Devoir surveillé de classification supervisée : apGIS4

Vincent Vandewalle

04/04/2019

Durée 2h, tous documents autorisés

Présentation des données

On souhaite prédire la souscription à une assurance lors d'un vol aérien. Les variables sont les suivantes : - Target: Claim Status (Claim) - Name of agency (Agency) - Type of travel insurance agencies (Agency.Type) - Distribution channel of travel insurance agencies (Distribution.Channel) - Name of the travel insurance products (Product.Name) - Duration of travel (Duration) - Destination of travel (Destination) - Amount of sales of travel insurance policies (Net.Sales) - Commission received for travel insurance agency (Commission) - Gender of insured (Gender) - Age of insured (Age)

La variable à prédire est la variable `Claim`

Importation des données et premières analyses

Q1 : Importer les données à partir du fichier `assurance.csv`. On nommera `don` le data.frame résultant. Le jeu de données comporte-t'il des valeurs manquantes ? Quel option doit-on préciser dans R pour préciser la chaîne de caractères associée aux valeurs manquantes ?

R1 :

```
don = read.csv("assurance.csv",na.strings = "")
head(don)
```

```
##   Agency   Agency.Type Distribution.Channel                  Product.Name
## 1    CBH Travel Agency              Offline             Comprehensive Plan
## 2    CBH Travel Agency              Offline             Comprehensive Plan
## 3    CWT Travel Agency               Online Rental Vehicle Excess Insurance
## 4    CWT Travel Agency               Online Rental Vehicle Excess Insurance
## 5    CWT Travel Agency               Online Rental Vehicle Excess Insurance
## 6    JZI      Airlines               Online                       Value Plan
##   Claim Duration   Destination Net.Sales Commision..in.value. Gender Age
## 1    No      186      MALAYSIA     -29.0                 9.57      F  81
## 2    No      186      MALAYSIA     -29.0                 9.57      F  71
## 3    No       65     AUSTRALIA     -49.5                29.70   <NA>  32
## 4    No       60     AUSTRALIA     -39.6                23.76   <NA>  32
## 5    No       79         ITALY     -19.8                11.88   <NA>  41
## 6    No       66 UNITED STATES    -121.0                42.35      F  44
```

```
summary(don)
```

```
##      Agency             Agency.Type    Distribution.Channel
##  EPX    :35119   Airlines     :17457   Offline: 1107
##  CWT    : 8580   Travel Agency:45869   Online :62219
##  C2B    : 8267
##  JZI    : 6329
##  SSI    : 1056
##  JWT    :  749
##  (Other): 3226
##                            Product.Name   Claim          Duration
##  Cancellation Plan               :18630   No :62399   Min.   :  -2.00
##  2 way Comprehensive Plan        :13158   Yes:  927   1st Qu.:   9.00
##  Rental Vehicle Excess Insurance: 8580               Median :  22.00
##  Basic Plan                      : 5469              Mean   :  49.32
##  Bronze Plan                     : 4049              3rd Qu.:  53.00
##  1 way Comprehensive Plan        : 3331              Max.   :4881.00
##  (Other)                         :10109
##     Destination      Net.Sales       Commision..in.value.  Gender
##  SINGAPORE:13255   Min.   :-389.00   Min.   :  0.00       F   : 8872
##  MALAYSIA : 5930   1st Qu.:  18.00   1st Qu.:  0.00       M   : 9347
##  THAILAND : 5894   Median :  26.53   Median :  0.00       NA's:45107
##  CHINA    : 4796   Mean   :  40.70   Mean   :  9.81
##  AUSTRALIA: 3694   3rd Qu.:  48.00   3rd Qu.: 11.55
##  INDONESIA: 3452   Max.   : 810.00   Max.   :283.50
##  (Other)  :26305
##       Age
##  Min.   :  0.00
##  1st Qu.: 35.00
##  Median : 36.00
##  Mean   : 39.97
##  3rd Qu.: 43.00
##  Max.   :118.00
##
```

```
# save(don, file = "don.Rda")
```

Par la suite on chargera le fichier `don.Rda` contenant le data.frame `don` pour être sûr de partir sur de bonnes bases.

Q2 : En quoi le problème qui vous est posé est-il un problème de classificiation supervisée ? Quel intérêt peut-il bien y avoir à prédire la variable `Claim` ?

R2 : Ici la variable à prédire est la variable `Claim` qui prend comme valeur soit `Yes` soit `No`, nous sommes donc bien dans le contexte de la prédiction d'une variable qualitative (binaire ici car uniquement deux modalités) à partir d'autres variables. Ici celà peut servir par exemple à identifier des clients plus succeptibles que d'autres de souscrire l'assurance puis de leur envoyer par exemple un courrier personnalisé.

Q3 : Dans vos données quelles sont les fréquences des différentes modalités de la variable `Claim` ? Dans quel ordre des modalités de la variable `Claim` sont-elles codées ?

R3 :

```
prop.table(table(don$Claim))
```

```
##
##         No        Yes
## 0.98536146 0.01463854
```

```
levels(don$Claim)
```

```
## [1] "No"  "Yes"
```

Le première modalité du facteur `Claim` est donc `Yes` et la deuxième est `No`, donc quand on ajustera la régression logistique le phénomène qu'on prédira est $P(Claim = "Yes"|X = x)$ où X représente l'ensemble des variables explicatives.

Q4 : Pouvez-vous donner une règle de classement qui a un taux de bon classement supérieur à 98%

R4 : Oui, il suffit de classer tous les individus dans la classe `No` (dans ce cas le taux de bon classement serait de 98,54%, la sensibilité de 0%, et la spécificité de 100%).

Q5 : En utilisant judicieusement les fonctions `sapply` et `nlevels` donner le nombre de modalités de chacune des variables. Que dire de la variable `Destination` ?

R5 :

```
sapply(don, nlevels)
```

```
##               Agency          Agency.Type Distribution.Channel
##                   16                    2                    2
##         Product.Name                Claim             Duration
##                   26                    2                    0
##          Destination            Net.Sales Commision..in.value.
##                  149                    0                    0
##               Gender                  Age
##                    2                    0
```

Ici la variable destination a 149 modalités, ce qui peut poser des problèmes ensuite dans les modèles prédictifs.

Q6 : Réaliser un test statistique permettant répondre à la question d'existance d'un lien entre la variable `Claim` et la variable `Agency`. Que conclure ?

R6 :

```
chisq.test(don$Claim, don$Agency)
```

```
## Warning in chisq.test(don$Claim, don$Agency): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  don$Claim and don$Agency
## X-squared = 1877, df = 15, p-value < 2.2e-16
```

On réalise un test du chi deux d'indépendance. Ici l'hypothèse nulle testée est l'indépendance des deux variables. Du fait de la sortie on rejetterai l'hypothèse nulle au risque $\alpha = 0,05$ (car p-value $<$ 2.2e-16). Les variables sont donc *significativement* dépendantes. Ce qui augure une prédictibilité de la variable `Claim` par rapport à la variable `Agency` :)

Q7 : Dans la continuité de la question précédente afficher le vecteur contenant `P(Claim = Yes |Agency = x)` pour chacune des valeurs `x` de la variable agency, et la stocker dans une variable nommée `vecteur_proba`.

R7 :

```
vecteur_proba = prop.table(table(don$Claim,don$Agency),2)[2,]
# save(vecteur_proba , file = "vecteur_proba.Rda")
```

Q8 : Les probabilités précédentes pourraient faire office de score, en associant à chaque individu la probabilité `P(Claim = Yes |Agency = x)` où `x` est la modalité dont dispose l'individu pour la variable

`Agency`. En utilisant de manière adéquate l'indexation par nom, et en lançant une commande du type `vecteur_proba[don$Agency]`, obtenir le vecteur contenant les probabilités pour chaque individu. Enfin traçer la courbe ROC associée à ce score.

R8 :

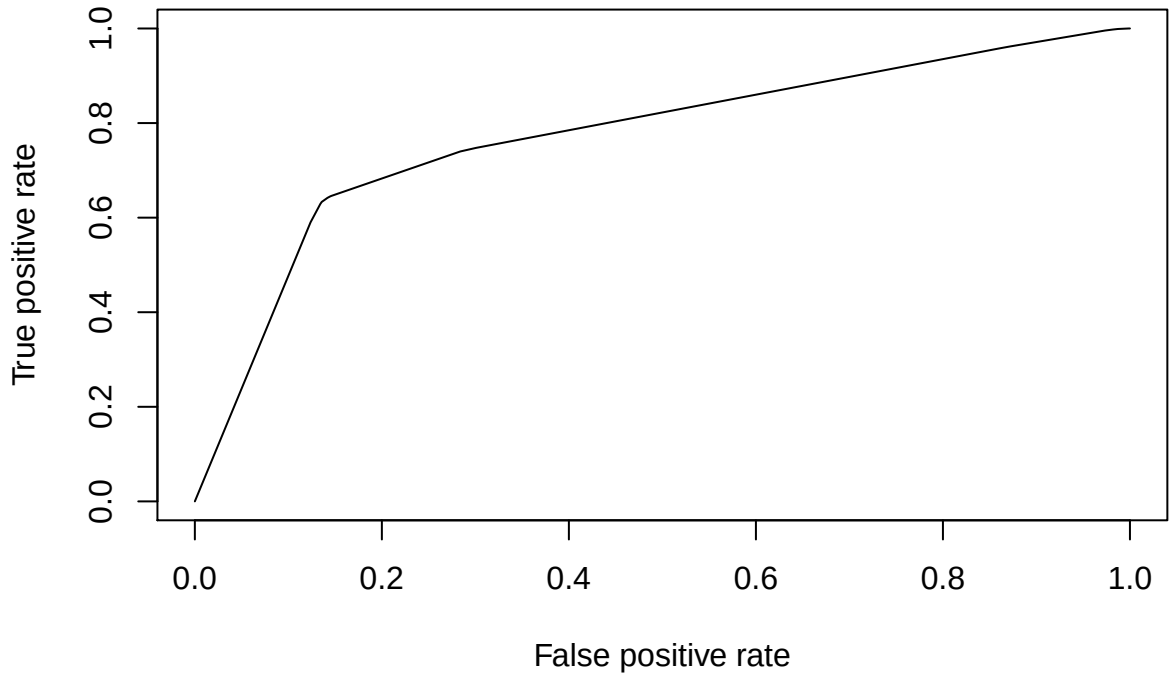
```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
Score = vecteur_proba[don$Agency] # Création du vecteur de score qui donne pour chaque individu P(Claim
pred = prediction(Score,don$Claim)
perf = performance(pred,"tpr","fpr")
plot(perf)
```



Q9 : A votre avis si on avait effectué une régression logistique en lançant : `glm(Claim ~ Agence, family = "binomial", data = don)`, aurait-on obtenu les mêmes résultats ? Justifier.

R9 : Oui en fait dans ce cas particulier d'une seule variable explicative binaire la régression logistique est équivalentes aux probabilités issue du tableau des profils colonnes.

Pour s'en convaincre on fait :

```
logit_agence = glm(Claim ~ Agency, family = "binomial", data = don)
vecteur_proba2 = predict(logit_agence,
                 newdata = data.frame(Agency = levels(don$Agency)),
                 type = "response")
library(ROCR)
rbind(vecteur_proba,vecteur_proba2)
```

```
##                          ADM         ART        C2B        CBH         CCR
## vecteur_proba  0.00000e+00 0.003021148 0.06616669 0.00990099 0.01546392
## vecteur_proba2 1.73677e-07 0.003021148 0.06616669 0.00990099 0.01546392
##                          CSR        CWT        EPX         JWT         JZI
## vecteur_proba  0.01162791 0.01002331 0.00555255 0.005340454 0.004898088
## vecteur_proba2 0.01162791 0.01002331 0.00555255 0.005340454 0.004898088
##                          KML        LWC        RAB         SSI         TST
## vecteur_proba  0.02040816 0.05224964 0.00137931 0.006628788 0.003787879
## vecteur_proba2 0.02040816 0.05224964 0.00137931 0.006628788 0.003787879
##                          TTW
## vecteur_proba  0.04081633
## vecteur_proba2 0.04081633
```

Ici les probabilité sont les mêmes à l'exception du cas ADM où on a 1.73e-07 contre 0, à relié peut-être à l'algorithme.

