Modélisation et apprentissage statistique appliqués à la tarification en assurance non vie

Yahia Chammami

2022-12-05

Introduction :

Les développements récents en tarification de l'assurance non-vie se concentrent majoritairement sur la maîtrise et l'amélioration des Modèles Linéaires Généralisés. Performants, ces modèles imposent à la fois trois méthodes d'apprentissage statistique :

- Les arbres de Classification • Régression, les forêts aléatoires • Les réseaux de neurones

Et les appliquent sur les données de plusieurs compagnies d'assurance automobile pour proposer La prime adaptée...

L'assurance non-vie, regroupant les assurances de responsabilité, de dommages et de personnes (exemples : contrats d'assurance automobile, habitation, santé, accidents de la vie.

L'assurance vie, regroupant les assurances de décès et les produits d'épargne à la fiscalité spécifique.

La tarification ,est le processus d'estimation de la prime d'assurance que l'assuré doit payer pour bénéficier d'une couverture en cas de sinistre.

Projet de Fin d'Etudes fait à Hydatis Engineering

Cadre du PFE: BOUGUILA Houcem Soutenu le 07 juillet 2021

Traitement des données L'assurance non-vie s'agit d'une police qui prévoit une indemnisation pour les pertes subies à la suite d'un sinistre. Cette police présente un document contractuel qui fixe les conditions d'un engagement entre une compagnie d'assurance et un assuré.

Cette étape est essentielle avant toute étude, étant donné l'importance de la qualité des données sur la modélisation, puis sur les résultats et les décisions. Il est donc important de choisir les informations à conserver et celles à rejeter et de gérer les valeurs manquantes et les erreurs de saisie.

La structure des variables qualitatives est la suivante :

```
var_quali<-c("sexe","statut","situation matrimniale","poste garantie","Classe")
modalite_porcentage <-c(("Femme: 51.24% Homme 48.76%"  ),
                        ("Principal 37.4% Conjoint 26.5% Enfant 36.1%"),
                        ("Divorcé_e 0.08% Marié_e 53.41% Célibataire 46.45%   Veuf_ve 0.06%"),
("Soins_courants 40.7% Dentaire 3.6% Pharmacie 42.2% Hospitalisation 11.5% Optique 2%" ),
("Classe1 50.86 ClasseA 49.14") )
d1<-data.frame(var_quali,modalite_porcentage)
d1
```

Postes de garanties	Exemples d'actes	Nombre d'actes
Soins Courants	- Consultation Médecin (généraliste ou spécialiste)	7188 (40.7%)
	- Analyses en laboratoires	
	- Centres de traitements et diagnostics	
	- Imagerie médicale	
Dentaire	- Consultation et visite chez les dentistes	636 (3.6 %)
	- Prothèses dentaires	
Pharmacie	- Médicaments	7433 (42.2 %)
Hospitalisation	- Frais d'interventions médicaux ou chirurgicaux.	2021 (11.5 %)
	- Frais de Séjour	
	- Maternité	
Optique	- Frais opticien (verres, lentilles, etc.)	350 (2%)

Table IV.1: Les différentes actes des postes des garantie

Figure 1: exemple de poste garentie.

```
##               var_quali
## 1                 sexe
## 2               statut
## 3 situation matrimniale
## 4        poste garantie
## 5                Classe
##                                                                  modalite_porcentage
## 1                                                      Femme: 51.24% Homme 48.76%
## 2                                  Principal 37.4% Conjoint 26.5% Enfant 36.1%
## 3                  Divorcé_e 0.08% Marié_e 53.41% Célibataire 46.45%   Veuf_ve 0.06%
## 4 Soins_courants 40.7% Dentaire 3.6% Pharmacie 42.2% Hospitalisation 11.5% Optique 2%
## 5                                                    Classe1 50.86 ClasseA 49.14
```

La structure des variables quantitatives est la suivante :

```
var_quanti<-c("Age","Nombre d'actes","Montant demandé","Montant remboursé")
minimum <-c(0,0,0,0)
Médiane<-c(34,0,45,40)
Moyenne<-c(29.96,0.19,80.33,64.51)
Maximum<-c(79,10,8441.52,3990.61)
d2<-data.frame(var_quanti,minimum,Médiane,Moyenne,Maximum)
d2
```

```
##          var_quanti minimum Médiane Moyenne Maximum
## 1               Age       0      34   29.96   79.00
## 2    Nombre d'actes       0       0    0.19   10.00
## 3   Montant demandé       0      45   80.33 8441.52
## 4 Montant remboursé       0      40   64.51 3990.61
```

