UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

MASTER 2 TECHNIQUES D'INFORMATION ET DATA-SCIENCE EN ENTREPRISE

SCORING - Mr Yves Péchiné, Guillaume Leorat, Pierre Bioche

PROJET DE SCORING SUR LES CONTRATS D'ASSURANCES AUTOMOBILES

Andres SOTO

2020-2021

Table des matières

1. Introduction

Les établissements d'assurance et les organismes financiers disposent de systèmes d'information où sont stockées les données de leurs clients ainsi que les contrats auxquels ils ont souscrit. Ces données sont utiles au pilotage de leurs activités, au suivi des offres ou à la réalisation d'études. Elles portent aussi bien sur les données des clients (Nom, prénom, téléphone, date de naissance etc.) que sur les données de leurs contrats, les données financières ou les données de risques. L'exploitation de ces données permet d'en faire un levier de création de valeur durable et ce en maîtrisant le risque grâce aux techniques de machine learning capables de détecter les fraudes, en améliorant l'efficacité commerciale à travers du scoring[1] ou des études de marché et enfin en créant de la pertinence relationnelle vis-à-vis des clients.

Dans notre cas d'étude, nous avons à notre disposition des données du marché de l'assurance automobile recueillant des données des clients et leurs contrats d'assurance automobiles. Parmi ces informations on trouve la marque de la voiture, l'ancienneté du client, son nombre de contrats actifs, son nombre de contrats résiliés, son sexe, sa date de naissance etc. L'idée est de déterminer le score d'attrition c'est-à-dire la probabilité que le client résilie ou non son contrat d'assurance automobile, en fonction de sa situation contractuelle et de ses antécédents.

Afin d'étudier plus en profondeur la problématique, on commence dans une première partie par réaliser une analyse exploratoire des données afin de comprendre la structure de notre base ainsi que les informations qu'elle présente. L'objectif de cette partie étant de comprendre les différents liens existants entre les variables, ainsi que de juger si un nettoyage de la base est requis, notamment en termes de valeurs manquantes ou aberrantes.

On applique ensuite dans une seconde partie, différents modèles capables de prédire si le client résilie son contrat. On compare ensuite ces modèles à travers différentes métriques pour déterminer lequel présente la meilleure performance et s'adapte le mieux à nos données.

[1] Le scoring renvoie au calcul de la probabilité d'occurrence d'un événement recherché.

2. Présentation de la base de données

La base de données utilisée, issue du marché de l'assurance automobile, comporte 90 247 observations et 58 variables relatives aux contrats d'assurance automobile et aux informations des clients. Nous l'avons séparé en deux bases : une base d'apprentissage composée de 63 172 observations (70% de la base initiale) sur laquelle on entraîne les modèles et une base de test de 27 075 observations pour effectuer des prédictions sur la variable cible et tester leur performance. Le jeu de données étant déséquilibré, c'est-à-dire que seulement 11% des observations concernent un contrat résilié, nous l'avons séparé de façon à ce qu'il y ait la même proportion des deux modalités (actif/résilié) aussi bien dans l'échantillon de test que de train. Nous reviendrons sur ce point de manière plus détaillée par la suite. Notre variable cible 'CONTRAT' est catégorielle et indique si le contrat est actif ou résilié.

Les tableaux 1 et 2 présentent certaines variables de notre base de données. Elles peuvent être regroupées selon deux catégories : des variables qui décrivent les caractéristiques du contrat et d'autres qui décrivent les caractéristiques du client.

Caractéristiques du contrat

Variable	Type de variable	Définition
CONTRAT	Catégorielle	Résilié/Non résilié
IDECON	Catégorielle	Numéro de contrat
DTDBUCON	Date	Date de début de contrat
MMJECHPP	Date	Date de d'échéance du contrat
CDMARVEH	Catégorielle	Marque de la voiture
DUSGAUT	Catégorielle	Code d'usage de la voiture
RN_VL_VH	Catégorielle	Rang valeur du véhicule
ETAT	Catégorielle	État du véhicule
DTPMRMCI	Date	Date de mise en circulation du véhicule
PUI_TRE	Catégorielle	Puissance fiscale du véhicule

DTEFTMVT	Date	Date du dernier mouvement réalisé sur le contrat
CD_AGT	Catégorielle	Agent qui a vendu le contrat
CDMCE	Catégorielle	Code marché professionnel/particulier

Table 1. Tableau des variables concernant les caractéristiques des contrats

Caractéristiques du client

Variable	Type de variable	Définition
DT_NAI	Date	Date de naissance
CD_SEX	Catégorielle	Sexe du client
DEPT	Catégorielle	Département de résidence
REGION	Catégorielle	Région de résidence
CDSITFAM	Catégorielle	Situation familiale
DTOBTPDC	Date	Date d'obtention du permis de conduire
NIVBM	Numérique	Niveau de bonus malus
ANCCLI	Date	Ancienneté du client (date du premier contrat)
AU4R	Numérique	Nombre de contrats actifs 4 roues
SA	Numérique	Nombre de contrats actifs santé
DI	Numérique	Nombre de contrats divers actifs
S_0_N	Numérique	Nombre de sinistres dans les 12 derniers mois (non responsable)
RESAU4R	Numérique	Nombre de contrats auto résiliés

Table 2. Tableau des variables concernant les caractéristiques des clients

Création de nouvelles variables

Nous disposons de plusieurs variables qui concernent des dates. Ces dates renvoient à la date de début du contrat, la date de naissance, la date du premier contrat, la date d'obtention du permis de conduire ou encore la date de mise en circulation du véhicule.
Nous devinons que ces dates ont un format spécifique à l'outil SAS[2]. Nous utilisons donc SAS pour afficher ces dates sous format JJ/MM/AAAA et ce dans le but de pouvoir exploiter ces données.

Cette manipulation nous permet la création de nouvelles informations. En effet, nous calculons les variables "Âge du client" et "Ancienneté du client". Ces variables sont calculées par rapport à la date de début du contrat, de ce fait elles correspondent à l'âge et l'ancienneté du client au moment où son contrat débute.

On remarque aussi que la base de données à notre disposition concerne des contrats dont la date de début ne dépasse pas le 30 Septembre 2000.

3. Analyse exploratoire

Dans cette partie, nous explorons les données afin d'avoir un meilleur aperçu des informations qu'elles contiennent et de les nettoyer si cela est nécessaire.
Notre variable cible étant catégorielle à 2 modalités ('ACTIF','RESIL') nous la transformons en variable dichotomique qui prend la valeur 1 si le contrat est résilié et 0 sinon. Cela nous permet de faciliter le processus de modélisation par la suite.

Comme indiqué précédemment, on observe (figure 1) que seules 11% des observations de la base de données concernent des contrats résiliés. Ce déséquilibre nous amène à introduire la notion de séparation stratifiée, qui consiste à séparer le jeu de données avec la même proportion de contrats réalisés dans la base de test et de train. Cela nous permettra par la suite d'éviter des situations de sur-apprentissage.

