REPUBLIQUE DE COTE D'IVOIRE

Union- Discipline-Travail

……

MINISTERE DU PLAN ET DU DEVELOPPEMENT

École Nationale Supérieure de Statistique et d'Économie Appliquée

ENSEA

RAPPORT PROJET DATA SCIENCE :

PREDICTION DE L'INTERET POUR UNE ASSURANCE AUTOMOBILE

PRESENTE PAR :

GANIYU ISAAC
HAMADOU ABDOU AZIZ
OKOUA YANN

SOUS LA SUPERVISION DE :

M. PIETTE PIERRICK

JUILLET 2022

Table des matières

I. PRETRAITEMENT DES DONNEES

1. FEATURE ENGINEERING

Nous commençons par visualiser la base de données qui nous a été soumise.

Le dictionnaire de données se présente comme suit

VARIABLE	TYPE	DEFINITION
Gender	Qualitatif binaire	Sexe du client $\begin{cases} Male: homme \\ Female: femme \end{cases}$
Age	Quantitatif discret	Age du client en années révolues
Driving_License	Quantitatif discret (catégoriel binaire)	$\begin{cases} 1: Le\ client\ a\ un\ permis\ de\ conduire \\ 0: Le\ client\ n'apas\ de\ permis\ de\ conduire \end{cases}$
Region_Code	Quantitatif discret	Code de la région du client compris entre 0 et 53
Previously_Insured	Quantitatif discret (catégoriel binaire)	$\begin{cases} 1: Le\ client\ a\ une\ assurance\ automobile \\ 0: Le\ client\ n'a\ pas\ d'assuranceautomobile \end{cases}$
Vehicle_Age	Qualitatif ordonné (catégoriel multinomial)	$\begin{cases} < 1\ Year: L'âge\ du\ véhicule\ du\ client\ est\ inférieur\ à\ 1\ an \\ 1-2\ Year: L'âge\ du\ véhicule\ du\ client\ est\ compris\ entre\ 1\ et\ 2\ ans \\ > 2\ Years: L'âge\ du\ véhicule\ du\ client\ est\ supérieur\ à\ 2\ ans \end{cases}$
Vehicle_Damage	Qualitatif (catégoriel binaire)	$\begin{cases} Yes: Le\ client\ a\ déjà\ endommagé\ son\ véhicule \\ No: Le\ client\ n'a\ jamais\ endommagé\ son\ véhicule \end{cases}$
Annual_Premium	Quantitatif discret	Le montant individuel de l'assurance premium du client s'il souscrit.
Policy_Sales_Channel	Quantitatif discret	Code désignant le canal de communication entre l'entreprise d'assurance et le client.
Vintage	Quantitatif discret	Nombre de jours d'ancienneté du client dans l'entreprise
Response	Quantitatif discret (catégoriel binaire)	$\begin{cases} 1: Le\ client\ est\ intéressé\ par\ l'offre \\ 0: Le\ client\ n'estpas\ intéressé\ par\ l'offre \end{cases}$

Pour une meilleure analyse de la base, nous décidons de transformer les variables qualitatives en variables quantitatives. Les variables qualitatives de notre base sont toutes catégorielles. Ainsi, nous transformons les variables « Gender » et « Vehicle_Damage » en quantitatives binaires puis nous décomposons la variable « Vehicle_Damage » en trois variables quantitatives binaires « inf_1 », « between_1_and_2 » et « sup_2 » représentant respectivement les modalités « < 1 Year », « 1 – 2 Year » et « > 2 Years ».

Ensuite nous décidons de supprimer la variable « Policy_Sales_Channel » car, intuitivement, elle ne nous semble pas pertinente dans la prédiction de la décision du client. De plus, il s'agit juste d'un code prenant des valeurs, qui selon nous, ne sont ni purement mathématiques, ni purement économiques. A la suite de ces opérations, le dictionnaire de données se présente comme suit :

VARIABLE	TYPE	DEFINITION
Gender	Quantitatif discret (catégoriel binaire)	Sexe du client $\begin{cases} 1: homme \\ 0: femme \end{cases}$
Age	Quantitatif discret	Age du client en années révolues
Driving_License	Quantitatif discret (catégoriel binaire)	$\begin{cases} 1: Le\ client\ a\ un\ permis\ de\ conduire \\ 0: Le\ client\ n'apas\ de\ permis\ de\ conduire \end{cases}$
Region_Code	Quantitatif discret	Code de la région du client compris entre 0 et 52
Previously_Insured	Quantitatif discret (catégoriel binaire)	$\begin{cases} 1: Le\ client\ a\ une\ assurance\ automobile \\ 0: Le\ client\ n'a\ pas\ d'assuranceautomobile \end{cases}$
inf_1	Quantitatif discret (catégoriel binaire)	$\begin{cases} 1: L'âge\ du\ véhicule\ du\ client\ est\ inférieur\ \text{à}\ 1\ an \\ 0: L'âge\ du\ véhicule\ du\ client\ n'est\ pas\ inférieur\ \text{à}\ 1\ an \end{cases}$
between_1_and_2	Quantitatif discret (catégoriel binaire)	$\begin{cases} 1: L'âge\ du\ véhicule\ du\ client\ est\ compris\ entre\ 1\ et\ 2\ ans \\ 0: L'âge\ du\ véhicule\ du\ client\ n'est\ pas\ compris\ entre\ 1\ et\ 2\ ans \end{cases}$
sup_2	Quantitatif discret (catégoriel binaire)	$\begin{cases} 1: L'âge\ du\ véhicule\ du\ client\ est\ supérieur\ \text{à}\ 2\ ans \\ 0: L'âge\ du\ véhicule\ du\ client\ n'est\ pas\ supérieur\ \text{à}\ 2\ ans \end{cases}$