Analyse par poste de garantie

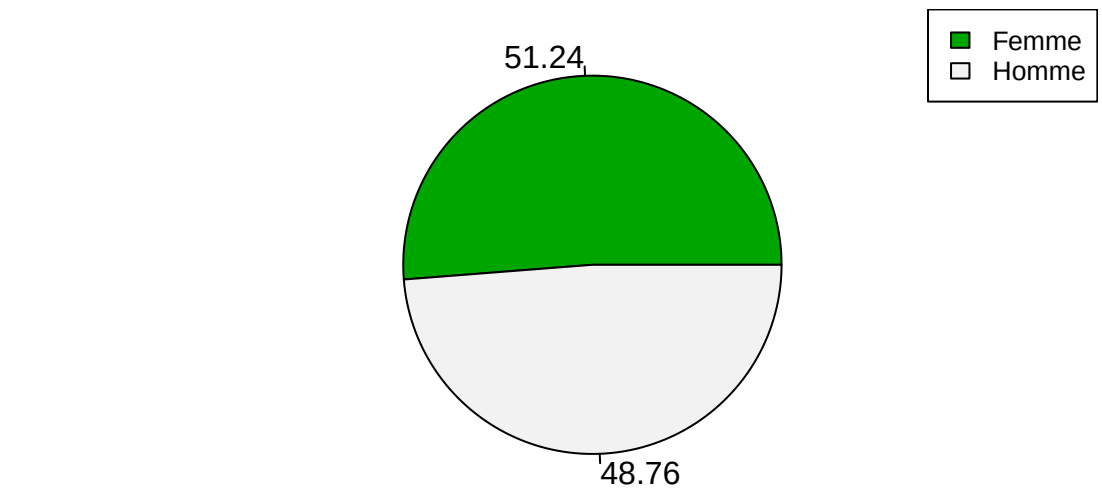
Une bonne pratique en matière de tarification d'assurance consiste à ne pas mélanger les sinistres de remboursement de différentes catégories de couverture, car elles se comportent différemment en termesde fréquence et de coût Une analyse par poste de garantie est alors nécessaire car nous essayons toujours de segmenter l'information pour affiner toujours plus les tarifs.

```
Postes_de_garanties<-c("Soins Courants","Dentaire","Pharmacie","Hospitalisation","Optique")
Exemples_d_actes<-c("Consultation Médecin Analyses en laboratoires Centres de traitements et diagnostics
                    "Consultation et visite chez les dentistes protheses dentaires ","Pharmacie",
                    "Frais d'interventions médicaux ou chirurgicaux Frais de Séjour Maternité ",
                    "Frais opticien ")
Nombre_d_actes<-c("7188 (40.7%)","636 (3.6 %)","7433 (42.2 %)","2021 (11.5 %)","350 (2%)")
d3<-data.frame(Postes_de_garanties,Nombre_d_actes,Exemples_d_actes)
d3
```

```
##   Postes_de_garanties Nombre_d_actes
## 1      Soins Courants   7188 (40.7%)
## 2            Dentaire    636 (3.6 %)
## 3           Pharmacie  7433 (42.2 %)
## 4     Hospitalisation  2021 (11.5 %)
## 5             Optique       350 (2%)
##                                                                                  Exemples_d_a
## 1 Consultation Médecin Analyses en laboratoires Centres de traitements et diagnostics Imagerie médica
## 2                                        Consultation et visite chez les dentistes protheses dentai
## 3                                                                                        Pharma
## 4                            Frais d'interventions médicaux ou chirurgicaux Frais de Séjour Matern
## 5                                                                                 Frais optic
```

Variable Sexe :

```
pourcen <- c(51.24, 48.76)
types <- c("Femme", "Homme")
pie(pourcen , labels = pourcen, main = "Répartition des assurés par sexe",
    col = terrain.colors(length(pourcen)))
legend("topright", types, cex = 0.8, fill = terrain.colors(length(pourcen)))
```



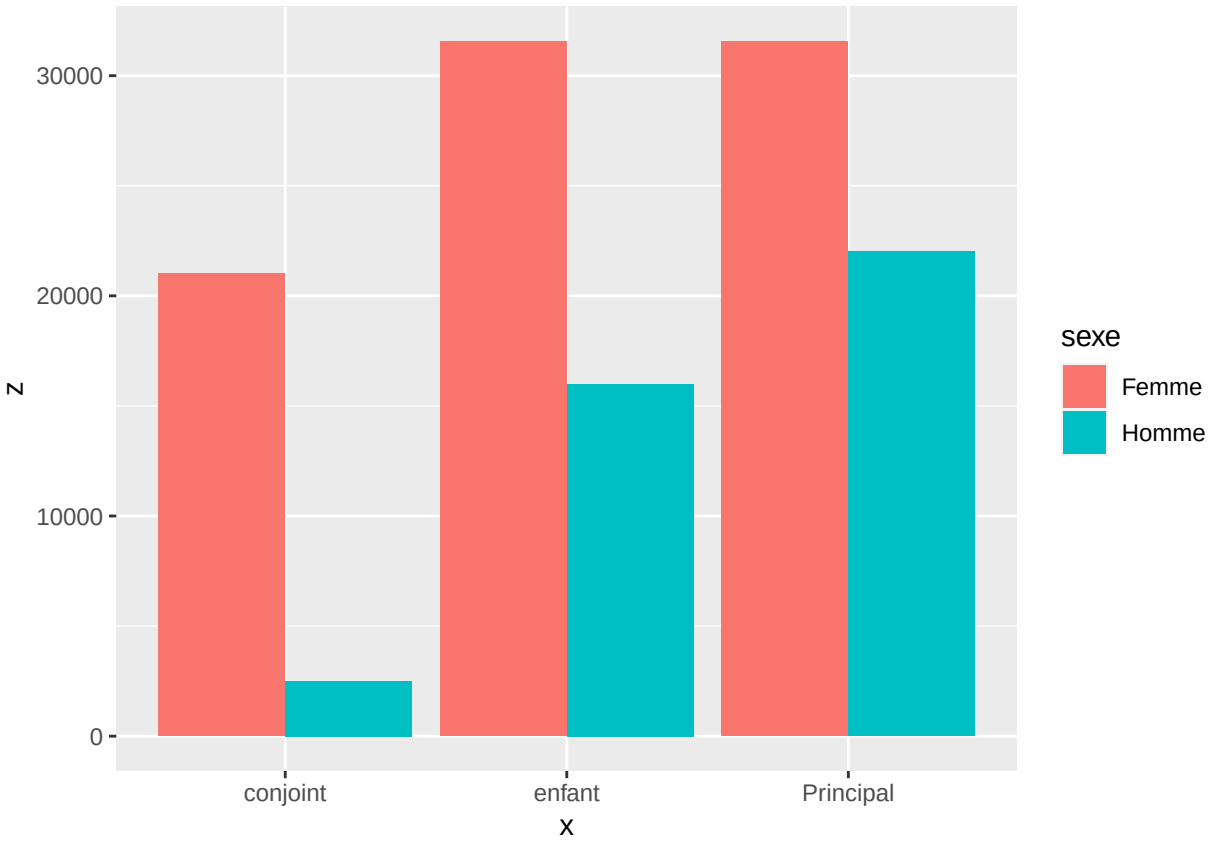
Répartition des assurés par statut et sexe