[2] Si aucun format n'est explicitement appliqué, SAS affiche les dates par défaut comme le nombre de jours écoulés depuis le 1er Janvier 1960.

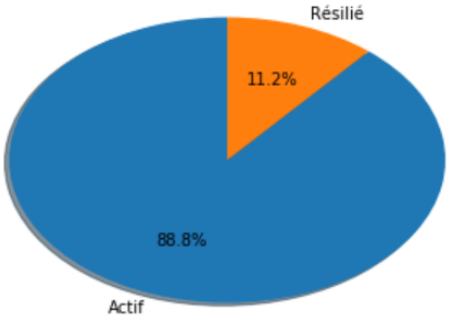

Figure 1. Proportion des contrats résiliés/Actifs

La distribution des variables (figure 2) à travers les boxplots met en évidence l'existence de variables avec des échelles très différentes. C'est le cas notamment de la variable NIVBM qui représente le niveau de bonus/malus, dont la valeur minimale commence à 50. Il est donc nécessaire de vérifier si le modèle de prédiction nécessite de normaliser les variables au préalable afin de pallier ce problème.

On constate aussi que les variables NIVBM et CRM présentent la même distribution. Leur coefficient de corrélation (présenté dans la matrice des corrélations figure 2) est égal à 1, cela signifie que les deux variables apportent exactement la même information. Il est donc judicieux de n'en garder qu'une seule afin d'éviter qu'il n'y ait de colinéarité[3] entre les variables. Nous choisissons de garder la variable NIVBM, puisque nous ne disposons pas de la définition de la variable CRM.

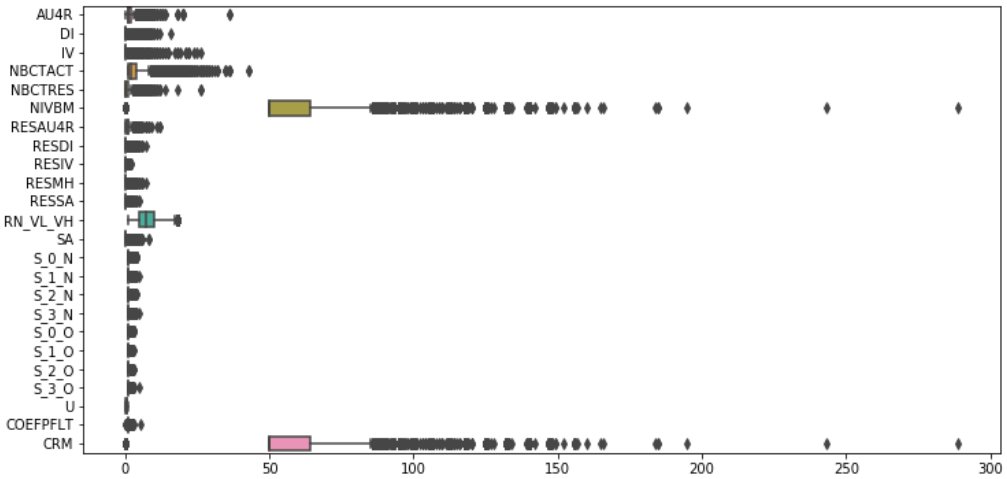

Figure 2. Distribution des variables numériques

[3] On parle de colinéarité parfaite si deux variables sont parfaitement dépendantes l'une de l'autre : $X2 = \lambda X1$ où $\lambda \neq 0$.Dans notre cas $\lambda = 1$.

Ainsi afin de détecter d'autres cas de colinéarité, nous traçons une matrice des corrélations (figure 3). On constate que la variable NBCTACT qui représente le nombre de contrats actifs présente une forte corrélation avec les variables DI , AU4R , MH , SA qui représentent respectivement le nombre de contrats divers actifs, le nombre de contrats actifs 4 roues, le nombre de contrats multirisques habitation actifs et le nombre de contrats actifs santé. Celle-ci est moyennement corrélée avec la variable IV qui représente le nombre de contrats individus-vie actifs. Ces corrélations sont cohérentes puisque la variable NBCTACT est une somme de tous les contrats actifs quel qu'en soit le motif. De même pour la variable NBCTRES représentant le nombre de contrats résiliés, qui est corrélée avec RESAU4R, RESDI ou encore RESMH et représentant le nombre de contrats résiliés pour les motifs auto, divers et multirisque habitation, respectivement. On décide donc de ne garder que les variables NBCTACT et NBCTRES.

On note aussi que la variable MTPAAREF (montant de la prime de référence) est corrélée à 80% avec la variable MTPAATTC (montant de la prime) et à 66% avec la variable RN_VL_VH (rang valeur du véhicule). Cela est cohérent puisque plus la valeur du véhicule (liée à sa cotation Argus[4]) augmente, plus la prime d'assurance à payer est élevée. On peut