Q10 : On décide maintenant d'ajuster une régression logistique en prenant en compte toutes les variables. Après avoir lancé la commande adaptée on obtient les messages : a. `glm.fit: l'algorithme n'a pas convergé`, b. `glm.fit: des probabilités ont été ajustées numériquement à 0 ou 1` De plus en analysant les sorties c. `45107 observations deleted due to missingness`

Pour le message b. il s'agit du fait que dans la formule du calcul `P(Y = 1|X = x)` l'ordinateur indique que pour certains individus cette probabilité est ajusté numériquement à 0 ou 1. Ce n'est pas un problème en soi, mais cela peut parfois être un signe d'overfitting.

Pour le message a. pourquoi parle-t'on d'algorithme ici ?

Pour le message c. expliquer précisement pourquoi 45107 observation on été supprimée. Cela est-il bien grâve ?

R10 : message a. : ici l'algorithme utilisé est une algorithme iteratif (algorithme de Newton-Raphson), ici les message nous indique que le critère de convergence de l'algorithme n'est pas vérifié une fois le nombre maximal d'itération atteint. On pourrait l'éviter en augmentant le nombre l'obtenir en passant `maxit` de 25 (sa valeur par défaut) à une valeur plus élévée (par exemple 100).

message c. : Cela est dû 45107 valeurs manquantes pour la variable `Gender`, du coup cela conduit à se priver d'une bonne partie des données (45107/63326 = 71,2%). Il vaut mieux garder ces données, les valeurs manquantes pouvant être vues comme des valeurs particulières.

Q11 : On décide de récoder la variable `Gender` comme suit :

```
don$Gender = as.character(don$Gender)
don$Gender[is.na(don$Gender)] = "UNKNOWN"
don$Gender = factor(don$Gender)
table(don$Gender)
```

```
##
##       F       M UNKNOWN
##    8872    9347   45107
```

Que fait le code précédent, et quel intêret pour la suite peut-il bien avoir à effectuer ce recodage ?

R11 : Le code précédant permet de recoder les valeurs manquantes comme une modalité particulière (modalité UNKNOWN) de la variable `Gender`. Cela permet par la suite des prendre en compte ces individus dans le modèle. Ce codage est tout à fait conforme puisque le fait de ne pas connaitre le Genre du passager peut être une information à par entière pour expliquer la souscription ou non de l'assurance.

Q12 : On relance maintenant l'ajustement du modèle complet (le code peut maintenant mettre un peu de temps à tourner au plus 2 min, sauvegarder avant de lancer ...). Tracer la courbe ROC associée et donner l'AUC.

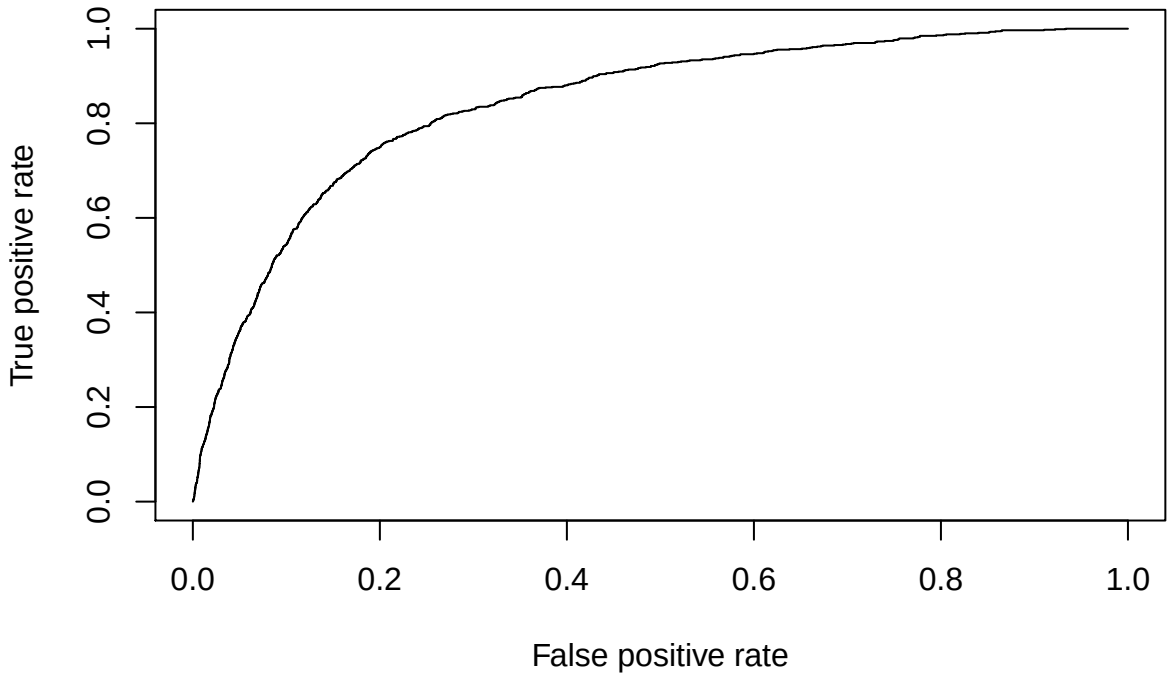
R12 :

```
mod_full = glm(Claim ~ ., family = "binomial", data = don)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
pred = prediction(mod_full$fitted.values,don$Claim)
perf = performance(pred,"tpr","fpr")
plot(perf) # Tracé de la courbe ROC
```




```
performance(pred,"auc")@y.values[[1]] # Valeur de l'AUC
```

```
## [1] 0.841652
```

Q13 : Ici les classes sont en effectifs très déséquilibrés. Les stratégies usuelles peuvent alors être : a. sous-échantillonner la classe en plus grand effectif b. sur-échantillonner la classe en plus faible effectif c. prendre directement en compte des poids dans l'ajustement du modèle, par exemple en prenant des poids inversement proportionnels à l'effectif des différentes classes

Les statégies a. et b. peuvent être mise en oeuvre en amont de l'ajustement du modèle, tandis que la stratégie c. nécessite l'utilisation de modèles capables de prendre en compte des poids.

D'un point de vue informatique la stratégie a. est la moins coûteuse en temps de calcul, mais elle conduit à perdre des données ... La stratégies b. conduirait à démultiplier le nombre de données ... Enfin la stratégie c. présenterait un coût informatique identique à l'ajustement du modèle classique (modula la prise en compte de poids)

En fait, quel intérêt peut-il bien avoir dans toutes ces stratégies par rapport à pas tenir compte du tout de ce déséquilibre ?

R13 : Ici l'intérêt peut être de repondérer l'influence des données du groupe `Yes` par rapport au groupe `No`. Pour que les données de la classe `Yes` ne se retrouvent pas noyées par rapport à la classe `No` dans la régle de décision.

Q14 : Ici, compte-tenu de la faible capacité informatique dont nous disposont pour aujourd'hui on opte pour

le sous-échantillonnage.

Fixer la graine du générateur de nombres pseudo-aléatoires à la valeur 1234.

Créer un nouveau jeu de données `don_us` comme "don under-sampled", constitué des 927 individus possédant la modalité `Yes` pour la variable `Claim` et de 927 individus tirés au hasard et sans remise parmi les 62399 individus possédant la modalité `No`. On pourra bien sûr, faire des selections pour distinguer les lignes avec `Yes` de celle avec `No`, utiliser la fonction `sample`, ainsi que la fonction`rbind` !

R14 :

```
set.seed(1234)
don_yes = don[don$Claim == "Yes",] # Individus avec Yes
don_no = don[don$Claim == "No",] # Ceux avec No
id_select_no = sample(nrow(don_no),nrow(don_yes)) # Choix au hasard des numéros d'individus parmi les `
# Autant que le nombre de `Yes`
don_us = rbind(don_yes, don_no[id_select_no,]) # Création du jeu de données don_us
save(don_us, file = "don_us.Rda")
```

Par la suite en cas de problème l'objet `don_us` est contenu dans le fichier `don_us.Rda` au besoin.

Q15 : Ajuster maintenant le modèle de régression logistique à partir de toutes les variables sur le data.frame `don_us`. En regardant maintenant la sortie on obtient le message : `Coefficients: (9 not defined because of singularities)`, cela indique en particulier que dans la matrice du modèle, certaines colonnes peuvent être déduites comme combinaison linéaire d'autres colonnes.

```
reg_us = glm(Claim ~ ., family = "binomial", data = don_us)
```

Si on stocke les résultats l'ajustement précédent dans le variable `reg_us`, on peut récupérer la matrice du modèle comme suit :

```
X = model.matrix(reg_us)
```

On s'intéresse aux valeurs propres de XX^T (matrice à inverser en régression linéaire, et à peu de chose prés celle qui nécessite d'être inversée en régression logistique ...)

```
eigen(t(X) %*% X)$values
```

```
##   [1]  5.567769e+07  3.578281e+06  1.788648e+06  4.245508e+05  2.009362e+03
##   [6]  8.382287e+02  3.706533e+02  3.289388e+02  3.064318e+02  2.051169e+02
##  [11]  1.964931e+02  1.305708e+02  1.142633e+02  1.006822e+02  9.213791e+01
##  [16]  7.669896e+01  7.302123e+01  7.024181e+01  6.520885e+01  5.804789e+01
##  [21]  5.473421e+01  4.908612e+01  4.266255e+01  3.914584e+01  3.625413e+01
##  [26]  3.382523e+01  3.341257e+01  3.129007e+01  3.061180e+01  2.668508e+01
##  [31]  2.422173e+01  2.197496e+01  1.867048e+01  1.763340e+01  1.745358e+01
##  [36]  1.640363e+01  1.592611e+01  1.519368e+01  1.439825e+01  1.362581e+01
##  [41]  1.228086e+01  1.187166e+01  1.074715e+01  1.034773e+01  1.001918e+01
##  [46]  9.406907e+00  9.011065e+00  8.093323e+00  7.192937e+00  6.996866e+00
##  [51]  6.669771e+00  6.092208e+00  5.959686e+00  5.527254e+00  5.277196e+00
##  [56]  5.041695e+00  4.173236e+00  3.959294e+00  3.761756e+00  3.607217e+00
##  [61]  3.164955e+00  2.991556e+00  2.658940e+00  2.386942e+00  2.306568e+00
##  [66]  2.100937e+00  2.020460e+00  1.999511e+00  1.986445e+00  1.977474e+00
##  [71]  1.945397e+00  1.881839e+00  1.361571e+00  1.275631e+00  1.000000e+00
##  [76]  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00
##  [81]  1.000000e+00  1.000000e+00  9.993129e-01  9.969264e-01  9.931176e-01
##  [86]  9.784362e-01  9.738629e-01  9.645210e-01  9.424387e-01  8.924343e-01
##  [91]  8.758045e-01  3.585840e-01  1.733535e-02  4.245747e-09  2.885059e-10
##  [96]  2.165821e-10  1.930250e-10  2.073652e-11 -9.333781e-11 -1.815484e-10
## [101] -1.560830e-09 -3.694498e-09
```

Que nous indique des valeurs propres trés faibles quand à notre objectif d'inversion de cette matrice ?