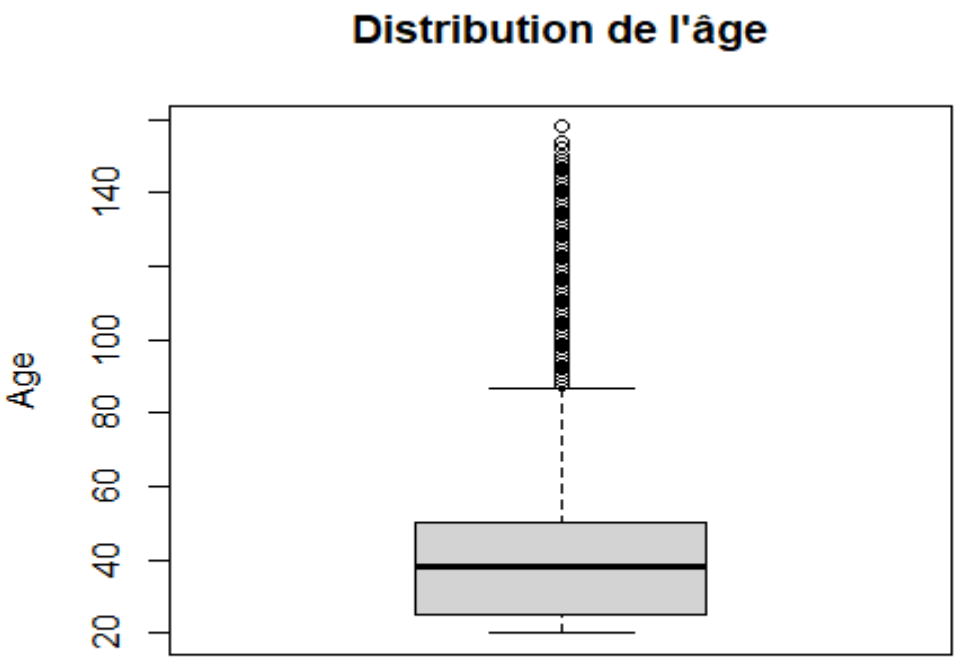
Vehicle_Damage	Quantitatif discret (catégoriel binaire)	$\begin{cases} 1: Le\ client\ a\ déjà\ endommagé\ son\ véhicule \\ 0: Le\ client\ n'a\ jamais\ endommagé\ son\ véhicule \end{cases}$
Annual_Premium	Quantitatif discret	Le montant individuel de l'assurance premium du client s'il souscrit.
Vintage	Quantitatif discret	Nombre de jours d'ancienneté du client dans l'entreprise
Response	Quantitatif discret (catégoriel binaire)	$\begin{cases} 1: Le\ client\ est\ intéressé\ par\ l'offre \\ 0: Le\ client\ n'estpas\ intéressé\ par\ l'offre \end{cases}$

2. PREMIERE ANALYSE EXPLORATOIRE

Nous étudions les variables de notre base de données une à une afin de détecter d'éventuelles améliorations ou corrections à faire.

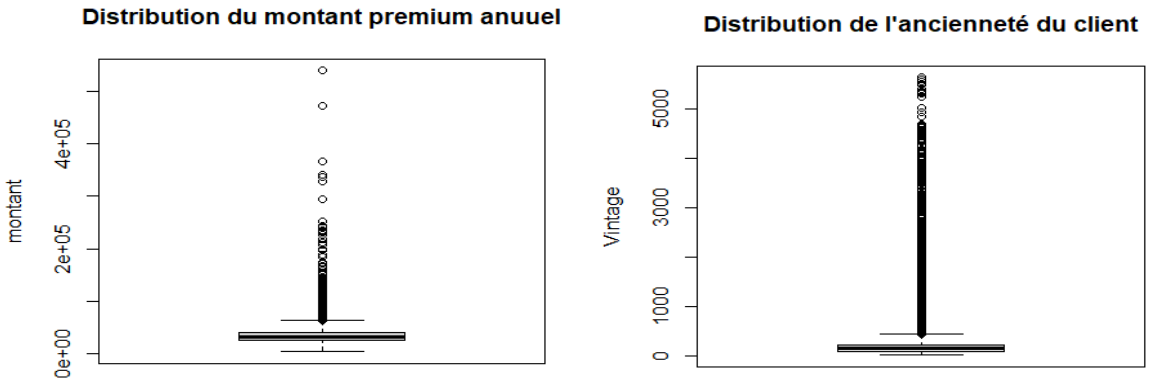
Nous disposons de 71 726 observations et de 12 variables. Il n'y a aucune valeur manquante dans nos données. La base est constituée de 54.7% d'hommes et 45.3% de femmes. L'âge moyen est de près de 40 ans. Les clients de l'entreprise sont localisés dans 53 régions différentes codées de 0 à 52. 99,81% des clients ont un permis de conduire. 4 clients sur 10 ont déjà une assurance automobile. 40,2% des clients ont un véhicule âgé de moins d'un an tandis que 54,9% d'entre eux ont un véhicule dont l'âge est compris entre 1 et 2 ans. Plus de la moitié des clients ont déjà endommagé leur véhicule. Le prix moyen de l'assurance premium est de 30697. 95% des clients ont moins de 288 jours de fidélité à l'entreprise. Enfin, plus de deux clients sur 10 se déclarent intéressés par l'offre d'assurance.

En inspectant de plus près la variable âge, on constate qu'elle contient des valeurs douteuses. Ses valeurs les plus hautes sont 149, 150, 152, 154 et 158. Ces valeurs d'âge nous laissent dubitatifs. La boite à moustache de la variable se présente comme suit :



On remarque une présence d'outliers dans ces données.

Nous représentons également les boites à moustaches des variables Annual_Premium et Vintage.



Là encore, nous remarquons la présence d'outliers pour ces variables.

3. DATA MANAGEMENT

A partir des diagrammes représentés plus haut, nous trouvons 562, 1929 et 600 outliers respectivement pour les variables « Age », « Annual_Premium » et « Vintage ». En considérant les outliers qui ne sont pas outlier de plus d'une de ces variables, nous trouvons en tout 3058 outliers dans ces données à partir de ces variables qui constituent les seules sur lesquelles nous pouvons repérer des outliers, les autres étant binaires ou codée (Region_Code).