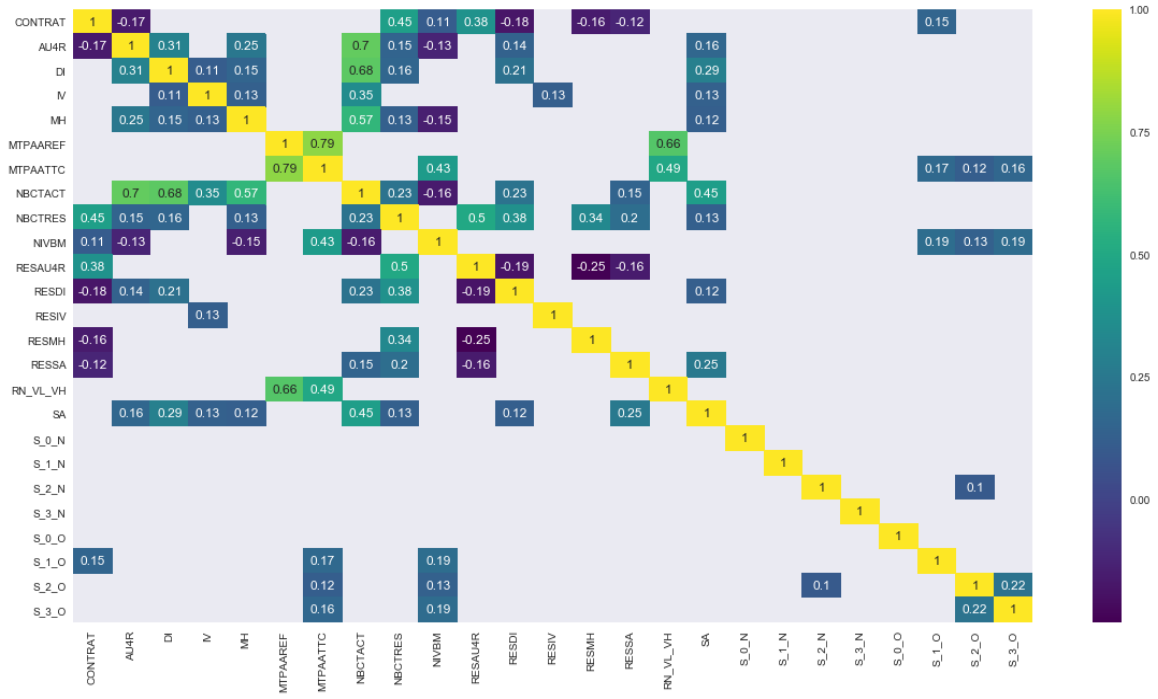

Figure 3. Matrice des corrélations

[4] La Cote Argus est une référence utile aux particuliers et professionnels qui s'en servent comme **base de négociation** pour l'achat d'une voiture d'occasion ou sa revente. Dans la plupart des cas, la Cote Argus est appréhendée comme une valeur de réserve. C'est-à-dire qu'elle correspond à un prix minimum.

visualiser cette relation dans la figure 4. Ainsi les véhicules ayant une bonne cotation ont en moyenne des primes d'assurance plus élevées. On choisit donc de ne garder que la variable MTPAATTC.

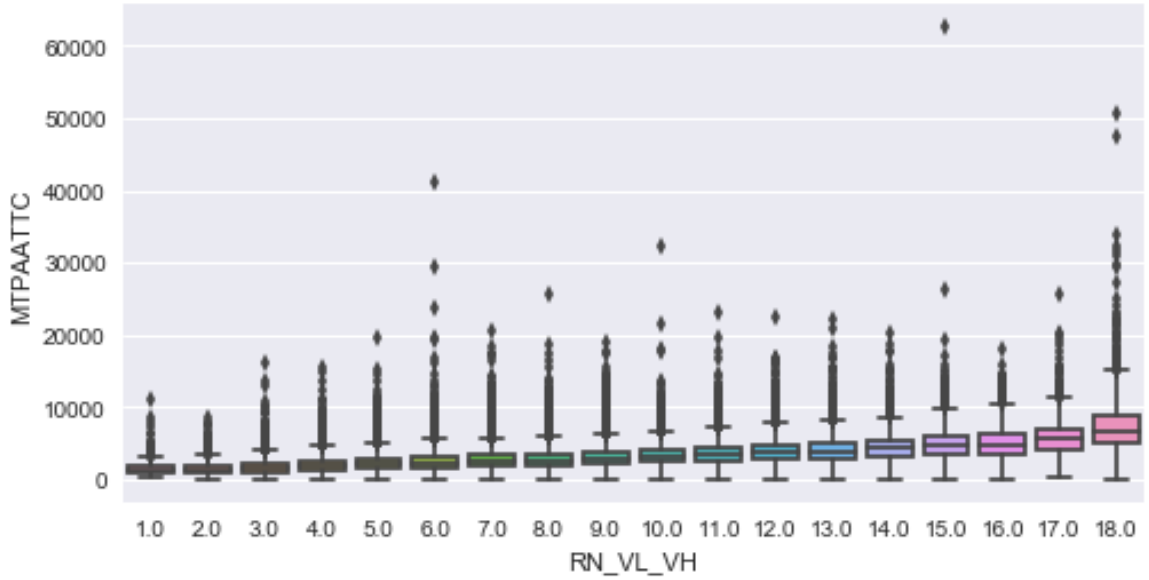

Figure 4. Distribution des montant de la prime d'assurance en fonction de la valeur du véhicule

4. Traitement des données et valeurs manquantes

Les variables catégorielles

Afin de mieux analyser nos variables, nous réalisons dans un premier temps un tableau qui contient pour chaque variable le nombre de valeurs distinctes, le nombre de valeurs manquantes ainsi que leurs pourcentages.

Variable	Nombre de valeurs distinctes	Nombre de valeurs manquantes	% de valeurs manquantes
NO_AFR	63170	0	0,00%
CD_AGT	1225	0	0,00%
CD_FML	25	366	0,58%
CDPRGES	6	0	0,00%

CDMARVEH	215	54	0,09%
LBMDLVH	5316	69	0,11%
NOTAREFF	17	183	0,29%
PUI_TRE	14	106	0,17%
CDMCE	2	0	0,00%
CD_SEX	2	66	0,10%
CDSITFAM	7	249	0,39%
DEPT	195	140	0,22%
REGION	11	140	0,22%
CLIACTIF	2	0	0,00%
CONTRAT	2	0	0,00%
ETAT	4	0	0,00%

Table 3. Nombre de valeurs distinctes et manquantes par variable

A l'issue des résultats du tableau, nous supprimons la variable d'identification NO_AFR.

Analyse des variables avec un fort nombre de modalités

Afin de réaliser cette analyse nous croisons les variables avec un fort nombre de modalités avec la variable cible (CONTRAT). Selon l'importance de la corrélation entre la variable et la variable cible; corrélation basée sur la part de chaque modalité de la variable et le pourcentage de contrat résilié pour chaque modalité; la variable est soit supprimée ou recodifiée. Ainsi les variables CD_AGT et LBMDLVH seront supprimées, tandis que les variables CDMARVEH et DEPT seront récodifiées en regroupant leurs modalités.

Les variables numériques

De même pour les variables numériques, nous créons la table des valeurs distinctes et des valeurs manquantes. Du constat des résultats, nous supprimons les variables d'identification

(IDECON, NUMFOY, NOCLIGES) et nous regroupons les modalités de la variable CDUSGAUT (Code d'usage de la voiture). Le regroupement des modalités est basé sur leur fréquence ainsi que le pourcentage de résiliation pour chacune des modalités.

On ne comptabilise pas plus de 0.1% de valeurs manquantes pour les variables catégorielles contre plus de 20% de valeurs manquantes pour les variables numériques.

Plusieurs de nos variables contiennent des valeurs manquantes. La figure 5 représente les 9 premières variables ayant un pourcentage de valeurs manquantes très élevé. Le reste des variables ne dépassent pas 5% de valeurs manquantes.

Ces 9 variables concernent la CSP soit la catégorie socio-professionnelle, ainsi que les variables relatives aux nombres de sinistres sur les années précédentes.