Le portefeuille de l'étude est caractérisé par une population féminine supérieure de 51,24% à la population masculine de 48,76%. Cette structure reste la même pour la catégorie des enfants, alors que dans la catégorie de l'assuré principal les hommes présentent la majorité de la population et inversement pour la catégorie du conjoint où le nombre de femmes est beaucoup plus important.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
x<-c("Principal","conjoint","enfant")
sexe<-c("Femme","Homme")
z<-c(31560,2500,31560,22000,21000,16000)
d<-data.frame(x,z,sexe)
ggplot(data=d, aes(x=x, y=z, fill=sexe)) +
geom_bar(stat="identity", position=position_dodge())
```

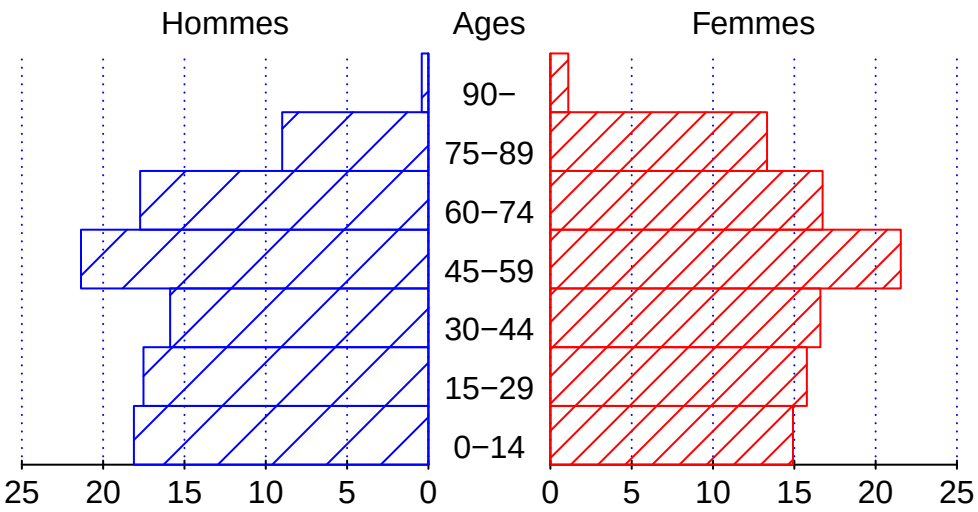


Variable Age

L'âge moyen des assurés est presque égal à 30 ans,cependant,on peut voir sur la pyramide des âgesque les groupes d'âge les plus représentés sont le groupe d'âge entre 45 et 59 ans pour les femmes et pour les hommes. Cela peut être dû au fait que cette catégorie est liée aux actes de santé relatifs à la maternité, d'où une souscription élevée de contrats pour cette catégorie

```
ages = c("0-14","15-29","30-44","45-59","60-74","75-89","90-")
males = c(707,684,620,834,692,351,16)
females = c(664,702,739,959,745,593,49)
males = males/sum(males)*100
females = females/sum(females)*100
names(males) <- ages


library(pyramid)
pyramids(Left=males,Llab="Hommes",Right=females,Rlab="Femmes",Laxis=c(0,5,10,15,20,25),main="Pyramide d
```



Nous affichons ci-dessus la pyramide des âges par sexe de l'assuré. La visualisation du graphique nous a conduit à construire la variable «Tranche d'âge »

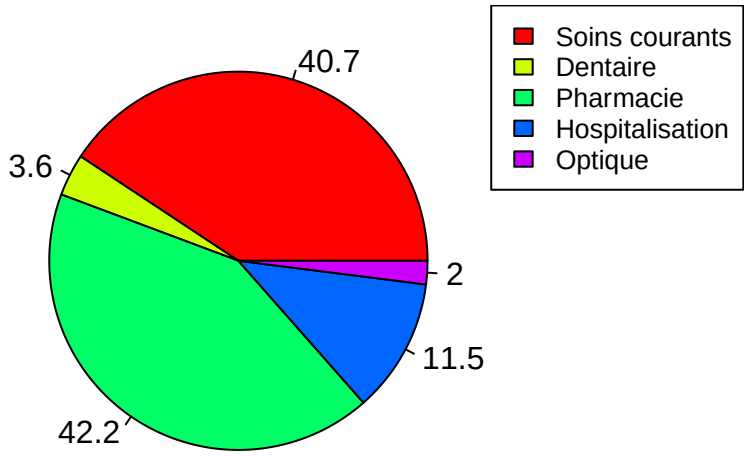
Test d'indépendance Khi-deux avec la variable Sinistre