Ces 3058 outliers représentent 4,26% de nos données initiales. Vu cette relativement faible proportion, nous décidons de supprimer ces valeurs aberrantes de notre base afin d'améliorer la qualité globale de nos modèles.

Après cette étape, nous jugeons que notre base de données est désormais prête à l'emploi.

4. DEUXIEME ANALYSE EXPLORATOIRE

a. STATISTIQUES UNIVARIEES

La nouvelle base de données compte 68668 observations de 11 variables.

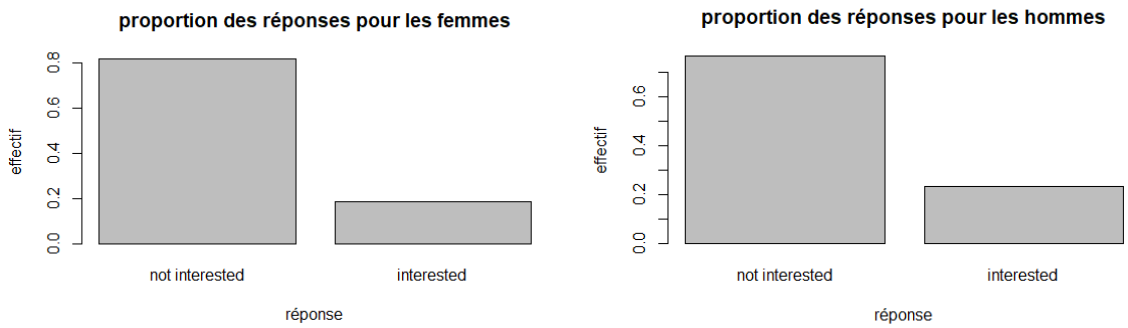
54,6% d'hommes et 45,4% de femmes. L'âge moyen est de 39,3 ans avec un âge maximum désormais de 87 ans. 99,81% des clients ont un permis de conduire. 41,2% des clients ont déjà une assurance automobile. 4 clients sur 10 ont un véhicule âgé de moins d'un an et plus de la moitié des clients ont un véhicule d'un âge compris entre 1 et 2 ans. 55,21% des clients ont déjà endommagé leur véhicule. Le montant moyen de l'assurance premium est de 31382. 95% des individus ont moins de 285 jours de fidélité à l'entreprise. Enfin, 21,1% des clients disent être intéressés par l'offre d'assurance.

Nous pouvons constater que la suppression des valeurs aberrantes n'a pas fondamentalement modifié ces statistiques sommaires.

b. STATISTIQUES BIVARIEES

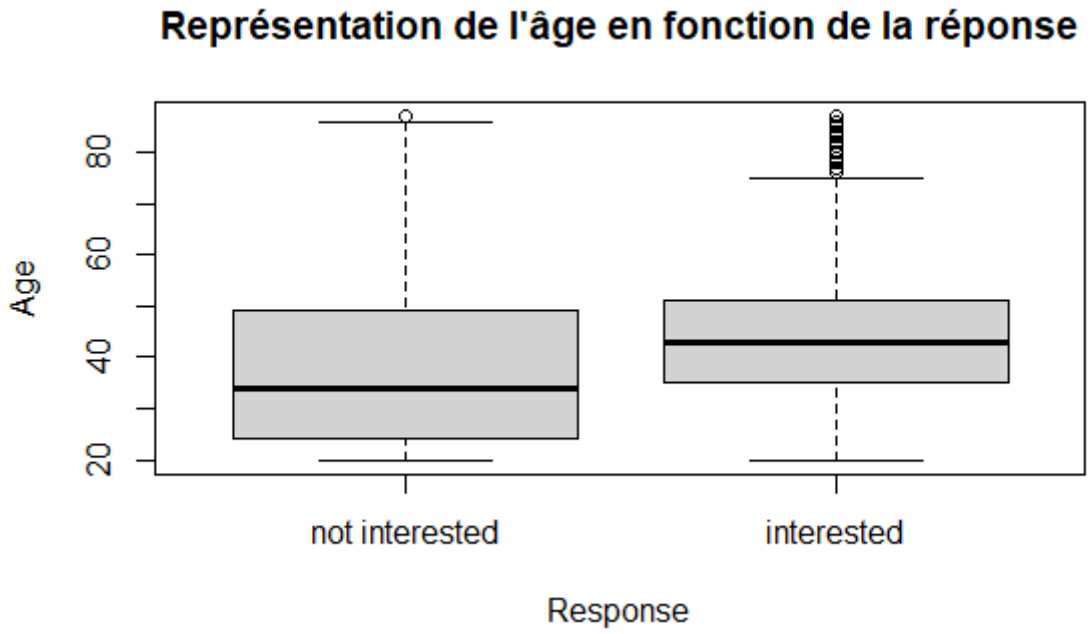
On croise la variable Response aux différentes variables de la base afin de suspecter d'éventuelles relations entre ces variables.

- **Relation entre la réponse et le sexe**



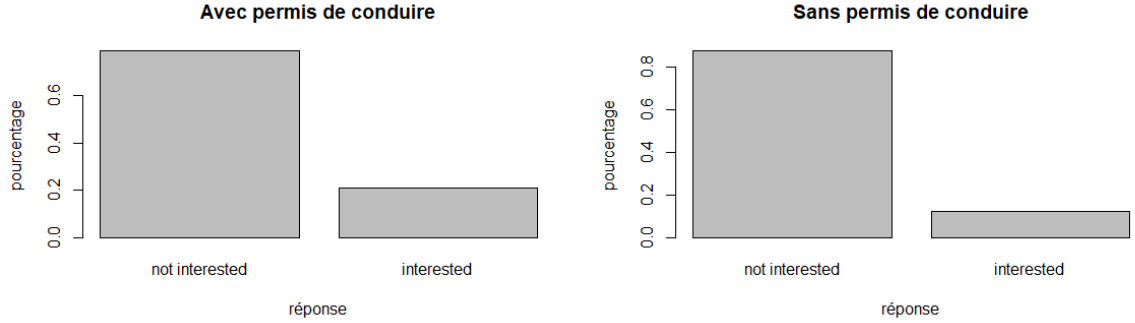
La réponse des clients ne semble pas être liée au sexe. Vu que les proportions par sexe se valent approximativement.