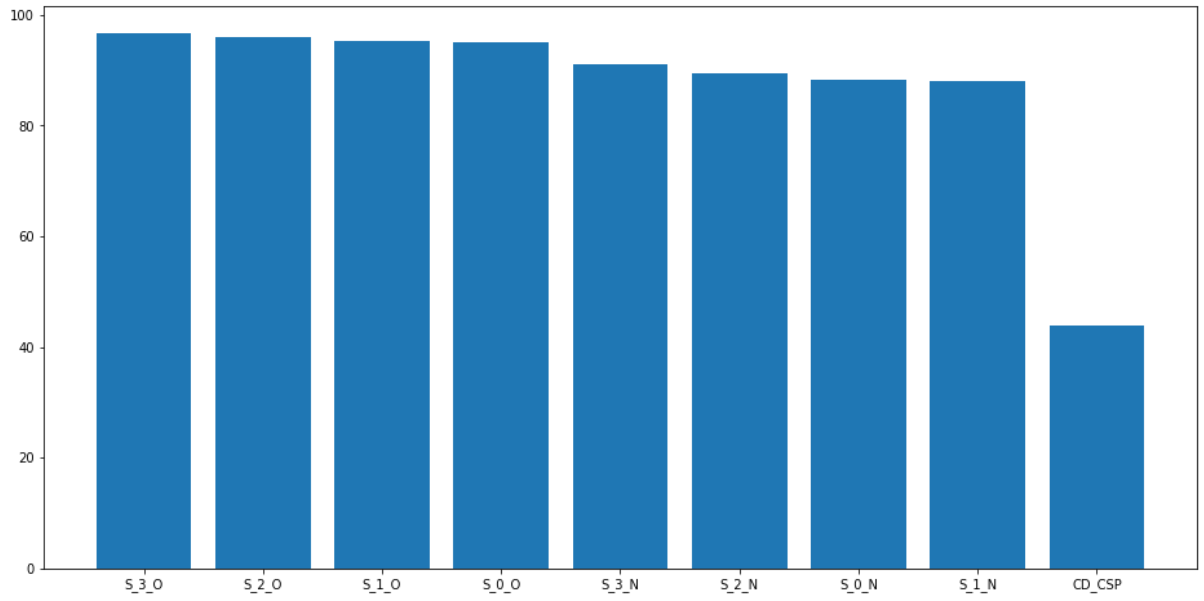

Figure 5. Pourcentage de valeur manquante par variable

Après vérification, on constate que les variables relatives au nombre de sinistres prennent des valeurs entre 1 et 5, et que dans 90% des cas ces valeurs ne sont pas renseignées. Or la probabilité d'occurrence d'un sinistre est faible. L'absence de la valeur 0 pour ces variables nous permet donc de déduire que le client ne remplit pas cette information dans le cas où il n'y aurait pas eu de sinistre, ainsi ces valeurs manquantes renvoient en réalité à 0 sinistres.

Nous remplaçons donc les valeurs manquantes de ces variables par des 0.

Concernant la variable CSP, celle-ci présente 44% de valeurs manquantes. Etant donné que nous ne disposons pas assez d'informations sur le client pour pouvoir réaliser des estimations sur ses valeurs, comme son salaire ou encore son domaine d'études, nous choisissons de nous passer de cette variable.

Imputation des valeurs manquantes par KNN

Concernant les autres variables ayant moins de 5% de valeurs manquantes, nous choisissons d'appliquer la méthode des KNN afin de remplacer ces valeurs.

L'algorithme des KNN est un algorithme d'apprentissage supervisé. C'est un algorithme de classification, mais il peut aussi être utilisé comme algorithme de régression. Cette méthode a pour but de classifier des points cibles (classe méconnue) en fonction de leurs distances par rapport à des points constituant un échantillon d'apprentissage (c'est-à-dire dont la classe est connue a priori).

Cet algorithme est efficace aussi bien sur les variables catégorielles que numériques. Nous décidons d'établir à 5 le nombre de clusters de départ.

Transformation des variables catégorielles en binaire

Nous transformons toutes nos variables catégorielles en variables dichotomiques, avec pour modalité 1 si l'événement se réalise et 0 sinon. Nous réalisons cette étape afin de pouvoir exploiter correctement ces variables dans les modèles. De plus, afin d'éviter le problème de multicolinéarité, nous enlevons pour chacune de ces variables 1 modalité.

5. Modélisation des données

Dans cette section, l'objectif est d'appliquer des algorithmes capables de prédire si le contrat a été résilié ou non, et par la suite comparer les performances de ces modèles.

5.1 Sélection des variables

Tests statistiques à grande échelle

Etant donné que nous disposons d'un nombre important de variables, nous étudions la relation entre notre variable cible et les autres variables par le biais de tests statistiques afin de ne sélectionner que les variables pertinentes dans nos modèles.

Variables catégorielles :

Afin de vérifier s'il existe des variables qui ont un effet sur la résiliation ou non d'un contrat, nous décidons de réaliser un **test de Khi2** entre notre variable cible et les autres variables. Pour rappel, les hypothèses du test de khi2 sont les suivantes :

Hypothèses du test :

- ➔ Hypothèse H0 : La distribution de la variable qualitative sachant que le contrat a été résilié(= 1) est la même que la distribution de la variable qualitative sachant que le contrat n'a pas été résilié (= 0)
- ➔ Hypothèse H1 : Les distributions sont différentes

Si la valeur de la p-value est inférieure à 5% on rejette l'hypothèse H0 avec un risque de 5%. On considère que les distributions ne sont pas identiques.

```
CDMCE: p-value test chi 2 = 2.1401321876929863e-20
CD_SEX: p-value test chi 2 = 7.66786727733557e-05
REGION: p-value test chi 2 = 9.54494197873799e-09
CLIACTIF: p-value test chi 2 = 0.0
ETAT: p-value test chi 2 = 0.0
Marque: p-value test chi 2 = 1.8807153980424385e-09
Serie: p-value test chi 2 = 4.354889725645666e-18
Code_Routier: p-value test chi 2 = 2.9351688655420885e-270
Code_PRGRES: p-value test chi 2 = 0.11356296157725719
Code_Tarif: p-value test chi 2 = 0.0
Code_Famille: p-value test chi 2 = 1.2107351488997388e-39
```

Figure 6. Résultats obtenus des tests de Khi2

Nous rejetons H0 si la p-value est inférieure à 5%, dans notre cas, seule la variable PRGRES n'a pas d'effet sur le fait de résilier ou non un contrat. Nous nous séparons donc de celle-ci. De plus, nous remarquons que certaines variables comme le tarif, CLIACTIF (le fait qu'un client soit actif ou non) ou encore ETAT (état de la voiture) possèdent une p-value égale à 0, et ont probablement une forte corrélation avec notre variable cible. Cela peut potentiellement provoquer du sur-apprentissage. Néanmoins, pour commencer, nous utiliserons les 10 variables catégoriques qui ne rejettent pas H0 selon le test du Khi2.

