personne de 25 ans, touchant 1800€ par mois et dépensant en moyenne 150€ lors de son passage en magasin serait susceptible d'acheter le nouveau téléviseur. Le Data Scientist se propose de construire un modèle pouvant répondre à la question.

Analyse des données

On commence par chargé les données sur R et d'en afficher les premières lignes:

Entrée [233]:

```
donnees = read.csv("C:/Users/vimon/Downloads/ForêtR.csv", header =T, sep = ":")
head(donnees)
```

Age	Dépense	Revenus	YesNo
37	191	4504	no
78	565	2316	no
54	938	4343	no
71	684	4640	no
46	246	1487	no
78	128	2060	no

Par exemple, on a une personne de 37 ans, dépensant en moyenne 191 euros lors de son passage en

magasin, gagnant 4504 euros par mois et n'ayant pas acheté le téléviseur, etc...

Dans un premier temps, on peut voir le nombre de personnes ayant acheté le téléviseur en faisant:

Entrée [235]:

```
table(donnees$YesNo)
```

```
  no  yes 
2755 1245 
```

Ainsi, 1245 personnes ont acheté le téléviseur et 2755 ne l'ont pas acheté.

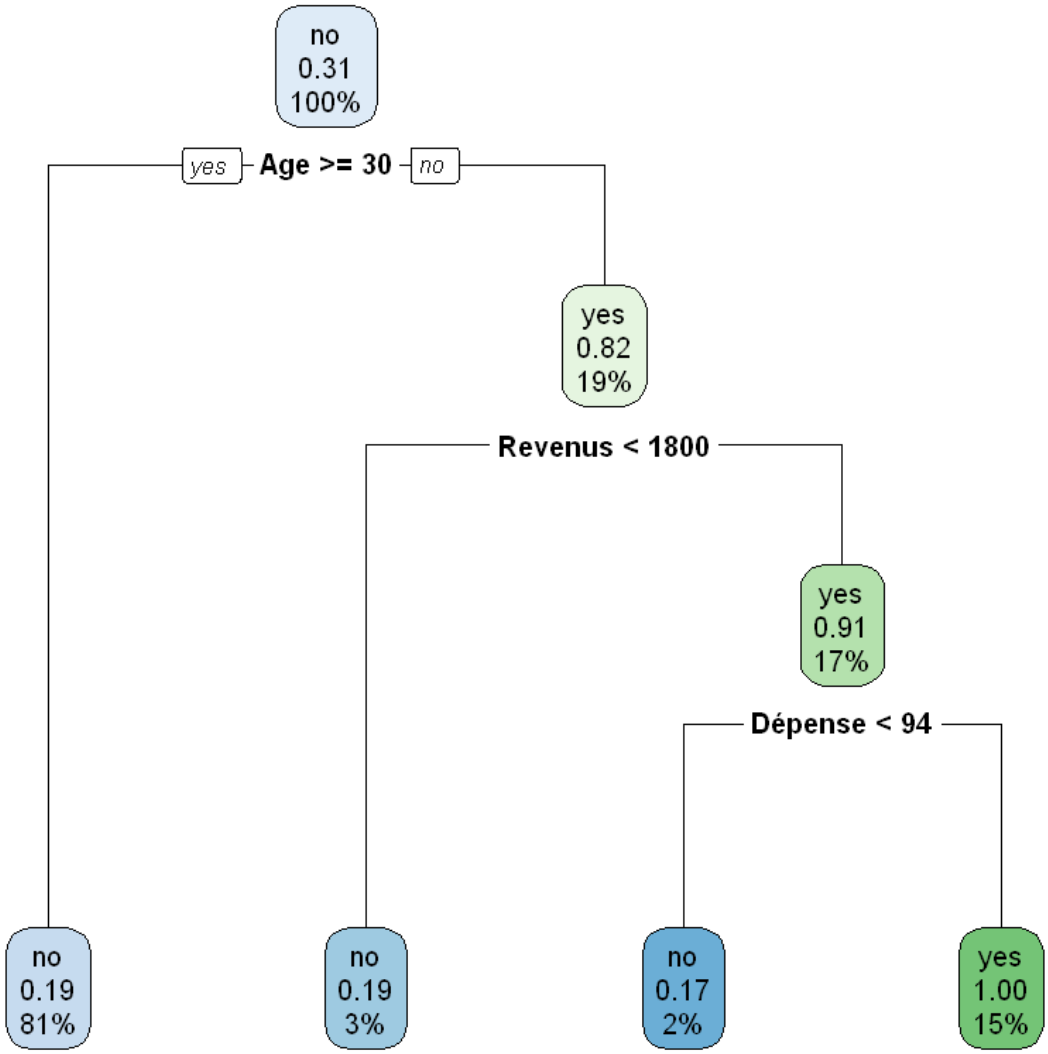
Arbre de classification CART

L'objectif est de savoir si un client dont on connait l'âge, la dépense moyenne lors du passage en magasin et le revenu peut être considéré comme un acheteur potentiel du nouveau téléviseur. On peut donc utiliser un arbre de classification de type CART.