- **Relation entre la réponse et l'âge**



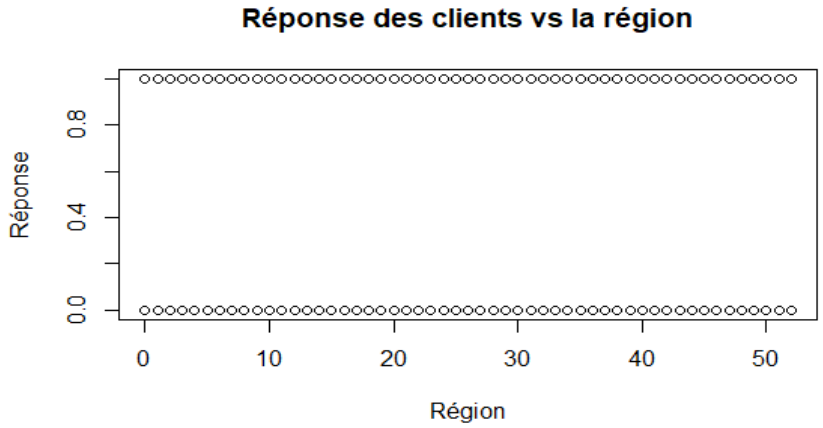
D'après cette représentation, les personnes intéressées ont tendance à être plus âgées que les personnes non intéressées par l'offre.

- **Relation entre la réponse et la possession d'un permis de conduire**



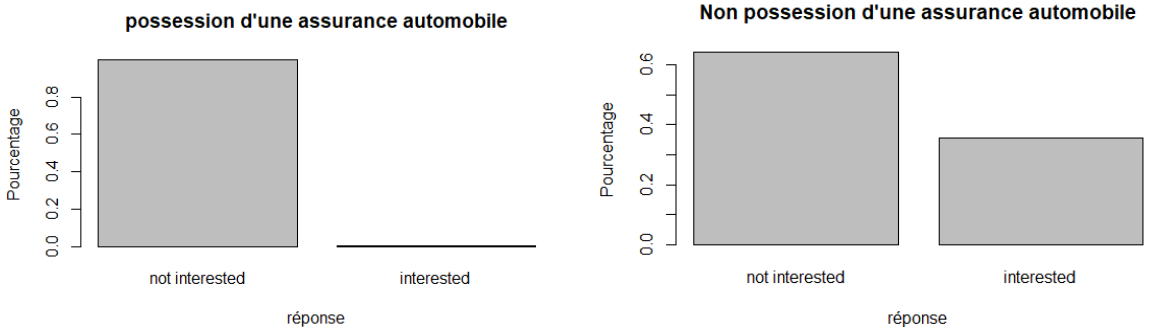
Les individus sans permis de conduire ont tendance à être moins intéressés par l'offre d'assurance que ceux avec permis de conduire. Cette variable semble donc avoir un effet positif sur la réponse.

- **Relation entre la réponse et la région**



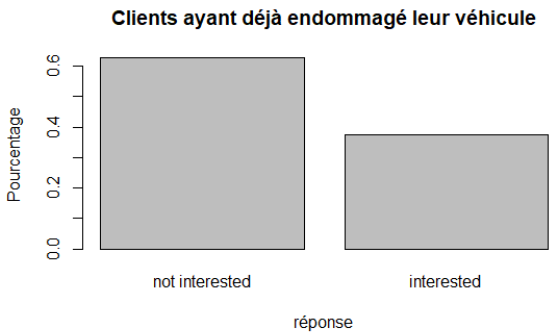
A partir de ce graphe, nous ne percevons pas d'effet de la région sur la réponse du client.

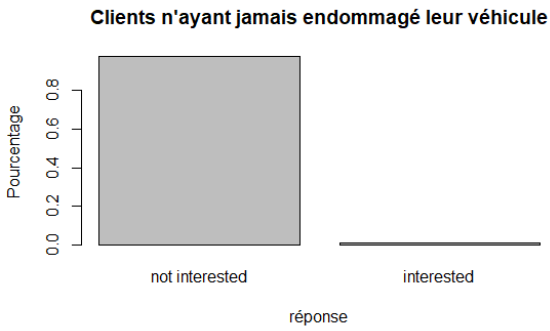
- **Relation entre la réponse et la possession d'une assurance automobile**



Les individus ayant déjà une assurance automobile sont largement moins intéressés par l'offre que ceux n'ayant pas d'assurance automobile. Ce graphique nous fait présager un effet négatif de la variable « Previously_Insured » sur la variable « Response ».

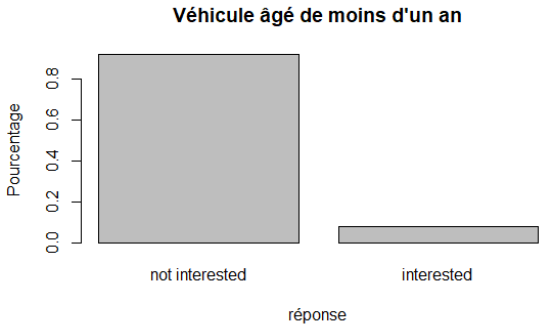
- **Relation entre la réponse et le fait d'avoir déjà endommagé son véhicule**