Variables numériques :

Pour les variables numériques, nous réalisons plusieurs **tests de student** entre notre variable cible et l'ensemble de nos variables explicatives. Les hypothèses sont les mêmes que celles du Khi2, sauf que cette fois-ci nous testerons des variables quantitatives.

```
MTPAATTC: p-value test Student = 5.80566510832367e-39
S_2_N: p-value test Student = 0.3960334158864298
S_3_N: p-value test Student = 0.0025708994488245576
S_1_N: p-value test Student = 0.006146989780707582
S_0_N: p-value test Student = 0.02206052435762694
S_3_O: p-value test Student = 0.000381125158713481
S_2_O: p-value test Student = 5.0822318297980345e-15
S_1_O: p-value test Student = 4.5796089723469486e-11
S_0_O: p-value test Student = 2.305478656406039e-16
NIVBM: p-value test Student = 2.2122350440677027e-159
NBCTRES: p-value test Student = 0.0
NBCTACT: p-value test Student = 5.730201245890376e-34
age: p-value test Student = 6.026837321568648e-28
anc_client: p-value test Student = 1.5534077247535359e-21
```

Figure 7. Résultats obtenus des tests de Student

Ici, seules les variables S_2_N (une des variables sur le nombre de sinistres dans les douzes derniers mois) et CAT_CDUSGAUT (les modalités les plus présentes dans le code d'usage de la voiture) ne rejettent pas l'hypothèse H0 au seuil de 5%.
Aussi, seule la variable NBCTRES (nombre de contrats résiliés) a une p-value de 0. Celle-ci semble jouer un rôle significatif dans le fait de résilier ou non un contrat d'assurance auto. Nous retenons au total 13 variables numériques.

5.2 Séparation du jeu de données et validation croisée

Comme cité précédemment, nous avons décidé de partir sur une proportion de 30% pour les données de la base test, car notre base contient assez peu de données. De plus, la variable à expliquer étant déséquilibrée (11% de contrat résiliés), nous introduisons la notion de séparation stratifiée, qui consiste à séparer le jeu de données tout en gardant les mêmes proportions de contrats réalisés dans les bases train et test. Ceci nous permettra d'éviter les situations de sur-apprentissage. Aussi, nous utilisons la technique de validation croisée. Elle est utile lorsque nous disposons de peu de données, car allouer une partie du dataset pour le test d'un modèle reviendrait à réduire la quantité déjà faible de données dont nous disposons.

Elle nous sera, par ailleurs, efficace pour le calcul de la métrique que nous utiliserons plus tard (l'AUC).

5.3 Algorithmes

Suite à l'introduction d'une variable binaire par modalité de variable catégorielle, nous disposons de 46 variables explicatives. Etant donné que le nombre de variables retenues est élevé, nous décidons de le réduire en effectuant un Random Forest (algorithme d'apprentissage supervisé détaillé par la suite) pour sélectionner les variables ayant un taux d'importance au-dessus de 0.05 (et donc différent de 0). Cela nous évitera de faire tourner le modèle avec des variables qui n'ont pas d'effet significatif dans la prédiction de résiliation d'un contrat, mais aussi d'identifier les variables les plus importantes.

```
4 features selected
NBCTACT 0.1023683864484297
NBCTRES 0.1029950191659264
CLIACTIF_1 0.13857689741385 36
ETAT_1 0.5995103291958848
```

Figure 8. Variables sélectionnées grâce au Random Forest

Ainsi, les 4 variables sélectionnées selon l'algorithme du Random Forest résumées dans le figure 8. L'une des modalités de la variable ETAT semble avoir une forte importance, nous décidons donc de ne pas la prendre en compte dans ce qui suit car celle-ci domine les autres variables.

```
9 features selected
MTPAATTC 0.08035312317141587
NIVBM 0.032399201143385715
NBCTRES 0.19689101988984847
NBCTACT 0.0586060439433601
age 0.0696284728439761
anc_client 0.05361860258546004
Marque_3 0.010638122975855108
CD_SEX_1 0.010264291550166957
CLIACTIF_1 0.30337467263257284
```

Figure 9. Variables sélectionnées après suppression de la variable "ETAT"

En supprimant la variable ETAT, nous remarquons cette fois-ci que 8 variables sont sélectionnées, dont une seule est catégorielle. Les plus "importantes" sont NBCTRES (le nombre de contrats résiliés) et CLIACTIF_1 (le fait qu'un client ne soit plus actif). Ces 8 variables seront celles que nous utiliserons dans nos modèles de prédiction.

Arbre de décision

Nous commençons par l'un des algorithmes de classification les plus simples : l'arbre de décision. Il fait partie des algorithmes d'apprentissage supervisé. Celui-ci est utilisable à la fois pour la classification et la régression. Il est réputé pour être assez simpliste et a tendance à réajuster, car tout est imbriqué en un seul arbre.

Concernant le choix des paramètres optimaux, nous décidons de partir sur 3 approches différentes pour commencer. Une première méthode dite "aléatoire" qui consiste à faire varier de manière aléatoire un certain nombre de paramètres, puis de sélectionner le meilleur ensemble de paramètres en se basant sur l'AUC.

La seconde méthode est le grid search. Il croise l'ensemble des paramètres et construit un modèle pour chacun de ceux-ci, de ce fait cette méthode est assez coûteuse en temps. Les meilleurs paramètres seront sélectionnés avec le meilleur score d'AUC.

Enfin, la dernière méthode consiste à effectuer un bayes search. Cette méthode consiste à construire un modèle probabiliste de la fonction qui relie les valeurs des hyperparamètres à l'objectif évalué sur un ensemble de validation en évaluant itérativement une configuration de paramètres prometteurs sur la base du modèle actuel, le but sera de chercher l'optimum. Les meilleurs paramètres seront sélectionnés avec le meilleur score d'AUC.

Afin de comparer les modèles obtenus à l'aide des 3 méthodes, nous utilisons la courbe ROC avec l'AUC et la courbe précision/rappel. La courbe ROC est une courbe représentant les performances d'un modèle de classification pour tous les seuils de classification. Elle trace le taux de vrais positifs en fonction du taux de faux positifs.