R15 : Ici on a certaines valeurs propres proches de 0, ce qui nous indique que la matrice est proche de la non-inversibilité, muticolinéarité des colonnes de X

Q16 : A l'aide de la commande suivante on réalise une sélection pas à pas :

```
min.model <- glm(Claim ~ 1, family = "binomial", data = don_us)
max.model <- glm(Claim ~ ., family = "binomial", data = don_us)
best.model = step(min.model, direction='both',
                  scope= list(lower = min.model,
                              upper = max.model))
```

```
## Start:  AIC=2572.19
## Claim ~ 1
##
##                         Df Deviance    AIC
## + Product.Name          23   1930.3 1978.3
## + Agency                14   2016.4 2046.4
## + Destination           56   2123.5 2237.5
## + Gender                 2   2270.1 2276.1
## + Net.Sales              1   2299.2 2303.2
## + Agency.Type            1   2325.1 2329.1
## + Commision..in.value.   1   2361.8 2365.8
## + Duration               1   2418.0 2422.0
## + Age                    1   2564.0 2568.0
## <none>                       2570.2 2572.2
## + Distribution.Channel   1   2569.0 2573.0
##
## Step:  AIC=1978.32
## Claim ~ Product.Name
##
##                         Df Deviance    AIC
## + Net.Sales              1   1885.5 1935.5
## + Commision..in.value.   1   1914.7 1964.7
## + Duration               1   1923.3 1973.3
## + Distribution.Channel   1   1927.2 1977.2
## <none>                       1930.3 1978.3
## + Agency.Type            1   1928.9 1978.9
## + Age                    1   1929.3 1979.3
## + Gender                 2   1927.7 1979.7
## + Destination           56   1822.0 1982.0
## + Agency                 6   1926.3 1986.3
## - Product.Name          23   2570.2 2572.2
##
## Step:  AIC=1935.55
## Claim ~ Product.Name + Net.Sales
##
##                         Df Deviance    AIC
## + Distribution.Channel   1   1881.6 1933.6
## + Gender                 2   1880.5 1934.5
## + Agency.Type            1   1883.3 1935.3
## + Duration               1   1883.4 1935.4
## <none>                       1885.5 1935.5
## + Age                    1   1884.1 1936.1
## + Commision..in.value.   1   1884.6 1936.6
```

```
## + Agency                6   1881.7 1943.7
## + Destination          56   1799.3 1961.3
## - Net.Sales             1   1930.3 1978.3
## - Product.Name         23   2299.2 2303.2
##
## Step:  AIC=1933.61
## Claim ~ Product.Name + Net.Sales + Distribution.Channel
##
##                        Df Deviance    AIC
## + Age                   1   1877.6 1931.6
## + Gender                2   1876.5 1932.5
## + Duration              1   1879.2 1933.2
## + Agency.Type           1   1879.3 1933.3
## <none>                      1881.6 1933.6
## + Commision..in.value.  1   1880.6 1934.6
## - Distribution.Channel  1   1885.5 1935.5
## + Agency                6   1877.8 1941.8
## + Destination          56   1795.2 1959.2
## - Net.Sales             1   1927.2 1977.2
## - Product.Name         23   2299.2 2305.2
##
## Step:  AIC=1931.56
## Claim ~ Product.Name + Net.Sales + Distribution.Channel + Age
##
##                        Df Deviance    AIC
## + Duration              1   1875.4 1931.4
## <none>                      1877.6 1931.6
## + Gender                2   1873.7 1931.7
## + Agency.Type           1   1876.0 1932.0
## + Commision..in.value.  1   1876.7 1932.7
## - Age                   1   1881.6 1933.6
## - Distribution.Channel  1   1884.1 1936.1
## + Agency                6   1873.9 1939.9
## + Destination          56   1792.0 1958.0
## - Net.Sales             1   1924.2 1976.2
## - Product.Name         23   2290.6 2298.6
##
## Step:  AIC=1931.39
## Claim ~ Product.Name + Net.Sales + Distribution.Channel + Age +
##     Duration
##
##                        Df Deviance    AIC
## <none>                      1875.4 1931.4
## - Duration              1   1877.6 1931.6
## + Gender                2   1871.7 1931.7
## + Agency.Type           1   1873.9 1931.9
## + Commision..in.value.  1   1874.5 1932.5
## - Age                   1   1879.2 1933.2
## - Distribution.Channel  1   1882.2 1936.2
## + Agency                6   1871.7 1939.7
## + Destination          56   1791.6 1959.6
## - Net.Sales             1   1917.0 1971.0
## - Product.Name         23   2287.6 2297.6
```

A quoi sert l'option direction = 'both' ? Quel intérêt dans la recherche peut-il y avoir à partir du modèle le plus simple ?

R16 : L'option direction = 'both' sert à faire la selection de variables forward-backward sur la base de l'optimisation du critère AIC. L'intérêt de commencer par le modèle le plus simple et d'éviter de passer par des modèles instables dans le processus d'optimisation de AIC (par exemple cas de variables avec trop de modalités).

Q17 : Interpréter le modèle retenu et évaluer ses performances (Courbe ROC, seuils effectuant le meilleur compromis sensibilité / spécificité, et sensibilité et spécificité associées)

R17 : Résumé du modèle : modèle retenu : `Claim ~ Product.Name + Net.Sales + Distribution.Channel + Age + Agency + Duration`

Résumé du modèle

```
summary(best.model)
```

```
##
## Call:
## glm(formula = Claim ~ Product.Name + Net.Sales + Distribution.Channel +
##     Age + Duration, family = "binomial", data = don_us)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1939  -0.8235   0.0000   0.7214   2.1958
##
## Coefficients:
##                                                Estimate Std. Error z value
## (Intercept)                                   3.338e-01  8.826e-01   0.378
## Product.Name2 way Comprehensive Plan          9.038e-01  3.935e-01   2.296
## Product.Name24 Protect                       -1.412e+01  7.183e+02  -0.020
## Product.NameAnnual Gold Plan                 -7.564e-01  9.956e-01  -0.760
## Product.NameAnnual Silver Plan                6.729e-01  7.023e-01   0.958
## Product.NameAnnual Travel Protect Gold        1.389e+01  4.510e+02   0.031
## Product.NameAnnual Travel Protect Platinum    1.299e+01  8.391e+02   0.015
## Product.NameAnnual Travel Protect Silver      1.433e+01  7.186e+02   0.020
## Product.NameBasic Plan                        2.771e-01  4.447e-01   0.623
## Product.NameBronze Plan                       2.838e+00  4.065e-01   6.981
## Product.NameCancellation Plan                -3.158e-01  4.117e-01  -0.767
## Product.NameComprehensive Plan               -8.259e-01  9.614e-01  -0.859
## Product.NameGold Plan                         2.051e+00  5.987e-01   3.426
## Product.NameIndividual Comprehensive Plan     1.371e+01  8.398e+02   0.016
## Product.NamePremier Plan                      1.314e+00  1.023e+00   1.285
## Product.NameRental Vehicle Excess Insurance   8.569e-01  4.078e-01   2.101
## Product.NameSilver Plan                       2.714e+00  4.256e-01   6.378
## Product.NameSingle Trip Travel Protect Gold   3.058e+00  8.656e-01   3.532
## Product.NameSingle Trip Travel Protect Platinum 1.700e+01  6.492e+02   0.026
## Product.NameSingle Trip Travel Protect Silver 1.464e+00  8.127e-01   1.802
## Product.NameSpouse or Parents Comprehensive Plan 1.386e+01  1.455e+03   0.010
## Product.NameTicket Protector                  9.300e-01  5.919e-01   1.571
## Product.NameTravel Cruise Protect            -2.049e+00  1.111e+00  -1.845
## Product.NameValue Plan                        6.718e-01  4.976e-01   1.350
## Net.Sales                                     1.177e-02  1.931e-03   6.095
## Distribution.ChannelOnline                   -1.949e+00  7.644e-01  -2.550
## Age                                          -8.935e-03  4.574e-03  -1.953
## Duration                                      1.924e-03  1.301e-03   1.479
```

```
##                                                Pr(>|z|)
## (Intercept)                                    0.705257
## Product.Name2 way Comprehensive Plan           0.021648 *
## Product.Name24 Protect                         0.984319
## Product.NameAnnual Gold Plan                   0.447410
## Product.NameAnnual Silver Plan                 0.338018
## Product.NameAnnual Travel Protect Gold         0.975427
## Product.NameAnnual Travel Protect Platinum     0.987649
## Product.NameAnnual Travel Protect Silver       0.984090
## Product.NameBasic Plan                         0.533275
## Product.NameBronze Plan                        2.93e-12 ***
## Product.NameCancellation Plan                  0.443115
## Product.NameComprehensive Plan                 0.390308
## Product.NameGold Plan                          0.000613 ***
## Product.NameIndividual Comprehensive Plan      0.986978
## Product.NamePremier Plan                       0.198772
## Product.NameRental Vehicle Excess Insurance    0.035621 *
## Product.NameSilver Plan                        1.79e-10 ***
## Product.NameSingle Trip Travel Protect Gold    0.000412 ***
## Product.NameSingle Trip Travel Protect Platinum 0.979112
## Product.NameSingle Trip Travel Protect Silver  0.071622 .
## Product.NameSpouse or Parents Comprehensive Plan 0.992404
## Product.NameTicket Protector                   0.116154
## Product.NameTravel Cruise Protect              0.065073 .
## Product.NameValue Plan                         0.176978
## Net.Sales                                      1.09e-09 ***
## Distribution.ChannelOnline                     0.010768 *
## Age                                            0.050761 .
## Duration                                       0.139102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2570.2  on 1853  degrees of freedom
## Residual deviance: 1875.4  on 1826  degrees of freedom
## AIC: 1931.4
##
## Number of Fisher Scoring iterations: 14
```

Problème, 8 valeurs estimées à NA pour les modalités de la variables Agency :

```
coeff_pb = names(which(is.na(best.model$coefficients)))
coeff_pb
```

```
## character(0)
```

On regarde si ces modalités sont présentes dans le jeu de données `don_us`:

```
table(don_us$Agency)
```

```
##
## ADM ART C2B CBH CCR CSR CWT EPX JWT JZI KML LWC RAB SSI TST TTW
##   0   6 663   4   8   2 206 710  15 134  15  42   8  24  13   4
```

Elles sont bien présentes !