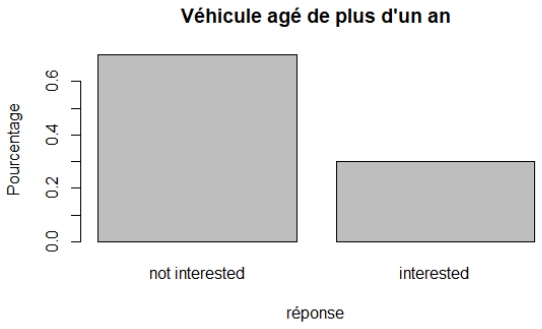


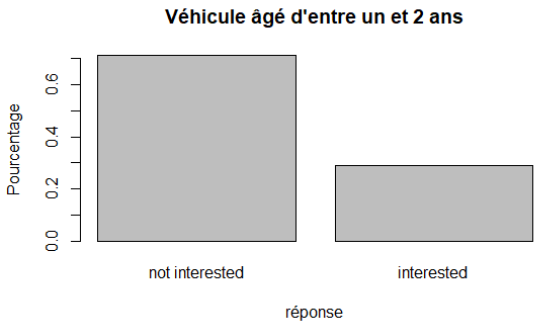


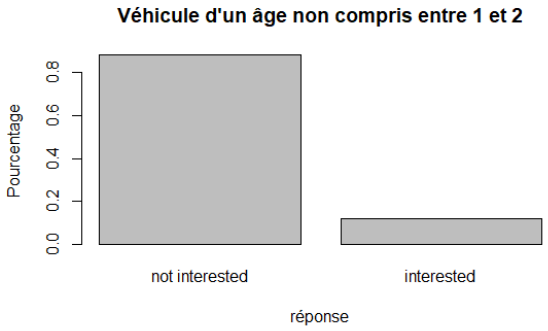
Les clients ayant une fois endommagé leur véhicule sont largement plus intéressés par l'offre que ceux n'ayant jamais endommagé leur véhicule. Ce graphique nous fait croire qu'il pourrait exister un effet positif de la variable « Vehicle_Damage » sur la variable « Response ».

- **Relation entre la réponse et l'âge du véhicule**



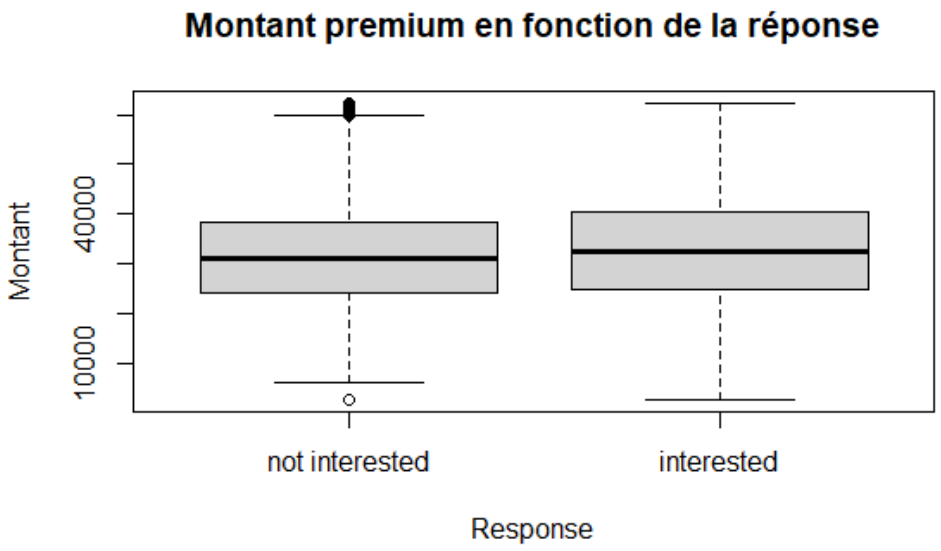






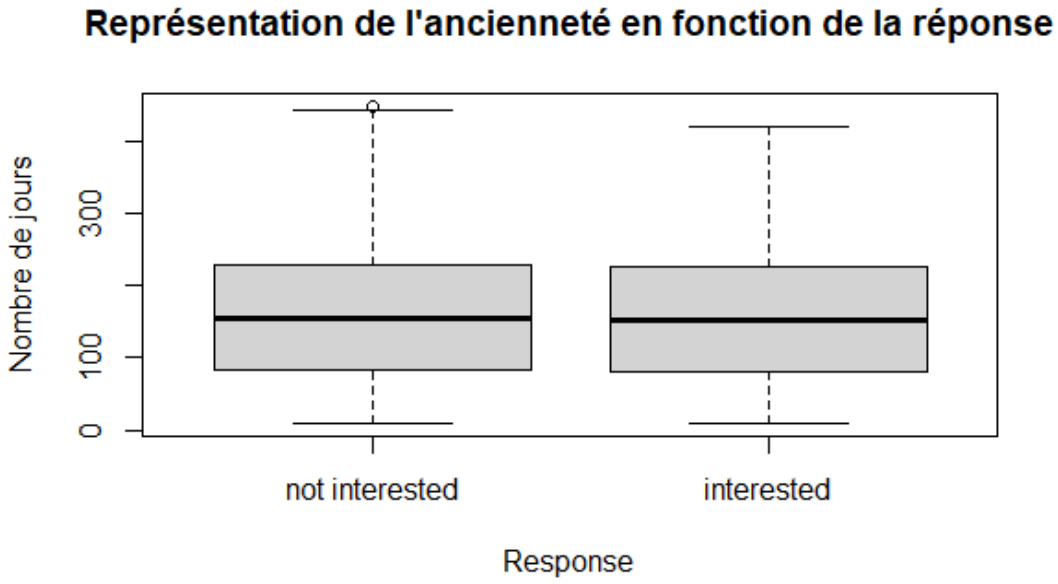
On peut lire sur ces graphiques que la réponse semble dépendre négativement du fait que le véhicule soit âgé de moins d'un an tandis qu'elle semble dépendre positivement du fait que le véhicule soit d'un âge compris entre un et deux ans.

- **Relation entre la réponse et le montant annuel de l'offre premium**



Ainsi la distribution du montant premium ne diffère pas grandement entre les intéressés et les non intéressés.