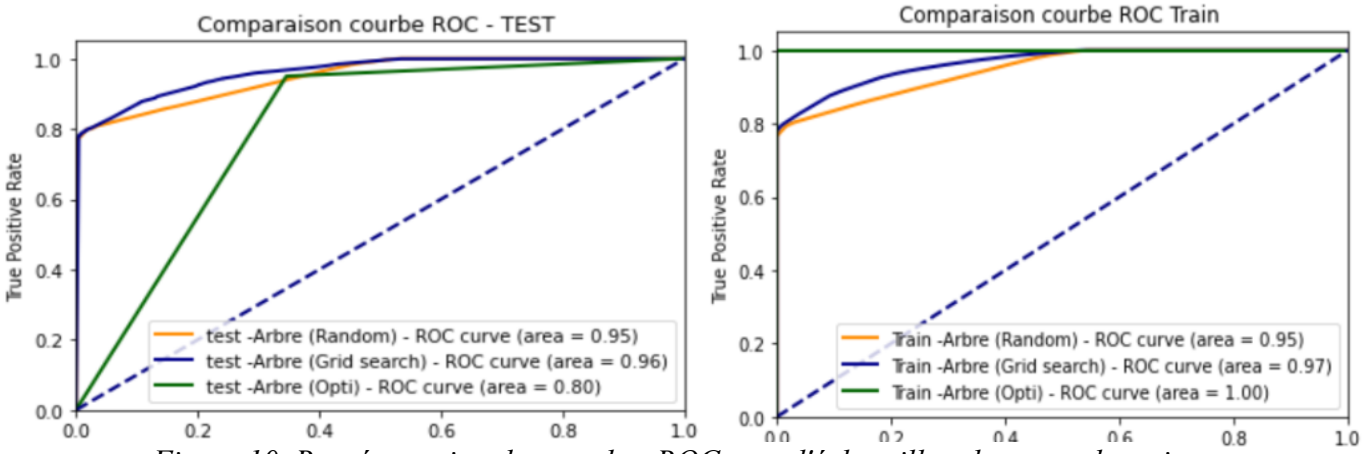

Figure 10. Représentation des courbes ROC pour l'échantillon de test et de train

La figure 10 représente la comparaison entre les courbes ROC sur le jeu de données d'entraînement et celles sur le jeu de données de test. La méthode ayant trouvé le modèle avec les paramètres les plus performants est celle du grid search. Le taux de vrais positifs en fonction du taux de faux positifs y est le plus élevé (l'AUC est de 0.96).

Nous retenons donc la méthode du grid search dans les algorithmes qui suivent, bien que celle-ci soit très coûteuse en temps. La méthode aléatoire et le bayes search procurent de moins bons résultats dans notre cas d'étude.

Enfin, nous remarquons que les modèles sont globalement très performants. Nous pensons que les variables que nous avons choisies (notamment le fait qu'un client soit actif) sont fortement corrélées avec la variable cible. Pour pallier ce problème, nous supprimons la variable CLIACTIF des modèles qui suivent.

La nouvelle liste de variables avec au moins 5 % d'importance selon le Random forest est la suivante :

```
5 features selected
MTPAATTC 0.10785345810974062
NBCTRES 0.28566527187681157
NBCTACT 0.0871992985548645
age 0.09484031987841544
anc_client 0.08361161124208556
```

Figure 11. Liste des variables retenues après suppression de CLIACTIF

Nous comparons dans ce qui suit l'ensemble des 6 modèles que nous avons appliqués pour sélectionner l'algorithme le plus performant dans notre cas d'étude.

Les 6 algorithmes sont les suivants :

- **L'arbre de décision**
- La **régression logistique** (algorithme d'apprentissage supervisé) : Il est utilisé en majorité pour des problèmes de classification binaire.
- **Random Forest** (algorithme d'apprentissage supervisé) : Tout comme le decision tree, il est utilisable pour les problèmes de classification et de régression. C'est un algorithme plus complexe que l'arbre de décision, qui réduit le surajustement causé

par un seul arbre. L'algorithme permet également de calculer l'importance des variables et est meilleur pour les problèmes de prédictions.

- ➢ **Bagging** : algorithme d'apprentissage qui combine plusieurs classifieurs et notamment l'arbre de décision, la régression logistique et le random forest dans notre cas.
- ➢ **Xgboost** : C'est une implémentation open source optimisée d'arbres de boosting de gradient (nous utiliserons comme classifieur l'arbre de décision, le random forest et la régression logistique)
- ➢ **Adaboost** : C'est un algorithme de Boosting qui combine plusieurs classifieurs peu performants pour amplifier leurs résultats (l'arbre à décision, la régression logistique et le random forest ici aussi).

La figure 12 représente la courbe ROC pour chacun des 6 modèles.

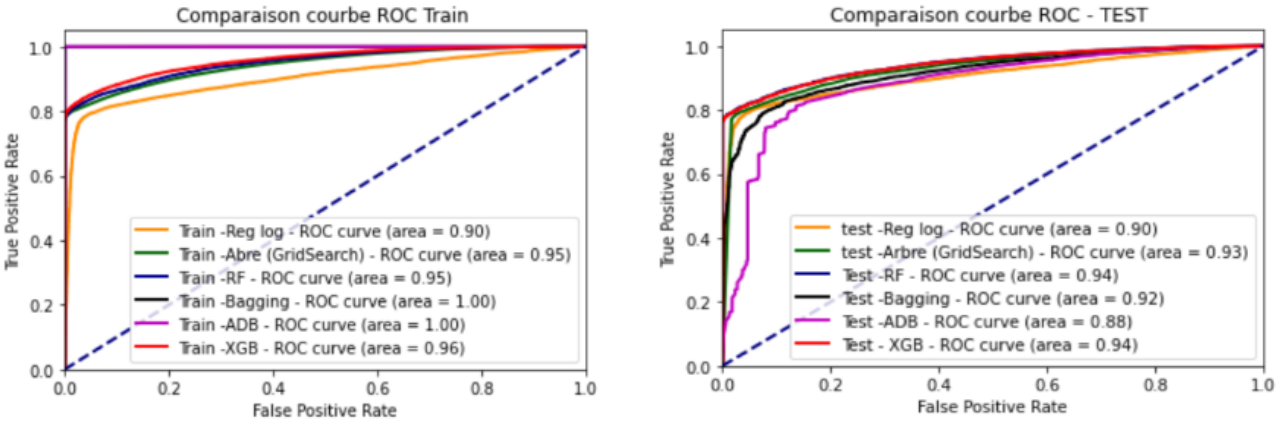

Figure 12. Courbes ROC des 6 modèles retenus pour l'échantillon d'apprentissage et de test

On constate que tous les modèles sont globalement très performants. L'AUC est supérieur à 0,90 pour presque tous les algorithmes. De plus, ceux-ci sont relativement stables car les courbes ROC sont comparables pour le train et le test. Les modèles sont donc globalement satisfaisants. Les 2 modèles qui semblent se distinguer sont le Xgboost et le Random Forest, avec un AUC de 0.94. Nous retiendrons le Xgboost, car celui-ci nous permettra de déterminer l'impact (positif/négatif) des variables explicatives sur le score d'attrition. De plus, selon la log-loss, le XGboost est le modèle se rapprochant le plus de 0.

	Arbre à décision	Régression logistique	Random Forest	Bagging	Adaboost	XgBoost
Log-loss	4.04	0.26	0.176	0.27	0.54	0.174

Table 4. Log-loss pour chacun des modèles

5.4 Performance du modèle retenu

Nous étudions dans cette partie les performances du modèle XGboost retenu.

Courbe précision/rappel

Nous commençons par établir un seuil de cut off. Le cut off est la valeur seuil telle que si le score est supérieur ou égal au cut off alors Ypred est égal à 1. Par défaut, dans Sckit-learn le cut off est à 0.5. Pour établir celui-ci, nous étudions la courbe précision/rappel du modèle.