```
Variable<-c("Tanche d'âge","Statut","Situation matrimoniale","Sexe","Classe")
P_Valeur<-c(0.3538,0.03383,0.1183,0.2129,0.8123)
d6<-data.frame(Variable,P_Valeur)
d6
```

```
##                  Variable P_Valeur
## 1           Tanche d'âge  0.35380
## 2                 Statut  0.03383
## 3 Situation matrimoniale  0.11830
## 4                   Sexe  0.21290
## 5                Classe  0.81230
```

Variable Poste garantie :

```
pourcen <- c(40.7, 3.6,42.2,11.5,2)
types <- c("Soins courants", "Dentaire","Pharmacie","Hospitalisation","Optique")
pie(pourcen , labels = pourcen, main = "Répartition des assurés par poste garentie",
    col = rainbow(length(pourcen)))
legend("topright", types, cex = 0.8, fill = rainbow(length(pourcen)))
```



Analyse de la fréquence

la modélisation de la fréquence est importante pour décrire sa structure dans le portefeuille et sa relation avec les autres variables.Parler de fréquence revient donc à parler du nombre de sinistres.

Il est clair qu'il y a une différence significative entre les valeurs prédites par ce modèle et les valeurs observées. On peut vérifier ces résultats par le recours aux tests d'ajustement.

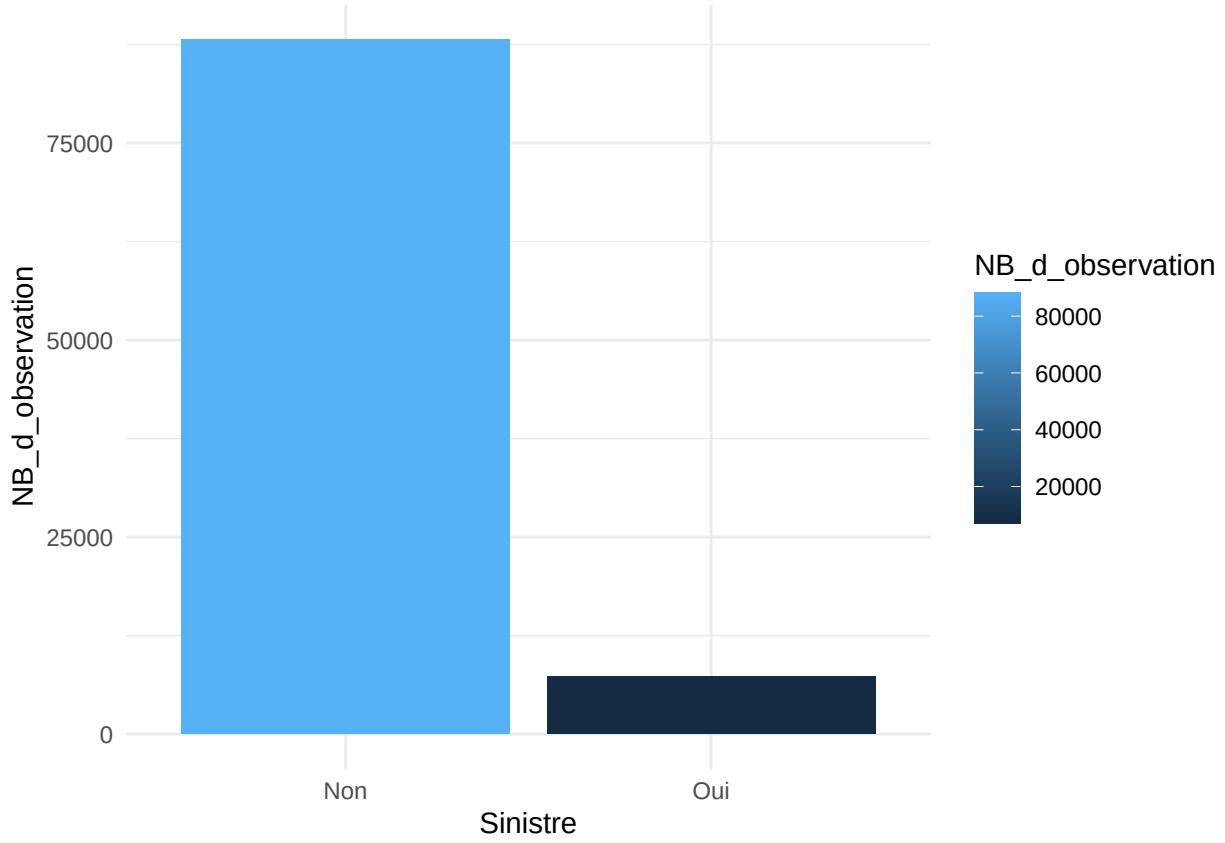
expresion fraction (fréquences= nb_des_sinistres/ exposition)

La variable du nombre de sinistres indique si le bénéficiaire a déclaré au moins un sinistre ou non. La structure de cette variable dans le portefeuille est la suivante :

```
Sinistre<-c("Oui","Non")
NB_d_observation<-c(7273,88157.8)
d4<-data.frame(Sinistre,NB_d_observation)
d4
```