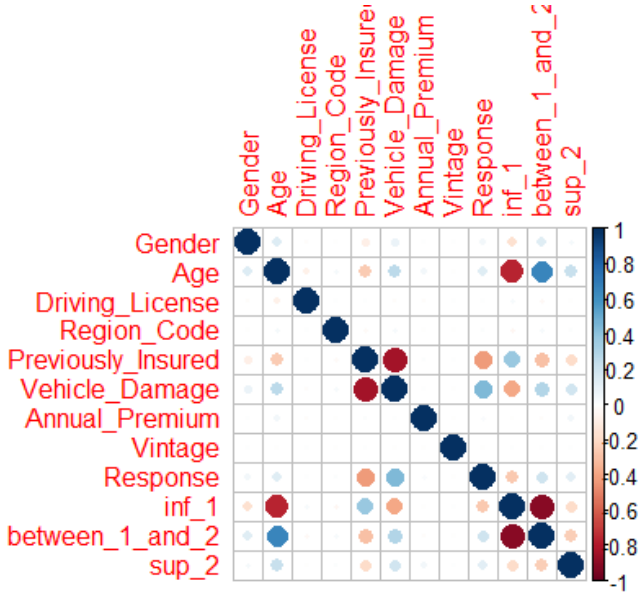
- **Relation entre la réponse et l'ancienneté**



Pour cette variable également, la distribution ne semble pas significativement différer de la catégorie des non intéressés à celle des intéressés. Il semnble donc que la variable « Vintage » n'ait pas d'effet important sur la réponse du client.

- **Corrélations entre les variables**

Les corrélations entre les différentes variables peuvent être résumées dans le graphe ci-dessous :



Les corrélations les plus fortes sont les corrélations négatives entre « Vehicle_Damage » et « Previously_Insured », entre « Age » et « inf_1 », entre « between_1_and_2 » et « inf_1 » et la corrélation positive entre « Age » et « between_1_and_2 ».

Les variables les plus corrélées à « Response » sont « Previously_Insured » (négativement) et « Vehicle_Damage » (positivement).

II. ESTIMATION DE MODELES SIMPLES

Dans cette partie et dans toute la suite du travail, nous scindons la base obtenue en une base train et une base test. La base train nous sert à entrainer le modèle tandis que la base test nous sert à le tester et à vérifier sa performance. La base test contient 80% de nos données soit 54935 observations. La base test en contient 13733.

1. LA REGRESSION LOGISTIQUE

Nous commençons par faire une régression logistique à partir des données de la nouvelle base. De cette régression, on obtient un modèle globalement significatif. Les variables « Region_Code » et « Vintage » ne sont pas individuellement significatives mais toutes les autres variables le sont.

Selon ce modèle, l'âge, la région, l'ancienneté, le fait d'être déjà assuré et l'âge du véhicule (caractérisé par deux variables, l'une étant retirée pour éviter la multi colinéarité) exercent une influence négative sur la réponse. Tandis que le fait d'être de sexe masculin, d'avoir un permis de conduire, d'avoir une fois endommagé son véhicule et le montant premium annuel exercent une influence positive sur la réponse. Ces relations ne sont pour la plupart pas surprenantes du fait de l'analyse exploratoire effectuée précédemment. Les quelques surprises proviennent des variables « Age » et « between_1_and_2 » qui ont une influence opposée à celle que nous laissaient présager l'analyse exploratoire.

Selon cette régression, la variable qui a le plus grand effet sur la variable réponse est la variable « Previously_Insured » qui a une influence fortement négative. Ensuite vient la variable « Vehicle_Damage » qui a une influence positive puis la variable « inf_1 » avec un effet négatif. De l'autre côté, les variables « Region_Code » et « Vintage » sont celles qui ont les plus faibles influences sur la variable réponse.

Nous aboutissons aux résultats de prévision suivants :

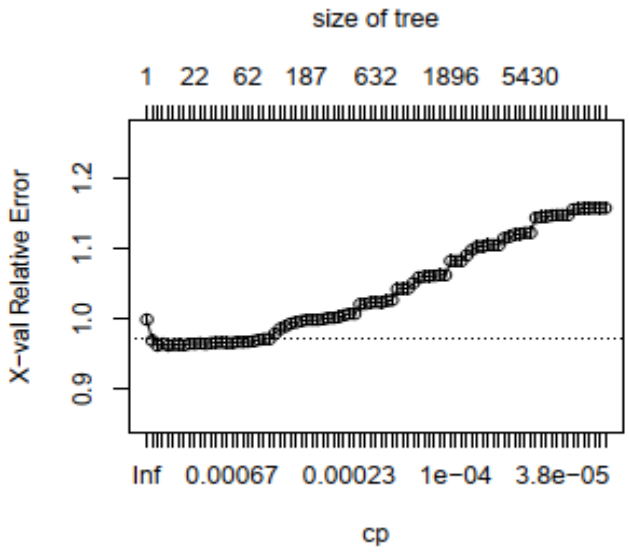
True class **Predictive Class**	**0**	**1**	**Total**
0	10513	2639	13152
1	284	297	581
Total	10797	2936	13733

Soit F1= 0.16889

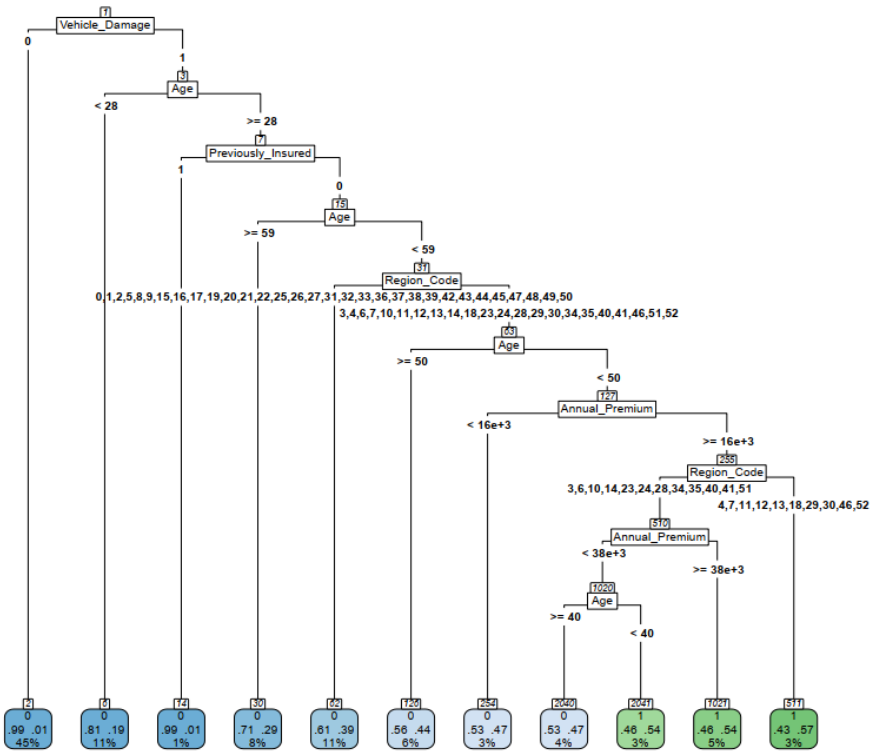
Le F1 score obtenu est très faible, on passe à l'estimation d'autres modèles