Pour cela, on fait:

Entrée [236]:

```
library(rpart)
arbre = rpart(YesNo ~., data = donnees)
library(rpart.plot)
rpart.plot(arbre)
```



En prenant Age=25, Revenus=1800 et Dépense= 150, on obtient le cheminement suivant:

1) Age>=30? No --> branche droite 2) Revenus<1800? No --> branche droite 3) Dépense<94? No --> branche droite

On arrive au nœud final "yes" représentant 15% des 4000 personnes interrogés.

On obtient le résultat en faisant:

Entrée [238]:

```
predict(arbre, newdata = data.frame(Age = 25, Dépense = 150, Revenus =
1800), type = "class")
```

1: yes
▶ **Levels**:

Ainsi, une valeur plausible $y*$ de yesno pour le client concerné est $y* = y$. Donc le client achètera probablement le téléviseur.

Pour aller un peu plus loin, on peut s'intéresser au taux d'erreur global de l'arbre de classification. On fait:

Entrée [220]:

```
pred = predict(arbre, newdata = donnees, type = "class")
mc = table(donnees$YesNo, pred)
t = (mc[1, 2] + mc[2, 1]) / sum(mc)
t
```

0.16025

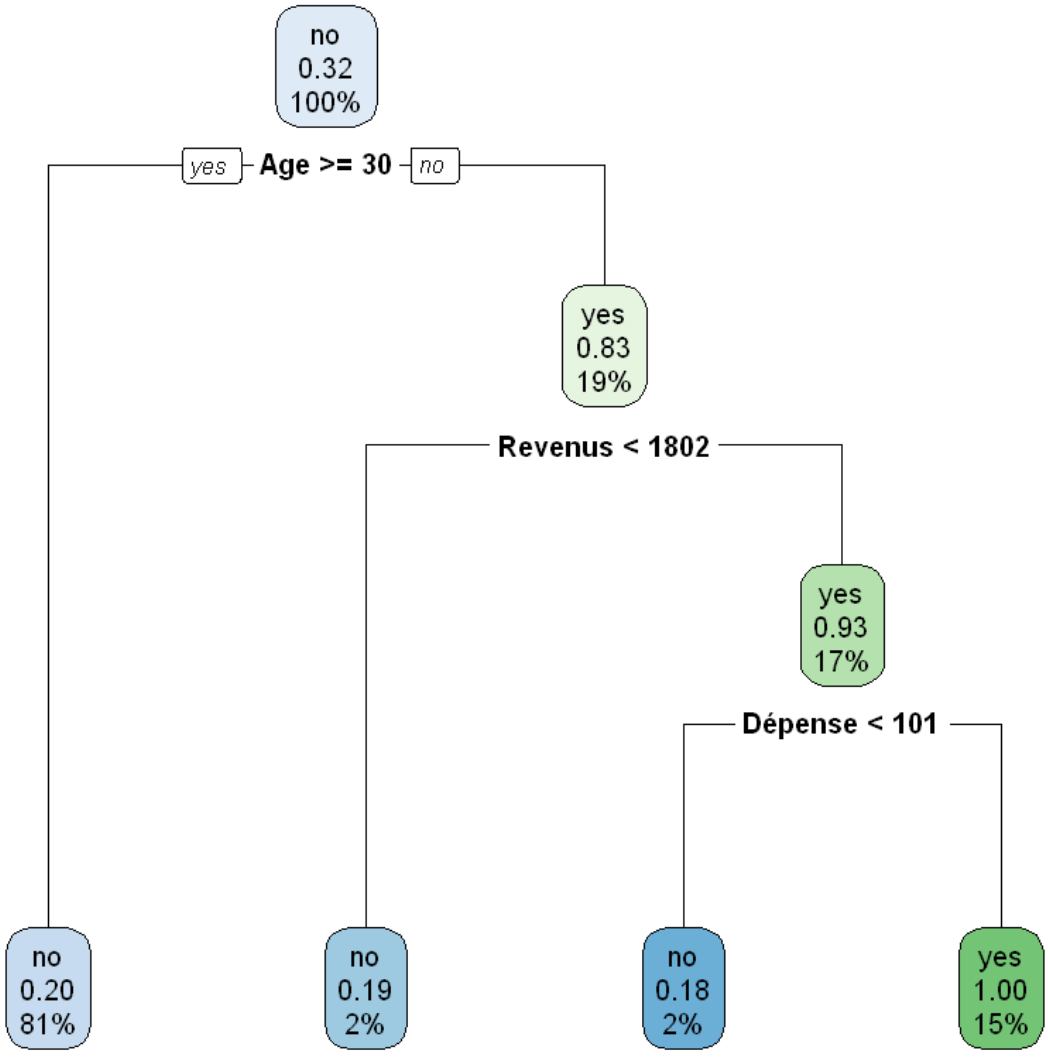
Ainsi, le taux d'erreur global est de 0.16025. Donc, pour les données considérées, l'arbre s'est trompé une fois sur 10 environ, ce qui est très acceptable.