On doit donc avoir un problème de multicolénarité entre cette varaible et d'autres variables du modèle. Par exemple, si on croise les variable Agency et Product.Name on obtient le tableau :

```
tab = table(don_us$Agency, don_us$Product.Name)
tab
```

```
##
##       1 way Comprehensive Plan 2 way Comprehensive Plan 24 Protect
##   ADM                        0                        0          0
##   ART                        0                        0          4
##   C2B                        0                        0          0
##   CBH                        0                        0          0
##   CCR                        0                        0          0
##   CSR                        0                        0          0
##   CWT                        0                        0          0
##   EPX                       50                      349          0
##   JWT                        0                        0          0
##   JZI                        0                        0          0
##   KML                        0                        0          0
##   LWC                        0                        0          0
##   RAB                        0                        0          0
##   SSI                        0                        0          0
##   TST                        0                        0          0
##   TTW                        0                        0          0
##
##       Annual Gold Plan Annual Silver Plan Annual Travel Protect Gold
##   ADM                0                  0                          0
##   ART                0                  0                          0
##   C2B               24                176                          0
##   CBH                0                  0                          0
##   CCR                0                  0                          0
##   CSR                0                  0                          0
##   CWT                0                  0                          0
##   EPX                0                  0                          0
##   JWT                0                  0                          0
##   JZI                0                  0                          0
##   KML                0                  0                          0
##   LWC                0                  0                         10
##   RAB                0                  0                          0
##   SSI                0                  0                          0
##   TST                0                  0                          0
##   TTW                0                  0                          0
##
##       Annual Travel Protect Platinum Annual Travel Protect Silver Basic Plan
##   ADM                              0                            0          0
##   ART                              0                            0          0
##   C2B                              0                            0          0
##   CBH                              0                            0          0
##   CCR                              0                            0          0
##   CSR                              0                            0          0
##   CWT                              0                            0          0
##   EPX                              0                            0          0
##   JWT                              0                            0          0
##   JZI                              0                            0        112
##   KML                              0                            0          0
```

```
##   LWC                          3                          4          0
##   RAB                          0                          0          0
##   SSI                          0                          0          0
##   TST                          0                          0          0
##   TTW                          0                          0          0
##
##      Bronze Plan Cancellation Plan Child Comprehensive Plan Comprehensive Plan
##   ADM           0                 0                        0                   0
##   ART           0                 0                        0                   0
##   C2B         264                 0                        0                   0
##   CBH           0                 0                        0                   4
##   CCR           0                 0                        0                   8
##   CSR           0                 0                        0                   2
##   CWT           0                 0                        0                   0
##   EPX           0               311                        0                   0
##   JWT           0                 0                        0                   0
##   JZI           0                 0                        0                   0
##   KML           0                 0                        0                   0
##   LWC           0                 0                        0                   0
##   RAB           0                 0                        0                   0
##   SSI           0                 0                        0                   0
##   TST           0                 0                        0                   0
##   TTW           0                 0                        0                   0
##
##      Gold Plan Individual Comprehensive Plan Premier Plan
##   ADM         0                           0            0
##   ART         0                           0            0
##   C2B        25                           0            0
##   CBH         0                           0            0
##   CCR         0                           0            0
##   CSR         0                           0            0
##   CWT         0                           0            0
##   EPX         0                           0            0
##   JWT         0                           0            0
##   JZI         0                           0            0
##   KML         0                           0            5
##   LWC         0                           0            0
##   RAB         0                           0            0
##   SSI         0                           0            0
##   TST         0                           0            0
##   TTW         0                           3            0
##
##      Rental Vehicle Excess Insurance Silver Plan
##   ADM                               0           0
##   ART                               0           0
##   C2B                               0         174
##   CBH                               0           0
##   CCR                               0           0
##   CSR                               0           0
##   CWT                             206           0
##   EPX                               0           0
##   JWT                               0           0
##   JZI                               0           0
##   KML                               0           0
```

```
##   LWC                                0           0
##   RAB                                0           0
##   SSI                                0           0
##   TST                                0           0
##   TTW                                0           0
##
##       Single Trip Travel Protect Gold Single Trip Travel Protect Platinum
##   ADM                                0                                    0
##   ART                                0                                    0
##   C2B                                0                                    0
##   CBH                                0                                    0
##   CCR                                0                                    0
##   CSR                                0                                    0
##   CWT                                0                                    0
##   EPX                                0                                    0
##   JWT                                0                                    0
##   JZI                                0                                    0
##   KML                                0                                    0
##   LWC                               12                                    5
##   RAB                                0                                    0
##   SSI                                0                                    0
##   TST                                0                                    0
##   TTW                                0                                    0
##
##       Single Trip Travel Protect Silver Spouse or Parents Comprehensive Plan
##   ADM                                0                                     0
##   ART                                0                                     0
##   C2B                                0                                     0
##   CBH                                0                                     0
##   CCR                                0                                     0
##   CSR                                0                                     0
##   CWT                                0                                     0
##   EPX                                0                                     0
##   JWT                                0                                     0
##   JZI                                0                                     0
##   KML                                0                                     0
##   LWC                                8                                     0
##   RAB                                0                                     0
##   SSI                                0                                     0
##   TST                                0                                     0
##   TTW                                0                                     1
##
##       Ticket Protector Travel Cruise Protect Travel Cruise Protect Family
##   ADM                0                    0                             0
##   ART                0                    0                             0
##   C2B                0                    0                             0
##   CBH                0                    0                             0
##   CCR                0                    0                             0
##   CSR                0                    0                             0
##   CWT                0                    0                             0
##   EPX                0                    0                             0
##   JWT                0                    0                             0
##   JZI                0                    0                             0
##   KML                0                    0                             0
```

```
##   LWC                0                    0                          0
##   RAB                0                    0                          0
##   SSI               24                    0                          0
##   TST                0                   13                          0
##   TTW                0                    0                          0
##
##       Value Plan
##   ADM          0
##   ART          2
##   C2B          0
##   CBH          0
##   CCR          0
##   CSR          0
##   CWT          0
##   EPX          0
##   JWT         15
##   JZI         22
##   KML         10
##   LWC          0
##   RAB          8
##   SSI          0
##   TST          0
##   TTW          0
```

On peut pour mieux comprendre les choses on peut afficher la liste des agences pour lesquelles la connaissance du produit implique la connaissance de l'agence (profils colonnes égaux à 1).

```
id = which((prop.table(tab,2) == 1), arr.ind = TRUE)
```

On obtient la liste ci-dessous :

```
cbind(rownames(tab)[id[,1]],colnames(tab)[id[,2]])
```

```
##        [,1]  [,2]
##  [1,] "EPX" "1 way Comprehensive Plan"
##  [2,] "EPX" "2 way Comprehensive Plan"
##  [3,] "ART" "24 Protect"
##  [4,] "C2B" "Annual Gold Plan"
##  [5,] "C2B" "Annual Silver Plan"
##  [6,] "LWC" "Annual Travel Protect Gold"
##  [7,] "LWC" "Annual Travel Protect Platinum"
##  [8,] "LWC" "Annual Travel Protect Silver"
##  [9,] "JZI" "Basic Plan"
## [10,] "C2B" "Bronze Plan"
## [11,] "EPX" "Cancellation Plan"
## [12,] "C2B" "Gold Plan"
## [13,] "TTW" "Individual Comprehensive Plan"
## [14,] "KML" "Premier Plan"
## [15,] "CWT" "Rental Vehicle Excess Insurance"
## [16,] "C2B" "Silver Plan"
## [17,] "LWC" "Single Trip Travel Protect Gold"
## [18,] "LWC" "Single Trip Travel Protect Platinum"
## [19,] "LWC" "Single Trip Travel Protect Silver"
## [20,] "TTW" "Spouse or Parents Comprehensive Plan"
## [21,] "SSI" "Ticket Protector"
## [22,] "TST" "Travel Cruise Protect"
```

Par exemple sachant que le produit est 1 way Comprehansive Plan, on sait qu'il s'agit de l'agence EPX.

On peut aussi afficher la liste des agences concernées :

```
sort(unique(rownames(tab)[id[,1]]))
```

```
##  [1] "ART" "C2B" "CWT" "EPX" "JZI" "KML" "LWC" "SSI" "TST" "TTW"
```

On retrouve bien la liste des agences concernée par le problème d'estimation des coefficients, modulo ART qui servait de modalité de référence pour la variable ART et qui n'apparait donc naturellement pas :

```
coeff_pb
```

```
## character(0)
```

Ainsi pour ces 8 agences, la connaissance du produit implique la connaissance de l'agence, l'information agence étant alors redondante par rapport à l'information produit. Dans ce cas R décide de ne pas estimer ces coefficients.

On pourrait aussi croiser maintenant la variable avec `Distribution.Channel` :

```
tab2 = table(don_us$Agency, don_us$Distribution.Channel)
tab2
```

```
##
##       Offline Online
##   ADM       0      0
##   ART       0      6
##   C2B       0    663
##   CBH       4      0
##   CCR       8      0
##   CSR       2      0
##   CWT       0    206
##   EPX       8    702
##   JWT       0     15
##   JZI       0    134
##   KML       1     14
##   LWC       0     42
##   RAB       0      8
##   SSI       1     23
##   TST      13      0
##   TTW       4      0
```

On remarque en fait qu'ici la connaissane de l'agence implique de façon quasi-déterministe la connaissance du réseau de distribution. Ainsi il est trés étonnant que cette variable ne soit pas supprimée . . . , peut être du au problèmes d'ajustement du modèle déjà rencontrés avant.

En pratique ce qui est fait lorsque deux variables sont trés dépendantes peut consister à créer une nouvelle variable à partir des deux variables précédentes :

```
library(tidyr) # pratique pour le nettoyage de données
don_us2 = don_us %>% unite(Agency.Product,Agency,Product.Name)
table(don_us2$Agency.Product)
```

```
##
##                           ART_24 Protect
##                                        4
##                           ART_Value Plan
##                                        2
##                     C2B_Annual Gold Plan
```

```
##                                     24
##                   C2B_Annual Silver Plan
##                                    176
##                          C2B_Bronze Plan
##                                    264
##                            C2B_Gold Plan
##                                     25
##                          C2B_Silver Plan
##                                    174
##                   CBH_Comprehensive Plan
##                                      4
##                   CCR_Comprehensive Plan
##                                      8
##                   CSR_Comprehensive Plan
##                                      2
##      CWT_Rental Vehicle Excess Insurance
##                                    206
##             EPX_1 way Comprehensive Plan
##                                     50
##             EPX_2 way Comprehensive Plan
##                                    349
##                    EPX_Cancellation Plan
##                                    311
##                           JWT_Value Plan
##                                     15
##                           JZI_Basic Plan
##                                    112
##                           JZI_Value Plan
##                                     22
##                         KML_Premier Plan
##                                      5
##                           KML_Value Plan
##                                     10
##           LWC_Annual Travel Protect Gold
##                                     10
##       LWC_Annual Travel Protect Platinum
##                                      3
##         LWC_Annual Travel Protect Silver
##                                      4
##      LWC_Single Trip Travel Protect Gold
##                                     12
##  LWC_Single Trip Travel Protect Platinum
##                                      5
##    LWC_Single Trip Travel Protect Silver
##                                      8
##                           RAB_Value Plan
##                                      8
##                     SSI_Ticket Protector
##                                     24
##                TST_Travel Cruise Protect
##                                     13
##        TTW_Individual Comprehensive Plan
##                                      3
## TTW_Spouse or Parents Comprehensive Plan
```

```
##                                         1
```