2. ARBRE ALEATOIRE (CART)

Pour la réalisation de notre CART, on met au point l'arbre maximal, qui sera ensuite élagué. L'arbre maximal obtenu contient **2144693077 nœuds** et à partir de celui-ci, on construit 67 sous arbres emboitées, associés respectivement à 67 hypers paramètres compris entre 0 et 2.469 e-03. Une représentation des différents hyper paramètres en fonction de la taille de sous l'arbre leur correspondant et de l'erreur relative est :



On détermine ainsi l'hyper paramètre optimal qui est **0.002589332.** Après quoi, on procède à l'élagage de l'arbre. L'arbre final obtenu, se présente comme suit :



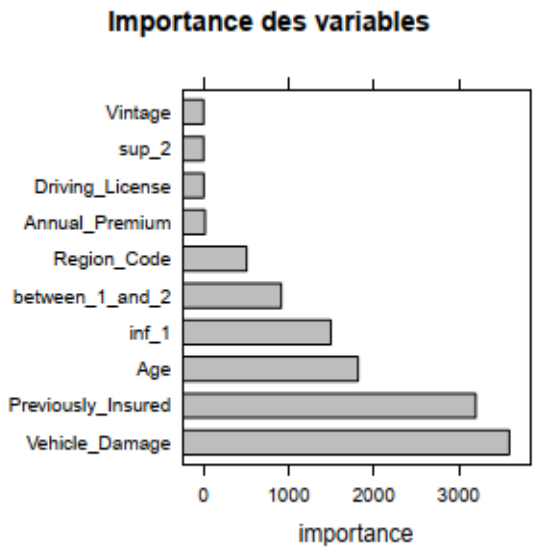
A partir du CART, on constate que conformément à l'analyse exploratoire, les individus n'ayant pas encore endommagé leur véhicule et ceux ayant déjà une assurance, n'ont pas tendance à être intéressé. Quand on passe aux prédictions, on obtient la matrice de confusion suivante :

True class **Predicted Class**	**0**	**1**
0	10115	2101
1	722	795

A partir de la matrice de confusion, on calcule les indicateurs que sont :

Accuracy : 79.44%
Precision : 52,40%
Recall : 27.45%
F1-Score : 36.03%

L'importance des différentes variables explicatives est présentée au travers du graphique ci-après :



On constate que les variables les plus importantes pour le pouvoir prédictif du modèle sont « Vehicle_damage », « Previously_Insured » et « Age ». Ceci est en accord avec l'arbre présenté, où les trois premières règles sont évaluées à partir de ces trois variables.

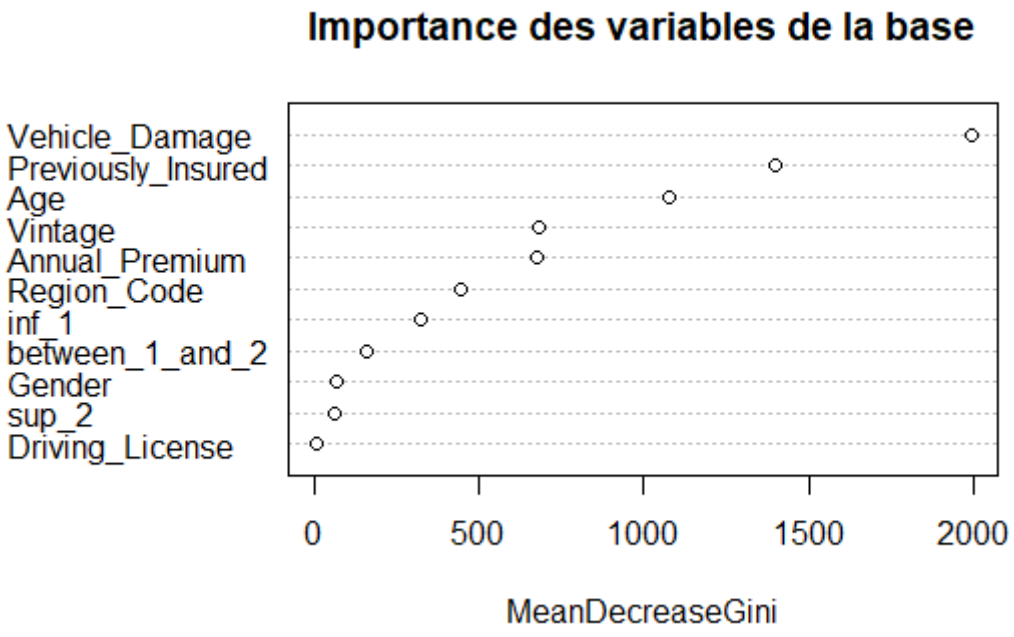
III. ESTIMATION DE MODELES COMPLEXES

On cherche désormais à obtenir de meilleures prédictions de la variable d'intérêt par l'estimation de modèles plus complexes. En effet les modèles d'ensemble learning étant généralement plus performants du point de vue de la prédiction nous nous attendons à obtenir de meilleurs résultats en procédant à l'estimation de ces modèles. Nous commençons par l'approche bagging avec le random forest puis nous nous intéresserons au boosting à travers le Adaboost et le XGboost.