```
##   Sinistre NB_d_observation
## 1      Oui           7273.0
## 2      Non          88157.8
```

```
library(ggplot2)
p<-ggplot(d4, aes(x=Sinistre, y=NB_d_observation, fill=NB_d_observation)) +
  geom_bar(stat="identity")+theme_minimal()
p
```



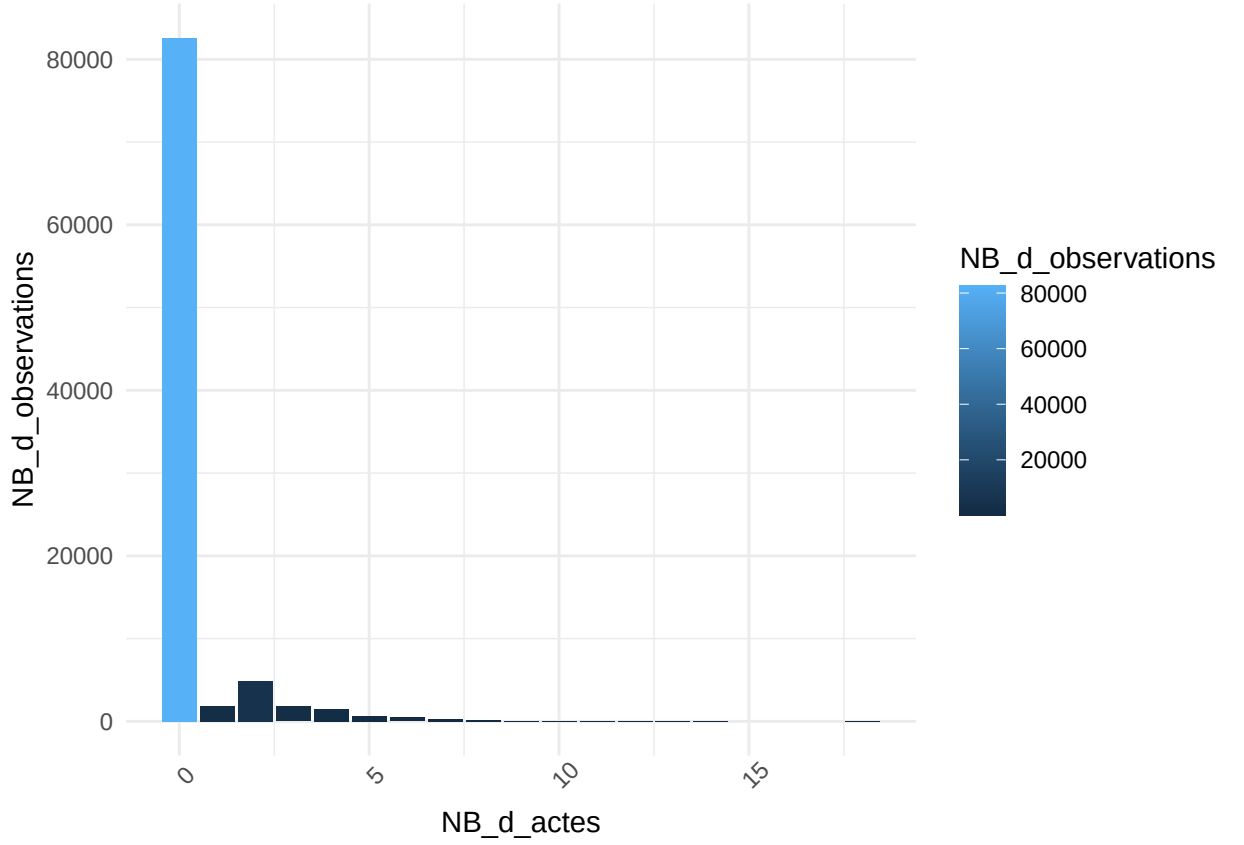
Nous passons ensuite à l'analyse de la variable du nombre de sinistres elle-même. Le tableau cidessous montre le nombre d'observations dans les données associées à chaque nombre de sinistres.

```
NB_d_actes<-c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,18)
NB_d_observations<-c(82610,1832,4916,1787,1523,607,453,262,122,86,78,13,4,6,4,1)
d5<-data.frame(NB_d_actes,NB_d_observations)
d5
```

```
##    NB_d_actes NB_d_observations
## 1           0             82610
## 2           1              1832
## 3           2              4916
## 4           3              1787
## 5           4              1523
## 6           5               607
## 7           6               453
## 8           7               262
## 9           8               122
## 10          9                86
## 11         10                78
## 12         11                13
## 13         12                 4
## 14         13                 6
## 15         14                 4
## 16         18                 1
```

Répartition de la fréquence des sinistres

```
library(ggplot2)
p<-ggplot(d5, aes(x=NB_d_actes, y=NB_d_observations, fill=NB_d_observations)) +
  geom_bar(stat="identity")+theme_minimal()+theme(plot.title = element_text(hjust = 0.5), axis.text.x =
p
```



Analyse du coût

Notre étude se concentre maintenant sur la consommation de l'assuré,et nous nous intéressons à la variable "Montant des sinistres" qui représente le coût réel des remboursements effectués par l'assurance nous devons donc créer une nouvelle variable "coût moyen" en divisant par le nombre d'actes consommés. la somme globale remboursée par l'assurance:

```
Poste_de_garantie<-c("Dentaire","Hospitalisation","Optique","Pharmacie","Soins courants")
Charge_totale_remboursée<-c(49626.4,183884.8,40183.5,285454,261440.4)
d7<-data.frame(Poste_de_garantie,Charge_totale_remboursée)
d7
```

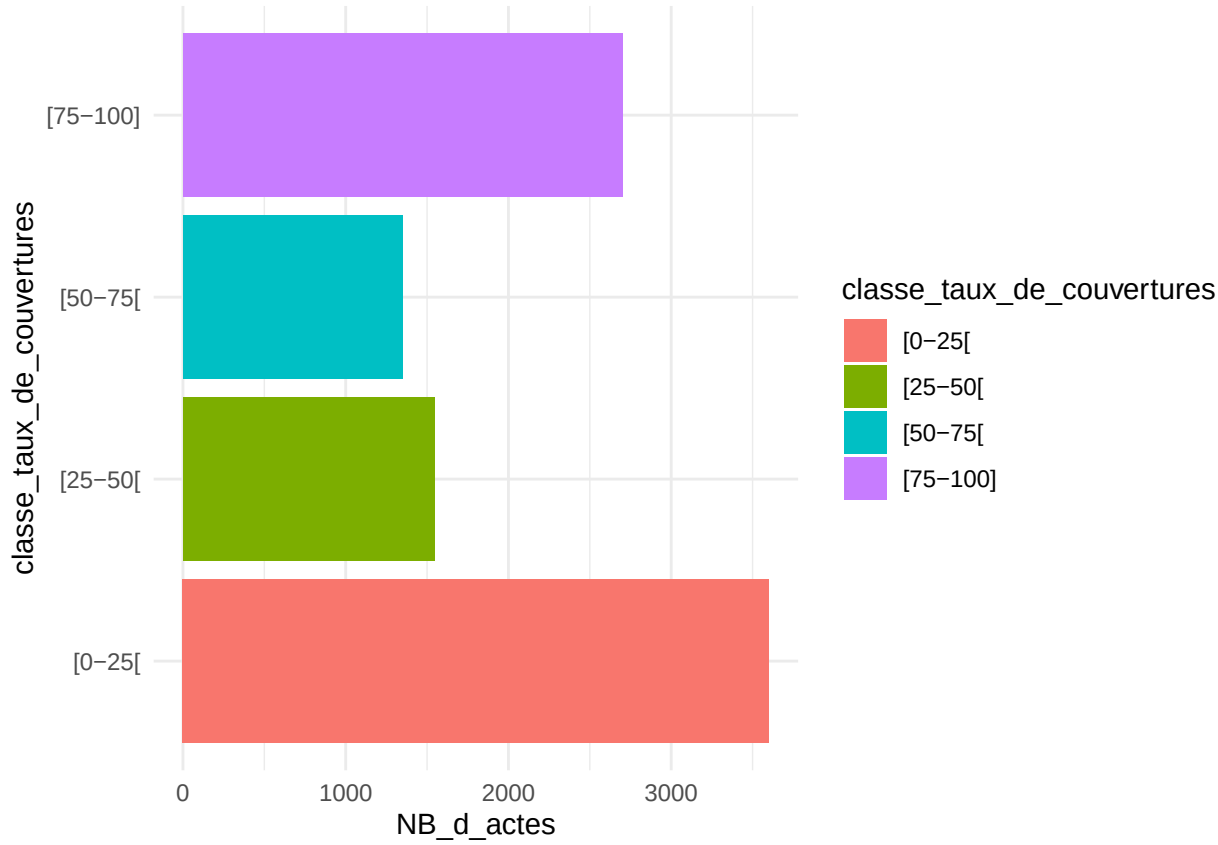
```
##   Poste_de_garantie Charge_totale_remboursée
## 1          Dentaire                 49626.4
## 2   Hospitalisation                183884.8
## 3           Optique                 40183.5
## 4         Pharmacie                285454.0
## 5    Soins courants                261440.4
```

Taux de couverture : est la capacité d'un assureur à couvrir le montant demandé par un assuré.La valeur du taux de couverture peut également être indiquée au moment de la souscription du contrat. Taux couverture=montant remboursé/montant demandé (frais réels).

```r
classe_taux_de_couvertures<-c("[0-25[","[25-50[","[50-75[","[75-100]")
NB_d_actes<-c(2200,1000,1300,2700,1400,550,50,5)
d8<-data.frame(classe_taux_de_couvertures,NB_d_actes)
d8
```