Le regression logistique ne posant maintenant plus de problème :

```
min.model <- glm(Claim ~ 1, family = "binomial", data = don_us2)
max.model <- glm(Claim ~ ., family = "binomial", data = don_us2)
best.model2 <- step(min.model, direction='both',
                    scope= list(lower = min.model,
                                upper = max.model))
```

```
## Start:  AIC=2572.19
## Claim ~ 1
##
##                         Df Deviance    AIC
## + Agency.Product        29   1926.3 1986.3
## + Destination           56   2123.5 2237.5
## + Gender                 2   2270.1 2276.1
## + Net.Sales              1   2299.2 2303.2
## + Agency.Type            1   2325.1 2329.1
## + Commision..in.value.   1   2361.8 2365.8
## + Duration               1   2418.0 2422.0
## + Age                    1   2564.0 2568.0
## <none>                       2570.2 2572.2
## + Distribution.Channel   1   2569.0 2573.0
##
## Step:  AIC=1986.26
## Claim ~ Agency.Product
##
##                         Df Deviance    AIC
## + Net.Sales              1   1881.7 1943.7
## + Commision..in.value.   1   1910.9 1972.9
## + Duration               1   1919.6 1981.6
## + Distribution.Channel   1   1923.1 1985.1
## <none>                       1926.3 1986.3
## + Age                    1   1925.4 1987.4
## + Gender                 2   1923.9 1987.9
## + Destination           56   1816.8 1988.8
## - Agency.Product        29   2570.2 2572.2
##
## Step:  AIC=1943.7
## Claim ~ Agency.Product + Net.Sales
##
##                         Df Deviance    AIC
## + Distribution.Channel   1   1877.8 1941.8
## + Gender                 2   1877.5 1943.5
## <none>                       1881.7 1943.7
## + Duration               1   1879.7 1943.7
## + Age                    1   1880.6 1944.6
## + Commision..in.value.   1   1880.8 1944.8
## + Destination           56   1794.2 1968.2
## - Net.Sales              1   1926.3 1986.3
## - Agency.Product        29   2299.2 2303.2
##
## Step:  AIC=1941.76
## Claim ~ Agency.Product + Net.Sales + Distribution.Channel
```

```
##
##                         Df Deviance    AIC
## + Age                    1   1873.9 1939.9
## + Duration               1   1875.5 1941.5
## + Gender                 2   1873.5 1941.5
## <none>                       1877.8 1941.8
## + Commision..in.value.   1   1876.8 1942.8
## - Distribution.Channel   1   1881.7 1943.7
## + Destination           56   1790.0 1966.0
## - Net.Sales              1   1923.1 1985.1
## - Agency.Product        29   2299.2 2305.2
##
## Step:  AIC=1939.88
## Claim ~ Agency.Product + Net.Sales + Distribution.Channel + Age
##
##                         Df Deviance    AIC
## + Duration               1   1871.7 1939.7
## <none>                       1873.9 1939.9
## + Gender                 2   1869.9 1939.9
## + Commision..in.value.   1   1873.0 1941.0
## - Age                    1   1877.8 1941.8
## - Distribution.Channel   1   1880.6 1944.6
## + Destination           56   1785.2 1963.2
## - Net.Sales              1   1920.0 1984.0
## - Agency.Product        29   2290.6 2298.6
##
## Step:  AIC=1939.7
## Claim ~ Agency.Product + Net.Sales + Distribution.Channel + Age +
##     Duration
##
##                         Df Deviance    AIC
## <none>                       1871.7 1939.7
## - Duration               1   1873.9 1939.9
## + Gender                 2   1867.9 1939.9
## + Commision..in.value.   1   1870.8 1940.8
## - Age                    1   1875.5 1941.5
## - Distribution.Channel   1   1878.7 1944.7
## + Destination           56   1784.9 1964.9
## - Net.Sales              1   1913.0 1979.0
## - Agency.Product        29   2287.6 2297.6
```

```
coefficients(best.model2)
```

```
##                                          (Intercept)
##                                          -13.65209379
##                           Agency.ProductART_Value Plan
##                                          16.05790853
##                        Agency.ProductC2B_Annual Gold Plan
##                                          13.33617653
##                      Agency.ProductC2B_Annual Silver Plan
##                                          14.76694590
##                             Agency.ProductC2B_Bronze Plan
##                                          16.94390989
##                               Agency.ProductC2B_Gold Plan
##                                          16.16824395
```

```
##                          Agency.ProductC2B_Silver Plan
##                                             16.81926271
##                    Agency.ProductCBH_Comprehensive Plan
##                                             12.57214111
##                    Agency.ProductCCR_Comprehensive Plan
##                                             13.36736296
##                    Agency.ProductCSR_Comprehensive Plan
##                                             13.87527954
##      Agency.ProductCWT_Rental Vehicle Excess Insurance
##                                             14.96234739
##              Agency.ProductEPX_1 way Comprehensive Plan
##                                             14.10469061
##              Agency.ProductEPX_2 way Comprehensive Plan
##                                             15.00796191
##                     Agency.ProductEPX_Cancellation Plan
##                                             13.78830827
##                            Agency.ProductJWT_Value Plan
##                                             14.91987099
##                            Agency.ProductJZI_Basic Plan
##                                             14.39015788
##                            Agency.ProductJZI_Value Plan
##                                             14.51826655
##                          Agency.ProductKML_Premier Plan
##                                             15.42031281
##                            Agency.ProductKML_Value Plan
##                                             15.50510008
##           Agency.ProductLWC_Annual Travel Protect Gold
##                                             27.98510152
##       Agency.ProductLWC_Annual Travel Protect Platinum
##                                             27.07815081
##         Agency.ProductLWC_Annual Travel Protect Silver
##                                             28.42267810
##      Agency.ProductLWC_Single Trip Travel Protect Gold
##                                             17.16773642
##  Agency.ProductLWC_Single Trip Travel Protect Platinum
##                                             31.11195548
##    Agency.ProductLWC_Single Trip Travel Protect Silver
##                                             15.57578988
##                            Agency.ProductRAB_Value Plan
##                                             13.89977201
##                      Agency.ProductSSI_Ticket Protector
##                                             15.04992798
##                Agency.ProductTST_Travel Cruise Protect
##                                             11.98667753
##        Agency.ProductTTW_Individual Comprehensive Plan
##                                             27.72520228
## Agency.ProductTTW_Spouse or Parents Comprehensive Plan
##                                             27.87724593
##                                               Net.Sales
##                                              0.01181936
##                               Distribution.ChannelOnline
##                                             -2.02669834
##                                                     Age
##                                             -0.01012792
```

```
##                                              Duration
##                                            0.00193801
```

```
formula(best.model2)
```

```
## Claim ~ Agency.Product + Net.Sales + Distribution.Channel + Age +
##     Duration
```

avec des coefficients ayants tous les coefficient avec des valeurs estimées différentes de NA.

On peut aussi regarder le résummé du modèle :

```
summary(best.model2)
```

```
##
## Call:
## glm(formula = Claim ~ Agency.Product + Net.Sales + Distribution.Channel +
##     Age + Duration, family = "binomial", data = don_us2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2020  -0.8139   0.0000   0.7203   2.1967
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                      -1.365e+01  7.183e+02
## Agency.ProductART_Value Plan                      1.606e+01  7.183e+02
## Agency.ProductC2B_Annual Gold Plan                1.334e+01  7.183e+02
## Agency.ProductC2B_Annual Silver Plan              1.477e+01  7.183e+02
## Agency.ProductC2B_Bronze Plan                     1.694e+01  7.183e+02
## Agency.ProductC2B_Gold Plan                       1.617e+01  7.183e+02
## Agency.ProductC2B_Silver Plan                     1.682e+01  7.183e+02
## Agency.ProductCBH_Comprehensive Plan              1.257e+01  7.183e+02
## Agency.ProductCCR_Comprehensive Plan              1.337e+01  7.183e+02
## Agency.ProductCSR_Comprehensive Plan              1.388e+01  7.183e+02
## Agency.ProductCWT_Rental Vehicle Excess Insurance 1.496e+01  7.183e+02
## Agency.ProductEPX_1 way Comprehensive Plan        1.410e+01  7.183e+02
## Agency.ProductEPX_2 way Comprehensive Plan        1.501e+01  7.183e+02
## Agency.ProductEPX_Cancellation Plan               1.379e+01  7.183e+02
## Agency.ProductJWT_Value Plan                      1.492e+01  7.183e+02
## Agency.ProductJZI_Basic Plan                      1.439e+01  7.183e+02
## Agency.ProductJZI_Value Plan                      1.452e+01  7.183e+02
## Agency.ProductKML_Premier Plan                    1.542e+01  7.183e+02
## Agency.ProductKML_Value Plan                      1.551e+01  7.183e+02
## Agency.ProductLWC_Annual Travel Protect Gold      2.799e+01  8.481e+02
## Agency.ProductLWC_Annual Travel Protect Platinum  2.708e+01  1.104e+03
## Agency.ProductLWC_Annual Travel Protect Silver    2.842e+01  1.016e+03
## Agency.ProductLWC_Single Trip Travel Protect Gold 1.717e+01  7.183e+02
## Agency.ProductLWC_Single Trip Travel Protect Platinum 3.111e+01  9.681e+02
## Agency.ProductLWC_Single Trip Travel Protect Silver 1.558e+01  7.183e+02
## Agency.ProductRAB_Value Plan                      1.390e+01  7.183e+02
## Agency.ProductSSI_Ticket Protector                1.505e+01  7.183e+02
## Agency.ProductTST_Travel Cruise Protect           1.199e+01  7.183e+02
## Agency.ProductTTW_Individual Comprehensive Plan   2.773e+01  1.105e+03
## Agency.ProductTTW_Spouse or Parents Comprehensive Plan 2.788e+01  1.623e+03
## Net.Sales                                         1.182e-02  1.947e-03
```

```
## Distribution.ChannelOnline                                  -2.027e+00  7.843e-01
## Age                                                         -1.013e-02  5.197e-03
## Duration                                                     1.938e-03  1.303e-03
##                                                             z value Pr(>|z|)
## (Intercept)                                                  -0.019  0.98484
## Agency.ProductART_Value Plan                                  0.022  0.98216
## Agency.ProductC2B_Annual Gold Plan                            0.019  0.98519
## Agency.ProductC2B_Annual Silver Plan                          0.021  0.98360
## Agency.ProductC2B_Bronze Plan                                 0.024  0.98118
## Agency.ProductC2B_Gold Plan                                   0.023  0.98204
## Agency.ProductC2B_Silver Plan                                 0.023  0.98132
## Agency.ProductCBH_Comprehensive Plan                          0.018  0.98603
## Agency.ProductCCR_Comprehensive Plan                          0.019  0.98515
## Agency.ProductCSR_Comprehensive Plan                          0.019  0.98459
## Agency.ProductCWT_Rental Vehicle Excess Insurance             0.021  0.98338
## Agency.ProductEPX_1 way Comprehensive Plan                    0.020  0.98433
## Agency.ProductEPX_2 way Comprehensive Plan                    0.021  0.98333
## Agency.ProductEPX_Cancellation Plan                           0.019  0.98468
## Agency.ProductJWT_Value Plan                                  0.021  0.98343
## Agency.ProductJZI_Basic Plan                                  0.020  0.98402
## Agency.ProductJZI_Value Plan                                  0.020  0.98387
## Agency.ProductKML_Premier Plan                                0.021  0.98287
## Agency.ProductKML_Value Plan                                  0.022  0.98278
## Agency.ProductLWC_Annual Travel Protect Gold                  0.033  0.97368
## Agency.ProductLWC_Annual Travel Protect Platinum              0.025  0.98044
## Agency.ProductLWC_Annual Travel Protect Silver                0.028  0.97768
## Agency.ProductLWC_Single Trip Travel Protect Gold             0.024  0.98093
## Agency.ProductLWC_Single Trip Travel Protect Platinum         0.032  0.97436
## Agency.ProductLWC_Single Trip Travel Protect Silver           0.022  0.98270
## Agency.ProductRAB_Value Plan                                  0.019  0.98456
## Agency.ProductSSI_Ticket Protector                            0.021  0.98328
## Agency.ProductTST_Travel Cruise Protect                       0.017  0.98668
## Agency.ProductTTW_Individual Comprehensive Plan               0.025  0.97998
## Agency.ProductTTW_Spouse or Parents Comprehensive Plan        0.017  0.98630
## Net.Sales                                                     6.069 1.29e-09 ***
## Distribution.ChannelOnline                                   -2.584  0.00977 **
## Age                                                          -1.949  0.05130 .
## Duration                                                      1.488  0.13682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2570.2  on 1853  degrees of freedom
## Residual deviance: 1871.7  on 1820  degrees of freedom
## AIC: 1939.7
##
## Number of Fisher Scoring iterations: 14
```

Ici les coefficients dans l'ensemble restent plutôt trés mals estimés, grande variance des estimateurs, du surement au grand nombre de modalités. Et donc peut de coefficients apparaissent comme significatifs.