Afin d'avoir un indicateur d'erreur plus solide, on peut appliquer la méthode apprentissage-test. Ainsi, on propose de sélectionner 1000 clients au hasard dans le jeu de données (pour qu'il en reste le chiffre rond de 3000), ce qui constituera le jeu de données de test. Puis, on utilise les clients non sélectionnés pour construire le jeu de données d'apprentissage avec lequel on va construire un nouvel arbre de classification, dont on calculera te taux d'erreur apprentissage-classification. On fait :

Entrée [230]:

```
n = nrow(donnees)
m = 3000
s = sample(1:n, m)
apprentissage = donnees[s, ]
test = donnees[-s, ]
arbre_new = rpart(YesNo ~
., data = apprentissage)
rpart.plot(arbre_new)
c = predict(arbre_new, test[1:3], type = "class")
sum(c != test[ ,4]) / (n - m)
```

0.135



On obtient un arbre de régression légèrement différent du premier et un taux d'erreur sensiblement inférieur taux d'erreur global.

Résolution du problème à l'aide du modèle

On peut prolonger la prédiction avec les clients allant de 23 à 30 ans:

Entrée [239]:

```
predict(arbre, newdata = data.frame(Age = 23, Dépense = 150, Revenus =
1800), type = "class")
predict(arbre, newdata = data.frame(Age = 24, Dépense = 150, Revenus =
1800), type = "class")
predict(arbre, newdata = data.frame(Age = 25, Dépense = 150, Revenus =
1800), type = "class")
predict(arbre, newdata = data.frame(Age = 26, Dépense = 150, Revenus =
1800), type = "class")
predict(arbre, newdata = data.frame(Age = 27, Dépense = 150, Revenus =
1800), type = "class")
predict(arbre, newdata = data.frame(Age = 28, Dépense = 150, Revenus =
1800), type = "class")
predict(arbre, newdata = data.frame(Age = 29, Dépense = 150, Revenus =
1800), type = "class")
predict(arbre, newdata = data.frame(Age = 30, Dépense = 150, Revenus =
1800), type = "class")
```

1: yes
▶ **Levels**:

1: yes
▶ **Levels**:

1: yes
▶ **Levels**:

1: yes
▶ **Levels**:

1: yes
▶ **Levels**:

1: yes
▶ **Levels**:

1: yes
▶ **Levels**:

1: no
▶ **Levels**:

L'ensemble des personnes de moins de 30 ans et respectant les critères de dépenses et de revenus définis précédemment seront susceptibles d'acheter le nouveau téléviseur. L'objectif de faire venir un public jeune avec le nouveau produit sera atteint.

Limites

Ce type d'étude peut comporter certaines limites:

1) Le sondage réalisé peut ne pas être représentatif de l'ensemble des clients du magasin puisqu'il se concentre uniquement sur ceux ayant une carte de fidélité. On peut vérifier la représentativité de l'échantillon en effectuant un t-test en prenant comme moyenne de référence la moyenne des dépenses de l'ensemble des clients du magasin.

2) Le résultat peut être biaisé du fait du délai trop court entre le lancement du nouveau téléviseur et de l'administration de l'enquête au client. On peut espacer ce délai de 3 mois pour atténuer l'effet "nouveauté" et ainsi obtenir de meilleures réponses.

3) Le lancement ne s'effectue que dans un seul magasin. Pour éviter les biais à la répartition géographique de la population, on peut lancer le produit dans 5 magasins situé dans des zones géographiques présentant des caractères socio-démographiques différents (population jeune-agé, aisé-moyenne,etc...).

Romain Vimont Octobre 2022