```
##   classe_taux_de_couvertures NB_d_actes
## 1                     [0-25[       2200
## 2                    [25-50[       1000
## 3                    [50-75[       1300
## 4                   [75-100]       2700
## 5                     [0-25[       1400
## 6                    [25-50[        550
## 7                    [50-75[         50
## 8                   [75-100]          5
```

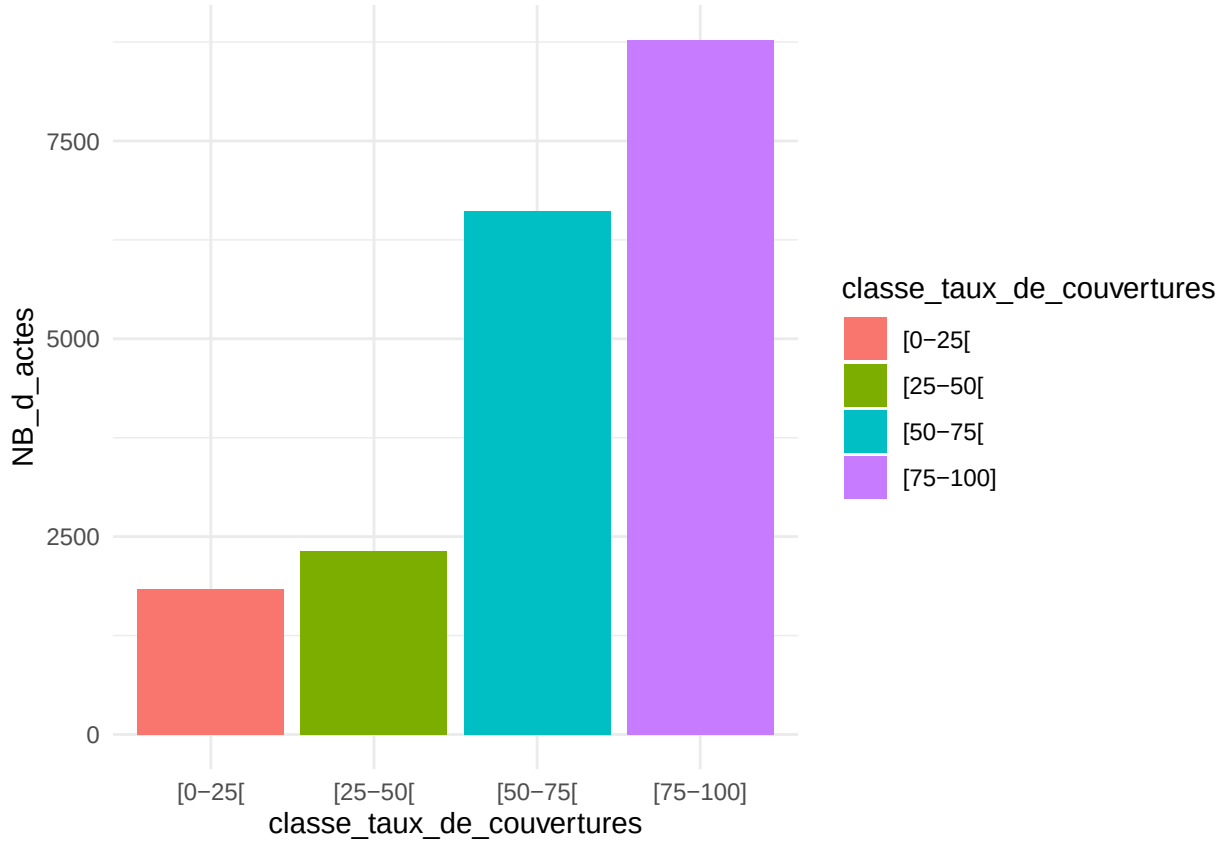
```r
library(ggplot2)
p<-ggplot(d8, aes(x=classe_taux_de_couvertures, y=NB_d_actes, fill=classe_taux_de_couvertures)) +
  geom_bar(stat="identity")+theme_minimal()+coord_flip()
p
```




```r
classe_taux_de_couvertures<-c("[0-25[","[25-50[","[50-75[","[75-100]")
c1<-c("Consultation Médecin Analyses en laboratoires Centres de traitements et diagnostics Imagerie médi
      "Frais d'interventions médicaux ou chirurgicaux Frais de Séjour Maternité ","Frais opticien ")
Exemples_d_actes<-c("c1","c1","c1","c1")
```

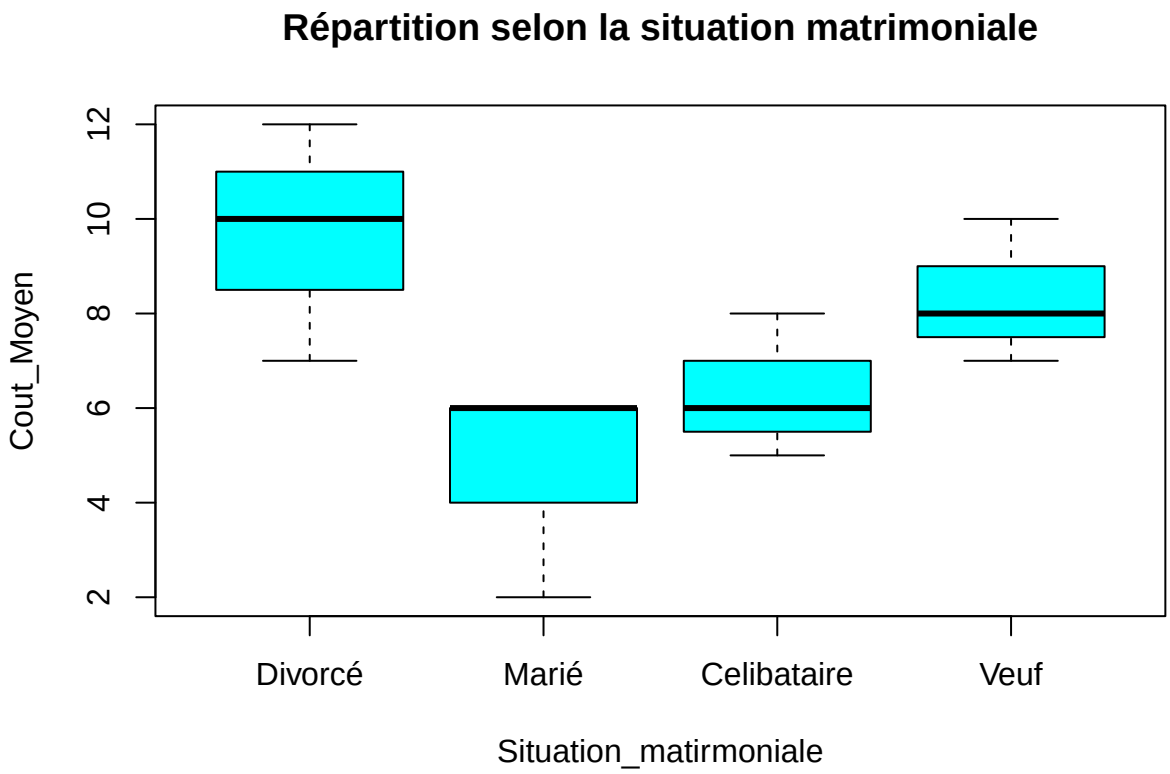
```
c2<-c(60,83,95.5,120,200)
c3<-c(70,85.3,90.14,112.6,260)
c4<-c(112,220,260,600,1200)
c5<-c(150,1200,1300,5120,8200)
NB_d_actes<-c(c2,c3,c4,c5)
d9<-data.frame(classe_taux_de_couvertures ,Exemples_d_actes,NB_d_actes)
```