Dans ce cas on pourrait essayer de réduire le nombre de modalités de la variable Agency.Product créée, par exemple sur la base de test du chi-2 (similaire aux arbre CHAID)

```r
library(CHAID)
```

```
## Loading required package: partykit
```

```
## Loading required package: grid
```

```
## Loading required package: libcoin
```

```
## Loading required package: mvtnorm
```

```r
# On crée un arbre CHAID avec uniquement la variable Agency.Product
# recodée en qualitative
chaid_tree <- chaid(Claim ~ as.factor(Agency.Product), don_us2,
      control = chaid_control(alpha2 = 0.05))
# On recupères les feuilles associées à chacune des données
feuilles_chaid <- names(predict(chaid_tree))
# Distribution des effectifs associés à chacun des classements
table(feuilles_chaid)
```

```
## feuilles_chaid
##  10  13  14  15  16  17   3   5   6   7   8   9
##  34 178 349 206 311 175 114 188 264  15  12   8
```

```r
# Croisement avec la variable Agency.Product
table(feuilles_chaid, don_us2$Agency.Product)
```

```
##
## feuilles_chaid ART_24 Protect ART_Value Plan C2B_Annual Gold Plan
##             10                0              0                   24
##             13                0              0                    0
##             14                0              0                    0
##             15                0              0                    0
##             16                0              0                    0
##             17                0              0                    0
##             3                 4              2                    0
##             5                 0              0                    0
##             6                 0              0                    0
##             7                 0              0                    0
##             8                 0              0                    0
##             9                 0              0                    0
##
## feuilles_chaid C2B_Annual Silver Plan C2B_Bronze Plan C2B_Gold Plan
##             10                      0               0             0
##             13                      0               0             0
##             14                      0               0             0
##             15                      0               0             0
##             16                      0               0             0
##             17                      0               0             0
##             3                       0               0            25
##             5                     176               0             0
##             6                       0             264             0
##             7                       0               0             0
##             8                       0               0             0
##             9                       0               0             0
##
## feuilles_chaid C2B_Silver Plan CBH_Comprehensive Plan CCR_Comprehensive Plan
##             10               0                      0                      0
```

```
##             13             174                     0                     0
##             14               0                     0                     0
##             15               0                     0                     0
##             16               0                     0                     0
##             17               0                     0                     0
##             3                0                     4                     8
##             5                0                     0                     0
##             6                0                     0                     0
##             7                0                     0                     0
##             8                0                     0                     0
##             9                0                     0                     0
##
## feuilles_chaid CSR_Comprehensive Plan CWT_Rental Vehicle Excess Insurance
##             10                       0                                   0
##             13                       0                                   0
##             14                       0                                   0
##             15                       0                                 206
##             16                       0                                   0
##             17                       0                                   0
##             3                        2                                   0
##             5                        0                                   0
##             6                        0                                   0
##             7                        0                                   0
##             8                        0                                   0
##             9                        0                                   0
##
## feuilles_chaid EPX_1 way Comprehensive Plan EPX_2 way Comprehensive Plan
##             10                            0                            0
##             13                            0                            0
##             14                            0                          349
##             15                            0                            0
##             16                            0                            0
##             17                           50                            0
##             3                             0                            0
##             5                             0                            0
##             6                             0                            0
##             7                             0                            0
##             8                             0                            0
##             9                             0                            0
##
## feuilles_chaid EPX_Cancellation Plan JWT_Value Plan JZI_Basic Plan
##             10                     0              0              0
##             13                     0              0              0
##             14                     0              0              0
##             15                     0              0              0
##             16                   311              0              0
##             17                     0              0            112
##             3                      0              0              0
##             5                      0              0              0
##             6                      0              0              0
##             7                      0             15              0
##             8                      0              0              0
##             9                      0              0              0
##
```

```
## feuilles_chaid JZI_Value Plan KML_Premier Plan KML_Value Plan
##             10               0                 0               0
##             13               0                 0               0
##             14               0                 0               0
##             15               0                 0               0
##             16               0                 0               0
##             17               0                 0               0
##             3               22                 5              10
##             5                0                 0               0
##             6                0                 0               0
##             7                0                 0               0
##             8                0                 0               0
##             9                0                 0               0
##
## feuilles_chaid LWC_Annual Travel Protect Gold
##             10                              10
##             13                               0
##             14                               0
##             15                               0
##             16                               0
##             17                               0
##             3                                0
##             5                                0
##             6                                0
##             7                                0
##             8                                0
##             9                                0
##
## feuilles_chaid LWC_Annual Travel Protect Platinum
##             10                                   0
##             13                                   0
##             14                                   0
##             15                                   0
##             16                                   0
##             17                                   0
##             3                                    0
##             5                                    0
##             6                                    0
##             7                                    0
##             8                                    3
##             9                                    0
##
## feuilles_chaid LWC_Annual Travel Protect Silver
##             10                                 0
##             13                                 0
##             14                                 0
##             15                                 0
##             16                                 0
##             17                                 0
##             3                                  0
##             5                                  0
##             6                                  0
##             7                                  0
##             8                                  4
```

```
##             9                                    0
##
## feuilles_chaid LWC_Single Trip Travel Protect Gold
##             10                                    0
##             13                                    0
##             14                                    0
##             15                                    0
##             16                                    0
##             17                                    0
##             3                                     0
##             5                                    12
##             6                                     0
##             7                                     0
##             8                                     0
##             9                                     0
##
## feuilles_chaid LWC_Single Trip Travel Protect Platinum
##             10                                        0
##             13                                        0
##             14                                        0
##             15                                        0
##             16                                        0
##             17                                        0
##             3                                         0
##             5                                         0
##             6                                         0
##             7                                         0
##             8                                         5
##             9                                         0
##
## feuilles_chaid LWC_Single Trip Travel Protect Silver RAB_Value Plan
##             10                                       0              0
##             13                                       0              0
##             14                                       0              0
##             15                                       0              0
##             16                                       0              0
##             17                                       0              0
##             3                                        8              0
##             5                                        0              0
##             6                                        0              0
##             7                                        0              0
##             8                                        0              0
##             9                                        0              8
##
## feuilles_chaid SSI_Ticket Protector TST_Travel Cruise Protect
##             10                     0                        0
##             13                     0                        0
##             14                     0                        0
##             15                     0                        0
##             16                     0                        0
##             17                     0                       13
##             3                     24                        0
##             5                      0                        0
##             6                      0                        0
```

```
##             7                     0                          0
##             8                     0                          0
##             9                     0                          0
##
## feuilles_chaid TTW_Individual Comprehensive Plan
##             10                                  0
##             13                                  3
##             14                                  0
##             15                                  0
##             16                                  0
##             17                                  0
##             3                                   0
##             5                                   0
##             6                                   0
##             7                                   0
##             8                                   0
##             9                                   0
##
## feuilles_chaid TTW_Spouse or Parents Comprehensive Plan
##             10                                          0
##             13                                          1
##             14                                          0
##             15                                          0
##             16                                          0
##             17                                          0
##             3                                           0
##             5                                           0
##             6                                           0
##             7                                           0
##             8                                           0
##             9                                           0
```

```
# On crée un nouveau data.frame en remplaçant Agency.Product par les modalités regroupées
don_us3 <- don_us2
don_us3 <- don_us3 %>% dplyr::select(-Agency.Product) %>%
  dplyr::mutate(Agency.Product.Grouped = factor(feuilles_chaid))
```

Enfin on peut réajuster le modèle de régression logistique :

```
min.model3 <- glm(Claim ~ 1, family = "binomial", data = don_us3)
max.model3 <- glm(Claim ~ ., family = "binomial", data = don_us3)
best.model3 <- step(min.model3, direction='both',
                    scope= list(lower = min.model3,
                                upper = max.model3))
```

```
## Start:  AIC=2572.19
## Claim ~ 1
##
##                           Df Deviance    AIC
## + Agency.Product.Grouped 11   1947.8 1971.8
## + Destination            56   2123.5 2237.5
## + Gender                  2   2270.1 2276.1
## + Net.Sales               1   2299.2 2303.2
## + Agency.Type             1   2325.1 2329.1
## + Commision..in.value.    1   2361.8 2365.8
```

```
## + Duration                1   2418.0 2422.0
## + Age                     1   2564.0 2568.0
## <none>                        2570.2 2572.2
## + Distribution.Channel    1   2569.0 2573.0
##
## Step:  AIC=1971.78
## Claim ~ Agency.Product.Grouped
##
##                          Df Deviance    AIC
## + Net.Sales               1   1904.9 1930.9
## + Commision..in.value.    1   1929.0 1955.0
## + Duration                1   1940.7 1966.7
## <none>                        1947.8 1971.8
## + Age                     1   1946.6 1972.6
## + Gender                  2   1944.8 1972.8
## + Distribution.Channel    1   1947.5 1973.5
## + Agency.Type             1   1947.8 1973.8
## + Destination            56   1841.8 1977.8
## - Agency.Product.Grouped 11   2570.2 2572.2
##
## Step:  AIC=1930.94
## Claim ~ Agency.Product.Grouped + Net.Sales
##
##                          Df Deviance    AIC
## <none>                        1904.9 1930.9
## + Agency.Type             1   1903.4 1931.4
## + Age                     1   1903.9 1931.9
## + Distribution.Channel    1   1904.2 1932.2
## + Duration                1   1904.8 1932.8
## + Commision..in.value.    1   1904.9 1932.9
## + Gender                  2   1903.8 1933.8
## + Destination            56   1816.8 1954.8
## - Net.Sales               1   1947.8 1971.8
## - Agency.Product.Grouped 11   2299.2 2303.2
```

```
summary(best.model3)
```

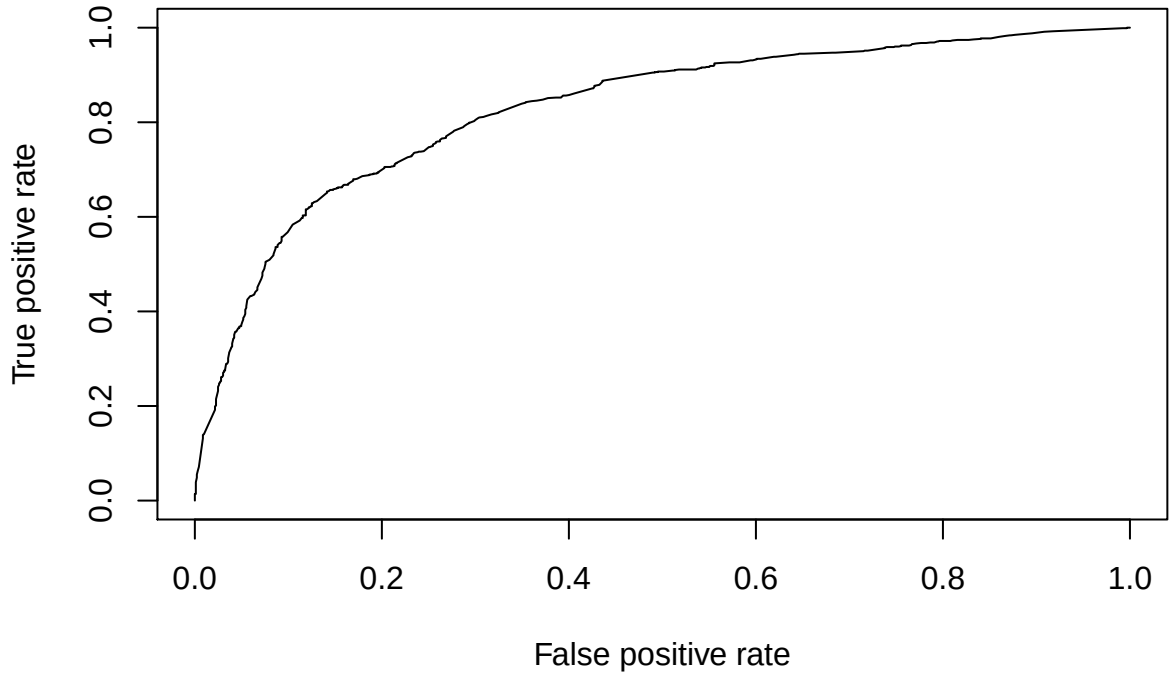
```
##
## Call:
## glm(formula = Claim ~ Agency.Product.Grouped + Net.Sales, family = "binomial",
##     data = don_us3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2640  -0.8935  -0.1630   0.7176   2.1465
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -0.735709   0.766862  -0.959   0.3374
## Agency.Product.Grouped13   1.662617   0.748228   2.222   0.0263 *
## Agency.Product.Grouped14  -0.119600   0.732249  -0.163   0.8703
## Agency.Product.Grouped15  -0.179195   0.729558  -0.246   0.8060
## Agency.Product.Grouped16  -1.361663   0.756155  -1.801   0.0717 .
## Agency.Product.Grouped17  -0.954924   0.766100  -1.246   0.2126
## Agency.Product.Grouped3    0.016720   0.747739   0.022   0.9822
```

```
## Agency.Product.Grouped5    0.709662   0.676936   1.048   0.2945
## Agency.Product.Grouped6    1.727810   0.751874   2.298   0.0216 *
## Agency.Product.Grouped7   -1.043352   0.924034  -1.129   0.2588
## Agency.Product.Grouped8   14.926250 377.805214   0.040   0.9685
## Agency.Product.Grouped9   -1.383217   1.306481  -1.059   0.2897
## Net.Sales                  0.010113   0.001603   6.308 2.83e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2570.2  on 1853  degrees of freedom
## Residual deviance: 1904.9  on 1841  degrees of freedom
## AIC: 1930.9
##
## Number of Fisher Scoring iterations: 14
```

On récupère maintenant des coefficients estimés bien définis (plus de variance démesurée).