Pour évaluer les performances d'un modèle de façon complète, il est nécessaire d'analyser à la fois la précision et le rappel. Malheureusement, précision et rappel sont fréquemment en tension. Ceci est dû au fait que l'amélioration de la précision se fait généralement au détriment du rappel et inversement.
La **précision** mesure le pourcentage de contrats résiliés ayant été classifié correctement, le **rappel** mesure le pourcentage de contrats non résiliés ayant été classifiés correctement.

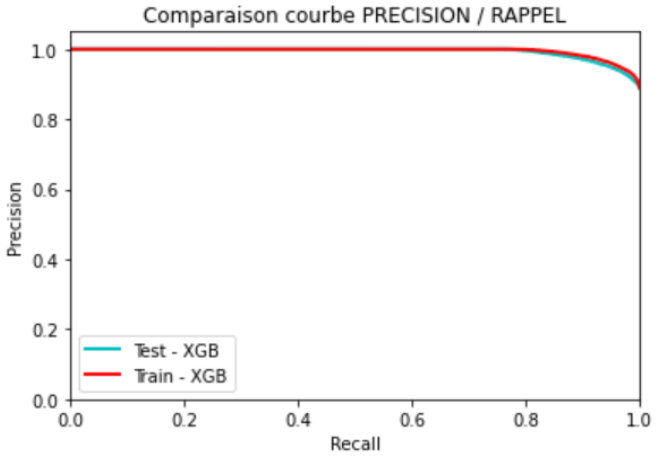

Figure 13. Comparaison courbe Précision/Rappel

La figure 13 présente la courbe de précision en fonction du rappel. Nous remarquons que la courbe correspond presque à une ligne droite. Ce qui ne s'agit pas d'une "bonne situation". Nous aurons presque toujours une précision parfaite (de 100%) mais en contrepartie le rappel sera beaucoup plus faible. En regardant la courbe test, et pour une précision de 100% nous

prendrons un rappel de 80%. C'est-à-dire que nous allons cibler 80% de la population des individus avec un contrat résilié pour 100% de précision.

Matrice de confusion

Pour évaluer la performance du modèle, nous traçons une matrice de confusion avec les prédictions du jeu de test et le choix du cut-off déterminé plus haut.

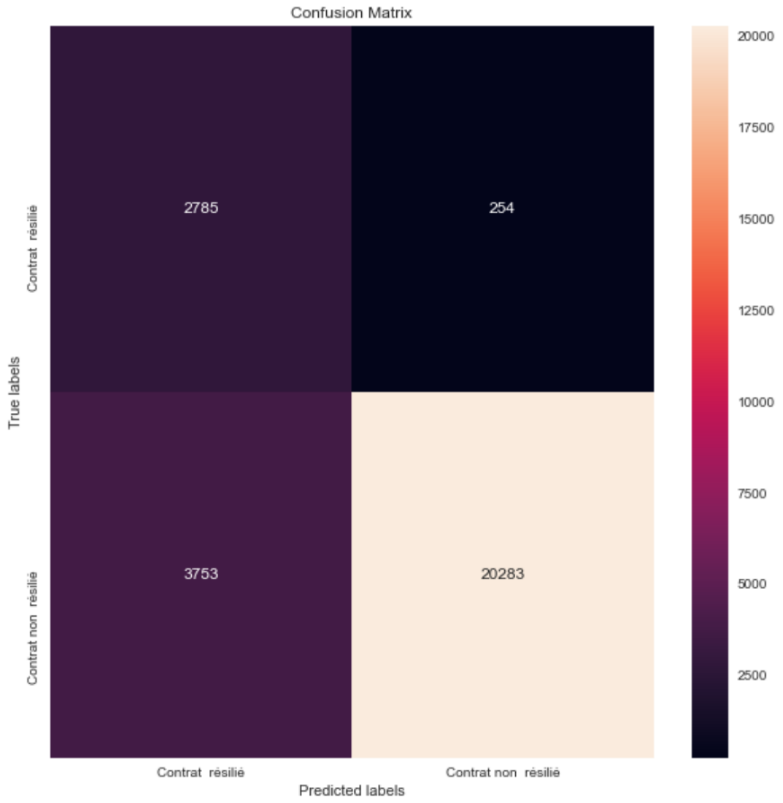

Figure 14. Matrice de confusion du modèle

Le modèle est globalement très performant. En effet, sur les 3039 contrats résiliés de la base de test celui-ci en a classé 2785 dans la bonne catégorie et 254 dans la mauvaise. Cela nous donne un taux d'erreur d'environ 8,5%. La précision n'est cependant pas équivalente à celle choisie grâce à la courbe précision/rappel. Pour ce qui est des contrats non résiliés, celui-ci en a classifié 20283 correctement et 3753 incorrectement, ce qui donne un taux de bonne prédiction de 86%. Le taux d'erreur (rappel) est d'environ 18%. Ce qui correspond presque parfaitement à ce que nous avions déterminé plus haut avec la courbe précision/rappel.

Variables les plus importantes

Nous représentons les 5 variables les plus significatives du modèle dans la prédiction de la résiliation ou non du contrat.

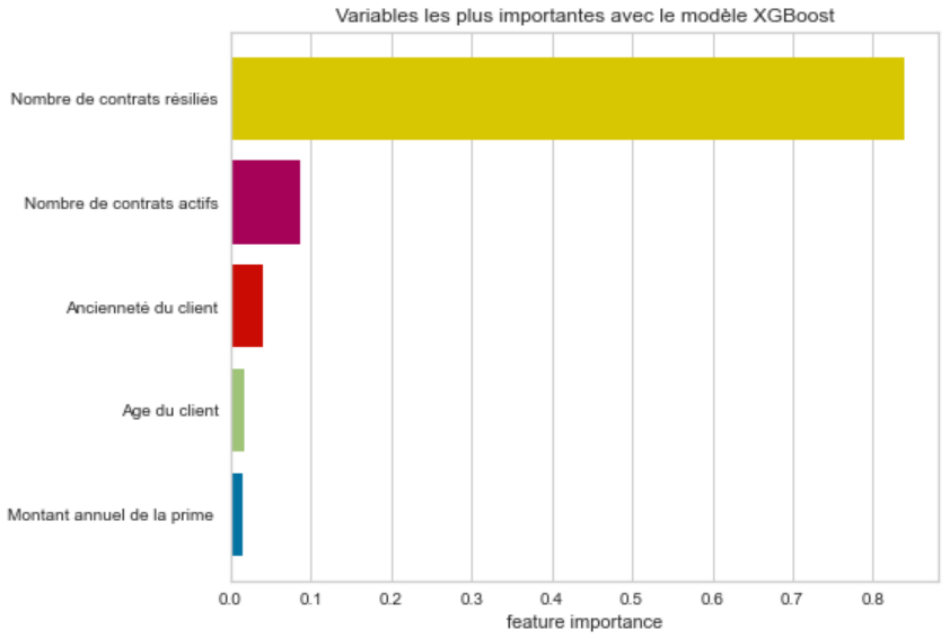

Figure 15. Les 5 variables les plus importantes du modèle XGBoost

Finalement, d'après le graphique ci-dessus, la variable NBCTRES (nombre de contrats résiliés) est celle ayant le plus d'importance (plus de 85%) dans la prédiction de résiliation d'un contrat d'assurance automobile. Celle-ci est suivie du nombre de contrats actifs, l'ancienneté du client, son âge et le montant annuel de la prime, qui ont toute une importance relativement moindre.