1. FORET ALEATOIRE (RANDOM FOREST)

On réalise d'abord une forêt aléatoire de type classification sur notre base de données.

L'importance des différentes variables se présente comme suit :



On constate que les variables les plus importantes pour le pouvoir prédictif du modèle sont « Vehicle_damage », « Previously_Insured » et « Age ». De l'autre côté, les variables les moins importantes sont « sup_2 » et « Driving_License ».

Le tuning nous permet d'obtenir le nombre optimal d'arbres de la forêt et le nombre optimal de covariables à tirer aléatoirement. Ces valeurs s'élèvent respectivement à 500 et 4.

On estime alors le modèle à partir de ces hyperparamètres. On obtient les résultats consignés dans le tableau suivant :

True class Predictive Class	0	1	Total
0	10204	2269	12473
1	593	667	1260
Total	10797	2936	13733

On obtient F1-score= 0.31792

2. ADABOOST

A la suite du bagging (caractérisé par la forêt aléatoire) qui représente une stratégie d'ensemble learning, nous nous intéressons au boosting. On commence par estimer un Adaboost. Pour la réalisation de notre Adaboost, on se sert de la fonction boosting du package Adabag. On fixe les paramètres « Boos » à False, « mfinal » (le nombre d'itération pour lequel le boosting est exécuté) à 1000 et « coeflearn » égal à « Breiman » où alpha=1/2ln((1-err)/err). La matrice de confusion obtenue est :

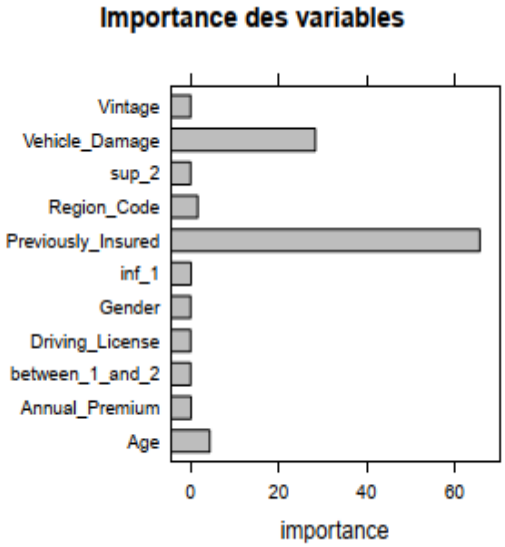
True class Predictive Class	0	1
0	10177	2228
1	660	668

Precision= 23.07%

Rappel = 50.30%

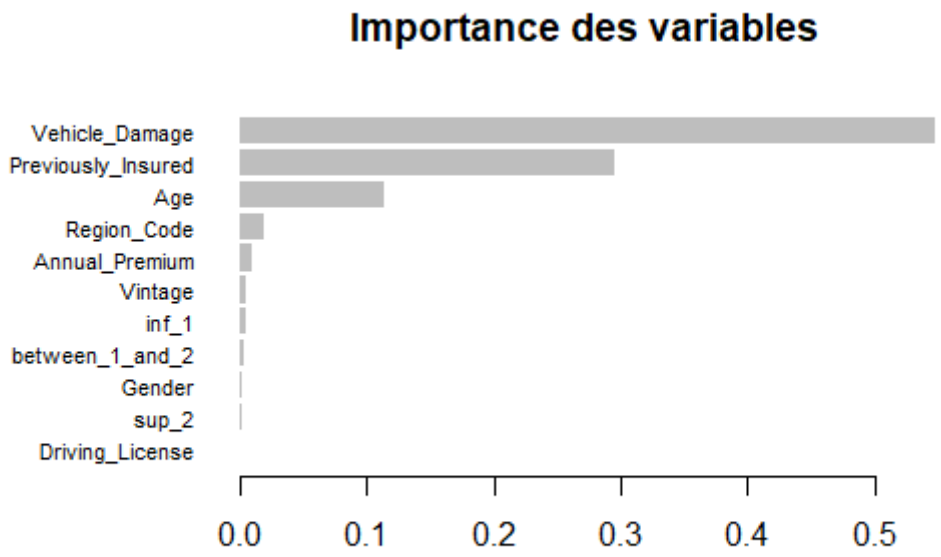
F1 score = 31.63%

Le graphe présentant l'importance des variables est le suivant :



3. XGBOOST

Nous estimons maintenant un XGboost. On utilise pour notre modèle une fonction objective logistique binaire qui correspond selon nous à notre situation, puisque notre variable d'intérêt est binaire. On procède donc à une classification selon cette variable. Après l'estimation d'un premier modèle, l'importance des variables se présente comme suit :



Ce résultat est semblable à celui obtenu précédemment.
Nous effectuons le tuning des hyperparamètres par cross-validation avec 5 folds. On obtient les valeurs suivantes :
taux d'apprentissage : 1/8
pénalisation : 1
ratio du nombre de covariables utilisées pour chaque nœud : 0.95
ratio du nombre d'observations utilisées par arbre : 0.95
Nombre maximal d'itérations : 1500

On obtient les résultats ci-dessous :

True class **Predictive Class**	**0**	**1**	**Total**
0	9654	1868	11522
1	1143	1068	2211
Total	10797	2936	13733

Soit F1= 0.41500

4. CHOIX DU MODELE

Le modèle final retenu est le XGboost. Nous choisissons ce modèle car il a le meilleur pouvoir de prédiction sur la base du F1-score.

Toutefois nous pouvons constater que ce score est relativement faible. Cela pourrait s'expliquer par le faible taux d'intéressés dans notre base de données d'entrainement. Le modèle aura donc tendance à prédire le non-intérêt plutôt que l'intérêt. De plus, l'analyse exploratoire ne nous a pas indiqué l'existence de plusieurs variables ayant un grand pouvoir discriminant.