Pas si simple que ça de développer un score ...

```
pred_best3 = prediction(best.model3$fitted.values,don_us3$Claim)
perf_best3 = performance(pred_best3,"tpr","fpr")
plot(perf_best3) # Tracé de la courbe ROC
```

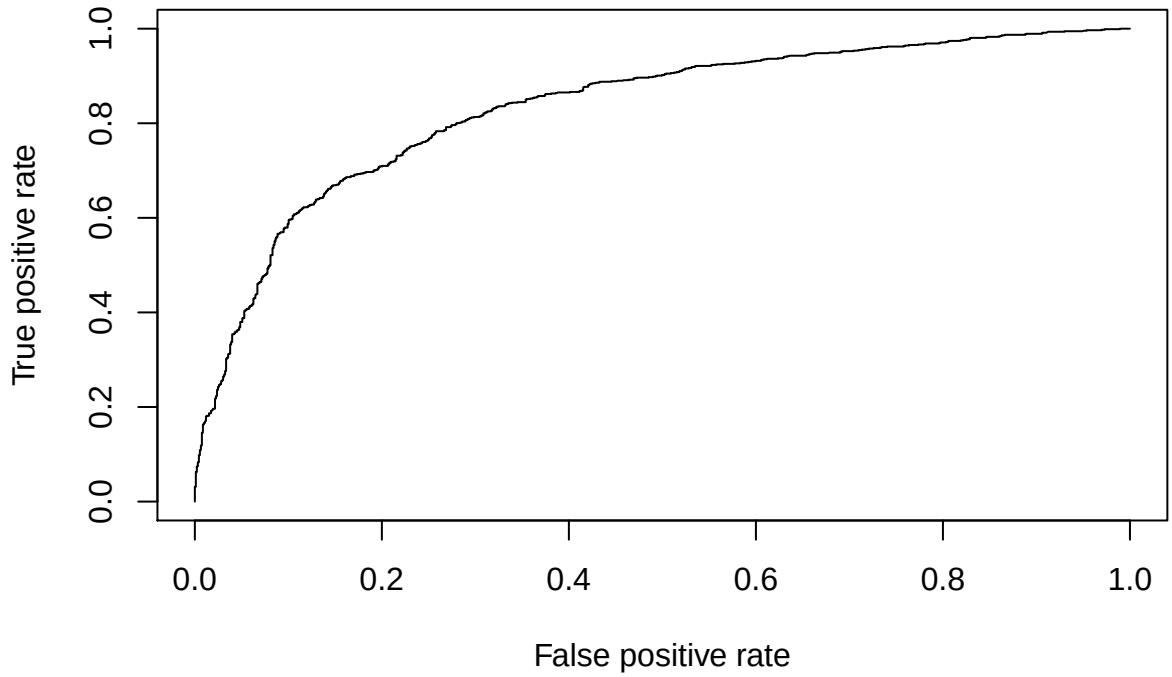



```
performance(pred_best3,"auc")@y.values[[1]] # Valeur de l'AUC
```

```
## [1] 0.8275626
```

Courbe qu'on aurait aussi pu comparer à celle du premier modèle :

```
pred_best = prediction(best.model$fitted.values,don_us$Claim)
perf_best = performance(pred_best,"tpr","fpr")
plot(perf_best) # Tracé de la courbe ROC
```




```
performance(pred_best,"auc")@y.values[[1]] # Valeur de l'AUC
```

```
## [1] 0.8323727
```

Les résultats sont somme toutes trés similaire que dans le cas du modèle 3, mais avec l'interprétabilité en plus.

Q18 : On souhaite ajuster un modèle de forêts aléatoires à l'aide de la commande :

```
library(randomForest)
randomForest(Claim ~ ., data = don_us)
```

On obtient le message d'erreur `Error in randomForest.default(m, y, ...) : Can not handle categorical predictors with more than 53 categories.`

Adapter le code pour pouvoir ajuster les random forests, et evaluer les performances de celles-ci.

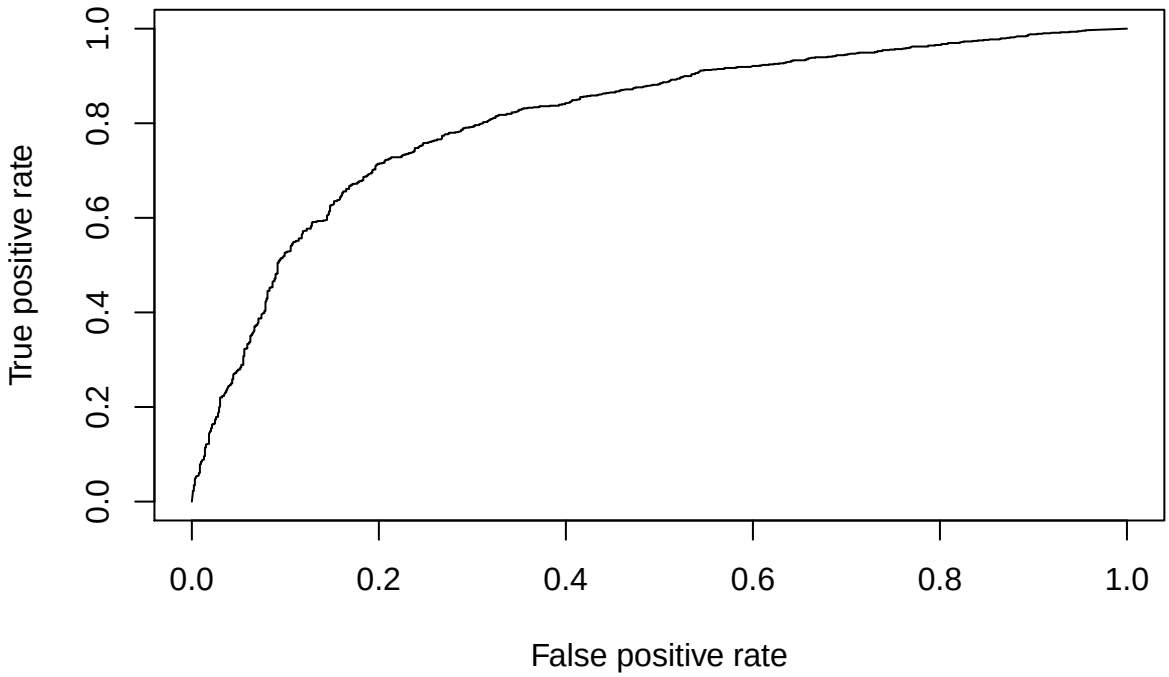
R18 : Il suffit simplement de supprimer la variable département (variable numéro 7)

```
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
rf = randomForest(Claim ~ ., data = don_us[-7])
pred_rf = prediction(rf$votes[,2],don_us$Claim)
perf_rf = performance(pred_rf,"tpr","fpr")
plot(perf_rf) # Tracé de la courbe ROC
```




```
performance(pred_rf,"auc")@y.values[[1]] # Valeur de l'AUC
```

```
## [1] 0.8116955
```

Résultats très similaire ici.

On peut s'intéresser à l'importance des différentes variables

```
rf$importance[order(rf$importance, decreasing = T),,drop = F]
```

```
##                       MeanDecreaseGini
## Duration                     137.246983
## Product.Name                 135.690757
## Net.Sales                    121.736231
## Agency                       102.530920
## Age                          100.760835
## Commision..in.value.          67.858412
## Gender                        41.787036
## Agency.Type                   14.613561
## Distribution.Channel           2.262595
```

Q19 : On aurait pu vouloir ajuster un modèle d'analyse discriminante probabiliste. Pourquoi ne peut-on pas utiliser ici la LDA ou la QDA ?

R19 : Ici LDA et QDA ne sont pas envisageables si toutes les variables ne sont pas quantitatives, donc si on veut faire usage de toutes les variables à notre disposition (qualitatives y compris) on ne peut pas faire usage de LDA et QDA.

Q20 : Ici on souhaite ajuster un classifieur de Bayes naïf à l'aide de la fonction `naiveBayes` du package `e1071`. Réaliser cet ajustement sur le jeu de données `don`, que remarquez vous sur le temps d'excusion par rapport au temps précédent ? Comment expliquez-vous celà ?

R20 :

```
library(e1071)
nb_don = naiveBayes(Claim ~ .,data = don)
```

Ici l'exécution de l'algorithme est instantanée, simples calculs de fréquences, variances, et moyennes.

*Q21 :** Le classifieur de Bayes naïf (cas particulier de méthode d'analyse discriminante probabiliste) est basé sur l'estimation des $P(Y = i) = \pi_i$ et $P(X = x|Y = i) = f_i(x)$, puis de l'application du théorème de Bayes.

En passant de `don` à `don_us`, comment l'estimation de π_0 et π_1 est-elle modifiée ? Les $f_0(x)$ et $f_1(x)$ estimés sont-ils foncièrement différents entre `don` et `don_us` ? Justifier.

R21 : Pour `don_us` les proportions estimée sont $\hat{\pi}_1 = \hat{\pi}_0 = 0,5$ tandis que pour `don` on a $\hat{\pi}_1 \simeq 0,01$, $\hat{\pi}_0 \simeq 0,99$. En fait ici les $f_1(x)$ estimés sont strictement les mêmes (strictement mêmes données pour l'estimation), pour $f_0(x)$ comme le tirage est fait selon un échantillonnage aléatoire simple, l'estimateur produit à partir du sous-echantillon à les mêmes propriétés asymptotique que l'estimateur à par de l'échantillon complet, par conséquent on devrait avoir des valeurs relativement simulaires pour `don` et don_us'.

```
nb_don_us = naiveBayes(Claim ~ .,data = don_us)
# Comparaison pour la variable Agency.Type
nb_don$tables$Agency.Type
```

```
##      Agency.Type
## Y      Airlines Travel Agency
##   No  0.2702928     0.7297072
##   Yes 0.6375405     0.3624595
```

```
nb_don_us$tables$Agency.Type
```

```
##      Agency.Type
## Y      Airlines Travel Agency
##   No  0.2793959     0.7206041
##   Yes 0.6375405     0.3624595
```

```
# Comparaison pour la variable Age
nb_don$tables$Age
```

```
##      Age
## Y         [,1]     [,2]
##   No  39.98982 14.01468
##   Yes 38.63430 14.11665
```

```
nb_don_us$tables$Age
```

```
##      Age
## Y         [,1]     [,2]
##   No  40.28479 14.51880
##   Yes 38.63430 14.11665
```

Q22 : En partant de la formule de Bayes, montrer que la probabilité $P(Y = 1|X = x)$ est une fonction croissante du rapport $f_1(x)/f_0(x)$.

R22 :

$$P(Y = 1|X = x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} = \frac{\pi_1 \frac{f_1(x)}{f_0(x)}}{\pi_0 + \pi_1 \frac{f_1(x)}{f_0(x)}} = 1 - \frac{1}{\pi_0 + \pi_1 \frac{f_1(x)}{f_0(x)}}$$

Ici comme $\frac{f_1(x)}{f_0(x)} > 0$, $\pi_1 > 0$, on en déduit facilement que $P(Y = 1|X = x)$ est une fonction croissante du rapport $\frac{f_1(x)}{f_0(x)}$

Q23 : En déduire que les valeurs estimées de π_0 et π_1 n'influent en rien sur l'ordre dans lequel seront rangés les individus en terme de probabilités.

R23 : Ici, dans l'expression de $P(Y = 1|X = x)$, la seule quantité qui dépend de x est le rapport $\frac{f_1(x)}{f_0(x)}$, et on sait d'après la question précédente que $P(Y = 1|X = x)$ est une fonction croissante de ce rapport. Donc π_0

et π_1 n'influent en rien sur l'ordre dans lequel seront rangés les individus en termes de probabilités.

Q24 : Conclure des question Q21 à Q23 que la courbe ROC est théoriquement inchangée entre les version répondérées ou non pour le cas de l'analyse discriminate probabiliste, dans le cas où les échantillons sont de grandes tailles (considérations asymptotiques oblige . . .)

R24 : Les individus devraient être rangés à peu près de la même manière avec ou sans sous-échantillonnage, d'après les questions précédentes (puisque moralement seuls les estimations de π_1 et π_0 différent entre l'échantillon `don` et `don_us`). Donc la ROC qui ne dépend que de la façons dont les individus sont ordonnés selon leur score est théoriquement inchangée en passant de de `don` à `don_us`.

Q25 : En conséquence expliquer pourquoi on a tout intérêt à plutôt utiliser `don` que `don_us`

R25 : On a tout intérêt à utiliser `don` tout entier puisque $f_0(x)$ sera mieux estimé avec plus de données.