```
library(ggplot2)
p<-ggplot(d9, aes(x=classe_taux_de_couvertures, y=NB_d_actes, fill=classe_taux_de_couvertures)) +
  geom_bar(stat="identity")+theme_minimal()
p
```



Coût Moyen en fonction de la variable Situation matrimoniale

```
a<-c("Divorcé","Marié","Celibataire","Veuf")
b<-matrix(c(12,6,5,10,10,2,6,8,7,6,8,7),nrow = 3,ncol = 4,byrow = TRUE)
boxplot(b,main="Répartition selon la situation matrimoniale",names=a,xlab="Situation_matirmoniale",ylab
```



Performance des modèles du coût moyen

```
Modèle<-c("GLM-Gamma","CART","Forêt aléatoire","GBM")
MAE <-c(11.824,11.732,11.688,11.69)
MSE<-c(261.895,259.691,259.516,260.41)
RMSE<-c(16.183,16.114,16.1,16.14)
t<-data.frame(Modèle,MAE,MSE,RMSE)
t
```

```
##            Modèle    MAE     MSE   RMSE
## 1       GLM-Gamma 11.824 261.895 16.183
## 2            CART 11.732 259.691 16.114
## 3 Forêt aléatoire 11.688 259.516 16.100
## 4             GBM 11.690 260.410 16.140
```

Le tableau ci-dessus contient les indicateurs d'écart entre les prédictions et les observations à savoir L'erreur absolue moyenne (MAE), le carré moyen des erreurs (MSE) et son racine carré RMSE.

Tous les indicateurs placent le modèle de Forêt aléatoire en premier, le modèle CART se place en deuxième position par rapport au critère MAE, cependant le modèle GBM possède une valeur MSE plus faible que le modèle CART ce qui les rend comparables en termes de performances.

Calcul de la prime d'assurance:

La prime d'assurance présente le montant qu'un assuré doit payer pour bénéficier de la couverture en cas du sinistre dans une période déterminée.

La valeur de la prime se compose de trois parties : une partie risque, une partie frais et une partie prestations.

La partie liée au risque est composé essentiellement de la prime pure qui présente le coût probable du sinistre, sa détermination représente le principal défi du processus de tarification car elle est directement liée à l'évolution du risque dans un sens ou dans l'autre (par exemple, augmentation ou diminution des prix aux soins de santé ou du prix des médicaments).

A cette composante du risque, nous ajoutons des frais de gestion pour couvrir les frais de fonctionnement de l'assureur ainsi que les taxes

Enfin, on ajoute au montant déjà établi, qui est entièrement technique, une part de bénéfice que l'assureur fixe en fonction de ses objectifs commerciaux.

La modélisation et la construction de modèles d'apprentissage automatique sont au coeur de l'analyse et de la prédiction des informations pour comprendre les phénomènes et prendre des décisions.

L'étape	Les fonctions développées
Traitement des données	- Importation des données - Mapping (adaptation des données à une structure spécifique) - Imputation des valeurs manquantes - Normalisation - Discrétisation - Echantillonnage (division train et test)
Construction des modèles	*) Modélisation de la fréquence - GLM - GBM - XGBOOST
	*) Modélisation du coût moyen - GLM - GBM - XGBOOST
Prédiction	- Affichage des prédictions - Analyse des performances
Interprétabilité	*) Techniques de l'intelligence artificielle explicable : - SHAP - Visualisation de l'importance des variables. - PDP
Tracking avec MLflow	- Suivi des paramètres des modèles. - Enregistrement des données d'apprentissage et test. - Enregistrement des modèles.

Table IV.29: Fonctions développées pour la partie déploiement

Figure 2: fts developpées.

Conclusion :

La tarification en assurance a été et sera toujours un sujet d'étude et d'amélioration étant donné son impact direct sur la compétitivité du marché. Dans ce projet, nous avons étudié la tarification sur un cas réel sur des données d'assurance santé. L'objectif était de rappeler les bases théoriques et d'appliquer les modèles linéaires généralisés présentant l'approche standard en matière de tarification, ainsi que certaines méthodes d'apprentissage qui intègrent de plus en plus le domaine ces derniers temps.