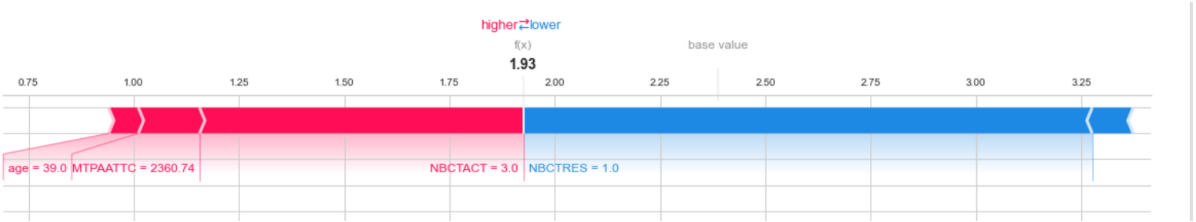

Figure 16 : Impact des variables explicatives sur la variable cible

En définitive, nous constatons sur la figure 16 que les variables qui ont un impact significatif sur la probabilité de résiliation d'un contrat d'assurance automobile, lorsque leurs valeurs augmentent sont l'âge, le montant annuel de la prime et le nombre de contrats actifs.

Courbe LIFT

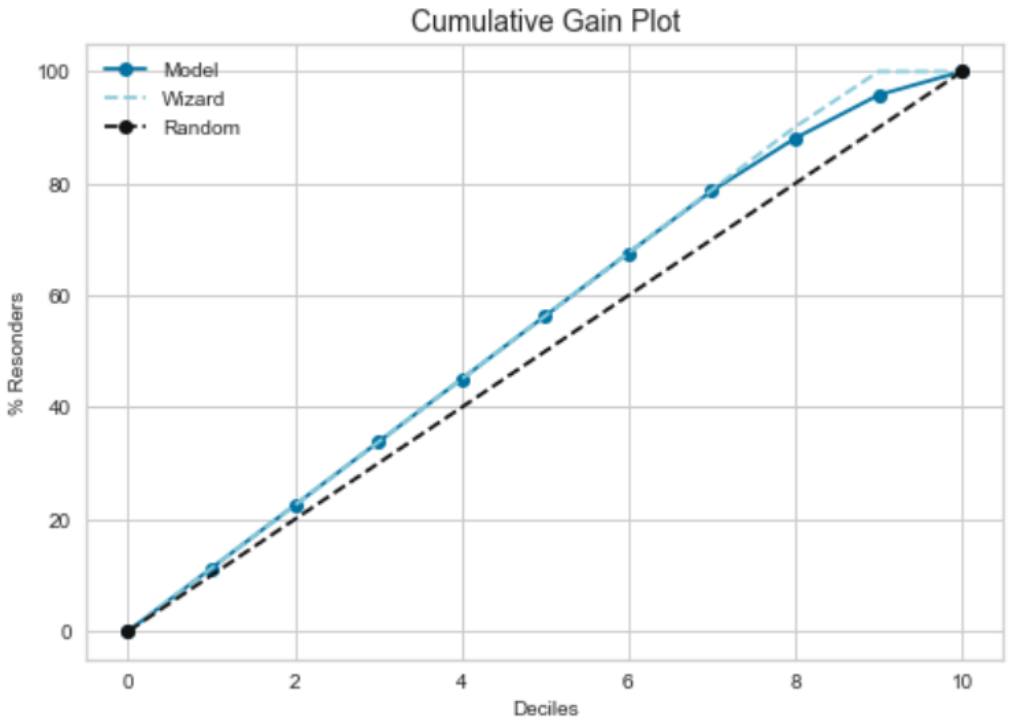

Figure 17 : Courbe LIFT de l'XGboost

La figure 17 présente la courbe LIFT du modèle Xgboost. Les lignes d'apprentissage (en pointillés bleu clair) et de test (trait bleu foncé) représentent la réponse attendue en utilisant le modèle prédictif. L'ensemble de données d'apprentissage entraîne le modèle et l'ensemble de données de test l'évalue. La ligne de référence pointillée noir représente une ligne avec une pente égale à 1, ce qui correspond à la réponse aléatoire attendue sans le modèle. Les gains supérieurs à 1 indiquent que les résultats du modèle prédictif sont meilleurs que les résultats aléatoires.

On peut considérer ici que le modèle est plutôt stable car la ligne de test se confond presque totalement à la ligne d'entraînement indiquant qu'il n'y a pas eu de surapprentissage.

6. Conclusion

Les scores d'attrition sont un moyen très efficace pour les entreprises d'identifier à l'avance les clients les plus à même de résilier leurs contrats et de partir à la concurrence. Nous avons à travers ce projet réalisé cet exercice sur des données assurantielles concernant des contrats automobiles.

Nous avons donc exploité des données clients qui concernent à la fois les contrats telles que la date de début du contrat, le nombre de contrats actifs ou résiliés ou encore le montant de la prime d'assurance ainsi que des données qui concernent les caractéristiques des clients comme leur âge, sexe, ancienneté, CSP etc. L'exploitation de ces données a nécessité un traitement préalable de celles-ci, notamment en termes de valeurs manquantes, de corrélations entre les variables, de regroupement des modalités ou encore de création de variables binaires à partir des modalités des variables catégorielles.

Le nombre important de variables à notre disposition nous a poussé à déterminer à travers des tests de Khi2 et de Student lesquelles sont statistiquement significatives au seuil de 5%, cette technique nous a permis de réduire les variables à prendre en compte dans les modèles afin de réduire le coût en termes de temps de calcul et de complexité.

Enfin, nous avons pu appliquer différents modèles afin de pouvoir prédire la probabilité qu'un contrat client soit résilié. Nous avons réussi à créer un modèle capable de prédire au mieux le score d'attrition. Celui-ci est globalement satisfaisant, à la fois performant et stable. Il nous a permis d'identifier les principaux déterminants de l'attrition. Cependant, la variable concernant le nombre de contrats résiliés semble tout de même avoir un fort pouvoir explicatif sur le modèle que nous avons retenu. Il est possible qu'à cause de celle-ci, le modèle soit légèrement biaisé.