Q26 : Parmis les évaluations réalisées pour les différents modèles, les classer de la plus sujette à une sur-évaluation des performances (biais d'optimisme), à celle la mois sujette à cette sur-évaluation. Préciser éventuellement si certaines ne sont pas sujettes du tout à ce biais.

R26 : Ici la méthode la plus sujette à la sur-évaluation des performance est la régression logistique, ensuite le classifieur de Bayes Naïf, et enfin les forêt aléatoires ne sont pas sujette à la sur-évaluation des performances puisque l'on utilise l'out-of-bag pour estimer les performances.

Q27 : Proposez une approche permettant de comparer "en toute honnêteté" les différents modèles proposés (implémentation non demandée)

R27 : Ici la bonne méthodes consiste à séparer l'échantillon entre un enchantillon d'apprentissage et un échantillon test, apprendre sur l'apprentissage et tester sur le test.

```
id_app = sample(nrow(don),0.7*nrow(don))
don_app = don[id_app,]
don_test = don[-id_app,]
```

Attention : ici on repart des données

On compare les modèles : - M1 : Régression logistique classique + stepwise - M2 : Régression logistique avec sous échantillonnage + stepwise - M3 : Régression logistique avec sous échantillonnage + croisement + regroupement de modalités + stepwise - M4 : Forêts aléatoires échantillon complet - M5 : Forêts aléatoires avec sous échantillonnage - M6 : Naive Bayes échantillon complet - M7 : Naive Bayes sous-échanillonnage

M1 :

```
min.model_M1 <- glm(Claim ~ 1, family = "binomial", data = don_app)
max.model_M1 <- glm(Claim ~ ., family = "binomial", data = don_app)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
M1 = step(min.model_M1, direction='both',
                  scope= list(lower = min.model_M1,
                              upper = max.model_M1))
```

```
## Start:  AIC=6730.99
## Claim ~ 1
##
##                         Df Deviance    AIC
## + Product.Name          25   5746.4 5798.4
## + Agency                15   5864.3 5896.3
## + Gender                 2   6268.7 6274.7
## + Net.Sales              1   6332.9 6336.9
## + Agency.Type            1   6343.7 6347.7
```

```
## + Destination          140   6120.8 6402.8
## + Commision..in.value.   1   6508.2 6512.2
## + Duration               1   6672.5 6676.5
## + Age                    1   6721.6 6725.6
## <none>                       6729.0 6731.0
## + Distribution.Channel   1   6728.9 6732.9
##
## Step:  AIC=5798.38
## Claim ~ Product.Name
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                         Df Deviance    AIC
## + Net.Sales              1   5697.4 5751.4
## + Commision..in.value.   1   5721.8 5775.8
## + Distribution.Channel   1   5742.3 5796.3
## <none>                       5746.4 5798.4
## + Age                    1   5744.4 5798.4
## + Agency.Type            1   5744.9 5798.9
## + Gender                 2   5743.3 5799.3
## + Duration               1   5745.4 5799.4
## + Agency                 8   5734.6 5802.6
## + Destination          140   5647.5 5979.5
## - Product.Name          25   6729.0 6731.0
##
## Step:  AIC=5751.36
## Claim ~ Product.Name + Net.Sales
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                         Df Deviance    AIC
## + Distribution.Channel   1   5693.3 5749.3
## <none>                       5697.4 5751.4
## + Age                    1   5695.7 5751.7
## + Agency.Type            1   5695.8 5751.8
## + Commision..in.value.   1   5696.0 5752.0
## + Gender                 2   5694.7 5752.7
## + Duration               1   5697.3 5753.3
## + Agency                 8   5686.9 5756.9
## - Net.Sales              1   5746.4 5798.4
## + Destination          140   5612.4 5946.4
## - Product.Name          25   6332.9 6336.9
##
## Step:  AIC=5749.33
## Claim ~ Product.Name + Net.Sales + Distribution.Channel
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                         Df Deviance    AIC
## + Age                    1   5691.0 5749.0
## <none>                       5693.3 5749.3
```

```
## + Agency.Type            1   5691.8 5749.8
## + Commision..in.value.   1   5692.1 5750.1
## + Duration               1   5693.3 5751.3
## + Gender                 2   5691.3 5751.3
## - Distribution.Channel   1   5697.4 5751.4
## + Agency                 8   5681.7 5753.7
## - Net.Sales              1   5742.3 5796.3
## + Destination          140   5608.6 5944.6
## - Product.Name          25   6332.7 6338.7
##
## Step:  AIC=5749.04
## Claim ~ Product.Name + Net.Sales + Distribution.Channel + Age
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                          Df Deviance    AIC
## <none>                        5691.0 5749.0
## - Age                     1   5693.3 5749.3
## + Agency.Type             1   5689.9 5749.9
## + Commision..in.value.    1   5689.9 5749.9
## + Gender                  2   5689.0 5751.0
## + Duration                1   5691.0 5751.0
## - Distribution.Channel    1   5695.7 5751.7
## + Agency                  8   5679.2 5753.2
## - Net.Sales               1   5739.6 5795.6
## + Destination           140   5606.6 5944.6
## - Product.Name           25   6320.6 6328.6
```

```
p_M1 <- predict(M1,don_app, type = "response")
```

M2 :

```
# Sous-échantillonnage
don_app_yes = don[don_app$Claim == "Yes",]
don_app_no = don[don_app$Claim == "No",]
id_select_app_no = sample(nrow(don_app_no),nrow(don_app_yes))
don_app_us = rbind(don_app_yes, don_app_no[id_select_app_no,])
# Fitting :
min.model_M2 <- glm(Claim ~ 1, family = "binomial", data = don_app_us)
max.model_M2 <- glm(Claim ~ ., family = "binomial", data = don_app_us)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
M2 = step(min.model_M2, direction='both',
                  scope= list(lower = min.model_M2,
                              upper = max.model_M2))
```

```
## Start:  AIC=248.63
## Claim ~ 1
##
##                         Df Deviance    AIC
## + Net.Sales              1   221.42 225.42
## + Gender                 2   233.34 239.34
## + Agency.Type            1   236.36 240.36
```

```
## + Commision..in.value.  1   237.12 241.12
## + Agency               15   211.47 243.46
## + Product.Name         23   196.85 244.85
## + Age                   1   242.62 246.62
## <none>                      246.63 248.63
## + Duration              1   245.43 249.43
## + Distribution.Channel  1   245.91 249.91
## + Destination          63   199.61 327.61
##
## Step:  AIC=225.42
## Claim ~ Net.Sales
##
##                        Df Deviance    AIC
## + Agency.Type           1   217.14 223.14
## + Gender                2   216.13 224.13
## + Age                   1   218.60 224.60
## + Duration              1   219.34 225.34
## <none>                      221.42 225.42
## + Commision..in.value.  1   220.64 226.64
## + Distribution.Channel  1   220.87 226.87
## + Agency               15   201.96 235.96
## + Product.Name         23   187.89 237.89
## - Net.Sales             1   246.63 248.63
## + Destination          63   184.12 314.12
##
## Step:  AIC=223.14
## Claim ~ Net.Sales + Agency.Type

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                        Df Deviance    AIC
## + Duration              1   213.97 221.97
## <none>                      217.14 223.14
## + Age                   1   215.57 223.57
## + Commision..in.value.  1   216.38 224.38
## + Distribution.Channel  1   216.72 224.72
## - Agency.Type           1   221.42 225.42
## + Gender                2   215.98 225.98
## + Agency               15   201.97 237.97
## + Product.Name         23   187.89 239.89
## - Net.Sales             1   236.36 240.36
## + Destination          63   183.80 315.80

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=221.97
## Claim ~ Net.Sales + Agency.Type + Duration

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                          Df Deviance    AIC
## <none>                       213.97 221.97
## + Age                     1   213.08 223.08
## - Duration                1   217.14 223.14
## + Commision..in.value.    1   213.38 223.38
## + Distribution.Channel    1   213.62 223.62
## + Gender                  2   212.25 224.25
## - Agency.Type             1   219.34 225.34
## + Agency                 14   197.13 233.13
## + Product.Name           23   186.68 240.68
## - Net.Sales               1   235.70 241.70
## + Destination            63   179.54 313.54
```

```
p_M2 <- predict(M2,don_app, type = "response")
```

M3 :

```
don_app_us2 = don_app_us %>% unite(Agency.Product,Agency,Product.Name)
app_chaid_tree <- chaid(Claim ~ as.factor(Agency.Product),
                        don_app_us2 ,
                        control = chaid_control(alpha2 = 0.000000001)) # seuil à fixer à la main
app_feuilles_chaid <- names(predict(app_chaid_tree))
#table(app_feuilles_chaid)
# table(app_feuilles_chaid, don_app_us2$Agency.Product)
don_app_us3 <- don_app_us2
don_app_us3 <- don_app_us3 %>% dplyr::select(-Agency.Product) %>%
  dplyr::mutate(Agency.Product.Grouped = factor(app_feuilles_chaid))
min.model_M3 <- glm(Claim ~ 1, family = "binomial", data = don_app_us3)
# Suppression de niveau avec une seule modalité
don_app_us3 = don_app_us3[sapply(don_app_us3,nlevels) != 1]
max.model_M3 <- glm(Claim ~ ., family = "binomial", data = don_app_us3)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
M3 = step(min.model_M3, direction='both',
                  scope= list(lower = min.model_M3,
                              upper = max.model_M3))
```

```
## Start:  AIC=248.63
## Claim ~ 1
##
##                          Df Deviance    AIC
## + Net.Sales               1   221.42 225.42
## + Gender                  2   233.34 239.34
## + Agency.Type             1   236.36 240.36
## + Commision..in.value.    1   237.12 241.12
## + Age                     1   242.62 246.62
## <none>                        246.63 248.63
```

```
## + Duration              1   245.43 249.43
## + Distribution.Channel  1   245.91 249.91
## + Destination          63   199.61 327.61
##
## Step:  AIC=225.42
## Claim ~ Net.Sales
##
##                        Df Deviance    AIC
## + Agency.Type           1   217.14 223.14
## + Gender                2   216.13 224.13
## + Age                   1   218.60 224.60
## + Duration              1   219.34 225.34
## <none>                      221.42 225.42
## + Commision..in.value.  1   220.64 226.64
## + Distribution.Channel  1   220.87 226.87
## - Net.Sales             1   246.63 248.63
## + Destination          63   184.12 314.12
##
## Step:  AIC=223.14
## Claim ~ Net.Sales + Agency.Type

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                        Df Deviance    AIC
## + Duration              1   213.97 221.97
## <none>                      217.14 223.14
## + Age                   1   215.57 223.57
## + Commision..in.value.  1   216.38 224.38
## + Distribution.Channel  1   216.72 224.72
## - Agency.Type           1   221.42 225.42
## + Gender                2   215.98 225.98
## - Net.Sales             1   236.36 240.36
## + Destination          63   183.80 315.80

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=221.97
## Claim ~ Net.Sales + Agency.Type + Duration

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                        Df Deviance    AIC
## <none>                      213.97 221.97
## + Age                   1   213.08 223.08
## - Duration              1   217.14 223.14
## + Commision..in.value.  1   213.38 223.38
## + Distribution.Channel  1   213.62 223.62
## + Gender                2   212.25 224.25
```

```
## - Agency.Type           1   219.34 225.34
## - Net.Sales             1   235.70 241.70
## + Destination          63   179.54 313.54
```

```
summary(M3)
```

```
##
## Call:
## glm(formula = Claim ~ Net.Sales + Agency.Type + Duration, family = "binomial",
##     data = don_app_us3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4954  -0.1709  -0.1144  -0.0978   3.4145
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -4.421063   0.384738 -11.491  < 2e-16 ***
## Net.Sales                 0.016741   0.004226   3.961 7.46e-05 ***
## Agency.TypeTravel Agency -1.080149   0.462707  -2.334   0.0196 *
## Duration                 -0.005641   0.003334  -1.692   0.0907 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 246.63  on 1813  degrees of freedom
## Residual deviance: 213.97  on 1810  degrees of freedom
## AIC: 221.97
##
## Number of Fisher Scoring iterations: 8
```

```
p_M3 <- predict(M3,don_app, type = "response")
```